THE INTELLIGIBILITY OF TRACHEOESOPHAGEAL SPEECH:

AN ANALYTIC AND REHABILITATION STUDY

This thesis came about with the support of:
The Netherlands Cancer Institute
Atos Medical AB, Sweden
De Nationale Hoorstichting

ISBN 10: 90-75575-15-7
ISBN 13: 978-90-75575-15- 6

Cover: Michel Nugteren (2Shine)
Lay-out: Legatron Electronic Publishing, Rotterdam
Printed by: PrintPartners Ipskamp, Enschede (www.ppi.nl)

THE INTELLIGIBILITY OF TRACHEOESOPHAGEAL SPEECH: AN ANALYTIC AND REHABILITATION STUDY

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor

aan de Universiteit van Amsterdam

op gezag van de Rector Magnificus

prof. dr. D.C. van den Boom

ten overstaan van een door het college voor promoties ingestelde

commissie, in het openbaar te verdedigen in de Aula der Universiteit

op woensdag 18 juni 2008, te 14.00 uur

door **Petra Jongmans**

geboren te Hendrik Ido Ambacht

Promotiecommissie

Promotores:	prof. dr. F.J.M. Hilgers
	prof. dr. ir. L.C.W. Pols
Copromotor:	dr. C.J. van As-Brooks
Overige leden:	prof. dr. P.P.G. Boersma
	prof. dr. U. Eysholdt
	prof. dr. V.J.J.P. van Heuven
	prof. dr. H.A.M. Marres
	dr. R. Speyer
	dr. I.M. Verdonck-de Leeuw

Faculteit der Geesteswetenschappen

This PhD project has been conducted at the Institute of Phonetic Sciences/ACLC and The Department of Head and Neck Oncology and Surgery of the Netherlands Cancer Institute and has been financed by a special grant from the Breuning ten Cate Foundation for Prof. Dr. Frans JM Hilgers of the Netherlands Cancer Institute.

Ter nagedachtenis aan mijn vader

Opgedragen aan mijn moeder

Tabel of Contents

Phonetic Symbols for Dutch

Vowels (monophthongs and diphthongs)

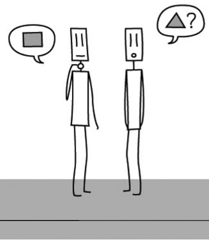
/ɪ/ as in kip
/ɛ/ as in tel
/ɑ/ as in tak
/ɔ/ as in top
/ʏ/ as in hut
/i/ as in kies
/eː/ as in keet
/aː/ as in taak
/oː/ as in pook
/u/ as in toen
/y/ as in fuut
/øː/ as in peuk
/ɛi/ as in lijst
/œy/ as in huis
/ʌu/ as in hout

Consonants

/b/ as in bad
/p/ as in pad
/d/ as in dat
/t/ as in tang
/k/ as in kat
/tj/ as in latje
/v/ as in vat
/f/ as in fit
/z/ as in zat
/s/ as in som
/h/ as in hand
/χ/ as in toch
/ʃ/ as in chocola
/m/ as in man
/n/ as in nat
/nj/ as in anjer
/ʋ/ as in wang
/j/ as in jong
/l/ as in lat
/r/ as in rat

1

General introduction



Abstract

For an adequate treatment of laryngeal cancer, total laryngectomy may be unavoidable. This is an operation whereby the entire larynx, including the vocal folds, is removed. After this surgical procedure, the anatomy and physiology of the whole speech channel have changed drastically, with the highly noticeable effect on speech quality. By applying a voice prosthesis, which enables the patient to use tracheoesophageal (TE) speech, speech quality is better than that of esophageal or electrolarynx speech, but still very deviant from normal laryngeal speech. Most studies so far have looked at *voice quality* rather than *speech intelligibility*. However, the various studies that have investigated the intelligibility of TE speech all show that intelligibility rates are lower than for normal, laryngeal speakers. In the Netherlands, only voice quality has been studied in detail. Considering the results of other studies, it is necessary to also study Dutch TE speech intelligibility. Once insight is gained in the typical problems of Dutch TE speakers, the next step is to improve TE speech intelligibility. The present study then is twofold: it is an assessment study and a therapy evaluation study. In the *assessment study*, perceptual and acoustic analyses have been employed to investigate TE speech intelligibility in more detail. For the *therapy evaluation study*, first a therapy program was developed based on typical TE speech problems. In order to evaluate the effects of the therapy, several *pre and post tests* were performed. In this introductory chapter, the necessary background information will be given as well as a brief overview of the literature. Also an outline of this thesis will be presented.

Part of this chapter appeared as:

Jongmans, P., Van As, C.J., Hilgers, F.J.M. & Pols, L.C.W. (2003). An introduction to the assessment of the intelligibility of tracheoesophageal speech. *IFA Proceedings 25*, 185-196.

1.1 Introduction

Communication is of the utmost importance in our lives, especially in the present day and age. In spoken language communication, the intelligibility of the speaker is very important for conveying the message correctly. When intelligibility is compromised, the communicative ability of the speaker will also diminish and with that his/her independence and ultimately the quality of life (Ackerstaff et al., 1994). Hence, it is necessary to try and improve the intelligibility of the speaker which, for instance in the case of laryngeal cancer, can be accomplished by surgical or therapeutic intervention. The current view is that these surgical and therapeutic interventions should be evidence-based. This means that the surgeon or speech language pathologist should search for, and examine, the evidence from clinical research, and then use his/her own experience and clinical judgment to decide if and how this evidence will be used (Sackett, 1996). A next step would be to measure and critically evaluate the effects of the therapy based on the results obtained.

There can be many causes for a diminished intelligibility. In this thesis, however, the focus will be on individuals who have had laryngeal cancer and have had a total laryngectomy in order to cure the disease. With a total laryngectomy, the entire larynx, including the vocal folds, is removed, with the most noticeable effect on voice and speech. There are three possible ways of voice rehabilitation after this surgical procedure: electrolarynx speech, esophageal speech or tracheoesophageal (TE) speech (details will be given in section 1.2). In this study, only TE speakers were included. Especially the voice quality of TE speakers has been studied quite extensively, also in Dutch (Van As, 2001). The choice for that topic was motivated by the fact that especially the voice source has altered dramatically. The question is if a loss in voice quality also entails a loss in intelligibility. However, TE speech intelligibility has received relatively little attention abroad and in the Netherlands. There are studies that have looked at TE speech intelligibility (for example Doyle et al., 1988; Hammarberg et al., 1990; Miralles & Cervera, 1995) and also for Dutch some pilot studies have been performed (Boon-Kamma, 2001; Roeleven & Polak, 1999; Oubrie, 1999), but unfortunately, most of these studies were limited in scope and most of them only focused on phoneme intelligibility. Nevertheless, these studies rendered interesting and valuable information on TE speech intelligibility and all showed that TE speech is less intelligible than normal laryngeal (NL) speech. Based on these results one would expect that more and structured attention would be paid to speech rehabilitation by the speech language pathologist. Unfortunately, to the best of our knowledge, no structured therapy program yet exists for TE speakers. Therefore, in this study, a therapy program has been developed and its effects have been evaluated by means of pre and post tests. In order to write such a program, first it needs to be investigated what the most typical problems are for Dutch TE speakers: it cannot be assumed that results from studies abroad will be the same for Dutch as each language has its own structure and phoneme inventory. In this study then, two main questions were formulated:

- What are the most typical speech problems TE speakers have?
- Can TE speech intelligibility be improved by means of a therapy program based on and addressing the typical problems found?

A third question related to the second one is:

- Is the set-up of the therapy program and are the pre and post tests used to measure effects appropriate?

In this introductory chapter, first it will briefly be explained what a total laryngectomy is and what kind of problems can be expected. After that a short overview of the literature will be given, both for intelligibility studies and therapy evaluation studies. This will be followed by an account of how intelligibility was investigated and how the therapy program was set up. This chapter will end with an overview of what is discussed in the separate chapters of this thesis.

1.2 Total laryngectomy and its consequences

In the Netherlands, around 600 individuals (with a male to female ratio of 5:1) are diagnosed with laryngeal cancer annually (Visser et al., 2003). When diagnosed at an early stage, these tumors can be treated well with radiotherapy or laser surgery. However, when the tumor is in an advanced stage, as well as in cases of recurrence of the tumor, a total laryngectomy is often necessary. Total laryngectomies occur in about 150 cases per year. During this operation the entire larynx is removed and the upper and lower airways become disconnected. The digestive tract is re-established and the trachea (wind pipe) is sutured into the skin at the base of the neck forming the tracheostoma (a permanent opening in the neck) through which the patient breathes (Figure 1.1).

One of the most obvious consequences of a total laryngectomy is the loss of natural voice. As was said in the introduction, over the years, different methods of voice rehabilitation have been developed, the two earliest being (i) the use of esophageal speech (speech whereby air is trapped in the esophagus and by ‘belching’ in a controlled way, this air can be used to produce short stretches of speech) and (ii) an electrolarynx (speech by means of a hand-held device with a vibrating membrane that acts as a

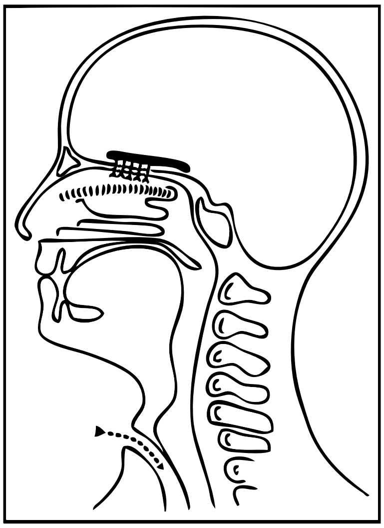


Figure 1.1: Schematic drawing of the situation after total laryngectomy with the patient breathing through a permanent stoma at the base of the neck.

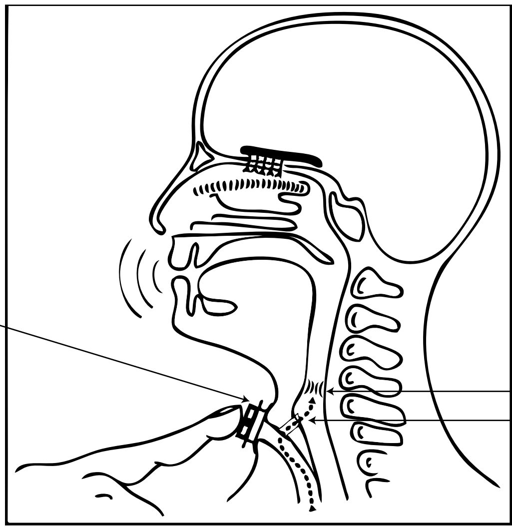


Figure 1.2: Schematic drawing of a speaker with a voice prosthesis, in particular of the new voice source (neoglottis) and the vocal tract. Names of the relevant (anatomical) parts are given; HME is heat and moisture exchanger, a device to improve the pulmonary climate.

primitive voice source, which is placed against the throat). This thesis will not cover these two methods of voice rehabilitation. From 1980 onwards, a voice prosthesis is often applied, enabling the patients to use tracheoesophageal (TE) speech (Singer & Blom, 1980). To make TE speech possible, a fistula (a small hole) is created between the trachea and the esophagus (digestive tract). The voice prosthesis (in this study the Provox® or Provox 2® prosthesis: Hilgers & Schouwenburg, 1990; Hilgers et al., 1997), inserted in the fistula to keep this tract open, acts as a one-way valve, which allows air to pass from the trachea into the esophagus, but prevents food from entering the trachea and the lungs (see Figure 1.2). Because the upper and lower airways are disconnected for breathing, the nasal functions during breathing (moisturizing, heating and filtering of the inhaled air) are lost. To compensate for this loss, and to maintain a proper pulmonary climate, a Heat and Moisture Exchanger (HME) was developed that covers the tracheostoma (Hilgers et al., 1996).

The pharynx with its mucosa overlying the pharyngeal constrictor muscles and/ or the cricopharyngeus muscle serves as a new voice source called the pharyngo-esophageal segment (PE segment; also called neoglottis or pseudo glottis). The great advantage of TE speech is that, like normal laryngeal (NL) voicing, it is pulmonary driven, i.e. air from the lungs is used to set the tissues of the neoglottis into vibration. This results in longer phonation times than achievable in esophageal speech and in a higher intelligibility rate than with esophageal or electrolarynx speech (Williams & Watson, 1987; Pindzola & Cain, 1988; Nieboer et al., 1988; Debruyne et al., 1994; Bertino et al. 1996; Max et al., 1996).

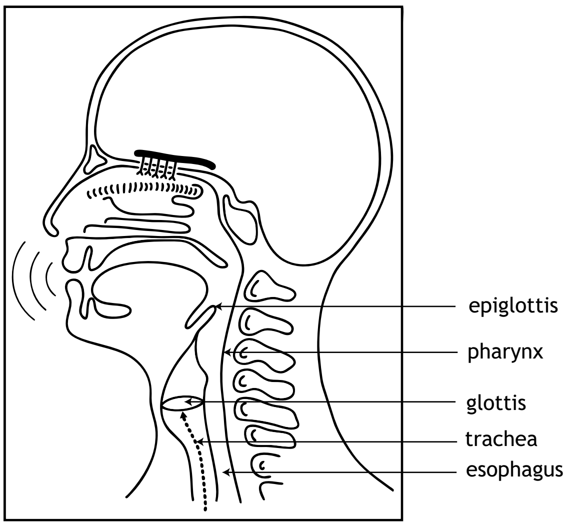


Figure 1.3: Schematic drawing of a normal laryngeal speaker, in particular the voice source and the vocal tract. Names of the relevant (anatomical) parts are given.

Even though TE speech is also pulmonary driven, it is still a very deviant way of speaking compared to NL speech. In Figure 1.3, it can be seen that, in the normal situation, the larynx and trachea are connected by the pharyngo-esophageal junction. The vocal folds (with the space between them called the glottis) are located in the larynx. The epiglottis, located at the point where larynx and base of tongue connect, helps to prevent food and liquids from entering the lungs. Phonation in NL speech is considered a myoelastic-aerodynamic event (meaning that both the elasticity of the laryngeal muscles and the airflow from the lungs are used for phonation) (Van den Berg, 1958). This myoelastic-aerodynamic theory of phonation implies that NL speech is initiated by pulmonary air from the lungs, setting the vocal folds (glottis) into vibration (aerodynamic component). The fundamental frequency of the vibration depends on the effective mass and stiffness of the vocal folds that are regulated by the sustained innervation of internal and external laryngeal muscles (the myoelastic component) and by the air stream from the lungs and the associated resonators. Speech production can be described according to the source-filter theory (Fant, 1960). This theory states that there is a source that produces sound. This sound is subsequently formed into speech by the resonators (oral cavities, nasal cavities) and the articulators (tongue, teeth and lips). The oral and nasal cavities are said to act as a *filter* that lets some frequencies pass and attenuates others. In the case of voiced sounds, for example, a periodic sound is produced in the glottis. The glottis is then called the *source*. In the case of voiced consonants, a second source is present at the place of constriction used to produce a particular consonant (for a more detailed account of laryngeal voicing, see for example Titze, 1994 and Liebermann, 1977). TE speakers do not have vocal folds, but they do have a new voice *source*: the neoglottis. The initiator of sound production remains the same (the lungs) as air is forced through the voice prosthesis from the trachea into the esophagus. The oral and nasal cavities *(filter)* and the articulators remain almost the same, even though the anatomy of the vocal tract changes slightly with the removal of the larynx, including the epiglottis. With a total laryngectomy, the hyoid bone is also removed and the pharyngeal mucosa, covered by the muscles at the base of the tongue and the constrictor pharyngeal muscles, is sutured in a tube-like fashion in order to re-establish the digestive tract. This procedure changes the shape of the pharynx and the range of motion of the base of the tongue might be influenced. In addition, the length of the vocal tract is more variable than in normal speakers due to the large variability in the cervical level of the neoglottis (Van As et al., 2001).

At this point, it is not clear how much control TE speakers still have over their neoglottis and whether their pseudo-phonation is myoelastic-aerodynamic or only aerodynamic. Some studies have investigated this issue (Moon & Weinberg, 1987; Mohri et al., 1994; Deschler et al., 1999). They all concluded that the myoelastic component is present, but that it is not used consistently and that it is not present in all speakers. It is assumed then that the neoglottis is less pliable than the vocal folds and that even if muscle activity is present, it is not used consistently.

Even though the greatest influence of a laryngectomy will be on voice quality, due to the extensive changes in the voice source, problems with intelligibility can also be expected based on the changes described above and on the existing literature:

Problems with the glottal phoneme /h/ *and the back phoneme* /k/: these phonemes are produced in the area where most changes have occurred. For /h/ also active adjustment of the neoglottis is needed, which is likely to be a problem for TE speakers.

Problems with voiced and voiceless plosives and fricatives: the production of a correct contrast between voiced and voiceless sounds requires active use of the neoglottis. As the neoglottis is assumed to be less pliable than the vocal folds, and the constrictor pharyngeal muscle activity is not used as consistently and controllably as are the laryngeal muscles, problems are expected. Also problems with the onset and offset of voicing and problems with stable voicing are expected.

Problems with vowels: the slightly altered vocal tract and the variable length of the vocal tract are likely to influence the acoustic properties of vowels, causing them to sound different, even though the articulation itself is not affected.

Problems with pitch: the instability of the neoglottic vibratory pattern causes undetectable or highly irregular pitch. Pitch as such will not be discussed in this thesis, as Van As (2001), Van Rossum et al. (2002) and Van Rossum (2005) studied pitch in alaryngeal speech in great detail. As Van As (2001) already showed in her study, the pitch of male and female laryngectomees is similarly low. The low voice causes additional emotional problems for female laryngectomees as their voice is often perceived to be a male voice. Due to the limited differences between male and female TE speech, only males have been included in this study. Even though only little attention will be paid to pitch in this thesis, pitch should not be ignored as irregular pitch patterns may also cause problems with accentuation in sentences, which can change the meaning of sentences and hence can diminish intelligibility.

Problems with phrasing: even though TE speakers also use pulmonary air for speaking, the timing of speech is assumed to be more difficult as speakers have to actively switch between breathing and speaking by closing off or keeping open the stoma with a finger, unless they use an automatic speaking valve, which obviates the need for this digital switching.

In the next section, a brief overview will be given of the main findings of TE speech intelligibility studies.

1.3 Tracheoesophageal speech intelligibility

In order to better understand the terms *voice quality*, *speech intelligibility* and *speech understanding*, short descriptions of these terms will be given first.

Abercrombie (1967) and Laver (1980) have both discussed voice quality extensively. The problem with the term *voice quality* is that it can be used for a variety of concepts. In this thesis, the definition of Laver (1980) is adopted: “Voice quality is conceived here in a broad sense as the characteristic auditory colouring of an individual’s voice, and not in the more narrow sense of the quality deriving solely from laryngeal activity. Both laryngeal and supralaryngeal feautures will be seen as contributing to voice quality”. Obviously, the terms laryngeal and supralaryngeal are not applicable to TE speakers. In this group, voice quality refers to neoglottal activity and vocal tract features.

With *speech intelligibility* the correct identification (production) of phonemes and words is meant. A good intelligibility is a prerequisite for a good *understanding* of what is said. Understanding is a higher level than intelligibility in the sense that only when a listener understands the speaker, the message of the speaker has been conveyed correctly. It is possible for speech to be perfectly intelligible without being understood (Hongyan, 2007).

So far most studies on TE speech have focused on *voice quality* rather than *speech intelligibility*. For Dutch, extensive research has been done by Van As (2001), who investigated TE voice quality in 40 male and female TE speakers. She used perceptual evaluations, acoustic analyses and imaging techniques to gain insight into the anatomical and morphological characteristics of the neoglottis in relation to TE voice quality. Her conclusion was that although TE speakers perform quite well considering the circumstances, TE voice quality is very deviant from normal speech. It is likely that a poor voice quality will influence intelligibility as well. Also, some changes in the vocal tract can be observed that might compromise intelligibility. Furthermore, certain TE speakers themselves complain about their poor intelligibility, especially in noisy situations, such as at parties and in shops. Unfortunately, intelligibility has received relatively little attention. It is, therefore, important to study TE speech intelligibility in more detail, so that oral communication after total laryngectomy can be optimized: a prosthetic voice should not only sound good, but, in the first place, should be well understood.

Several studies on TE speech intelligibility have been performed, either abroad or in the Netherlands. Main findings from these studies were that initial consonants are much harder than final consonants, that the voiced-voiceless distinction is difficult to produce for TE speakers and that the glottal phoneme /h/ is often omitted. In addition, the fricatives and plosives cause more problems than nasals and approximants, and also vowel intelligibility is compromised.

Although the abovementioned results provide valuable information on TE speech intelligibility, the problem is that most of the studies that investigated TE speech are limited in scale and/or investigate only a few aspects of speech. Phoneme intelligibility is by far the most frequently studied topic in this area. It is for this reason that in the present thesis extensive research has been conducted on the kind of problems TE speakers have, whereby it was tried to include almost all levels of speech: from phoneme level (consonants and vowels) to discourse level (monologues). Conversations and other communicative tasks, for example, will not be discussed in this thesis. Before the assessment of TE speech problems is discussed in more detail, an overview is given of the main findings of studies on therapy for alaryngeal speakers.

1.4 Therapy techniques/programs for alaryngeal speakers

As was mentioned in the introduction of this chapter, very little attention has been paid to the improvement of TE speech intelligibility in literature. Esophageal speech has received more attention, but most of the techniques used are *experience-based* (no effect measurements have been performed, but experience suggests that the technique(s) might help) rather than *evidence-based* (technique has been tested and found to be useful). As the goal in this thesis is to develop an evidence-based therapy program, we preferably used techniques that have been found to be successful in other studies. A systematic literature search was performed to investigate what kind of programs exist for the improvement of alaryngeal, and in particular TE speech, if any. We found only fifteen studies, seven of which were evidence-based and eight of which were experience-based. Only one study addressed TE speech, the other studies addressed esophageal speech. The most relevant of these studies focused on the improvement of consonants by using self-administered home practice, and on the improvement of voiceless plosives by using the so called ‘push harder’ technique, whereby pressure is built up and the plosive is released with more force. The other evidence-based studies looked at the effect of noise on the intensity of esophageal speech, and the effect of a psychological approach (for more detail, see chapter 4). If we could not find suitable training techniques in evidence-based literature for a certain problem included in the therapy program, techniques were based on the experience-based literature and on the experience of the speech language pathologists that designed the therapy program. The fact that many speech problems are found with TE speakers, both in Dutch and other languages, and the fact that only very little attention has been paid to solving these problems shows the necessity for a detailed assessment of TE speech, and, if TE speech indeed proves to be compromised, a well-structured evidence-based training program with pre and post tests. In the next section it will be described which methods can be used for assessing TE speech, followed by an account of how the therapy program was developed and how effects of the therapy were measured.

1.5 Assessment of TE speech intelligibility

Based on the literature and on the anatomical and physiological changes after total laryngectomy, it is strongly suspected that TE speech intelligibility will also be affected. If this is indeed the case, a therapy program should be available to improve intelligibility. In the introduction it was stated that it should first be investigated what the typical problems are that TE speakers have. When assessing speech intelligibility, a multidimensional approach is best, as speech itself is multidimensional: not only articulation is important, but also, for example, prosody and voice quality. Especially when one wants to know the possible cause of the problem, objective measurements are necessary. There are several possible assessment tests. In the original project proposal, four tests were described: perceptual analyses, acoustic analyses, videofluoroscopy and digital high-speed imaging. All four tests would provide valuable information and would complement each other rather than replace each other. A short description of the tests will be given below, but due to the breadth of topic, in the actual study, we had to limit ourselves to perceptual and acoustic analyses, leaving the two imaging techniques for a subsequent project.

1.5.1 Perceptual analyses

A perception experiment is still considered the gold standard in order to find out what kind of difficulties TE speakers have, as communication is mostly a perceptual matter. It should be noted though that communication is per definition a process between at least two parties. When problems in communication arise, both the speaker and receiver can be the cause. Even though it is a subjective method of research, it gives a good overview of how speech is perceived and what the difficulties are. The results from a perception experiment can be used as a starting point for further analyses. We included almost all levels of speech in our *speech material*: all Dutch consonants and vowels, consonant clusters, sentences and discourse. Originally, speech material intended to investigate the production of pitch was also included. However, at a later stage this material was omitted from further analyses as this subject has been discussed extensively by Van Rossum (2005).

The *choice of raters* is always an important issue in a perception experiment: experienced listeners will score differently from inexperienced listeners and naïve listeners will score differently from listeners acquainted with TE speech. In this study, naïve listeners (representing the population that laryngectomized patients encounter in daily life) were chosen as raters, since their ratings will best reflect how well a TE speaker is perceived (understood) in daily life. Also it was found in literature that naïve listeners are to be preferred when the research goal is the communicative function of (TE) speech (Van As et al., 1998) and that the intelligibility data of naïve listeners are of greater utility when establishing clinical goals (Doyle et al., 1989).

The results of the perceptual analyses, discussed in chapter 2, were used as the basis for the therapy program.

1.5.2 Acoustic analyses

In this thesis, the results of the perceptual analyses were especially important in order to decide on the contents of the therapy program. However, perceptual analyses only show *what* the problem is (with non manipulated speech) and not necessarily *why* this problem might occur. Acoustic analyses can show why something is perceived the way it is. Several studies have looked at the acoustics of TE speech. Sometimes these studies were voice-quality related (Van As et al, 1998), but also studies investigating pitch (Van Rossum, 2005; Most et al., 2000), vowel formant frequencies and vowel durations exist (Cervera et al., 2001; Oubrie, 1999; Robbins et al., 1986). However, most studies, including the study in this thesis (Jongmans et al., 2006; 2007), have investigated the voiced-voiceless contrast, as confusions between voiced and voiceless plosives and fricatives are the most common confusions in TE speech (for more detail, see chapter 3). The problem with voicing was to be expected as voicing is produced at (neo)glottal level and it is assumed TE speakers have less control over their neoglottis than NL speakers have over their glottis. In this thesis, the voicing contrast has also been studied, specifically what acoustic correlates are used to produce this contrast. There is another reason for studying the voiced-voiceless contrast: the (in)ability to produce the voicing contrast correctly and consistently can potentially provide information on whether TE speech is myoelastic-aerodynamic or only aerodynamic. If, for example, TE speakers use vibrations of the neoglottis to distinguish between voiced and voiceless, it is likely that a myoelastic component is present. Results of the acoustic analyses are useful for therapy purposes: once it is known what TE speakers do when producing voice, techniques in the training program can focus more specifically on certain problems.

1.5.3 Videofluoroscopy

Videofluoroscopy is a commonly used clinical imaging tool used for the visualization of the neoglottis, mostly in lateral view. The anatomy and morphology of the neoglottis as observed with this method appears to predict tracheoesophageal voice quality in a rather consistent way (Van As et al., 2001, see also Op de Coul et al., 2003). With videofluoroscopy, x-ray images of the neoglottis allow to visualize the anatomy/morphology of the neoglottis and the vocal tract. The patients swallow barium sulphate, a contrast fluid that enables visualization of the surface of the mucosa. For investigation of tracheoesophageal voice production a sustained vowel is most often used. For intelligibility research different vowels could be used to observe the changes in the neoglottis and vocal tract and how these changes influence intelligibility. Observations and quantitative measures of the neoglottis and the vocal tract for different vowels may help to find out why certain vowels cause difficulties. However, as already mentioned, it was not possible to include this imaging of the neoglottis and vocal tract in this study and hence we have to leave this technique for the next study project.

1.5.4 Stroboscopy, digital high-speed imaging and videokymography

In laryngeal voicing, *stroboscopy* is often used to visualize vocal fold vibration. With stroboscopy a virtual slow-motion picture of the rapid vibration is acquired by subsequent light flashes that are slightly out of phase with the fundamental frequency of the vocal fold vibration. However, as the neoglottis works differently from the vocal folds and often does not show a stable fundamental frequency, stroboscopy is not a very useful instrument to study vibrations in the neoglottis. Other instruments are available that are frequency-independent, such as *digital high-speed imaging* (Wittenberg et al., 1995) and *videokymography* (Švec & Schutte, 1996). These methods have been used to study pathological laryngeal voices and have also proven to be suitable to visualize vibrations of the neoglottis (Van As et al., 1999). The latter authors were the first to use digital high-speed imaging to investigate whether this instrument would be useful in TE speech research. They showed that useful images, giving realistic visual information about the vibration of the neoglottis, can be collected.

Lundström & Hammarberg (2004) later used digital high-speed imaging to study the voiced-voiceless distinction in an esophageal and a tracheoesophageal speaker and they conclude that high-speed recordings used together with voice signals can answer important questions about the functions of the pharyngo-esophageal segment. In this thesis, a start was made with investigating the role of the neoglottis in the production of voiced and voiceless plosives. Initial trials were promising (Jongmans et al., 2006), but due to technical issues with the equipment, it was not possible to continue this line of research given the time limits of this project.

1.6 The therapy program

Several studies have found that TE speech intelligibility is overall much lower than NL speech intelligibility. It is expected that these results will be confirmed for Dutch TE speech. If this is indeed the case, these results indicate the necessity for a speech training program focusing specifically on this group. However, as already mentioned above (section 1.4), an extensive literature study showed that no therapy program exists yet for Dutch or other languages. Therefore, in this study, we developed our own comprehensive therapy program. The contents of this program were based on the problems found in the perception experiment and on some additional analyses of spontaneous speech. The techniques used were based, where possible, on techniques described in evidence-based studies. With the program, we hope to achieve improvements at the phoneme level, for both consonants and vowels and, very importantly, an improvement of the intelligibility of spontaneous speech. Also, the self-satisfaction of the participants is something we hope to improve. In order to measure if our goals have been met, several pre and post tests were performed. By using the same tests before and after the therapy, differences between the two moments in time can be calculated. Each test that was used measured a different aspect of speech and each test was rated by different raters. The different tests with the different raters were chosen to ensure that all levels of speech were evaluated by the raters most suitable for that aspect. Speijer (2003) wrote that “the only way to gain insight into the complete effects of voice therapy is to design a study in which changes of different aspects on the voice are examined by a diversity of evaluation instruments”. The same will hold for speech intelligibility. The above mentioned imaging techniques (1.5.4) would also provide very valuable and objective information on changes after therapy. However, for the same reasons as those mentioned for the assessment study, no imaging studies could be performed.

The therapy program and its evaluation are the main topics of this thesis and will be discussed extensively in chapters 4, 5, 6 and 7.

1.7 Outline of this thesis

As was already stated in the introduction, this study wishes to answer the following questions:

1. What are the typical intelligibility problems TE speakers have?
2. Can TE speech intelligibility be improved with a structured therapy program?

3. Do the structure of the program and the pre and post tests used prove to be suitable for the purpose?

In *chapter 2* the perceptual assessment of TE speech is discussed. Results are given per type of speech material and a comparison with literature is made. As the voiced-voiceless distinction causes most problems, this topic is discussed in more detail. Correlations between the different tests are calculated in order to investigate whether some tests measure the same aspects on intelligibility. Based on the results, implications for the therapy evaluation study are described.

In *chapter 3*, the acoustic analyses of the voiced-voiceless distinction are discussed at great length. This is the only study in this thesis with a control group of NL speakers. Main questions in this chapter are what correlates TE speakers and NL speakers use to convey a ***correct*** voiced-voiceless contrast and whether the two groups of speakers differ in the kind of correlates used to make the distinction. If it would appear from the analyses that successful TE speakers mainly use compensatory strategies, this would mean that in the therapy program the participants could be trained to strengthen these compensatory techniques.

In *chapter 4* the development of the training program is discussed. The results of the systematic literature search on evidence-based studies are given and the decisions for contents and structure of the program are explained.

In *chapter 5* the first overall and individual results of the pre and post tests used to evaluate the therapy effects are presented. The tests discussed in this chapter are the phoneme identification and sentence intelligibility tasks for phonetically trained, but naïve listeners where TE speech is concerned, and for speech language pathologists.

In *chapter 6* the overall and individual results of the other pre and post tests are given: the study-specific, structured questionnaires for speakers and their relatives, the Voice Handicap Index for the speakers, and the semantic scaling experiment for inexperienced, naïve listeners.

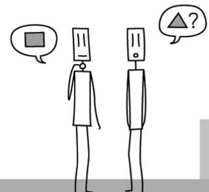
In *chapter 7* the therapy program as a whole is evaluated. Chapters 5 and 6 are compared and the suitability of the pre and post tests is discussed. Based on the results and the opinions of the researcher and the participating speech language pathologists, recommendations are made to improve the therapy program.

In *chapter 8* the thesis is summarized and evaluated. General conclusions and directions for future research will also be given.

The thesis will end with a summary in English and Dutch.

2

Perceptual assessment of tracheoesophageal speech, in terms of phoneme and sentence intelligibility, and spontaneous speech judgments



Abstract

Due to the changes in the anatomy and physiology of the voice source, problems with tracheoesophageal (TE) speech intelligibility are to be expected. Our goal is to find out what kind of speech problems TE speakers have on almost all speech levels. This knowledge is necessary for the development of a therapy program aiming to improve TE speech intelligibility. A wide range of speech material was chosen, testing all Dutch consonants and vowels, consonant clusters, sentences and spontaneous discourse. A listening experiment was performed with naïve, untrained listeners. Results show that, overall, TE speakers perform worse than normal laryngeal speakers. This suggests that a training program should be developed. Even though individual variations are large, some common phoneme and feature confusions are found. The voiced-voiceless distinction proves to be difficult; nasals, fricatives and approximants also cause many problems. Correlations between the different subtests of the listening experiment are mostly quite low, which indicates that the speakers' problems are limited to only certain categories that can differ per speaker. The most common confusions will be included in the training program.

Parts of this chapter have been published in:

Jongmans, P., Hilgers, F.J.M., Pols, L.C.W. & van As-Brooks, C.J. (2005). The intelligibility of tracheoesophageal speech: first results. *Proceedings Interspeech 2005,* 1749-1752.

Jongmans, P., Hilgers, F.J.M., Pols, L.C.W. & Van As-Brooks, C.J. (2006). The intelligibility of tracheoesophageal speech, with an emphasis on the voiced-voiceless distinction. *Logopedics. Phoniatrics. Vocology, 31*, 172-181.

2.1 Introduction

Total laryngectomy has far-reaching effects on vocal tract anatomy and physiology, as has been discussed in chapter 1. The most noticeable consequence of these effects is the loss of natural voice. Even though the voice quality of tracheoesophageal (TE) speakers is considered better than the voice quality of esophageal and electrolarynx speakers (Williams & Watsons, 1987; Pindzola & Cain, 1988; Nieboer et al., 1988; Debruyne et al., 1994; Bertino et al., 1996; Max et al., 1996;), it is still deviant from that of normal laryngeal speech (Robbins, 1984; Robbins et al., 1984a; Robbins et al., 1984b; Van As et al., 1998). At present, voice rehabilitation after total laryngectomy focuses mainly on the prosthesis and troubleshooting, on stoma occlusion and on first sound production and optimizing voice quality. However, changes in the anatomy of the voice source and the vocal tract also compromise speech intelligibility. Research has shown that TE speakers consistently score lower on intelligibility tests than normal laryngeal speakers (Doyle et al., 1988; Hammarberg et al., 1990; Miralles & Cervera, 1995; Searl et al., 2001), and some (small-scale) studies on Dutch TE speech confirm these results (Roeleven & Polak, 1999; Oubrie, 1999; Boon-Kamma, 2001). These results seem to indicate the necessity of including intelligibility in the rehabilitation program. Before therapeutic therapy can be started, however, it is necessary to study what typical intelligibility problems TE speakers, and Dutch TE speakers in particular, have. As languages differ in their phoneme inventories, prosody and other aspects such as the grammatical structure, it cannot be assumed that problems found in one language will also occur in the other. It is therefore important to conduct research for all languages separately. However, there are certain expectations as to what kind of problems occur in Dutch based on literature and the changed anatomy and physiology of voice source and vocal tract. Problems with the voiced-voiceless distinction are expected as the new voice source does not have the intrinsic musculature of the larynx. This problem has been reported as the most prominent one in several languages (e.g. Hammarberg et al., 1990; Nord et al., 1992 – both for Swedish; Saito et al., 2000 for Japanese; Searl et al., 2001 for English) Problems with glottal and velar sounds are also expected as these phonemes are produced at places of articulation where most anatomical changes have occurred (Miralles & Cervera, 1995). Due to the changes in the vocal tract, vowels are also expected to sound different. Several studies have found that TE speakers show deviant formant values (for example Oubrie, 1999; Cervera et al., 2001). If indeed these and other problems are found for Dutch TE speech, speech training will need to focus on these aspects as well.

In order to find out what kind of difficulties TE speakers have, a perception experiment is still considered the gold standard, as communication is mostly a perceptual matter.

Even though it is a subjective method of research, it shows better how intelligibility is perceived than objective measurements of which the perceptual relevance is not known, such as articulatory measurements and certain acoustic analyses, and methods based on automatic speech recognition technologies.

There is much discussion in literature about the type of raters to choose for a perception task. It has been found that the reliability of raters' perception is influenced by the type of raters, varying from naïve to professional. Experienced listeners, for example, tend to display higher intelligibility scores than naïve listeners (Bassich & Ludlow, 1986; Kreiman et al., 1992). It also matters whether the raters are trained for a particular voice pathology and for particular speech material. It has been found that agreement between listeners can be improved by training (Bassich & Ludlow, 1986; Kreiman et al., 1993). As far as the difference between trained and naïve listeners is concerned, it was also found that most naïve listeners use similar perceptual strategies and that overall the inter-rater reliability (Cronbach's) of naïve listeners is sufficiently high, whereas speech language pathologists differ in parameters they consider important (Kreiman et al., 1992). Several studies state that naïve listeners are to be preferred when the research goal is the communicative function of (TE) speech (e.g. Van As et al., 1998). It can also be concluded that naïve listeners are less biased than experienced listeners and that the intelligibility data of naïve listeners are of greater utility when establishing clinical goals (Doyle et al., 1989). It is for this reason that naïve listeners were chosen in our perception experiment.

Perception tasks can vary according to the type of information wanted, such as voice quality or phoneme, word or sentence intelligibility. Many voice quality studies have made use of semantic scales to investigate how the quality of the voice is perceived, focusing on specific characteristics of the voice. Using rating scales increases the difficulty in interpreting the results of a perception task as the terms used on either side of the scales do not mean the same for all listeners. Even though scale judgments can provide valuable information, they do not provide a detailed view on the specific difficulties TE speakers have. Scales were only used to test the general impression of TE speech intelligibility in this investigation.

Another method used is the identification test. In such a test the listener is asked to identify a phoneme, a whole word or a whole sentence. This test allows for a more detailed analysis of the problems. In TE speech intelligibility studies identification experiments are often used (Doyle et al., 1988; Hammarberg et al., 1990; Miralles & Cervera, 1995; Searl et al., 2001). For Dutch, two studies (MA papers) investigated consonant intelligibility (Roeleven & Polak, 1999; Boon-Kamma, 2001) Another Dutch study (MA paper) looked at TE vowels, mainly focusing on formant values, but also describing perceptual vowel confusions (Oubrie, 1999).

Although the studies mentioned above provide valuable information that was used as a starting point for this study, they are all limited in scope as, mostly, only one kind of identification test is used and only a limited number of phonemes or words were included in the study. The aim of the present study is to create as complete a picture of TE speech intelligibility as possible by including phonemes (both consonants and vowels) and spontaneous speech and by using the entire Dutch phoneme inventory. Therefore, all Dutch consonants in initial, medial and final position, consonant clusters in initial, medial and final position, and all Dutch vowels, as well as sentence intelligibility were tested. Free monologues were also recorded for judging overall intelligibility in spontaneous discourse. The results of this perception experiment and the implications for the therapy evaluation study will be discussed in this chapter. The therapy program based on the results of the perception experiment is discussed in chapter 4. In chapter 3 the acoustic aspects of the voiced-voiceless distinction in plosives and fricatives will be discussed in more detail first as it was found that voicing errors were the most common errors in the perception experiment.

2.2 Subjects and methods

2.2.1 Subjects

Subjects were 11 laryngectomized patients using TE speech by means of an indwelling (Provox®) voice prosthesis (Hilgers et al., 1997). All of them had undergone a standard total laryngectomy. Five speakers used hands-free speech (Hilgers et al., 2003), the other six used digital occlusion of the stoma on top of a Heat and Moisture Exchanger (Hilgers et al., 1996). All subjects were male, with a mean age of 66;9 years (age range 44 to 78 years). Post-operation times ranged from 2;2 to 17;5 years (mean time 9;4 years). Ten patients had received irradiation. Subjects were obtained via the records of the Netherlands Cancer Institute. The study was approved by the Protocol Review Board of the Netherlands Cancer Institute and all patients gave informed consent.

2.2.2 Recordings

Recordings were made in a sound-treated room with a Marantz CDR 770 audio CD recorder. A Sennheiser MD421 microphone was placed at a microphone-to-mouth distance of 30 cm. Before the actual recording was made, the sound level was optimally adjusted for that particular subject and then fixed. Also a calibration signal was recorded onto CD.

2.2.3 Speech material

The speech material consisted of syllables (testing both consonants and vowels, as well as consonant clusters), nonsense sentences and spontaneous speech (see appendix 2.1 at the end of this chapter for a specification of the speech material used).

2.2.3.1 Syllables

Syllables testing consonants had the following structures: C(onsonant)V(owel) (n=18x3), V_iCV_i (n=21x3), and VC (n=11x3). The syllables consisted of all the possible native consonants in Dutch meaning that loan phonemes were excluded. At a later stage /ŋ/ was omitted from further study as most subjects had pronounced this sound as it is spelled in Dutch (ng). The consonants were combined with the vowels /i/, /u/ and /aː/ as these are the corner vowels and hence maximally distinguishable. It also meant there were three realizations of each consonant.

Consonant clusters were also tested in initial, medial and final position and were also combined with /i/, /u/ and /aː/. Per position, 5 clusters x 3 vowels were included. The clusters for initial and medial position were the same (/pr sl tr kr fl/) whereas the clusters in final position were different (/lp lf mst mt rk/). The choice for the included clusters was based on how many difficulties TE speakers were expected to have with the phonemes in the cluster, especially with switching between voiced and voiceless phonemes within the cluster. Only a limited set of clusters was chosen as the amount of speech material was already quite large. It was also not considered to be the main focus of the study.

The vowels were tested in h-V-t syllables, using all 15 Dutch vowels (12 monophthongs and 3 diphthongs). /h/ was chosen as initial phoneme as it least interferes with a following vowel and the plosive /t/ in final position was chosen to create a clear ending of the vowel.

The syllabic structures were chosen so as not to provide phonological or semantic cues. Nonsense syllables also avoid learning effects.

2.2.3.2 Semantically Unpredictable Sentences

Semantically Unpredictable Sentences (SUS) were included to test sentence intelligibility (Benoit, 1990; Benoit et al., 1996). These sentences are grammatically correct, but semantically anomalous. This way, listeners cannot rely on the (semantic) context when identifying the sentences. An example of such a sentence is “Het boek loopt door de smalle boom” (the book walks through the narrow tree). Even though these sentences were developed originally to test intelligibility of speech synthesis systems, the fact that the listener cannot rely on context makes them very suitable for the purposes of this study as well. The sentences are generated randomly using a fixed vocabulary of high frequency words per word position. The fact that no sentence will occur more than once reduces the risk of learning effects. To reduce the use of cues even further, five different simple syntactic structures are generated by the program. No sentence exceeds eight words to avoid short term memory problems. We generated sets of five sentences representing all five syntactic structures for each subject. Five practice sentences were included at the beginning of the experiment. These were the same as the last five sentences of the experiment. The practice items were not used for further analysis. For a complete list of all sentences used, see appendix 2.1.

2.2.3.3 Spontaneous speech

The spontaneous speech samples consisted of 2 to 3 minutes of free monologues by the speakers.

2.2.4 Listening Task

The listeners consisted of 10 naïve listeners, with no prior experience of TE speech. Mean age was 43 years (age range 27-59). None of the listeners had any known hearing problems. The listening task was performed online, so that listeners could be flexible in where and when they could participate in the experiment (the online experiment was developed by R.J.J.H. van Son, Institute of Phonetic Sciences, Universiteit van Amsterdam). All listeners received the same headphones (Pro Luxe, Pro10).

All the speech material described above was included in the listening experiment. Per type of speech material, a separate experiment was set up, resulting in 7 sub experiments (1. CV, 2. V_iCV_i, 3. VC, 4. h-V-t, 5. clusters, 6. SUS, 7. spontaneous speech).

The listeners did not receive training prior to the experiment. However, at the start of each listening experiment, an explanation of the test was given with instructions on how to perform the tests. This explanation also contained a sound file with samples from all speakers, so the listeners could hear the whole range of speakers before they started. This had two advantages: firstly they could get used to the sound of TE speech and secondly it allowed them to compare the speakers better. The stimuli were randomized for each listener so that no listener would listen to the same order of speakers. The material was blocked per speaker. The last five items of the experiment were also included as practice items at the beginning of the experiment. These items were not used for further analysis.

2.2.4.1 Phoneme and sentence level

The listeners listened to the phoneme samples and for each syllable (or word) they typed what they perceived in normal spelling, which is unambiguous in Dutch. Software was available that automatically generated files from these responses, these files then were used for further data analysis. The SUS sentences were also typed by the listener in normal spelling. Initial studies by Benoit (1990) and Benoit et al. (1996) showed that word correct scores correlate highly with sentence correct scores. In our experiment, we chose to use the percentage of sentence correct score.

2.2.4.2 Spontaneous speech

Where spontaneous speech is concerned, having listeners transcribe literally what they perceive is not an option as it is then very hard to check whether the listener's perception matches what was actually said by the TE speakers. In this case semantic scales are a realistic option as the outcome gives a general judgment on several aspects of TE speech intelligibility without having to transcribe speech.

The use of scales was first described by Osgood et al. (1957) for English. In this thesis, the scales were based on particular semantic scaling studies that have been performed for Dutch, because the scales of these studies have been used earlier for the judgment of alaryngeal speech. Fagel et al. (1983) came up with a list of 14 scales to rate voice quality. Nieboer et al. (1988) later used this list (except for two scales) for investigating the difference in voice quality between esophageal and tracheoesophageal speakers. Verdonck-de Leeuw (1998) and Van As et al. (2003) based their set of scales on Fagel et al. (1983) and Nieboer et al. (1988), but added their own scales and came up with their own subset for voice quality. All these studies looked at voice quality rather than speech intelligibility. For that reason, in the present study only some of the existing scales were used and some new ones were added. Both intelligibility scales and voice quality scales were included as voice quality was expected to influence speech intelligibility. Listeners were instructed on the screen what the difference between the sets of scales was. The Dutch scales used and their English translations are listed in Table 2.1. It can be seen that one scale (normal-deviant) occurs both under the intelligibility scales and the voice quality scales. It was hoped that the speakers would be able to distinguish between normal or deviant articulation versus normal or deviant voice quality. All scales were seven-point scales except one scale on the overall impression. This scale consists of three options: good-moderate-poor, where good is defined as most similar to normal speech, poor as least similar and moderate is used for speakers between those extremes (van As et al., 2003). It should be noted that the English translations can have different connotations than the original Dutch terms.

Table 2.1: The semantic scales in Dutch as used in the rating experiment and their English translations.

Dutch term	English term
Intelligibility scales	
1. makkelijk verstaanbaar-moeilijk verstaanbaar	easy to understand-difficult to understand
2. snel-langzaam	fast-slow
3. precies-slordig	precise-sloppy
4. normaal-afwijkend	normal-deviant
5. natuurlijk-niet natuurlijk	natural-unnatural
6. duidelijk-binnensmonds	clear speech-mumbled speech
7. moeiteloos te verstaan-te verstaan met veel inspanning	takes no effort to understand-takes effort to understand
Voice quality scales	
8. normaal-afwijkend	normal-deviant
9. mooi-lelijk	beautiful-ugly
10. laag-hoog	low-high
11. diep-schril	deep-shrill
Overall judgment	
12. goed-matig-slecht	good-moderate-poor

2.2.5 Statistical analyses

Inter-rater reliability measurements were performed for all perception experiments using Cronbach's alpha. This test tests the homogeneity of the listeners and thus how reliable responses are. A high value (>.070) indicates a good reliability, a low value indicates poor reliability.

Chi-squared tests were used to calculate differences between hands-free speakers and speakers that close off the stoma digitally, and also between the different consonant categories, and the different vowel categories. It was also used to investigate the confusions between voiced and voiceless plosives and fricatives. Chi-squared tests establish the association between two (or more) qualitative variables. When an association is established, however, it does not say how strong the association is. Measures for strength of association do exist and in this research Cramer's V (Φ) is preferred: with more complex tables, Cramer's V measure can still, as in a 2x2 case, achieve its maximum value of unity (Kinnear & Gray, 1999). Bonferroni corrections were also used where multiple testing was involved. Descriptive results are expressed in percentages correct or incorrect.

The results of the semantic 7-point scales rated for spontaneous speech were only included for further analysis when Cronbach's alpha was greater than 0.70. A Principal Component analysis (PC) was then performed on the correlation matrix of those semantic scales. A PC analysis investigates the relations between different rating scales and can be used as a multivariate technique in order to investigate to what extent the scales measure the same underlying factors or variables (Rietveld & Van Hout, 1993). Only components with an eigenvalue>1 were taken into account. After the PC analysis a VARIMAX rotation was performed, which maximizes the variance of the component loadings per factor. Factor loadings above .50 are considered relevant.

An ANOVA was used to investigate the relation between the overall judgment of intelligibility (good, moderate, poor) and the semantic scales (with a confidence level of .05). A post hoc Tukey test was performed to gain insight into the differences between the separate subgroups. Lastly, a discriminant analysis was used to see if group membership could be predicted from the different variables. A step-wise method was used, with Wilk's Lambda. With a step-wise discriminant analysis, the effect of the addition or removal of an independent variable is monitored by a statistical test and the result is used as a basis for the inclusion of the variable in the final analysis. Wilk's Lambda (Λ) is the statistic used for weighing up the addition or removal of variables from the analysis. The significance of the change in Λ is obtained from an F test. At each step of adding a variable to the analysis, the variable with the largest F is included and variables that do not contribute anymore to maximizing the assignment of cases to the correct groups are excluded. Once the process is completed, the variables remaining in the analysis are the ones used in the discriminant analysis (Kinnear & Gray, 1999).

Correlations were calculated between the phoneme and sentence identification tasks and the semantic scales. The idea behind this was to investigate if scores on one or more of the phoneme and sentence identification tasks can predict scores of spontaneous speech intelligibility. A Spearman rank correlation was used. Simply said, this test investigates whether the rank order of cases in one test is similar to the rank order in another test, or in this study, whether speakers that score poorly in one test, will also score poorly in the other.

2.2.6 Praat

The program Praat (Boersma & Weenink, 1996) was used to adapt the audio material to make it suitable for use in the listening experiment.

As the recordings made by the Marantz CD recorder contained noise, a noise reducer script was developed to improve the general quality of the recordings.

As manual segmentation of the speech signals would be too time consuming, an automatic word extraction script was composed and used to extract the words from the recordings and save them as wav files. (All praat scripts were developed by A.G. Wempe, Institute of Phonetic Sciences, University of Amsterdam).

2.3 Results

In the following section, the results of the listening experiment will be discussed per category, starting with the syllables (consonants and consonant clusters first, followed by vowels) and then the SUS sentences and finally spontaneous speech.

2.3.1 Syllables - Consonants

2.3.1.1 Overall results

The inter-listener reliability for consonants measured with Cronbach's alpha was .883 for initial position, .858 for medial position, and .861 for final position. Since these values are sufficiently high, averages across the listeners can be used for further analysis. As two different speech modes were used during the experiment, the difference between hands-free and digital occlusions was calculated for all three positions separately and for all positions together. The results are shown in Table 2.2. Total numbers are based on the number of speakers x number of stimuli x number of vowels x number of listeners. For example, with speech mode digital, the total consists of 6 speakers x 18 consonants x 3 vowels x 10 listeners = 3240.

Table 2.2: Numbers and percentages of correctly and incorrectly perceived consonants for digital and hands-free speakers in initial, medial and final positions as well as all positions together (overall).

Speech mode	Position	Correct (%)	Incorrect (%)	Total
Digital (N=6)	*Initial*	2070 (64)	1170 (36)	3240
	Medial	2541 (71)	1059 (29)	3600
	Final	1478 (82)	323 (18)	1801
	Overall	6089 (71)	2552 (30)	8641
Hands-free (N=5)	*Initial*	1870 (69)	830 (31)	2700
	Medial	2285 (76)	715 (24)	3000
	Final	1221 (82)	278 (19)	1499
	Overall	5376 (75)	1823 (25)	7199

For initial and medial position a difference was found between digital and hands-free speakers, with a better intelligibility for hands-free speakers (χ^2=19.019, df=1, p<.01; Φ=.057, p<.01 and χ^2=25.955, df=1, p<.01; Φ=.063). No difference was found in final position. Overall, the difference between the groups is maintained, with hands-free speakers being perceived correctly more often than people who digitally close off their stoma (χ^2=34.83, df=1, p<.01; Φ=.047, p<.01).

In the following table, the individual scores per speaker and per position are given.

Table 2.3: Average scores in percentages per speaker for all 3 consonant positions as well as overall scores, with the first row showing speaker number. Underlined scores indicate the speaker with the highest score in the different positions; bold and italicized scores indicate the speaker with the lowest score in the different positions.

	1	2	3	4	5	6	7	8	9	10	11	Overall
Initial	83	70	66	71	69	66	***52***	60	55	71	69	66
Medial	87	80	75	78	65	66	***63***	77	68	69	77	73
Final	91	91	87	88	78	76	77	80	79	82	***71***	82
Overall	86	78	74	77	69	68	***62***	71	65	72	73	72

Table 2.3 shows that the range of scores averaged over the three positions together is 62%-86%, indicating large differences in performance between individual speakers. The same can be observed for each consonant position. Speaker 7 shows the lowest score for initial and medial position and also for the overall scores. Speaker 11 shows the lowest score in final position. Speaker 1 shows the highest overall score and the highest score for each position separately. For the other speakers, no real consistency can be found between their mean score and the scores per position.

The consonants were divided into four manner-of-articulation categories: plosives, fricatives, nasals and approximants, as we expect different problems for the different categories. In Table 2.4 the mean scores for all manners of articulation per consonant position are given.

Table 2.4: Mean scores in percentages for all manners of articulation per consonant position, plus overall scores. The range and the phonemes included per category are given between brackets.

Manner	Initial (%)	Medial (%)	Final (%)	Overall (%)
Plosive	76 (57-93) (/b p d t k/)	79 (61-91) (/b p d t k tj/)	90 (73-99) (/p t k/)	80
Fricative	52 (34-83) (/v f z s h x ʃ/)	56 (42-89) (/v f z s h x ʃ/)	88 (67-100) (/f s x/)	60
Nasal	77 (52-92) (/m n/)	86 (70-98) (/m n nj/)	62 (23-80) (/m n/)	77
Approximant	75 (30-96) (/ʋ j l r/)	85 (70-99) (/ʋ j l r/)	79 (53-97) (/l r/)	80
Overall score	66	73	82	72

Table 2.4 shows an overall score of 72% (all positions together) which is markedly lower than the percentage correct score of 94% in normal laryngeal speakers as found by Pols

(1983) for CVCVC words (also all positions together). Other studies that investigated American-English TE speech actually found similar low intelligibility scores (Doyle et al., 1989; Doyle, 1988; Searl et al., 2001).

When the mean scores per position are considered, the initial position scores are significantly lower than the scores in other positions (χ^2=256.673, df=2, p<.01; Φ=.127, p<.01) whereas final position scores are highest. This division was also found by others (Doyle et al., 1988; Hammarberg et al., 1990; Roeleven & Polak, 1999). The reason that final position shows the highest score is related to the fact that there are fewer consonants possible in Dutch in final position than in initial or medial position and hence fewer errors are possible. In addition, Dutch final position cannot contain voiced plosives and fricatives, or the phoneme /h/. This means that the most common source for errors is not present in final position. Initial position may also be more difficult due to the fact that the consonant starts abruptly. As a TE speaker needs to actively close off his stoma before speaking, it is likely that TE speakers need some time to start up phonation and as a result the initial consonant may not sound as clear.

Significant differences were also found between manners of articulation for all consonant positions. In initial and medial position, the fricatives score lowest (χ^2=375.399, df=3, p<.01; Φ=.251, p<.01 and χ^2=521.659, df=3, p<.01; Φ=.295, p<.01 respectively). In final position, the approximants score lower than plosives and fricatives and the nasals score even lower than approximants, plosives and fricatives (χ^2= 254.088, df=3, p<.01; Φ=.277, p<.01). In American-English (Doyle et al., 1988) and Spanish (Miralles & Cervera, 1995) problems with fricatives were also found, just as problems with plosives. For Swedish (Hammarberg et al., 1990), however, no problems with fricatives were found. The problems with nasals in final position confirm earlier results for Dutch (Boon-Kamma, 2001).

The following bar graphs show the mean scores per speaker for manner of articulation in initial (Figure 2.1a-d), medial (Figure 2.2a-d) and final position (Figure 2.3a-d).

As can be seen in these bar graphs, overall, speakers show similar scores around the mean. With the fricatives in initial position, the scores for speaker 1 are much higher than the rest. However, without this speaker, the score would be even lower, which confirms that this category is very difficult. With the approximants in initial position, speaker 7 shows a markedly lower score. Again, leaving this patient out does not influence the overall result. For medial position, speaker 1 shows the same high score for fricatives. No other outliers were found. For final position, only one outlier was found: speaker 11 shows a markedly lower score for the nasals, which, however, does not change the tendency of the score. The results show that the few outliers that exist do not greatly influence the overall results.

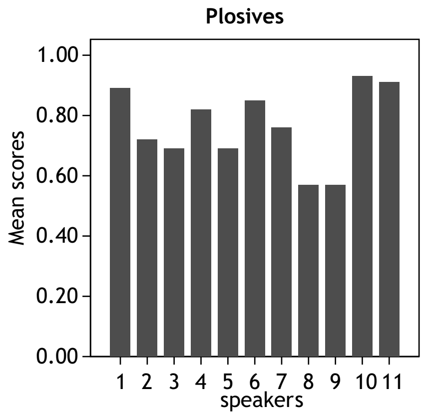


Figure 2.1a: Mean scores per speaker for initial /b p d t k/

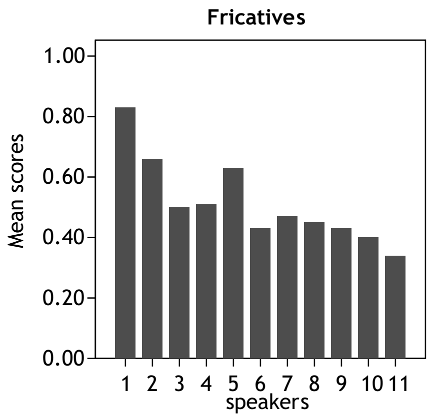


Figure 2.1b: Mean scores per speaker for initial /v f z s h x ʃ/

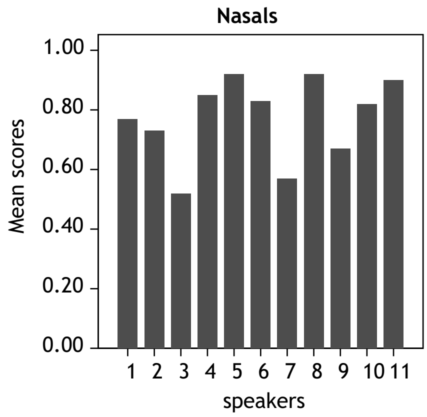


Figure 2.1c: Mean scores per speaker for initial /m n/

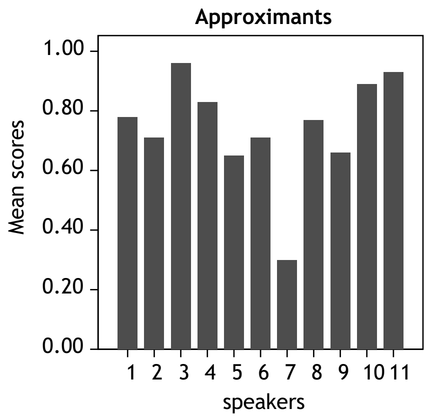


Figure 2.1d: Mean scores per speaker for initial /ʋ j l r/

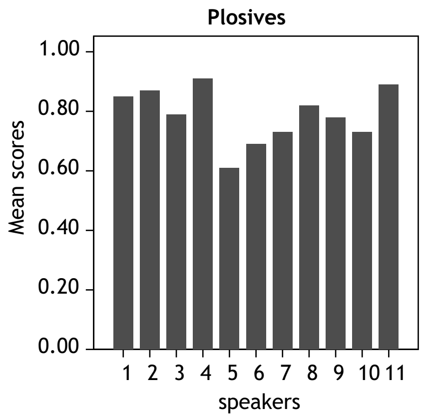


Figure 2.2a: Mean scores per speaker for medial /b p d t k tj/

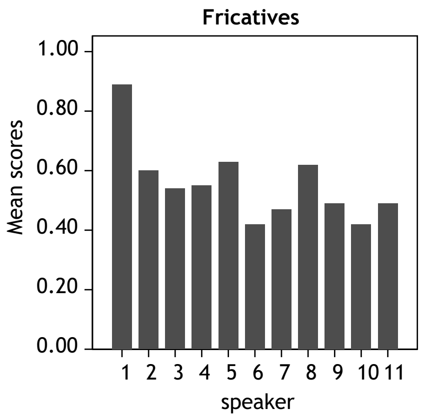


Figure 2.2b: Mean scores per speaker for medial /v f z s h x ʃ/

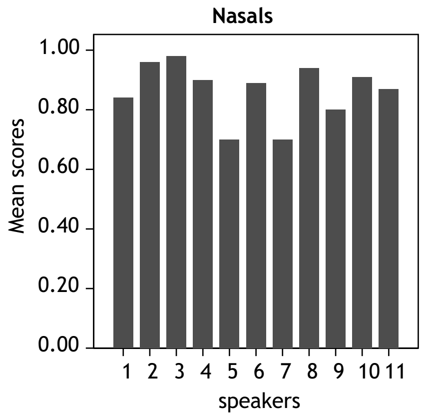


Figure 2.2c: Mean scores per speaker for medial /m n nj/

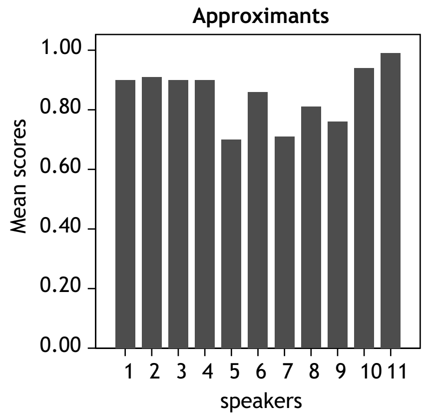


Figure 2.2d: Mean scores per speaker for medial (/ʋ j l r/)

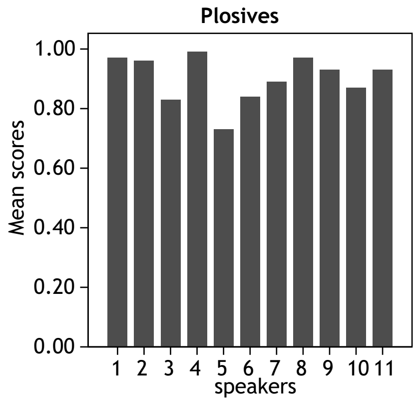


Figure 2.3a: Mean scores per speaker for final /p t k/

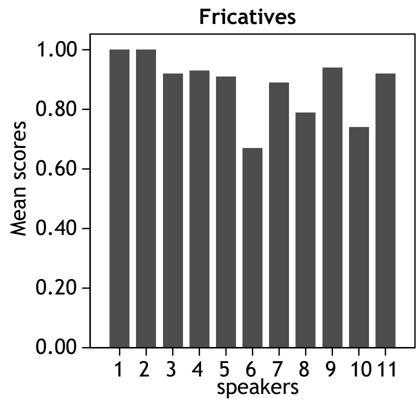


Figure 2.3b: Mean scores per speaker for final /f s x/

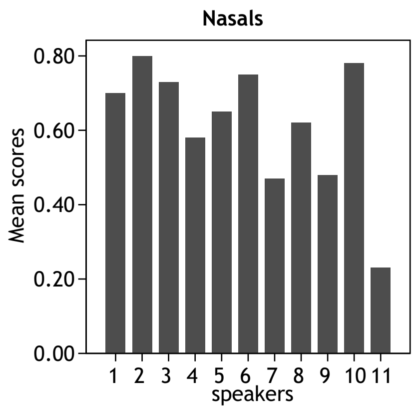


Figure 2.3c: Mean scores per speaker for final /m n/

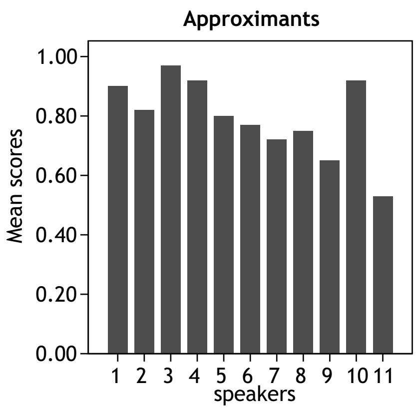


Figure 2.3d: Mean scores per speaker for final /l r/

In the next section, confusions per manner of articulation will be discussed for each position in more detail.

2.3.1.2 Phoneme confusions per position

With regards to the therapy program to be developed in later chapters, this study is particularly concerned with the kind of confusions made. Tables 2.5, 2.8 and 2.11 show the phoneme confusions for initial, medial and final position, respectively. The phonemes are grouped in blocks per manner of articulation and will be discussed accordingly.

Initial position

Table 2.5: Phoneme confusions for **initial** position. Numbers in percentages multiplied by 10. 'Missing category' consists of scores deleted or uninterpretable items, 'other' category consists of all other responses, like clusters, that we are not concerned with now. Note that the actual total number of observations per phoneme is 330 (3 realizations x 11 speakers x 10 listeners).

	p	b	t	d	k	x	h	f	v	s	z	ʃ	j	l	ʋ	r	m	n	ʒ	tj	ts	tʃ	missing	other
p	**700**	230	3		9					12					9								6	31
b	39	**782**	3	21	3				3	3					27		59	15					18	27
t	24	12	**673**	225	9					6			9										12	30
d		15	37	**879**		3			9	3				3			9	12					24	6
k	49	43	39	18	**767**	21				6					6								33	18
x		6			72	**624**	27	18	18		3		3		27	55							129	18
h	3	18	3		6		**377**	46	48	6		12	42	15	45	30	6	3	3		3	6	325	3
f							3	**558**	339	3			3		52								18	24
v						3		379	**482**	12	6				70		3						33	12
s						3			3	**561**	312	12	3						3	3	70	9	15	6
z			6			3		9	9	243	**609**	21	3							3	55	9	24	6
ʃ			3		3	12	6			167	36	**389**	55		6				139	15	12	85	60	12
j							27	3		27	3	6	**795**	70	3	3		18					27	18
l	6	18				3		33	12	3				**716**	27	9	24	68					57	24
ʋ	3	24	3	3				18	82	15	6		30		**689**	3	24	9					55	36
r	15	9	9	6		3	9	18	18					15	6	**783**		3					54	52
m		67		3				9						9	24		**785**	76					24	3
n		3		12		9		9	3	3	3			48	6		124	**759**					15	6

In describing the results the focus is only on the confusions observed. The percentages given below relate to the total number of confusions for that phoneme (or group of phonemes) excluding the numbers for correctly perceived phonemes. It should be kept in mind that there are great differences in the number of confusions across all phonemes, which is important when comparing percentages.

Apart from looking at the individual phoneme confusions in the matrices, it is also useful to look at confusions between manners of articulation and confusions between places of articulation. Certain places of articulation may be more difficult for TE speakers than others. The following places of articulation are distinguished: labial, alveolar, palatal, velar and glottal. The phoneme /r/ is not placed in any category as this phoneme can have different places of articulation. In Table 2.6 the confusions between the four manners of articulation are given and in Table 2.7 the confusions between places of articulation are given. The percentages are all based on the total number of confusions in that matrix. That number can be calculated from Table 2.4 and the confusion matrices. Table 2.4 shows a mean consonant correct score of 76% (76.3) for plosives in initial position. This means that 24% (23.7) of the plosives in initial position is incorrectly perceived. There are five plosives in initial position x 330 realizations = 1650. 24% (23.7) of 1650 = 392.

Table 2.6: Manner confusions for initial consonants (in percentages), with the 'missing' category consisting of deletions and unintelligible responses. N gives the actual total number of confusions. Highlighted cells indicate that manner was perceived correctly.

	Plosive	Fricative	Nasal	Approximant	Missing	N
Plosive	69	10	9	5	8	392
Fricative	12	58	0	12	18	1120
Nasal	19	11	44	19	9	151
Approximant	13	31	16	21	19	337

Table 2.7: Place confusions for initial consonants (in percentages), with the 'missing' category consisting of deleted or unintelligible responses. /r/ is mentioned separately as it can have different places of articulation. Highlighted cells indicate that place was perceived correctly.

		Labial	Alveolar	Palatal	Velar	Glottal	R	Missing	N
Plosive	*Labial*	78	15	0	2	0	0	5	166
	Alveolar	25	3	2	3	0	0	8	147
	Velar	45	30	0	12	0	1	13	78
Fricative	*Labial*	91	2	0	1	0	0	6	315
	Alveolar	6	84	5	1	0	0	5	274
	Palatal	1	50	34	4	1	0	10	201
	Velar	22	1	1	23	7	14	34	124
	Glottal	28	5	9	1	0	4	52	206
Nasal	*Labial*	9	39	0	0	0	0	11	71
	Alveolar	61	29	0	4	0	0	6	80
Approximant	*Labial*	54	18	10	0	0	1	17	103
	Alveolar	52	19	0	1	0	7	20	94
	Palatal	9	62	3	0	12	2	13	68
	R	51	19	0	1	3	0	25	72

From Table 2.6 it can be seen that for plosives and fricatives, most confusions are with members of their own group (69% and 58%, respectively). This means that manner was mostly perceived correctly for both categories. When place is also taken into account, it is found that within plosives and fricatives place is mainly perceived correctly for the labial and alveolar sounds. From this it may be inferred that for these consonant categories the feature voice is the main cause for confusion, which means that cognates (a pair of phonemes that differs in one feature only, in this case the voicing feature), are often confused. Similar results were found in research by Pols (1983) who looked at the intelligibility of consonant features of normal speakers and found the same clustering of responses for plosives and fricatives. He also found that in the confusions for fricatives especially the feature voice played a role.

/k/ and /x/ are slightly different as they barely have a voiced counterpart in Dutch. /x/ has a high percentage of deletions and /k/ mainly shows confusions with the other plosives. That velar sounds cause problems for TE speakers can be expected as the tongue base is somewhat retracted after a total laryngectomy. This will audibly influence the production of these sounds (Miralles & Cervera, 1995). /h/ also deserves some individual attention as it shows a wide range of confusions (especially within the approximant group) and has the highest percentage of deletions (33%). This is in line with results from other studies on TE speech intelligibility (Roeleven & Polak, 1999; Hammarberg et al., 1990). It seems that the glottal sound /h/ is difficult to produce, which confirms expectations. In normal speech, voiceless /h/ is produced by increasing the tension of the vocal folds so that no vibration can take place and the sound is generally perceived as voiceless friction. However, in TE speech, when the air passes the neoglottis, the chance that this results in vibration is rather high, because control over the tension of the neoglottis is diminished. For the phonemes /s z ʃ/ many confusions with /ts tj tʃ/ are found. Listeners apparently hear an inserted /t/.

The nasals show a different confusion pattern. Less than half of the confusions are with the other nasal phoneme and many confusions are found with plosives (19%) and approximants (19%). Place is also difficult, especially for /n/ which is often confused with /m/. Other common confusions are between /m/ and /b/ and between /n/ and /l/.

For the approximants, the majority of the confusions concerns (at least) manner of articulation: only in 21% of the confusions manner is perceived correctly. As much as 31% of the confusions falls within the fricative group and 19% in the 'missing' group. The most common confusions are that /j/ is perceived as /l/ or as a fricative, /l/ is often perceived as a nasal, especially /n/, /ʋ/ is frequently perceived as a fricative, especially as a /v/ and finally /r/ is often perceived as a plosive or a fricative.

Medial position

In Table 2.3 the mean scores per speaker were given for all positions separately and in Table 2.4 the mean scores per manner of articulation were given for all positions separately. In Table 2.8, the confusion matrix for medial position is shown.

Table 2.8: Phoneme confusions for **medial** position. Numbers in percentages multiplied by 10. ‘Missing’ category consists of scores deleted or incomprehensible items, ‘other’ category consists of all other responses, like clusters, that are not of concern now. Note that the actual total number of observations per phoneme is 330 (3 realizations x 11 speakers x 10 listeners).

	p	b	t	d	k	tj	x	h	f	v	s	z	ʃ	j	l	ʋ	r	m	n	ɳ	ʒ	dj	tʃ	missing	other
p	**870**	112			3				3															6	6
b	70	**761**		12	6				3							67		42	9					6	24
t	30	3	**840**	85	21		3																		18
d		15	24	**846**	9		15			3				15	15				15					28	15
k	39		27		**849**		24		3															49	9
tj			85	6	33	**570**	3						3	58						12		152	27	18	33
x			3		27		**686**	33	9	6	3			3	3	6	155							39	27
h	6	3		3	3		9	**553**		3				76	27	48	94		3					163	9
f	15	3		3					**595**	345						3								18	18
v	12	3	3						449	**470**						15			3					24	21
s									9	3	**628**	207	48	3	3						18			33	48
z							3		6	15	452	**412**	27	9							9			9	58
ʃ						3	12		3		124	30	**525**	64							115	18		21	85
j								27						**859**	15	6	12		27	15				24	15
l														30	**822**	18	43		24					30	33
ʋ		9						3		67				3	9	**828**	30	9						27	15
r								30							24	9	**910**							12	15
m		9		3												6		**879**	79					12	12
n															6	6		42	**916**					12	18
ɳ				3	3	3							3	91				9	18	**795**		6	3	15	51

Just as in initial position, it can be seen in Table 2.8 that the responses in medial position are mainly clustered within the same manner of articulation. Table 2.9 shows the confusions between the different manners of articulation in percentages. Table 2.10 shows the confusions between the different places of articulation.

Table 2.9: Manner confusions for medial consonants in percentages, with the ‘missing’ category consisting of deletions and unintelligible responses. N gives the actual total number of confusions. Highlighted cells indicate that manner was perceived correctly.

	Plosive	Fricative	Nasal	Approximant	Missing	N
Plosive	67	6	6	12	8	419
Fricative	5	68	1	17	10	1024
Nasal	9	7	48	27	10	136
Approximant	4	29	13	38	16	195

Just as in initial position, Table 2.9 shows that manner of articulation is mostly perceived correctly for both the plosives and the fricatives (67% and 68%, respectively). There is, however, slightly more variation in the responses, both consonant categories showing quite some confusions with other categories as well. Table 2.8 shows, for example, that /b/ is often confused with /m/ and /ʋ/. /h/ shows the highest number of missing responses and is confused relatively often with approximants. /x/ is often perceived as /r/.

Table 2.10: Place confusions for medial consonants in percentages, with the 'missing' category consisting of deleted or unintelligible responses. /r/ is mentioned separately as it can have different places of articulation. Highlighted cells indicate that place was perceived correctly.

		Labial	Alveolar	Palatal	Velar	Glottal	R	Missing	N
Plosive	*Labial*	88	6	0	3	1	0	3	122
	Alveolar	17	51	6	15	1	1	9	104
	Palatal	1	62	21	12	0	0	4	143
	Velar	28	20	0	20	0	0	32	50
Fricative	*Labial*	94	1	0	0	0	0	5	309
	Alveolar	5	76	14	0	0	.6	5	310
	Palatal	4	47	41	4	0	0	5	153
	Velar	12	3	1	13	11	49	13	104
	Glottal	16	7	17	3	0	21	37	148
Nasal	*Labial*	20	68	0	3	0	0	10	40
	Alveolar	79	7	0	0	0	0	14	28
	Palatal	10	30	50	3	0	0	7	68
Approximant	*Labial*	51	7	2	4	2	19	16	57
	Alveolar	24	15	17	2	0	25	17	59
	Palatal	13	40	2	2	19	8	17	48
	R	16	29	7	0	32	0	16	31

Where place of articulation is concerned, Table 2.10 shows that wrongly identified labial plosives and fricatives are mostly confused with other labial phonemes (88% and 94%, respectively). For the alveolar plosives and fricatives it can be seen that the aspect alveolar is perceived correctly 51% of the time for plosives and 76% for fricatives. The other place categories of the plosives and fricatives cause more problems. Overall, most of the confusions for plosives and fricatives still concern cognates, just as in initial position.

Nasals show the same confusion pattern in medial position as in initial position. Where manner is concerned, 48% of the confusions is with other nasals whereas 27% of the confusions is with approximants. Place is also often misperceived. Problems with place were also found for Swedish nasals (Hammarberg et al., 1990). /ɳ/ shows a large number of confusions with /j/, which indicates that at least part of the phoneme has been heard correctly.

For the approximants in medial position there is a lower range of confusions than in initial position. Manner still seems difficult to perceive as only 38% of the confusions falls within the approximant category. Many confusions are with fricatives, but this is mainly caused by confusions between /ʋ/ and /v/. Confusions with nasals (13%) are also observed just as many missing values. In the majority of the confusions, place causes problems as well (see Table 2.10).

Final position

In Table 2.11 the confusion matrix for all individual final phonemes is given, followed by Table 2.12 giving the confusions between manners of articulation. In Table 2.13 the confusions between the different places of articulation are given.

Table 2.11: Phoneme confusions for **final** position. Numbers in percentages multiplied by 10. ‘Missing’ category consists of scores deleted or incomprehensible items, ‘other’ category consists of all other responses, like clusters, that are not of concern now. Note that the actual total number of observations per phoneme is 330 (3 realizations x 11 speakers x 10 listeners).

	p	t	k	f	x	s	m	n	l	r	d	v	ʋ	b	mp	ŋ	ʃ	z	missing	other
p	**931**		21	3										33					12	
t	55	**827**	49		3	3		3		3	36								12	9
k	6	30	**946**		6														6	6
f	9	6		**913**						9	3	30							27	3
x	3		27	3	**846**				3	46		3							45	24
s				12		**895**											24	12	21	36
m	45			9			**583**	194		6			3	6	61	21			72	
n	18	9	3	6			143	**656**	9	9	3			6	45	15			66	12
l	9				6		12	3	**671**	30			36						176	57
r	9	3							21	**922**				3					24	18

Table 2.12: Manner confusions for final consonants in percentages, with the ‘missing’ category consisting of deletions and unintelligible responses. N gives the actual total number of confusions. Highlighted cells indicate that manner was perceived correctly.

	Plosive	Fricative	Nasal	Approximant	Missing	N
Plosive	80	5	2	3	10	98
Fricative	15	28	0	30	27	115
Nasal	12	2	64	4	18	252
Approximant	7	2	8	35	49	136

Table 2.13: Place confusions for final consonants in percentages, with the 'missing' category consisting of deleted or unintelligible responses. /r/ is mentioned separately as it can have different places of articulation. Highlighted cells indicate that place was perceived correctly.

		Labial	Alveolar	Palatal	Velar	R	Missing	N
Plosive	*Labial*	52	0	0	30	0	17	23
	Alveolar	32	28	0	30	4	7	57
	Velar	11	56	0	17	6	11	18
Fricative	*Labial*	45	10	0	3	10	31	29
	Alveolar	11	17	37	0	14	20	35
	Velar	6	2	2	20	40	30	50
Nasal	*Labial*	30	46	0	5	1	17	138
	Alveolar	63	9	1	5	3	19	114
Approximant	*Alveolar*	21	11	0	4	11	53	109
	R	15	30	0	0	26	30	27

For the wrongly identified plosives in final position manner is mostly perceived correctly (80%). For fricatives, the opposite is found: manner is perceived correctly only 28% of the time. Even though confusions from voiceless to voiced were not expected to be found for plosives and fricatives in final position, as voiced plosives and fricatives do not exist in Dutch, it can be seen in Table 2.11 that voiced responses nevertheless do occur, although not very often. In initial and medial position, cognate confusions were very common. The fact that these confusions hardly occur in final position, is one of the explanations for the high final consonant correct score.

Nasals are misperceived most often in final position: 38% of the time (see Table 2.4). 64% of these confusions falls within the nasal category. Approximants are mostly confused with other manners of articulation. Only 35% of these confusions falls within the approximant category. Place of articulation is a problem for all four consonant categories. All consonant categories show a high number of missing values, consisting of deletions or incomprehensible sounds. Especially for /l/ high missing values are observed. For nasals, this was also found in Spanish (Miralles & Cervera, 1995). A possible explanation is that at the end of a word, intensity drops and articulation is not as precise as in initial and medial position. Also interesting are the confusions from /n/ and /m/ with /mp/ and /p/. An explanation might be that the speaker abruptly stops articulating, resulting in a build-up of pressure, sounding somewhat like a /p/.

2.3.2 Voiced-voiceless contrast

The confusions discussed in the previous section showed that most confusions of plosives and fricatives concern the feature voice, at least in initial and medial position (see Tables 2.5 and 2.8). This was not an unexpected finding, as the most common confusions found in literature are between voiced and voiceless plosives and fricatives (Doyle et al., 1988; Hammarberg et al., 1990; Nord et al., 1992; Miralles & Cervera, 1995; Roeleven & Polak, 1999; Boon-Kamma, 2001; Searl et al., 2001;). These confusions deserve further investigation as the (in)ability to produce a correct voiced-voiceless distinction may provide more information about the amount of control TE speakers have over their neoglottis.

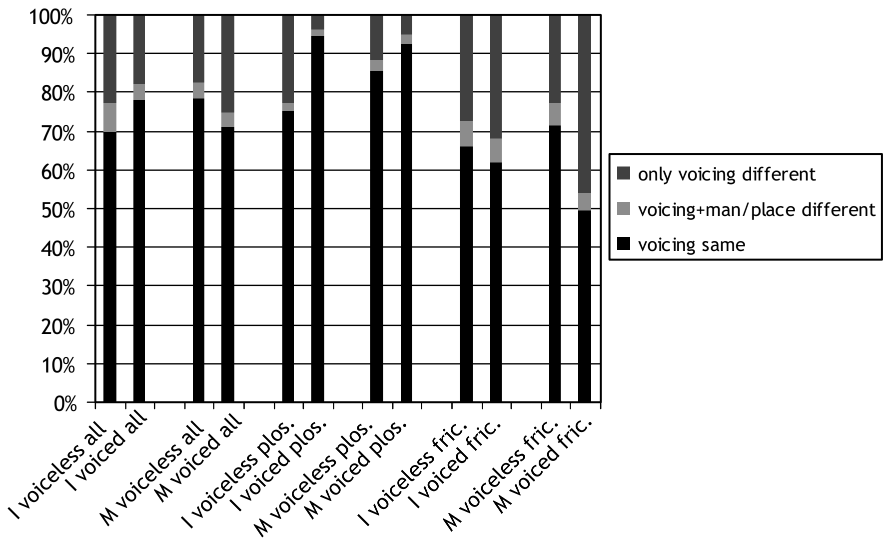


Figure 2.4: The perception of voice in voiced and voiceless plosives and fricatives together (voiceless/voiced all) as well as for plosives and fricatives separately, both in initial (I) and in medial (M) position. 'Only voicing different' means that the phoneme has been confused with its cognate, 'voicing + man/place different' means that in the response the feature voice has been changed as well as place and/or manner. 'voicing same' means that the feature voice has been perceived correctly. Correctly perceived phonemes were also included in this figure as the feature voice is then automatically also perceived correctly. 100% is the total number of stimuli. Phonemes included were /p b t d tj f v s z ʃ/.

Figure 2.4 shows the overall voicing errors of plosives and fricatives in initial (I) and medial (M) position. The final position was excluded from further analysis as in Dutch voiced plosives and fricatives in word-final position are not allowed. As most voiced-voiceless confusions seem to concern cognates, only plosives and fricatives with a cognate were included (/p b t d tj f v s z ʃ/), N=2902 for initial position and N=3246 for medial position) for further analysis. The phonemes /tj/ and /ʃ/ were included as well, because they showed many confusions with the rare loan phonemes /ʒ/ and /dj/.

In Figure 2.4, three different categories are used. For example, a confusion from /p/ with /k/ is marked as 'voicing same' as the feature voice has been retained. It should be noted, that in this figure, correct phoneme perceptions were also included as the feature voice is then automatically also perceived correctly; a confusion from /p/ with /b/ is categorized as 'only voicing confused', meaning cognates are confused; and a confusion from /p/ with /m/ is marked as 'voicing + man/place different' whereby manner and/or place can be confused as well.

Figure 2.4 shows that for 'voiceless all' and 'voiced all' in initial position, voiceless sounds become voiced more often (30%) than voiced sounds become voiceless (22%) (χ^2= 26.208, df=1, p< .01; Φ=.095, p<.01). For medial position, the opposite effect is found: the voiced sounds become unvoiced slightly more often (29%) than the other way round (22%) (χ^2= 23.775, df=1, p< .01; Φ= -.086, p<.01). When looking at plosives and fricatives separately, it can be seen that in initial position voiceless plosives become voiced more often (25%) than voiced plosives become voiceless (5%) (χ^2=95.986, df=1, p<.01; Φ=.272, p<.01) whereas for the fricatives no significant effect is found (34% vs. 38%). In medial position, a mix is seen. Again the plosives mostly show changes from voiceless to voiced (14% vs. 7%) (χ^2=18.489, df=1, p<.01; Φ=.106, p<.01) whereas the fricatives mostly show changes from voiced to voiceless (51% vs. 29%) (χ^2=79.212, df=1, p<.01; Φ= -.221). Overall, it can be seen that voicing cognates are most often confused. When taking only confusions with cognates into account (rather than 'only voicing different' + 'voicing + man/place different' as was done above), the same results are found as described above, except that now for fricatives also a significant effect is found in initial position (with more confusions from voiced to voiceless). Most studies found confusions from voiceless to voiced both for plosives and fricatives, but Nord et al. (1992), for example, found more confusions from voiced to voiceless. In the present study, a mix was found with plosives mainly showing confusions from voiceless to voiced and the fricatives mainly showing confusions from voiced to voiceless. This result is different from findings in other studies. As yet it is not clear why these two types of confusions occur. Confusions from voiceless to voiced are usually attributed to the assumed slow decay of vibrations of the neoglottis, especially in medial position. However, as Searl (2001) states, PE segment features would not fully explain why stops are less affected than fricatives and why the expected error is reversed for some phonemes. A possible explanation for Dutch would be that, especially in the Amsterdam region, there is a tendency to devoice fricatives and listeners may be inclined to perceive fricatives as voiceless. This may explain some of the confusions found. The findings may also well be related to the morphology of the neoglottis. According to van As (2001) the finding that not only voiceless sounds are produced as voiced, but also vice versa, might be explained by the fact that when only aerodynamic forces are assumed, this should be related to the closure of the neoglottis. For patients with a closed neoglottis, it might be more likely that voiceless sounds are produced as their voiced counterparts, while in patients with an incomplete neoglottic closure it is more likely that voiced consonants are produced as their voiceless counterparts. This topic deserves further research, whereby the anatomy and physiology of the neoglottis and vocal tract of individual speakers must also be taken into account.

2.3.3 Consonant clusters

2.3.3.1 Overall results

Inter-listener reliability for consonant clusters was measured using Cronbach's alpha which was .94, .88 and .92 for initial, medial and final position, respectively. Just as with the consonants, the difference between digital and hands-free was investigated. Results are given in Table 2.14.

Per position, 1650 responses were available (5 clusters x 3 vowels x 11 speakers x 10 listeners). In medial position, some stimuli have been removed, as some of the speakers either inserted long pauses between the sounds of the cluster, thus producing different individual phonemes, or inserted a pause after the vowel thereby turning a medial cluster into an initial one. One speaker consistently did this and was omitted from the sample of medial clusters. 1471 responses remained.

Table 2.14: Absolute and relative frequency of correctly and incorrectly perceived consonant clusters for digital and hands-free speakers in initial, medial and final positions as well as all positions together (overall).

Speech mode	Position	Correct (%)	Incorrect (%)	Total
Digital	*Initial*	604 (67)	296 (33)	900
	Medial	685 (79)	185 (21)	870
	Final	775 (86)	125 (14)	900
	Overall	2064 (77)	606 (23)	2670
Hands-free	*Initial*	520 (69)	230 (31)	750
	Medial	524 (87)	77 (13)	601
	Final	619 (83)	131 (18)	750
	Overall	1663 (79)	438 (21)	2101

No significant difference between digital and hands-free speakers could be established for any of the positions.

Table 2.15: Mean percentage correct score for the consonant clusters per position with the ranges between brackets. Also the clusters that were tested are included between brackets.

	Initial (%)	Medial (%)	Final (%)	Overall (%)
Clusters	68 (49-86) (/pr sl tr kr fl/)	82 (67-96) (/pr sl tr kr fl/)	85 (68-95) (/lp lf mst mt rk/)	78

Table 2.15 shows that correct scores in initial position are significantly lower than in medial and final position (χ^2= 149.854, df=2, p<.01; Φ=.177, p<.01), which is the same result as we found for single consonants.

The individual scores of the speakers were again investigated to make sure that the scores were not influenced by only a few of the subjects. The bar graphs in Fig. 2.5 show the correct scores for initial, medial and final position of the consonant clusters per speaker.

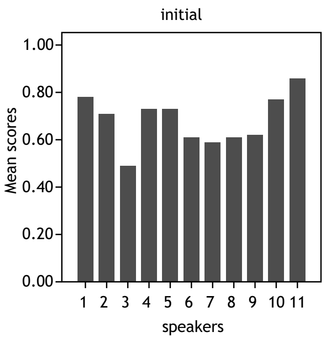


Figure 2.5a: Mean consonant cluster scores per speaker for initial position

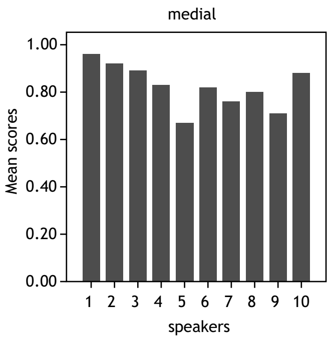


Figure 2.5b: Mean consonant cluster scores per speaker for medial position

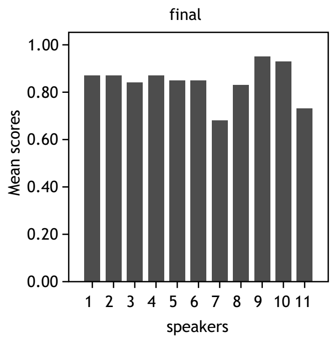


Figure 2.5c: Mean consonant cluster scores per speaker for final position

The bar graphs show no real outliers. Speaker 3 shows the lowest score in initial position, but elimination of his result does not change the tendency of the overall results.

Confusions of clusters are hard to analyze as more than one phoneme of the cluster may be affected. For this reason, it was decided to analyze two aspects specifically: can TE speakers produce clusters, i.e. can they produce two or three consonants next to each other, or do they reduce them, and secondly, can TE speakers make a correct transition from voiced to voiceless or the other way round within the cluster? It is expected that a rapid transition from voiced to voiceless or the other way round will be difficult for a TE speaker.

Just as with the consonants discussed before, the percentages given relate to the total number of confusions found within a category.

2.3.3.2 Cluster length

Table 2.16 shows in what percentage of cases the length of the cluster (number of phonemes) was perceived correctly and whether it was increased or decreased in length, unintelligible or completely deleted when it was not perceived correctly.

Table 2.16: Percentage correct score for the length of the consonant cluster in initial, medial and final position, plus the scores for the deviant responses.

Length	Initial	Medial	Final
Equal	79	89	49
Decreased	12	9	26
Increased	7	2	23
Unintelligible	1	0	0
Complete deletion	1	0	2
N	526	267	299

From Table 2.16 it can be seen that in initial and medial position, the length of the cluster is often perceived correctly. When this is not the case, usually the length of the cluster is decreased (‘cluster simplification’). The cases where the number of phonemes increased in initial position were caused mostly by an insertion of /s/ in front of the cluster. It might be the case that stoma noise, sometimes present at the start of articulation, is perceived as a fricative.

In final position, more problems concerning length are found than in initial and medial positions. These problems include both a decrease and an increase in number of phonemes. In this case, the increase is mostly caused by a /p/ or /b/ insertion between /m/ and /t/ in V-mt. This confusion can be explained by the fact that V-mt and V-mpt are perceptually very similar and they are both existing clusters in Dutch. It is also not completely unexpected to find that many clusters are decreased in length when the results are compared with final single consonants that were also frequently deleted. Again, in final position, consonants are less clearly articulated.

2.3.3.3 Voiced-voiceless distinction in consonant clusters

The other question under investigation is whether TE speakers can make a correct transition from voiced to voiceless or from voiceless to voiced within a consonant cluster. In the next table, the type of transition found with the number of occurrences in percentages is given.

Table 2.17: The occurrence of transitions in the responses per consonant cluster position in percentages. The set of initial and medial clusters as presented in the experiment always have a transition from voiceless to voiced (UV), whereas all transitions in final position are from voiced to voiceless (VU). Other occurrences for responses are VV for completely voiced, UU for completely voiceless, V for voiced and U for voiceless in cases of reduction. Highlighted cells indicate that the transition was perceived correctly.

Transition	Initial (UV)	Medial (UV)	Final (VU)
VU	1	1	68
UV	30	36	0
VV	56	54	2
UU	0	0	6
V	5	0	14
U	8	9	9
N	513	259	245

Table 2.17 shows that in initial and medial position, the transition from voiceless to voiced (UV) is most often replaced by a completely voiced cluster (VV) (56% and 54%, respectively). This was not an unexpected finding, since it had been established in the section on single consonants that voiceless sounds are often perceived as voiced. In final position, the transition is mainly perceived correctly. Since Dutch does not have any voiced plosives or fricatives in syllable final position the listener will be inclined to perceive the final element as voiceless.

2.3.4. Syllables - Vowels in h-V-t

2.3.4.1 Overall results

The inter-listener reliability for vowel identification in h-V-t syllables, measured using Cronbach's alpha, was .823 which allowed the use of the means over the ten naïve listeners again. Table 2.18 gives the numbers for correctly and incorrectly perceived vowels for hands-free and digital speakers.

Table 2.18: Numbers and percentages of correctly and incorrectly perceived vowels for digital and hands-free speakers (15 vowels x 11 speakers x 10 listeners = 1650).

Speech mode	Correct (%)	Incorrect (%)	Total
Digital	664 (74)	236 (26)	900
Hands-free	560 (75)	190 (25)	750
Overall	1224 (74)	426 (26)	1650

No significant differences between hands-free and digital occlusion could be observed.

A common division of vowels is into short vowels, long vowels and diphthongs. The results will be discussed accordingly. In Table 2.19 percentage correct scores are shown for these three subgroups. Each vowel had 110 responses (Vowel x 11 speakers x 10 listeners).

Table 2.19: Percentage correct score for vowels with the ranges given between brackets. The subsets of vowels involved are specified as well.

Stimulus	Short vowels (%)	Long vowels (%)	Diphthongs (%)	Overall (%)
Vowels	80 (56-96) *(ɪ ɛ ɑ ɔ ʏ)*	64 (47-79) *(i eː aː oː u y øː)*	87 (63-100) *(ɛi œy ʌu)*	74

The bar graphs in Figure 2.6a-c below show the individual scores per speaker on the three vowel categories. Although variation between speakers is apparent, the mean scores are not influenced by particular speakers.

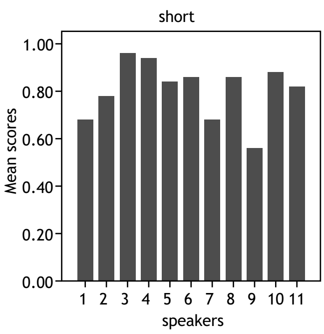


Figure 2.6a: Mean vowel scores per speaker for initial position

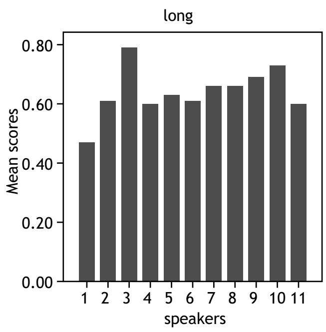


Figure 2.6b: Mean vowel scores per speaker for medial position

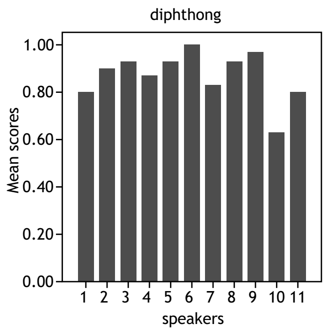


Figure 2.6c: Mean vowel scores per speaker for final position

The overall score of 74% shown in Table 2.19 is lower than the 84% score found in literature for normal speakers (Koopmans-van Beinum, 1980). Klein et al. (1970) found the same percentage of 74% for normal laryngeal speakers, but they used isolated vowels, which is the most difficult task for vowels. For Dutch, Oubrie (1999) has looked at the intelligibility of vowels and their formant values for TE speakers. She found an overall score correct of 49%, which is considerably poorer than the scores obtained in this investigation. Other studies have also looked at TE vowels. Miralles & Cervera (1995), for example, found a vowel correct score of 90% for Spanish. This seems much higher, but it needs to be taken into account that Spanish only has few vowels and the fewer vowels there are, the fewer errors can be made. In addition, they also used existing Spanish words making the listening task easier than with nonsense h-V-t syllables.

When the different categories of vowels are considered, it can be seen in Table 2.19 that long vowels are misperceived significantly more often than short vowels and diphthongs (χ^2= 82.624, df=2, p<.01; Φ=.224, p<.01). The ranges show that individual speakers show a large variation in performance as well.

2.3.4.2 Vowel confusions

It is not only interesting to look at individual vowel confusions, but also to see if certain features are misperceived more often than others. In this study, phonetic features are used that combine acoustic and articulatory features of vowels. Two dimensions are used: front-back and high-low. The front-back dimension pertains to the place where the upper surface of the tongue causes a constriction in the mouth. The high-low dimension describes how close the upper surface of the tongue is to the roof of the mouth. The shape of the lips is also important: the lips can be spread, rounded or unrounded. The following figure shows the phonetic features for the Dutch steady state vowels.

Place of constriction	**Front**		**Central**		**Back**	
Lip shape ⟶ Degree of ↓ constriction	*Spread*	*Unrounded*	*Rounded*	*Unrounded*	*Rounded*	*Unrounded*
High	i k*ie*s		y f*uu*t		u t*oe*n	
Mid		ɪ eː k*i*p k*ee*t	ʏ øː h*u*t p*eu*k		ɔ oː t*o*p p*oo*k	
Low		ɛ t*e*l		aː t*aa*k		ɑ t*a*k

Figure 2.7: Dutch vowel system for monophthongs, divided according to constriction place, degree of mouth opening and lip shape (from Pols, 2002: p. 253)

The diphthongs (ɛi œy ʌu) are not present in Figure 2.7, as they are not produced at a fixed position, but have a low starting point and contain a glide to a higher (more closed) position, which is more fronted for ɛi (l*ij*st) and œy (h*ui*s) and more back for ʌu (h*ou*t).

Table 2.20 shows the confusion matrix of the vowels. In this matrix the individual phoneme confusions are shown.

Table 2.20: Phoneme confusions for vowels. Numbers in percentages multiplied by 10.

	i	eː	aː	oː	u	y	øː	ɪ	ɛ	ɑ	ɔ	ʏ	ɛi	ʌu	œy	ai
i	**800**	150						45							9	
eː	9	**646**						64	9				250		27	
aː			**955**							45						
oː				**326**							55	55		560	9	
u		9			**882**	27	9	9	9						55	
y		9			160	**654**	64	9	9						91	
øː		9			170	64	**209**								550	
ɪ	55	73				9		**764**	81				9		9	
ɛ					36	9		18	**901**						36	
ɑ			250		9				18	**682**	36			9		
ɔ				9	18					140	**800**	9		27		
ʏ				91								**891**		18		
ɛi		27							9	9			**846**		91	18
ʌu		9								9	9		82	**891**		
œy			27		18		9		9	18			9	27	**883**	

As stated earlier, it is not only interesting to look at the individual vowel confusions, but also to see if certain features change more often than others. Therefore Table 2.21 shows the confusions concerning the features height (or degree of constriction), place of constriction and tenseness. Table 2.22 shows in more detail what happens with the feature height when it is confused.

Table 2.21: Feature confusions for short and long vowels and diphthongs, with *P* being place, *H* height, *R* rounding and *L* length. *Diph* stands for diphthongization, *- glide* for monophthongization and *missing* for unintelligible or deleted phonemes. Scores are in percentages related to the number of confusions. *N* shows the actual total number of confusions. Due to rounding, added percentages might not total 100%.

	PHRL	RH	PR	L	P	H	R	Diph	-glide	Missing	N
Short	8	18	1	35	2	20	2	15		1	107
Long	1		1	14		26		60			277
Diphthong			14		67		2		17		42
Total	3	4	2	17	7	22	1	42	2	2	427

Table 2.22: Confusions concerning height of the vowel in actual numbers for short and long vowels separately. 'x' means no confusion is possible in that cell.

	Short vowels			Long vowels		
	low	mid	high	low	mid	high
low	x	11	1	x	0	0
mid	25	x	11	1	x	14
high	0	0	x	1	58	x

When the short vowels are considered, it can be seen that the most common problem is height. In 46% of the confusions, the feature height is changed (PHRL+RH+H).

It can be seen that half of these responses is higher (has the tongue higher up towards the roof of the mouth) than the intended vowel and half is lower than the intended vowel (Table 2.22). Also common (Table 2.21) are confusions between long and short vowels (35%). This is, however, almost entirely caused by the phoneme /ɑ/. 15% of de short vowels are diphthongized. Hardly any errors concerning place were made.

Long vowels were misperceived in 36% of the cases (277 / total number of long vowels (770)). The most common confusion is diphthongization of the vowel (60%). Especially the vowels /e oː øː/ are diphthongized. This is not an unexpected error, as it also happens fairly often in normal laryngeal speech as was described in Koopmans-van Beinum (1969). Height also causes many confusions (27%). Contrary to the short vowels, a general lowering (position of tongue is lower than it should be) of the vowels can be observed (see Table 2.22). When short and long vowels are taken together, lowering vowels is twice as common as raising them. Once again, hardly any place confusions are found, as was the case with short vowels.

Diphthongs were only misperceived in 13% of the cases (42/330). Slightly more than half of these confusions concerns a confusion between diphthongs, as can be seen in the confusion matrix (Table 2.20). In the other cases, the diphthong is perceived as a steady state vowel. No pattern can be discovered in the confusions.

Oubrie (1999) found the same types of confusions and the same diphthongization for the vowels /oː/ /eː/ and /øː/ as in this study, even though the number of confusions per vowel differs from the present study. As Oubrie also used h-V-t stimuli to test vowels, the lower percentage correct scores as found by Oubrie may have been caused by the fact that her speaker group contained several people with a reconstruction, whereas in our experiment only speakers with a standard laryngectomy were included. Only the best two and the worst two perceived vowels are the same in Oubrie's and the present data. Where the lowering of sounds is concerned, perception might be influenced by higher formant values that Oubrie found in her study for TE speakers.

Another factor is the generally lower F0 (pitch) that was not actually measured by us in these syllables, but is known to exist in TE speech. A large difference between an F0 and an F1 (first formant) will also influence the judgment of a listener in such a way that a higher F1 than is actually present in the signal will be perceived (Traunmüller, 1981; Hoemeke & Diehl, 1994). These findings need to be investigated in more detail. Oubrie (1999) also gives a possible explanation for the diphthongization of /oː/ /eː/ and /øː/ which is stronger than the diphthongization in normal laryngeal speakers. She states that /oː/ /eː/ and /øː/ lie near each other in the vowel diagram at almost the same height and also near the short vowels /ɔ/, /ɪ/ and /ʏ/. In normal speech, it can be seen that formants change during the articulation of the vowel (so there is actually also slight diphthongization), which is called coloring of the formants. It might be that for TE speakers this coloring is too large, moving the formant values too strongly into the direction of /u/, /iː/ and /y/, resulting in the diphthongs /ʌu/, /ɛi/ and /œy/. An explanation for this strong formant coloring might be that TE speakers have little control over the dorsal part of their tongue.

2.3.5 Semantically Unpredictable Sentences

The inter-listener reliability (Cronbach's alpha) for Semantically Unpredictable Sentences (SUS) was .856. Like in the phoneme assessment, in this test hands-free speakers score significantly higher than the speakers in this study who digitally close off the stoma (χ^2=13.805, df=1, p<.01) as can be seen in Table 2.23.

Table 2.23: Absolute numbers and percentages of incorrectly and correctly perceived Semantically Unpredictable Sentences for digital and hands-free speakers.

Speech mode	Incorrect (%)	Correct (%)	Total
Digital	147 (51)	143 (49)	290
Hands-free	87 (35)	163 (65)	250

The scoring system of SUS sentences is strict: even one word (or phoneme) wrong will lead to the score 'sentence incorrect'. The next table shows the scores per speaker. Scores are based on 50 responses (5 sentences x 10 listeners), except for speaker 6, where one sentence could not be used for analysis due to a technical failure, leaving 40 responses.

Table 2.24: Sentence correct scores in percentages per speaker. The underlined score indicates the speaker with the highest score; the bold and italicized score indicates the speaker with the lowest score.

speaker	1	2	3	4	5	6	7	8	9	10	11	overall
score %	<u>94</u>	78	82	74	68	55	32	26	***12***	46	56	57

It is apparent from Table 2.24 that there is much individual variation, ranging from only a few sentences correct (12%) to almost all sentences correct (94%). No difference between the types of sentences was found, which means that mean scores can be used. The mean score for all subjects together (57%) is much lower than the mean phoneme scores discussed in the previous section.

Unfortunately, no data are available on SUS sentences used with normal laryngeal speakers of Dutch. Studies have been performed for English and French (Hazan et al., 1989). For English, a word score was found of 99%. Sentence scores are generally lower, but it was found that word scores correlate highly with sentence scores. For French, a word score of 98% was found against 90% for sentences. SUS sentences were designed to be comparable across languages, which is why it can be assumed that sentence scores will be similar for Dutch (Benoit et al., 1996). As the best subject score in this study is 94% this seems to be a reasonable assumption. Comparison of scores shows that sentence intelligibility of TE speakers is much lower than for normal speakers. Two Dutch master's theses (Roeleven and Polak, 1999 and Boon-Kamma, 2001) also studied TE sentence intelligibility. Roeleven & Polak used nonsense sentences from the Swinging Speech Test for aphasic speakers (Bocca, 1963) and found very similar results to the present study with an overall correct score of 60%. This is only slightly higher than the overall score of the present study. The slightly higher score may be caused by the fact that vowel confusions were not noted as mistakes in Roeleven & Polak's study. Boon-Kamma found a much higher sentence intelligibility score of 73%. However, compared to this study, she was less strict in her evaluation. She also did not use SUS sentences, but 9-syllable sequences of spontaneous speech of her subjects. Even though not much context was available in the samples, the sequences themselves were not semantically anomalous, which probably makes the sequences easier to perceive.

2.3.6 Spontaneous speech

2.3.6.1 Overall results

Spontaneous speech was evaluated using semantic bipolar 7-point scales (see Table 2.1). Seven scales were used for the evaluation of intelligibility and four scales were included to evaluate voice quality. One scale for overall judgment was included as well, consisting of three choices: good-moderate-poor (see also section 2.3.4.3). In Table

2.25, the mean scores plus ranges are given, together with the standard deviation and the inter-rater reliability value (Cronbach's alpha).

Table 2.25: Scores per scale, with the ranges given between brackets, the standard deviation (SD) and the reliability value (α).

Scales	Mean scores (range)	SD	α
Intelligibility scales			
1. easy to understand-difficult to understand	3.35 (2.0-4.9)	1.84	.954
2. fast-slow	3.45 (3.2-6.3)	1.14	.775
3. precise-sloppy	3.36 (1.4-6.0)	1.40	.889
4. normal-deviant	4.43 (2.6-4.4)	1.64	.936
5. natural-unnatural	4.57 (3.0-6.2)	1.61	.901
6. clear speech-mumbled speech	3.30 (1.6-5.4)	1.62	.928
7. takes no effort to understand-takes effort to understand	3.50 (1.6-6.2)	1.94	.958
Voice quality scales			
8. normal-deviant	4.50 (2.9-5.9)	1.49	.927
9. beautiful-ugly	4.44 (3.1-6.0)	1.49	.927
10. low-high	3.26 (2.8-3.9)	1.19	.476
11. deep-shrill	3.59 (3.1-4.2)	1.39	.576
Overall Judgment (3 points)			
12. good-moderate-poor	1.67 (1-2.5)	.721	.918

When the scores per scale are considered, it can be seen that mean scores are mostly in the mid range. However, the individual speakers show large variation, meaning there are great differences between speakers. This is further illustrated in Table 2.26.

A low score in general means a positive judgment and a high score a negative one. However, two scales show a deviant pattern: low-high and deep-shrill. Both these scales have negatives on either side of the scale. In retrospect, they should have been divided into two separate scales, once with the first term on the right hand side and once with the second term on the right hand side. Strictly speaking, they are not very useful in our further analyses, but for now they will be taken into account.

For each scale, inter-rater reliability was calculated. Scales with a score lower than .70 were excluded from further analysis. Except for scale 10 and 11, concerning the pitch of the voice, reliability is high. Hands-free speakers again score significantly higher than non-hands-free speakers (a mean score of 3.24 versus 4.25 respectively; Mann Whitney, $p<.01$).

Table 2.26: Mean scores for each of the 12 7-point scales and for each of the 11 speakers.

	Scales											
Speakers	1	2	3	4	5	6	7	8	9	10	11	12
1	2.1	2.9	2.6	3.1	3.2	2.2	2.3	2.9	3.1	3.1	3.1	1.2
2	4.5	4.4	4.3	5.6	5.8	3.6	4.5	5.6	5.7	3.2	4.2	2.2
3	2.9	2.6	3.2	3.7	3.7	2.9	3.2	3.7	3.8	2.8	3.2	1.4
4	2.7	3.0	3.1	3.7	4.1	2.9	2.8	4.0	3.9	2.8	3.2	1.5
5	1.4	2.9	2.0	3.0	3.4	1.6	1.6	3.6	3.4	3.2	3.1	1.0
6	1.9	3.0	2.4	3.7	3.8	1.9	1.7	3.8	3.6	3.1	3.4	1.1
7	2.5	3.5	2.6	4.0	4.3	2.8	2.4	4.1	3.6	3.9	4.0	1.2
8	3.5	3.7	4.3	4.8	4.7	4.1	3.8	4.9	5.2	3.7	4.2	1.9
9	6.0	4.0	4.9	6.2	6.3	5.4	6.2	5.9	6.0	3.2	3.9	2.5
10	3.6	3.7	3.4	4.8	5.2	4.1	4.0	5.1	5.0	3.0	3.3	1.9
11	5.7	4.3	4.2	6.0	5.8	4.8	5.9	5.9	5.5	3.9	3.9	2.4

2.3.6.2 Factor analysis

A Principal Component analysis was performed on the correlation matrix of the mean values for each speaker on each scale to investigate relations between the different rating scales that were applied. Results are shown in Table 2.27.

Only one component with eigenvalues>1 could be extracted. This component explained almost 74% of the variance. It is not possible to perform a VARIMAX rotation on one component only. This poses an interpretation problem. It seems that listeners have not been able to distinguish between voice quality and speech intelligibility scales, since both sets of scales are represented in component 1. This is similar to what Van As (2001) found in her study on TE voice quality. Van As states that naïve listeners do not overcome the deviancies of these voices and that once they hear a deviant voice they are not able to hear the specific aspects represented by the scales. Van As found only two components for naïve listeners against four for experienced listeners. As only one component was found in the present study, it seems that in this study all scales measure the same underlying variable for the naïve listener. Van As (2001) also found that only the scales related to pitch (deep-high, low-shrill) were judged independently of the other scales. Again this is similar to the results in this study where these two scales were the only scales with a low reliability and thus were not included in the Principal Component analysis.

The fact that intelligibility scales and voice quality scales are taken as one group by the listeners was, besides the present study and the study by van As (2001) also reported earlier by Nieboer et al. (1988). This finding may imply that the decreased intelligibility of TE speakers is mainly related to voice quality. However, it is very difficult to interpret listeners' behavior and more research in this area is needed.

Table 2.27: Results of the Principal Component analysis. Only one component with eigenvalues >1 could be extracted. In the left column the scale names are given, in the second column the unrotated factor loadings and in column three the communalities (the proportion of the test variance that is common factor variance).

Scales	Component 1	Communality
1. easy to understand-difficult to understand	.876	.767
2. fast-slow	.629	.396
3. precise-sloppy	.841	.708
4. normal-deviant	.920	.846
5. natural-unnatural	.899	.808
6. clear speech-mumbled speech	.843	.710
7. takes no effort to understand-takes effort to understand	.902	.813
8. normal-deviant	.881	.776
9. beautiful-ugly	.898	.807
Percentage of variance explained	73.68%	

2.3.6.3 Scale judgments versus overall judgment

A judgment of overall speech intelligibility (good-moderate-poor) was given by all raters. Even though it is likely that naïve listeners use normal speech as their internal standard (Van As, 2001) it is still interesting to see if the overall judgment is related to the other scale judgments. To investigate this, an ANOVA was performed, with the overall judgment as factor and the scales as dependent variables. Results are shown in Table 2.28. Superscript numbers indicate to which significant subset the scale belongs. The scale *easy to understand – difficult to understand*, for example, shows significantly different scores for 'good', 'moderate', and 'poor'. The scale *fast – slow*, however, shows only two subsets: good + moderate, and poor.

Except for the scale low-high, all results were significant. Post hoc tests show that the scales *easy to understand-difficult to understand, precise-sloppy, normal-deviant, natural-unnatural, clear speech-mumbled speech, takes no effort to understand-takes effort to understand, normal-deviant (voice),* and *beautiful-ugly* differ on all three judgments. The scales *fast-slow* and *deep-shrill* show significant differences between good and poor. All scales show that the average scale values increase with a decrease of intelligibility (the higher the score, the poorer the intelligibility). The overall judgment scale then seems to distinguish well between the speakers.

A discriminant analysis was performed to see if it could be predicted to which group (good, moderate or poor) a speaker belonged, based on the different scale judgments. The stepwise method, with the default Wilks' Lambda was performed. Only one variable (scale) seems to be important in predicting class membership, namely:

easy to understand - difficult to understand. The average correct classification score for this variable is 79.6% whereby it is easiest to classify speakers as good or poor, and less easy as moderate.

Even though the PC analysis found only one component, it seems that the scale *easy to understand-hard to understand* is the most important scale for intelligibility research in TE speakers.

Table 2.28: Results for ANOVA and post hoc Tukey tests for the 11 7-point semantic scales relative to the overall speech intelligibility judgment (good – moderate – poor). In the second column, the p-value of the ANOVA is shown and in columns 3-5 the mean values per scale are given. Superscript numbers indicate to which significant subset the mean value belongs.

	P-value	good	moderate	poor
Intelligibility scales				
1. easy to understand-difficult to understand	<.001	1.90^{1}	4.12^{2}	6.13^{3}
2. fast-slow	<.001	2.88^{1}	3.88^{2}	4.31^{2}
3. precise-sloppy	<.001	2.37^{1}	3.93^{2}	5.25^{3}
4. normal-deviant	<.001	3.25^{1}	5.20^{2}	6.25^{3}
5. natural-unnatural	<.001	3.52^{1}	5.22^{2}	6.38^{3}
6. clear speech-mumbled speech	<.001	2.02^{1}	4.17^{2}	5.25^{3}
7. takes no effort to understand-takes effort to understand	<.001	1.78^{1}	4.56^{2}	6.25^{3}
Voice quality scales				
8. normal-deviant	<.001	3.54^{1}	5.10^{2}	6.19^{3}
9. beautiful-ugly	<.001	3.40^{1}	5.12^{2}	6.13^{3}
10. low-high	.207	3.15^{1}	3.22^{1}	3.75^{1}
11. deep-shrill	<.05	3.33^{1}	3.63^{1-2}	4.38^{2}

2.4 Correlations between different speech tasks

It is important to know whether phoneme or SUS intelligibility tests correlate well with spontaneous speech judgments. The main goal of therapy is to improve people's functioning in daily life, meaning that spontaneous speech intelligibility needs to be as good and easy as possible. The problem with spontaneous speech is that it is hard to test for intelligibility. It is much easier to obtain phoneme or SUS sentence intelligibility scores. If, therefore, a phoneme score or a SUS sentence score correlates highly with spontaneous speech scores, we may use these simpler tests to predict intelligibility of spontaneous speech. For this reason, the correlations between the phoneme scores and the SUS sentence scores and the semantic scales that were used to judge spontaneous speech were looked at by means of a Spearman rank correlation. This renders the results presented in Table 2.29.

Table 2.29: Correlations between the 12 scale judgments and the various phoneme and SUS sentence identification tasks, with the scales as rows (1=easy to understand-difficult to understand; 2=fast-slow; 3=precise-sloppy; 4=normal-deviant; 5=natural-unnatural; 6=clear speech-mumbled speech; 7=takes no effort to understand-takes effort to understand; 8=normal-deviant; 9=beautiful-ugly; 10=low-high; 11=deep-shrill; 12=good-moderate-poor) and the other tests in the columns. * indicates a significance of p<.05, ** indicates a significance of p<.01.

Scale (down)	Initial cons.	Medial cons.	Final cons.	Vowels	Initial clusters	Medial clusters	Final clusters	SUS
1	-.572	-.538	-.510	-.362	-.169	-.503	-.041	-.891**
2	-.399	-.603*	-.211	-.041	-.297	-.581	-.313	-.783**
3	-.471	-.549	-.551	-.301	-.161	-.354	-.088	-.781**
4	-.572	-.694*	-.507	-.225	-.245	-.665*	-.035	-.906**
5	-.585	-.685*	-.411	-.096	-.352	-.644*	-.037	-.879**
6	-.667*	-.545	-.542	-.315	-.294	-.559	-.109	-.877**
7	-.572	-.538	-.510	-.362	-.169	-.503	-.041	-.891**
8	-.615*	-.676*	-.411	.115	-.357	-.681*	.088	-.902**
9	-.537	-.580	-.457	-.161	-.284	-.474	-.002	-.843**
10	-.356	-.771**	-.319	.012	-.320	-.654*	.259	-.314
11	-.203	-.720*	-.492	.074	-.058	-.587	.206	-.610*
12	-.511	-.454	-.384	-.411	-.192	-.389	-.048	-.815**

It should be noted that the negative values are caused by the fact that a higher scale judgment denotes a lower quality. It can be seen in Table 2.29 that many of the tests do not correlate in any significant way with the scale judgments (consonants in final position, vowels, and clusters in initial and final position). Consonants in initial and medial position, and clusters in medial position do correlate quite well with some, but not all, of the scales. The Semantically Unpredictable Sentences, however, correlate highly with all scales except low-high (scale 10). This scale behaves differently from the other scales, which was also seen in the Principal Component analysis and the reliability measures. A likely explanation for the high correlation between SUS sentence scores and spontaneous speech judgments is that sentences are closer to real speech than the other tests. However, care should be taken when interpreting results. The correlations are based on only 10 speakers. A significant correlation indicates that the rank order of speakers is roughly the same for both tests, but due to the low number of cases, it cannot be concluded that one of the tests is redundant. Also, a non-significant correlation can mean that the tests involved measure different aspects of speech, but low correlation can also be caused by the low number of cases.

2.5 Implications for therapy

The results from the perception experiment have clearly shown that TE speech intelligibility is severely compromised when compared to normal laryngeal speech. This finding strengthens the belief that speech training may be beneficial for this group of speakers. Doyle (1988) also stated that "even those TE speakers judged to be highly proficient in discourse exhibit substantial reductions in intelligibility at a consonantal level. As such, regular clinical measures of intelligibility may expose specific deficits that may be masked in general assessments of TE speakers. Consequently, some TE speakers may require more individualized and more comprehensive therapeutic programs prior to discharge. That is, treatment can be designed to focus on those consonants that are most problematic for a given patient." Much variation between speakers has also been observed in this study. Even though it is not known at this stage why some speakers are so much better than others, we assume that similarly high scores can just as well be obtained for the poorer speakers by specific speech training. Based on the present results it was decided what should be included in the therapy program which is discussed in chapter 4.

The consonants are discussed first. In the following figure the percentage correct scores are given per consonant per position.

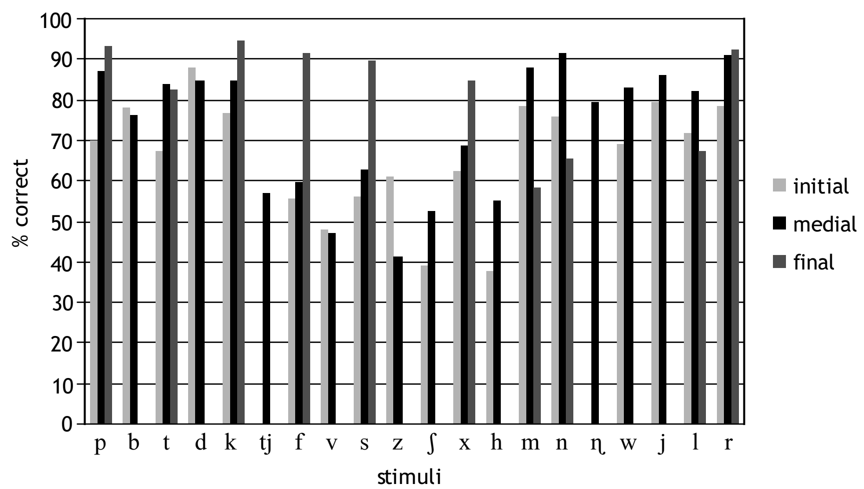


Figure 2.8: Percentage correct score per consonant and per consonant position.

As already discussed in section 2.3.1, the fricatives show the lowest scores of all manner categories, at least in initial and medial position. Hence, the phonemes /f v s z/ should be included in the therapy program. Yet, the phonemes /x/ and /h/ have even poorer scores; these phonemes therefore also have to be included. Even though plosives are on the whole not the most difficult phonemes, they should be included as they also show many confusions between cognates, just as the fricatives. The consonant clusters showed voiced-voiceless confusions as well. A large part of the therapy program then should consist of training a correct voiced-voiceless distinction. Finally, the nasals /m/ and /n/ and the liquid /l/ should be incorporated in the program.

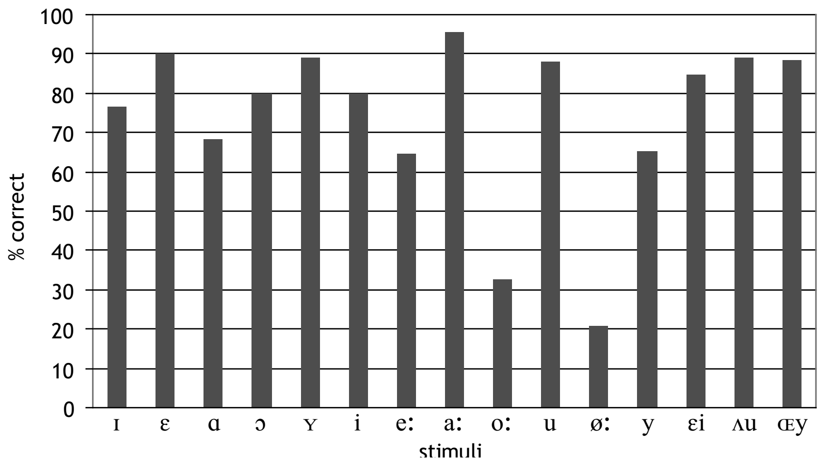


Figure 2.9: Vowel correct scores in percentages

The following figure shows the percentage correct scores for the <u>vowels</u>. The problem with vowels is that, contrary to the consonants, it is hard to train the particular articulatory features of vowels. It is physically not possible to teach speakers to raise their tongue one or two millimeters to produce a more closed vowel for example. This is the reason why vowels are not included as individual phonemes in the program, but are passively trained by using clear speech (Krause & Braida, 2002), which basically means a clear and deliberate articulation of all sounds, which also tends to improve the acoustic qualities of the vowels. This technique then is also beneficial for the consonant clusters that can both be trained at phoneme level and at sentence level, as a deletion of part of the cluster was often seen.

The perception experiment has also shown that ***sentence*** intelligibility and intelligibility of ***spontaneous speech*** are compromised, which is why training at sentence and discourse level also should be included in the program, making use of clear speech again. A full account of the therapy program, its structure and techniques will be given in chapter 4.

2.6 Discussion and conclusion

Overall results

The main statement in the introduction, based on the changed anatomy and physiology, was that TE speakers would show a diminished intelligibility. To find out whether this was indeed the case, a comprehensive perception experiment was performed, which, together with the results, was discussed in this chapter. It was found in this study that TE speech intelligibility was mildly to severely compromised at all levels of speech when compared to data available for normal laryngeal speakers. This finding confirms earlier studies for Dutch and other languages and clearly shows the necessity for specialized speech therapy.

Another interesting finding was that hands-free speakers scored significantly higher on several speech tasks (consonants in initial and medial position, SUS sentences and spontaneous speech) than speakers who digitally closed off their stoma. A number of studies have compared hands-free versus occlusion by finger, and have shown that the hands-free method is slightly less favorable, as more effort is required to initiate and sustain voicing. Op de Coul et al. (2005) studied the difference between digital occlusion and hands-free speech (with two types of automatic stoma valves), in particular in relation to voice quality. They found that hands-free speakers scored lower on maximum phonation time, and dynamic loudness range. The speakers involved in Op de Coul's study also filled out a questionnaire and results indicated that hands-free speakers were less satisfied with their intelligibility than the speakers who digitally closed off their stoma. This is in contrast to our findings where intelligibility rates were higher for hands-free speakers. However, in view of the problem patients and naive listeners have in distinguishing between voice quality and speech intelligibility, the contrast might not be as deep as it seems. One explanation for the better intelligibility of hands-free speech is that although the hands-free method requires more speaking effort, it more closely resembles normal speech than occlusion of the stoma valve by finger. In the latter condition, the coordinated movement of a finger or thumb (to occlude the stoma) becomes a crucial, but unnatural, part of the speech act. It is expected that this will interfere with the natural rhythm and timing of speech, and thus with phrasing (personal communication with M. van Rossum, 2007). This study has indeed shown that hands-free speakers score better, not only on phoneme level but also on conversational level. Another possible explanation for the difference found is that usually the better adjusted speakers benefit from an automatic speaking valve. As such, there seems to be a pre-selection in the type of speakers that use a hands-free valve. More research is needed on this topic, but the fact that there is a difference between the two speech modes means drawing conclusions has to be carefully done when both types of speakers are included in a study. This warning was already given by Bridges (1991) who stated it was probably unjust not to make a distinction between both types of speakers.

TE speakers also show a wide variation in scores. It is not completely clear why some speakers are better than others. For therapy purposes it is important to find out whether differences are caused by differences in anatomy between subjects, meaning training may not be effective, whether some speakers are just more comfortable in using TE speech than others, or whether it is a combination of factors. When performing experiments, it is difficult to pick a random sample of subjects as variables should be controlled so as to avoid difficulties in interpreting results. However, it is very difficult to control all variables.

Phoneme level

Despite the fact that naïve listeners had no prior experience with TE speakers and that the speakers themselves show large variations, inter-listener reliability was quite high, with the highest agreement for clusters (.94, .88 and .92) and the lowest (.823) for the vowels. These high reliability scores confirm the findings in other studies (Doyle et al, 1989; Van As et al, 1999; Cullinan et al., 1986) that naïve listeners are capable of evaluating TE speech reliably.

With the consonants, the most consistent errors were found for plosives and fricatives, with the fricatives being the most difficult phonemes in initial and medial position. The errors consisted mostly of confusions between voiced and voiceless phonemes and between cognates in particular. With the consonant clusters, the same errors were observed, meaning that the voiced-voiceless distinction within a cluster was often not maintained. Even though problems with voicing were expected, as it is assumed that the neoglottis is less pliable than the vocal folds are, the pattern of confusions was different than expected. The plosives mainly showed confusions from voiceless to voiced, whereas the fricatives mainly showed confusions from voiced to voiceless. The assumption that TE speakers have less control over their neoglottis than over their larynx cannot explain both types of confusions. The confusions from voiced to voiceless in the fricatives may partly be explained by a Dutch listener bias towards voiceless fricatives, but more research will need to be done on this topic whereby the anatomy and physiology of the neoglottis are also taken into account. In the next chapter, acoustic analyses will be discussed that were performed to investigate which acoustic correlates TE speakers employ to convey a *correct* voiced-voiceless distinction and whether these correlates differ from normal laryngeal speakers. Even though only correctly perceived plosives and fricatives were included, it may shed some light on this problem.

In the introduction it was hypothesized that vowel intelligibility would be compromised as well. Results have shown that the vowels do have intelligibility scores lower than published for normal laryngeal speakers. Care has to be taken though to compare results from different studies as vowel environment can influence perception profoundly. However, scores in the present study are persistently lower than scores found for normal laryngeal speakers and therefore the assumption that TE speakers perform worse than normal laryngeal speakers seems valid. It is likely that the perception of the vowels is influenced by changes in the vocal tract. The PE segment is not necessarily at the same height as the larynx was and due to the laryngectomy and the simultaneous removal of the hyoid bone, the base of the tongue is somewhat 'retracted' or at least in a different position.

The studies that investigated vowels in alaryngeal speech mainly focused on formant values. All acoustic vowel studies found that formant values of alaryngeal speakers differed from normal laryngeal speakers and many studies concluded that based on the formant values it was assumed that the vocal tract of alaryngeal speakers is shorter (Miralles & Cervera, 2001; Christensen & Weinberg, 1976; Sisty & Weinberg, 1972). This study does not concur with that statement. Isman and O'Brien (1992) say there are small differences between speakers concerning the shape, length and exact position of the PE segment. Van As (1999) also found that the position of the neoglottis varies between speakers and that the position can vary some 12 to 50 mm at the level of the fourth to the sixth cervical vertebra. Based on these findings the assumption seems valid that the vocal tract will have changed after the operation, but that the vocal tract can be either longer or shorter than before. This fact may explain, at least partly, the variation between the speakers. Section 2.3.4 described the difficulty in perceiving the height of the vowel with a general tendency of a lowering of the vowel. A possible explanation that was offered was the distance between F0-F1 that might confuse the listener. Another explanation for the confusions found is that if listeners do not know who the speaker is and whether the speaker is male or female (and hence if the vocal tract is long or short), listeners make more mistakes in perception than when they would know the speaker (described in Lieberman & Blumstein, 1988 and Van Bergem et al., 1988). The different position of the tongue, lacking the attachment to the hyoid bone, and hence the diminished control over the dorsal part of the tongue seems to explain the strong diphthongization of certain vowels. However, vowel confusions are still hard to explain, especially for TE speakers. There are many variables that influence the perception, and articulatory movements are not necessarily the most important variable. As Lindau et al. (1972) state: "The nature of some vowel targets is much more likely to be auditory than articulatory. The particular articulatory mechanisms that a speaker makes use of to attain a vowel target is of secondary importance only". There are different strategies to obtain the same vowel sound. This area deserves detailed research whereby anatomy, physiology, articulation and perception are all taken into account.

Sentence level and spontaneous speech

To test sentence intelligibility, quite a difficult task was chosen. The great advantage of using Semantically Unpredictable Sentences, however, is that listeners cannot rely on semantic information and that they really have to focus on every individual sound. This in its turn also makes it a difficult task as there is a greater cognitive load involved in correctly identifying a complete string of semantically unconnected words (Benoit, 1996). It is also a very strict test: all words need to be identified correctly instead of only key words. The difficulty of the test explains the low scores, but when compared to English and French SUS sentences spoken by normal laryngeal speakers, the low scores are realistic, especially when also considering the low scores at phoneme level.

Intelligibility of spontaneous speech is hard to evaluate. It is not possible to ask listeners to write down exactly what they heard, as it cannot be checked whether it was exactly what was being said. Semantic scales are a good way of obtaining general judgments on specific aspects of spontaneous speech. However, there are some drawbacks. Firstly, semantic scales have mainly been used to evaluate voice quality instead of speech intelligibility and secondly it is quite hard for naïve listeners to use semantic scales as they do not always know the meaning of the terms used on either side of the scale. However, in the present study, naïve listeners did rather well, with high reliability scores for all but two scales. The problem with these two scales may be explained in two ways. Both scales were pitch related, which is an aspect of speech that is difficult to evaluate for untrained listeners. Another problem with these scales is that both sides were negatives. In retrospect, they should have been divided into four scales, each having a positive term on the left hand side and a negative term on the right hand side. In fact, listeners scored in the middle for both scales, which suggests that they did not think pitch was very deviant for male voices.

In this experiment both voice quality scales and intelligibility scales were included, but clearly marked as such. A factor analysis revealed only one component with all scales included having high factor loadings. This shows that listeners do not make a difference between voice quality and intelligibility ratings. Having only one component made it impossible to reduce the number of scales. However, the discriminant analysis, which was used to investigate which scale best predicted membership of the groups poor, moderate and good speech, revealed only one scale, namely *easy to understand-hard to understand*. It seems that this is the most useful scale when evaluating spontaneous speech.

Correlations were calculated to see which of the intelligibility tests at phoneme and sentence level correlated highest with spontaneous speech ratings. It was found that the Semantically Unpredictable Sentences correlated highly with all scales except the two pitch scales. Although this seems to indicate that this test is a good predictor of spontaneous speech intelligibility, it was argued in the results that due to the low number of cases, interpretation of the results should be done with great care.

Naturally, we could have made better comparisons with normal laryngeal speakers if we had had a control group in this study. As we did not have a control group, results were compared with the literature. For the consonants, the study by Pols (1983) was chosen. Although his and our experiment are different, the experiments can be compared as both experiments used nonsense syllables, tested consonants in initial, medial and final position and in both experiments the listeners typed in what they perceived. For the vowels, the study by Koopmans-van Beinum (1980) was used. She measured vowels in isolated monosyllabic words. The structure of the listening experiment was similar. Even though comparing with literature is not as ideal as a control group, we believe that results can be compared due to the similarities in the experimental set-up.

Conclusion

In this chapter a comprehensive perception experiment of Dutch TE speech has been described. The results of these experiments will be used to compose a therapy program. It was found that TE speech intelligibility is severely compromised, with scores well below the average scores for normal laryngeal speakers in comparable experiments. It was also discovered that TE speakers form a heterogeneous population and that although some common errors such as the inability to produce a voiced-voiceless distinction are found, great inter- and intra speaker variability exists. Even speakers that have control over their neoglottis do not show this control consistently. It is unclear why some TE speakers are better than others. It might be that the neoglottis of intelligible speakers has different characteristics from the neoglottis of unintelligible speakers. Good speakers might also use compensating strategies and in that case these strategies may be useful for poorer speakers.

In the next chapter, we discuss the acoustic analyses of the voiced-voiceless distinction correctly produced by speakers which will show whether compensating strategies are used or not. The main aim of this thesis is to investigate whether it is possible to improve speech intelligibility by specialized speech rehabilitation based on the research findings from this chapter and chapter three. In chapter 4, the development of the therapy program will be discussed, followed by the results of the therapy evaluation studies in chapters 5 and 6.

Appendix 2.1 Speech material

Phonemes

Consonants in initial position (CV)

PAA	DAA	VAA	GAA	RAA	LAA
PIE	DIE	VIE	GIE	RIE	LIE
POE	DOE	VOE	GOE	ROE	LOE
BAA	KAA	SAA	MAA	WAA	HAA
BIE	KIE	SIE	MIE	WIE	HIE
BOE	KOE	SOE	MOE	WOE	HOE
TAA	FAA	ZAA	NAA	JAA	SJAA
Tie	FIE	ZIE	NIE	JIE	SJIE
TOE	FOE	ZOE	NOE	JOE	SJOE

Consonants in medial position (CVC)

AAPAA	IEDIE	OEVOE	AAMAA	IERIE	OELOE	AANJAA
IEPIE	OEDOE	AASAA	IEMIE	OEROE	AAHAA	IENJIE
OEPOE	AAKAA	IESIE	OEMOE	AAWAA	IEHIE	OENJOE
AABAA	IEKIE	OESOE	AANAA	IEWIE	OEHOE	
IEBIEE	OEKOE	AAZAA	IENIE	OEWOE	AASJAA	
OEBOE	AAFAA	IEZIE	OENOE	AAJAA	IESJIE	
AATAA	IEFIE	OEZOE	AANGAA	IEJIE	OESJOE	
IETIE	OEFOE	AAGAA	IENGIE	OEJOE	AATJAA	
OETOE	AAVAA	IEGIE	OENGOE	AALAA	IETJIE	
AADAA	IEVIE	OEGOE	AARAA	IELIE	OETJOE	

Consonants in final position (VC)

AAP	OET	IES	AAG	OEM	IENG	AAL
IEP	AAK	OES	IEG	AAN	OENG	IEL
OEP	IEK	AAF	OEG	IEN	AAR	OEL
AAT	OEK	IEF	AAM	OEN	IER	
IET	AAS	OEF	IEM	AANG	OER	

Consonant clusters

initial	medial	final
FLIE	IEFLIE	AMST
FLA	AAFLAA	IEMST
FLOE	OEFLOE	OEMST
PRIE	IEPRIE	AMT
PRA	AAPRAA	IEMT
PROE	OEPROE	OEMT
KRIE	IEKRIE	ALF
KRA	AAKRAA	IELF
KROE	OEKROE	OELF
TRIE	IETRIE	ARK
TRA	AATRAA	IERK
TROE	OETROE	OERK
SLIE	IESLIE	ALP
SLA	AASLAA	IELP
SLOE	OESLOE	OELP

Vowels

HAAT/aː/	HOT /ɔ/	HEUT /øː/
HAT /ɑ/	HUUT /y/	HOET /u/
HIET /iː/	HUT /ʏ/	HIJT /ɛi/
HIT /ɪ/	HEET /eː/	HAUT /ʌu/
HOOT /oː/	HET /ɛ/	HUIT /œy/

Nonsense sentences (SUS sentences)

subject 1:
Het land huilt langs de vieze naam
Het scherpe boek vormt de tuin
Wek de pijn en de stoel
Waar streelt het feit de smalle oom?
De zee prijst het dier dat smeekt

subject 2:
Het beeld stapt voor de dunne nacht
De lege krant vult de heer
Schenk het bed of de zoon
Wanneer wenst de staat het lange plan?
Het bos wast het geld dat praat

subject 3:
De taal komt na de volle lijn
De grote prijs slacht het been
Eis de hand of de vraag
Wanneer haat de rest de dikke vrouw?
De brief zoek de geest die gaat

subject 4:
De mens woont bij het lage woord
De diepe stem leent het volk
Blus het paar of de kerk
Waarom boeit de taak de trage maand?
Het blad schept de raad die kruipt

subject 5:
De lucht klimt naar de warme droom
De mooie school temt het jaar
Huur de wind en de broer
Waarom poetst de kaart de lauwe bank?
Het uur doodt de strijd die schrikt

subject 6:
De borst zit op de korte zak
De hete macht noemt het werk
Kweek het raam of de bloem
Waarom troost de tand de ruime weg?
De wijn scheert het paard dat zwijgt

subject 7:
Het huis loopt in het rare dorp
De vuile kracht neemt de fles
Zaai het hart of de kunst
Waarom groet de hond het gladde doel?
De dienst krijgt de eeuw die stijgt

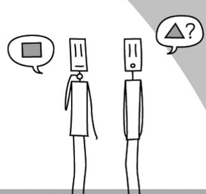
subject 8:
De film rent uit de schone rug
De snelle grens meldt de stad
Keur het spel en de hoek
Waar waagt de arm de hoge muur?
De gang kookt de vorm die trilt

subject 9:
Het licht zwemt om het leuke vuur
Het koude glas ploegt de kop
Meng de deur en het ding
Waarom likt het lid de dure grond?
De week sloopt het schip dat zweeft

Subject 10:
De boom groeit tot de strakke man
De brede angst eert de neus
Was de hulp of de trap
Waar vindt de dag de kleine vriend?
Het bloed proeft de plaats die springt

3

Acoustic analysis of the voiced-voiceless distinction in Dutch tracheoesophageal speech



Abstract

Confusions between voiced and voiceless plosives and between voiced and voiceless fricatives (especially cognates) are the most common confusions in Dutch tracheoesophageal (TE) speech. The problem is attributed to the working of the new voice source: the pharyngo-esophageal segment, or neoglottis. In order to learn how TE speakers convey a *correct* voiced-voiceless distinction, detailed analyses are necessary. In this study, *acoustic* analyses are performed on plosives and fricatives to investigate which acoustic correlates are used by TE speakers to convey a *correct* voicing contrast and to see if TE speakers use different correlates than normal laryngeal (NL) speakers. Eleven TE speakers and five NL speakers were included in this study. Fifteen acoustic correlates (and a subset of six for the fricatives) were selected and analyzed. Results show that TE speakers do not differ much from NL speakers, except where pitch-related correlates are concerned. For the plosives all correlates distinguish between voiced and voiceless for both speaker groups. For the fricatives, not all correlates contributed to the distinction. However, only very few fricatives could be analyzed. The main distinguishing correlate for the plosives was relative phonation time in the closure and for fricatives the consonant duration.

A slightly adapted version of this chapter has been submitted for publication:

Jongmans, P, Wempe, A.G., Van Tinteren, H., Hilgers, F.J.M., Pols, L.C.W. & Van As-Brooks, C.J. The acoustic analysis of the voiced-voiceless distinction in Dutch tracheoesophageal speech.

Part of this chapter has been published in:

Jongmans, P., Hilgers, F.J.M., Pols, L.C.W. & Van As-Brooks, C.J. (2006). The intelligibility of tracheoesophageal speech, with an emphasis on the voiced-voiceless distinction. *Logopedics. Phoniatrics. Vocology, 31*, 172-181.

Jongmans, P., Wempe, A.G., Hilgers, F.J.M., Pols, L.C.W. & Van As-Brooks, C.J. (2007). Acoustic correlates of the voiced-voiceless distinction in Dutch normal and tracheoesophageal speakers. *Proceedings ICPhS XVI*, 1997-2000.

3.1 Introduction

As has been discussed in the two previous chapters, despite the fact that speech intelligibility has improved with the introduction of voice prostheses, tracheoesophageal (TE) speech intelligibility is still severely compromised when compared to normal laryngeal (NL) speech intelligibility. Phoneme intelligibility experiments whereby phoneme confusions in Dutch TE speech at syllable level, as judged by naive listeners, were analyzed, showed that one of the most common confusions found was between voiced and voiceless sounds. The difficulty in making a good distinction between voiced and voiceless sounds exists within the plosive and fricative groups. It is argued that the production of the voicing contrast is difficult for a TE speaker as it is assumed that the neoglottis is less pliable than the vocal folds and cannot be easily controlled. Perceptual data seem to confirm this assumption. It is a relevant area to study in more detail because of its importance for intelligibility and because it may help to understand the underlying principles of neoglottic voice production. Consequently, the results of the present research may be relevant for therapeutic purposes as well. To understand how voiced and voiceless plosives and fricatives are produced, acoustic analyses can be performed to complement perceptual data.

In other languages, several acoustic studies have looked at the production of the voiced-voiceless distinction in TE speech (Robbins et al., 1986; Nord et al., 1992; Saito et al., 2000; Searl & Carpenter, 2002; Searl & Ousley, 2004; Lundström & Hammarberg, 2004). The acoustic correlates they investigated were closure duration, phonation onset and offset during the closure duration, VOT, consonant SPL, and vowel duration. Intra-oral pressure was also investigated. Nord et al. (1992), Lundström & Hammarberg (2004) and Saito et al. (2000) found that highly intelligible TE speakers show longer closure duration and longer VOT for voiceless plosives and often continuous vibration during voiced plosives. Saito et al. (2000) also found that values for closure duration and VOT were longer for TE speakers than for NL speakers. This was also found by Robbins et al. (1986), who also looked at preceding vowel length. Searl & Ousley (2004) found no difference for phonation offset between TE voiced and voiceless plosives, but did establish a difference between TE and NL speakers. Searl & Carpenter (2002) only compared TE speakers with NL speakers and found that consonant SPL, closure duration and preceding vowel length were longer/higher for TE speakers than for NL speakers.

Dutch is a language that contrasts prevoiced and voiceless unaspirated plosives (Lisker & Abramson, 1964). There has been some debate in literature as to what the main acoustic cue is that listeners use when classifying plosives as voiced or voiceless. Slis & Cohen (1969) stated that the presence of a voice lead is an important cue to perceive a sound as voiced, with a length of 80 ms giving the best perceptual results.

Lisker & Abramson (1964) found similar results. Later, Slis (1970) stated that when the continuation of phonation after the moment of oral closure (=voice tail) is <50 ms a plosive is perceived as voiceless, whereas a voice tail >50 ms prompts a voiced response. In 2004, Van Alphen & Smits found that the presence/absence of prevoicing (VOT) is the most important cue, but that it is not used consistently by all Dutch speakers in the experiment. For fricatives, the duration and intensity of the friction sound is generally considered the most important cue (Lisker & Abramson, 1964; Slis & Cohen, 1969). What all studies have in common though, is that voicing is considered to be the most important cue. Though the cues described above were found to be perceptually most relevant, other acoustic correlates also distinguish between voiced and voiceless plosives and fricatives, which Slis & Cohen (1969) also mentioned. As it is not known which acoustic correlates will be used by TE speakers to convey the voiced-voiceless distinction, in this study the acoustic correlates used in previous studies will be combined and will be complemented with other acoustic correlates described in literature on normal laryngeal voicing. In this study, only the plosives and fricatives that were perceived more than eighty percent correct have been included. The first interest of this study thus goes towards a better understanding of the importance of specific acoustic features once utterances are perceived correctly. This set-up also allows us to compare TE speakers with NL speakers. The hypothesis is that TE speakers will exaggerate certain correlates, especially durational ones, to convey a correct voiced-voiceless distinction and that TE speakers will show more problems with actual voicing (pitch) than NL speakers.

In this chapter, the data from chapter 2 relevant for the acoustic study will be summarized first so that perceptual findings can be related to acoustic findings. This will be followed by a discussion on the acoustic analyses of the voiced and voiceless plosives and fricatives.

3.2 Subjects and methods

3.2.1 Subjects

Eleven laryngectomized subjects participated in the phoneme intelligibility study. All of them underwent a total standard laryngectomy. They were all using TE speech by means of an indwelling (Provox2) voice prosthesis (Hilgers et al., 1997). All subjects were male, with a mean age of 66;9 years (age range 44 to 78 years). Further characteristics of the TE group can be found in chapter 2. Furthermore, 5 normal laryngeal speakers (NL) were included in order to compare outcomes. This group consisted of 5 male speakers, with a mean age of 56 (range 45;9-72;3). These speakers were matched for age range, but have a lower mean age.

3.2.2 Recordings

For TE speakers, recording details can be found in chapter 2. For the laryngeal speakers recordings were made in a sound studio with a Pioneer PDR-555 RW CD recorder and a Sennheiser MKH 105T microphone, with a microphone-to-mouth distance of 30 centimeters. A microphone amplifier, made by the Institute of Phonetic Sciences Amsterdam was also used.

3.2.3 Speech material

The speech material for the acoustic analyses consisted of medial plosives and fricatives only as they are most relevant for the acoustic study (for a full account of the perception experiment, see Jongmans et al., 2006).

3.2.4 Listening Task

The listening task, described in more detail in chapter 2, was performed by ten naïve listeners with no prior experience of TE speech. The listening task was performed online, so that listeners could be flexible in where and when they could perform the experiment. The listeners listened to the speech samples in a quiet environment and all used headphones. For each syllable they typed the consonant they perceived in normal spelling, which is unambiguous in Dutch. When listeners perceived a cluster, the first phoneme of that cluster was used to determine manner, place and voicing. Software was available that automatically generated response files that were used to investigate which phonemes were misperceived and which confusions were made between phonemes.

3.2.5 Acoustic analyses

Acoustic analyses were performed on the VCV utterances of the cognate pairs /p-b, t-d, f-v, s-z/. The reason for this is that acoustic analyses are slightly easier to perform on medial plosives and fricatives and in the literature these analyses have also mainly been restricted to medial position. The total number of utterances was 11 speakers x 8 phonemes x 3 vowels = 264 for TE speakers and 5 speakers x 8 phonemes x 3 vowels = 120 for laryngeal speakers.

3.2.6 Software

For all acoustic analyses the program Praat was used (www.praat.org) (Boersma & Weenink, 1996). Especially its scripting function was very useful here as it allowed us to automate the analyses and calculations.

3.2.7 Statistics

In section 3.3.1 the results of the perception experiment are discussed. The descriptive results are expressed in percentages correct or incorrect. Chi-squared tests are used to investigate differences between categories. Chi-squared tests establish the association between two (or more) qualitative variables. When an association is established, however, it does not say how strong the association is. Measures for strength of association do exist and in this study Cramer's V (Φ) is preferred: with more complex tables, Φ measure can still, as in a 2x2 case, achieve its maximum value of unity (Kinnear & Gray, 1999).

For the acoustic analyses (section 3.3.2) it was studied which acoustic correlates differ significantly for voiced and voiceless plosives and fricatives for TE speakers and NL speakers and if the use of the acoustic correlates differs for the two speaker groups. The statistical model used for the analyses is called a 'two-level conditional hierarchical linear model'. This is a type of 'linear mixed effects model' whereby the observations are tested within speakers and whereby speakers (random effect) are classified according to type (TE or NL speaker) (fixed effect). Use has been made of PROC MIXED (SAS 9.1.3). Four factors were included in the model: type of speaker, voiced or voiceless, type of vowel, and place of articulation.

In order to study which of the acoustic correlates distinguishes best between voiced and voiceless plosives and fricatives, conditional inference trees are used. Conditional inference trees, implemented in the PARTY package (Hothorn et al., 2006; 2006a), available for R statistical software (http://www.r-project.org/), estimate a regression relationship by binary recursive partitioning in a conditional inference framework. Roughly, the algorithm works as follows: 1) Test the global null hypothesis of independence between any of the input variables and the response (in this case voiced/voiceless). Stop if this hypothesis cannot be rejected at a pre-specified nominal level of α (in our study 0.05). Otherwise select the input variable with the strongest association to the response. This association is measured by a p-value corresponding to a test for the partial null hypothesis of a single input variable and the response. 2) Implement a binary split in the selected input variable. A permutation test framework is applied to find the optimal binary split in one selected covariate and the goodness of a fit is evaluated by two-sample linear statistics. 3) Steps 1) and 2) are recursively repeated.

The implementation utilizes a unified framework for conditional inference, or permutation tests, developed by Strasser and Weber (1999). The accuracy was calculated for each tree separately and confidence levels are given for each accuracy score.

3.3 Results and discussion

3.3.1 Perception experiment

3.3.1.1 Overall results

Inter-listener reliability was measured using Cronbach's alpha for each consonant position individually (nasals and approximants included). For medial position a value of .996 was found. This value indicates that inter-listener reliability is sufficiently high and that averages over the listeners can be used for further analysis. Table 3.1 shows the average scores per consonant category in medial position.

Table 3.1: Average scores in percentage correctly perceived consonants for all 4 manners of articulation. The range and the phonemes included are given between brackets. A star (*) indicates that this manner of articulation was perceived significantly worse than the other manners of articulation.

Manner	Medial (%)
Plosive	79 (61-91) (/b p d t k tj/)
Fricative	56* (42-89) (/v f z s h x ʃ/)
Nasal	86 (70-98) (/m n nj/)
Approximant	85 (70-99) (/ʋ j l r/)
Overall score	76

From the scores in Table 3.1, it can be seen that the lowest percentages correctly perceived consonants are found for the plosives and fricatives, whereby the fricatives score lowest (χ^2=375.399, df=3, p<.01; Φ=.251, p<.01 and χ^2=521.659, df=3, p<.01; Φ=.295, p<.01).

Table 3.2 provides a detailed overview of the confusions made with plosives and fricatives in medial position. Only the phonemes that were analyzed acoustically are included in this table. As can be seen, manner is mostly perceived correctly. In addition, especially for the fricatives but also for the plosives, the most common confusions are between the voiced and voiceless feature. Of these confusions, cognate pairs are most often confused. The term ‘cognates’ in this thesis refers to phonemes that differ only in the feature voice. In the next paragraph, the voiced-voiceless confusions will be discussed in more detail.

Table 3.2: Phoneme confusions for medial position. Numbers in percentages multiplied by 10. ‘Missing’ category consists of scores deleted or incomprehensible items, ‘other’ category consists of all other responses, like clusters, that are not of concern now. The grey areas indicate the same manner of articulation; the bold numbers denote confusions between cognates.

	p	b	t	d	k	tj	x	f	v	s	z	ʃ	j	l	ʋ	m	n	ʒ	missing	other
p	870	**112**			3			3											6	6
b	**70**	761		12	6			3							67	42	9		6	24
t	30	3	840	**85**	21		3													18
d		15	**24**	846	9		15		3				15	15			15		28	15
f	15	3		3				595	**345**						3				18	18
v	12	3	3					**449**	470						15		3		24	21
s								9	3	628	**207**	48	3	3				18	33	48
z							3	6	15	**452**	412	27	9					9	9	58

3.3.1.2 Voiced-voiceless contrast in plosives and fricatives

Since the interest of the current chapter is to investigate the most common confusions in the voiced-voiceless distinction in TE speech in relation to their acoustic correlates, only plosives and fricatives with a cognate (/p b t d f v s z/) were included for further analysis. Having only cognate pairs allows for better comparison between voiced and voiceless pairs, both perceptually and acoustically.

Figure 3.1 shows the overall voicing confusions for the 8 plosives and fricatives in medial position. In this figure, three different categories of confusions are used. For example, a confusion from /p/ with /k/ is marked as ‘voicing same’ as the feature voice has been retained; a confusion from /p/ with /b/ is categorized as ‘only voicing confused’, meaning cognates are confused; and a confusion from /p/ with /m/ is marked as ‘voicing + man/place different’ whereby manner and/or place can be confused as well. For the perception experiment it is important to take all confusions into account. For the acoustic analyses only the cognate confusions are taken into account, as they are minimally different pairs for which only the feature voice differs. This allows for a better investigation of the relation between perception and acoustics.

It should be kept in mind that only the *confusions* found in Table 3.2 are taken into account in Figure 3.1. The actual number of confusions is given in Figure 3.1 (N). There were 330 responses per phoneme. This means, for example, that there were 660 responses for the voiceless plosives. 94 of the 660 responses were incorrect, which is a percentage of 14%. This percentage matches the findings in Table 3.2. The fact that there are large differences in the number of confusions found requires careful interpretation of the results.

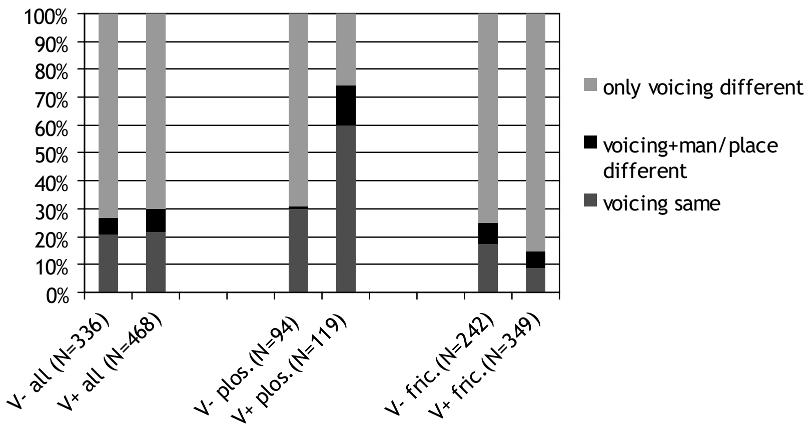


Figure 3.1: Confusions between voiced (V+) and voiceless (V-) plosives and fricatives together (voiceless/voiced all) as well as for plosives and fricatives separately in medial position. 'Only voicing different' means that the phoneme has been confused with its cognate, 'voicing + man/ place different' means that in the response the feature voice has been changed as well as place and/or manner. 'voicing same' means that the feature voice has been perceived correctly. Phonemes included were /p b t d f v s z/. The total number of confusions for the different categories is given between brackets (N). 100% indicates the total number of confusions for that particular category.

For medial position it can be seen that most confusions are between cognates. For the plosives and fricatives together, there is no difference between voiced (V+) and voiceless (V-) for the type of confusions. When looking at plosives and fricatives separately, it can be seen that voiceless plosives become voiced more often than voiced plosives become voiceless (70.2% vs. 40.3%, χ^2=18.844, df=1, p<.01; Φ=.297), whereas the fricatives mostly show changes from voiced to voiceless (91.4% vs. 82.6%, χ^2=10.249, df=1, p<.01; Φ=.132). Of these voicing confusions, the absolute majority concerns confusions between cognates.

Confusions for the feature voice are discussed in many papers on TE speech intelligibility (Robbins et al., 1986; Doyle et al., 1988; Hammarberg & Lundström, 1990; Nord et al., 1992; Miralles & Cervera, 1995; Roeleven & Polak, 1999; Saito et al., 2000; Boon-Kamma, 2001; Searl & Carpenter, 2002; Lundström & Hammarberg, 2004; Searl & Ousley, 2004; Jongmans et al., 2005; 2006) and were also clearly present in this experiment. However, it is important to look not only at the perceptual results but also at the acoustic signals to see what acoustic correlates TE speakers use to produce a voiced or voiceless sound correctly and whether these cues are different from normal speakers. In the following paragraphs the acoustic analyses of the plosives and fricatives will be discussed. After that, so called ‘conditional inference trees’ are used to investigate which of the acoustic correlates are most relevant for the production of a correct voiced-voiceless distinction.

3.3.2 Acoustic analyses

For this study, an extensive list of acoustic correlates was investigated. In the next section, it will first be discussed how the measurements were performed and then what the outcomes of the measurements were.

3.3.2.1 Measurements

In speech, several acoustic correlates are used to distinguish between a voiced and a voiceless sound. In literature these correlates have been investigated both for laryngeal speakers (NL) and TE speakers. For Dutch normal speakers, extensive research has among others been done by Lisker & Abramson (1964) and Slis & Cohen (1969). The common way to investigate these correlates is by segmenting the acoustic speech signal and then studying the segmental characteristics. This segmentation is already a difficult task for normal speech, but even more so in the case of TE speech. As periodicity is often lacking and noise is often present, this task is at times even impossible. It is therefore very important to properly define the correlates and describe the segmentation criteria. Existing segmentation criteria were mainly used, but sometimes small alterations were made. In Figures 3.2a and 3.2b, an example of a segmented voiced and voiceless plosive is shown that was produced by two different TE speakers. The phonemes in the example were perceived correctly.

The following correlates for medial consonants were included for investigation. All 14 were used for the plosives and six were used for the fricatives (1, 4, 8, 11-13) as not all correlates related to fricatives.

Durational correlates

1. Preceding vowel duration
2. Phonation offset time after the first vowel
3. Phonation time in closure
4. Closure duration (plosives) or consonant duration (fricatives)
5. Closure duration + burst
6. Burst duration
7. Phonation onset time
8. Following vowel duration

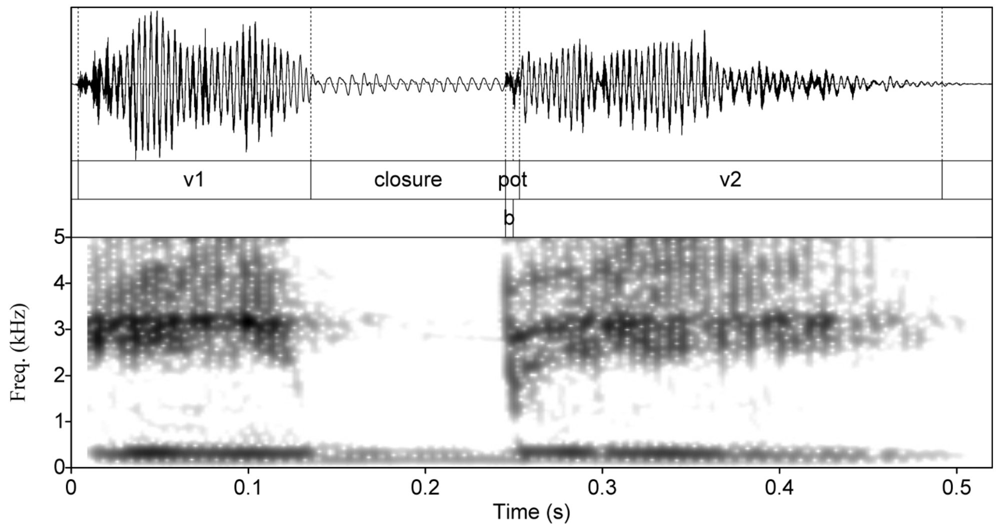


Figure 3.2a: Example of segmentation for /ibi/ that was identified correctly, produced by a TE speaker, with below the oscillogram preceding vowel (*V1*), closure, phonation onset (time) (*pot*) and following vowel (*V2*) on the first tier. On the second tier the burst (*b)* is marked. Also the spectrogram is shown.

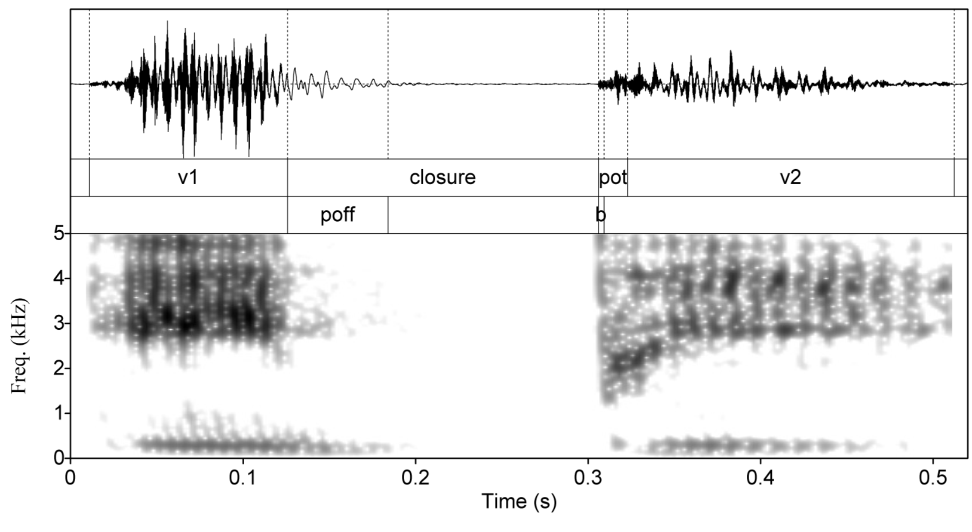


Figure 3.2b: Example of segmentation for /ipi/ that was identified correctly, produced by another TE speaker, with below the oscillogram preceding vowel (*V1*), closure, phonation onset (time) (*pot*) and following vowel (*V2*) on the first tier. On the second tier the phonation offset after V1 (*poff*) and the burst (*b*) are marked. Also the spectrogram is shown.

Calculated correlates/derived measures

9. Phonation offset after first vowel as percentage of closure duration
10. Phonation as percentage of closure duration
11. Percentage of voiced frames in the closure or consonant duration (fricatives)
12. HNR of voiced frames in the closure or consonant duration (fricatives)
13. HNR of closure duration or consonant duration (fricatives)
14. Relative burst intensity

Durational correlates

Preceding vowel length (V1) in ms: based on literature (Slis & Cohen, 1969; Searl & Carpenter, 2002) the preceding vowel was expected to be longer before voiced consonants. For TE speakers the preceding vowel was expected to be even longer than laryngeal preceding vowels in both voiced and voiceless consonants. The first vowel has its first segment boundary at the start of the VCV utterance (all utterances were recorded in isolation and started from silence). The second boundary was more difficult to determine as the vowel usually fades out and does not have a clear-cut ending. Therefore changes in the periodic structures were observed first and if no clear periodic structure was present, the formant structures in the spectrogram were used as an additional cue (see also Searl & Carpenter, 2002).

Phonation offset after preceding vowel (poffvowel) in ms: it is suspected that the length of the phonation offset may be a cue to perceive a sound as voiced or voiceless (Robbins et al., 1986; Saito et al., 2000; Searl & Ousley, 2004). Searl & Ousley (2004) further predicted that TE voiceless fricatives and plosives would show a long phonation offset, due to the slower decay of vibrations of the preceding vowel and that this may cause a listener to perceive the sounds as voiced. They studied this phenomenon in more detail by measuring phonation offset in voiced and voiceless plosives produced by TE and NL speakers. As onset boundary they used the end of the preceding vowel and as offset boundary they chose the point where the glottal pulses during the stop gap of the plosives ended. They found that 'phonation offset' and 'phonation offset as percentage of the closure duration' were significantly longer for the TE speakers as compared to laryngeal speakers, but that it was not used as a salient cue to differentiate between TE voiced and voiceless plosives. Searl & Ousley (2004) used pitch (the end of glottal pulses) to determine the boundaries of the phonation offset. Pitch however, is a problematic feature in TE speech. Completely voiceless speech signals as well as irregularly voiced speech signals are common in TE speech. For this reason, it was decided to adopt a different strategy for determining the phonation offset segment boundary. As onset boundary the preceding vowel border was still used but as offset boundary the point was used where the amplitude of the signal is 20 dB lower than the average amplitude of the preceding vowel. Figure 3.2b shows the onset and offset boundaries of the *poffvowel*. The average amplitude of the preceding vowel was calculated over the most representative (middle) part of the vowel (with a window length of 32 ms). The offset value of 20 dB was based on trial and error: a value higher than 20 dB would hardly ever be met, whereas a lower value hardly ever corresponded to the true cessation of the vowel in the cases where this was clearly visible in the acoustic signal. As this value is somewhat arbitrary, it could be possible that in very loud vowels, the signal is still quite strong even after the offset boundary. However, it appears to be a good compromise in our data and a relative point in the signal was needed that could be compared across speakers. Where no *poffvowel* can be found, the value of this variable is set to zero. In Figure 3.2a, for example, the intensity of the closure did not decrease by more than 20 dB relative to the vowel. This might be caused by the fact that the vibrations of the vowel are masked by the vocal murmur of the voiced plosive. *Poffvowel* was thus considered to be 0 here.

Longer *poffvowel* durations are expected for the TE speakers as compared to the laryngeal speakers and differences are also expected to be found between voiced and voiceless plosives for both speaker groups.

Phonation time in closure (phonclosure) in ms: due to the irregularities in TE speech and the frequent absence of voice, the following definition of phonation is adopted: measurable sound, which can be both voiced and voiceless. Hence whispering, for example, is also considered phonation. This particular correlate resembles the *poffvowel* and is actually completely identical for voiceless sounds (unless no offset can be found). It is different in the sense that in this correlate the time of phonation in the closure is measured, whether this phonation is caused by the preceding vowel or by the vocal murmur in the consonant. The offset is again the 20 dB below average intensity marker. However, when no offset time is found, phonation was assumed to be present during the whole closure and the value of this variable is then equal to the value found for the closure duration. NL speakers are expected to be able to control better the duration of phonation needed to make a correct voicing distinction than the TE speakers, who are expected to be less consistent in the use of phonation. Differences between voiced and voiceless phonemes within the groups are also expected.

Closure/consonant duration (cd) in ms: the term closure is used for plosives. For fricatives the term consonant duration is used. Based on literature (Slis & Cohen, 1969) the *cd* is expected to be longer in voiceless sounds for both speaker groups. The closure/consonant starts right after the preceding vowel and stops when either the phonation onset time with the burst (if present) begins or when the second vowel begins, which can be indicated by more noise and a greater amplitude than the closure phase, or by the start of the periodic structure of the vowel, respectively (see also Lisker & Abramson, 1964; Saito et al., 2000; Searl & Carpenter, 2002).

Closure duration + burst duration (cd+durb) in ms: The closure plus the burst form the total plosive duration.

Burst duration (durb) in ms: based on literature (Slis & Cohen, 1969) values for voiceless plosives are expected to be higher than for voiced plosives. The boundaries of the burst are determined via the waveform and spectrogram: they are the points where a sudden increase can be seen (onset) and then a decrease (offset) of amplitude and noise (often visible as a noise bar at the end of the closure). The first boundary is the end of the closure and the second boundary is after the noise bar.

Phonation onset time (pot) in ms: this is a parameter very similar to voice onset time but is not identical. As the following vowel is not always voiced in the case of TE speakers, the term *phonation*, rather than *voice* onset time is preferred. In this study the segment concerns the stretch of speech after the *cd* boundary and before the start of the second vowel, the latter being determined by the first periodic structures or clear formant structures. This *pot* may or may not contain a burst. The same definition, but then for VOT, has been given by Saito et al. (2000) and specifically for Dutch by van Alphen & Smits (2004). The *pot* contains too much noise to be part of the vowel. As aspirated stops do not exist in Standard Dutch, it is highly unlikely that this noise belongs to the consonant. It is more likely that this is an attempt of starting up phonation after the silent interval of the plosive.

Following vowel length (V2) in ms: this is the duration of the second vowel, which starts either after the closure and the burst (if present) or after the phonation onset (if present) and ends when the vowel shows no periodic structure anymore and/or when no clear formant structure is visible in the spectrogram.

Calculated correlates/derived measures

Phonation offset time of the preceding vowel as percentage of closure duration (relpoff) in %: this percentage value is defined as the length of the phonation offset divided by the length of the closure, times 100. This relative value was used also to accommodate for variation in speaking rate.

Phonation time as percentage of the closure (relphonclosure) in %: this percentage value is defined as the phonation offset time divided by the closure duration, times 100.

Percentage of voiced frames in the closure (percvoicedframes) in %: in NL speech, the presence or absence of voicing during the closure is one of the main cues for the correct perception of voiced and voiceless phonemes (Lisker & Abramson, 1964; Slis & Cohen, 1969; Van Alphen & Smits, 2004). Although the amount of voicing (pitch) is extremely hard to measure (correctly) in pathological voices like TE voices, it is still important to know up to what level TE speakers make use of voicing as a cue to differentiate between voiced and voiceless sounds. A Praat script was used to determine the amount of voicing. First, the optimal pitch setting for TE speech had to be found, since the default values appeared not to be appropriate. The pitch range was set from 65 to 250 Hz. Pitch values over 250 Hz are not usually found for TE voices. In the few cases where the algorithm found a higher pitch, it always appeared to be a measurement error. It is well known that TE speakers can have a very low pitch (Van As et al., 1998). It is, however, not possible to go far below 65 Hz because of the related window length. As the amount of voicing in the closure is of particular interest, a window length is needed which is shorter than the duration of the closure. Having set the pitch range, other parameters were additionally changed. The silence threshold was set to 0.01 instead of 0.03. This setting means that frames that do not contain amplitudes above this threshold (relative to the global maximum amplitude) are considered to be silence. The better the quality of the recordings, the lower this threshold value can be. The voicing threshold (the strength of the voiceless candidate, relative to the maximum possible autocorrelation) was changed from 0.45 to 0.35. Such a lower value is often needed in pathological voices to get pitch values even in irregular parts of the signal. The value in this study was determined by first considering a speech signal that had no perceptible voicing. 0.35 proved to be the minimum value above which no voicing would be found in these voiceless signals. After that, two people listened to ten speech signals that had voicing in it. The value of 0.35 again proved to be the best value where the amount of periodicity found matched the perception of pitch of the two listeners. With the settings discussed above, pitch was measured over the whole VCV utterance. However, only voicing in the closure was of interest in this case. Therefore, the frame numbers at the start and end time of the closure were established and frame by frame voicing was investigated. The total number of voiced frames found was then divided by the total number of frames within the closure.

For the NL speakers, the default setting of 0.45 was kept for the voicing threshold. An unrepresentative amount of voicing was found for this group if the value for this group was lowered to 0.35 as well.

For the NL speakers higher percentage of voiced frames are expected to be found for the voiced than for the voiceless plosives and fricatives. For the TE speakers, problems with voicing are expected, with irregular voicing patterns and inconsistency within speakers. Where voicing is employed correctly by speakers, this correlate will be very similar to *relphonclosure*.

Harmonics-to-noise ratio of the voiced frames in the closure (hnrvoicedframes) in dB: the HNR is normally calculated in sustained vowels as one of the measures of voice quality. It is not commonly used on consonants, but it may give valuable information on the quality of the voicing in the voiced frames during the closure. The HNR that is used here is based on pitch and therefore the settings of the pitch measurements discussed above have been maintained here. This type of HNR is only calculated over the voiced parts in the closure and can in this case be linked to the voiced frames found. It is expected that TE speakers show a poorer HNR for the voiced parts in the closure than NL speakers and that the HNR will be higher in voiced than in voiceless plosives and fricatives.

Harmonics-to-noise ratio during closure (hnrcd) in dB: This HNR is measured differently from the HNR discussed above. This measurement is not limited to the voiced parts only but measures HNR of the whole segment as long as it is above the silence threshold. The *hnrcd* is meant to measure the amount of voicing during closure just like the *percvoicedframes*, but its outcome will only say how much voicing was present and not what percentage in time of the closure was voiced. It needs to be kept in mind that the HNR values are not representative of the amount of noise and harmonics truly present in the consonant, but values are expected to differentiate between voiced and voiceless sounds, with higher values for the voiced sounds. Again it is expected that TE speakers show lower HNR scores than the NL speakers.

Relative burst intensity (relbint) as ratio: a higher intensity is assumed to be found for the voiceless plosives (Slis & Cohen, 1969). The intensity of the burst is expressed as the ratio of the maximum intensity of the signal. A *relative* burst intensity makes it easier to compare different speakers with each other.

3.3.2.2 Plosives: results and discussion

For the TE speakers, 126 out of a possible 132 plosives (11 speakers x 4 plosives x 3 vowels) were segmented using the described criteria (62 voiceless and 64 voiced). Six plosives were omitted as these could not be segmented properly due to the low quality of the signal. For the five NL speakers, all 60 plosives were segmented (30 labial and 30 alveolar). After segmenting the speech signals by hand, all parameters were calculated using a Praat script. As was discussed before, only stimuli that were perceived more than 80 percent correctly were included for further analysis as this study is interested in which correlates are used to convey a correct voiced-voiceless distinction. It also allows for a comparison of TE speakers with NL speakers. The total number of stimuli included for the TE speakers was 98 (47 voiceless and 51 voiced) and for NL speakers 60 (30 voiceless and 30 voiced)

The problem with including only stimuli that were perceived correctly more than 80 percent is that not every subject may contribute to the pool of stimuli. In this study, for instance, there was one subject who did not contribute to [p], one that did not contribute to [b] and one that did not contribute to [t]. This implies, though, that all subjects were represented in the acoustic study.

It was discussed in the section on statistics (3.2.7) that the linear mixed effect model used to calculate differences included four factors: TE speakers vs. NL speakers, voiced vs. voiceless, vowel type, and place of articulation. Results will be discussed per factor. Tables 3.3a and 3.3b show the differences between voiced and voiceless plosives both for TE and NL speakers. Due to limited space, the mean scores in these tables are 'raw' values, meaning that effects for vowels and place of articulation have not been corrected for yet. The mean numbers are, however, illustrative for the differences between voiced and voiceless and between TE and NL speakers.

TE speakers versus NL speakers

Only a few variables showed an effect for speaker group. For the segmental variables (Table 3.3a), only phonation onset time (*pot*) showed an effect ($p<.01$), with longer *pot's* for TE speakers in the voiceless sounds. This finding is similar to what Saito et al. (2000) found for VOT that was longer in voiceless sounds for the TE than the NL speakers. They explain this by proposing that the high intra-oral pressure during [p] production induces an elongation of the burst and the volume of the neoglottis requires a longer duration to revibrate after the offset of vibration. The only difference between their data and the data from this study is that even though the TE speakers in this study show a longer burst in the voiceless sounds, it is not significantly longer than the burst of NL speakers.

As far as the calculated variables are concerned (Table 3.3b), an effect was found for the percentage of voiced frames (*percvoicedframes*) ($p<.01$) for the voiced sounds, HNR in the closure (*hnrcd*) ($p<.01$ for voiced and $p<.05$ for voiceless) and the HNR in the voiced frames (*hnrvoicedframes*) ($p<.01$ for voiced and voiceless). Both groups used *percvoicedframes* to distinguish between voiced and voiceless plosives, but TE speakers showed a significantly lower percentage for the voiced plosives (77% versus 98%), implying that TE speakers use voice to a lesser extent than NL speakers.

For the *hnrcd* TE speakers showed lower values for both the voiced and the voiceless plosives. This means that the quality of the voicing found is lower for the TE speakers than for the NL speakers. This is not unexpected, as it is well known that TE speech contains more noise than NL speech (e.g. Van As et al., 1998). For *hnrvoicedframes* TE speakers show lower values in both the voiced and the voiceless sounds. This is not surprising, considering the values for *hnrcd* are almost exactly the same as the values for *hnrcdvoiced*.

Table 3.3a: Mean values of the acoustic durational correlates for TE and NL speakers for the voiced (V+) and unvoiced (V-) plosives. Significance is given for the voiced-voiceless distinction (Sign. V+/V- in the last row of each block) and for the speaker groups (Sign. TE/NL in the last column).

Correlate	Voicing	TE mean (N=98)	SD	NL mean (N=60)	SD	Sign. TE/NL
durational correlates						
V1 (ms)	V+	206	56	189	58	NS
	V-	165	47	160	51	NS
Sign. V+/V-		*p<.01*		*p<.01*		
poffvowel (ms)	V+	11	21	0	0	NS
	V-	25	23	22	9	NS
Sign. V+/V-		*p<.01*		*p<.01*		
phonclosure (ms)	V+	59	37	99	39	NS
	V-	31	34	22	9	NS
Sign. V+/V-		*p<.01*		*p<.01*		
cd (ms)	V+	87	45	99	39	NS
	V-	135	59	143	43	NS
Sign. V+/V-		*p<.01*		*p<.01*		
cd+durb (ms)	V+	92	45	103	40	NS
	V-	149	58	153	44	NS
Sign. V+/V-		*p<.01*		*p<.01*		
durb (dB)	V+	5	6	5	5	NS
	V-	14	13	10	8	NS
Sign. V+/V-		*p<.01*		*p<.01*		
pot (ms)	V+	13	13	8	6	NS
	V-	33	17	19	9	p<.01
Sign. V+/V-		*p<.01*		*p<.01*		
V2 (ms)	V+	249	52	243	46	NS
	V-	219	51	240	55	NS
Sign. V+/V-		*p<.01*		*p<.01*		

Voiced versus voiceless

As can be seen in Table 3.3a+b all variables showed a significant difference between the voiced and voiceless plosives.

V1 was longer before voiced sounds, which is in line with the literature, both for TE and NL speakers. NL speakers show a difference of 29 ms which is in accordance with Slis & Cohen (1969). TE speakers show a difference of 41 ms. Slis & Cohen (1969) described the difference in length for voiced and voiceless plosives as length compensation, i.e., a preceding vowel and the closure together should be equally long for voiced and voiceless sounds. The shorter closure duration (*cd*) for voiced sounds is then compensated by a longer vowel duration. Slis & Cohen (1969) found this consistently in their data for Dutch.

Table 3.3b: Mean values of the acoustic calculated correlates/derived measures for TE and NL speakers for the voiced (V+) and unvoiced (V-) plosives. Significance is given for the voiced-voiceless distinction (Sign. V+/V- in the last row of each block) and for the speaker groups (Sign. TE/NL in the last column).

Correlate	Voicing	TE mean (N=98)	SD	NL mean (N=60)	SD	Sign. TE/NL
calculated correlates						
relpoffvowel (%)	V+	13	21	0	0	NS
	V-	20	23	17	10	NS
Sign. V+/V-		*p<.01*		*p<.01*		
relphonclosure	V+	72	37	100	0	NS
(%)	V-	24	27	17	10	NS
Sign. V+/V-		*p<.01*		*p<.01*		
percvoicedframes	V+	77	32	98	2	p<.04
(%)	V-	40	35	43	19	NS
Sign. V+/V-		*p<.01*		*p<.01*		
hnrvoicedframes	V+	7	4	17	6	p<.01
(dB)	V-	3	3	9	3	p<.01
Sign. V+/V-		*p<.01*		*p<.01*		
hnrcd (dB)	V+	7	5	17	6	p<.01
	V-	3	4	8	3	p<.05
Sign. V+/V-		*p<.01*		*p<.01*		
relbint (ratio)	V+	.22	.37	.17	.16	NS
	V-	.43	.48	.15	.07	NS
Sign. V+/V-		*p<.05.*		*p<.05*		

A significant effect was also found for the phonation offset of the vowel (*poffvowel*) and relative phonation offset of the vowel (*relpoffvowel*). However, these findings need more discussion: in the literature, phonation offset after a preceding vowel has been discussed both for TE and NL speakers. Even though NL speakers can adjust their vocal fold behavior at will, it is still common to find vibrations after a vowel. As Lisker & Abramson (1964) state: "Spectrograms suggest that the laryngeal oscillations of a preceding voiced environment may simply continue for a while even after the glottis has begun to open for a voiceless stop; these vibrations are so low in intensity that any auditory effect they might have by themselves seems to be marked out by the stop burst and the noise of turbulent air rushing through the glottis." It may be the case that when a TE speaker produces not so prominent a burst, the phonation offset may influence perception.

The fact that no difference was found between speaker groups for the preceding vowel duration (*V1*) is in contrast with the findings of Robbins et al., (1986) and Searl & Carpenter (2002) who did find a significant difference, although the TE speakers in this study do show higher values for *V1* than the NL speakers.

The problem with *poffvowel* is that it is hard, if not impossible, to measure in voiced consonants as the vibrations of the vowel are masked by the vocal murmur of the voiced consonant. This was clearly the case for NL speakers, where no offset point could be detected in the signal. Also some of the TE speakers did not decrease their intensity with more than 20 dB, but for the other TE speakers offset points were found for the TE speakers in voiced plosives. This can mean two things: there is no vocal murmur during the closure of voiced consonants or the preceding vowel was so loud, that a decrease of 20 dB still results in an audible signal. Strangely, TE speakers show a shorter *poffvowel* for voiced than for voiceless plosives, which is confirmed by *relpoffvowel*, which also shows a lower percentage for voiced plosives. Normally, the voiced consonant would be expected to be so loud that the 20 dB cut off point is never reached. As discussed earlier, however, part of this finding may be explained by a very high intensity of the preceding vowel.

In retrospect, Searl and Ousley's (2004) method of using the cessation of glottal pulses as the phonation offset point seems a better option, at least in theory. In practice, the problem remains that glottal pulses could often not be detected in the signal of TE speakers.

Both TE and NL speakers show (relatively) longer phonation times (*phonclosure*) in voiced plosives. Both groups use it to distinguish between voiced and voiceless plosives, but TE speakers show a lower percentage of phonation in the voiced plosives and a longer phonation in the voiceless plosives when compared to NL speakers. This suggests that on the one hand TE speakers have problems sustaining phonation in voiced plosives and on the other hand they have problems ending phonation in voiceless consonants.

The closure duration *(cd)* showed *a* difference between the voiced and voiceless sounds with a mean difference of 46 ms. This is a larger difference than the 28 ms Slis & Cohen (1969) described in literature for Dutch. Unfortunately not much information is given by the authors on the effect of place of articulation or on the vowels used, so an explanation of this difference cannot be given at the moment. The durations we found, are, however, comparable to durations found for English TE speech. Searl & Carpenter (2002) found a mean difference of 40 ms between voiced and voiceless sounds.

For the burst, longer burst durations (*durb*) were found for the voiceless sounds, which confirms literature (Slis & Cohen, 1969; Ernestus, 2000). The low mean duration of the burst in the voiced sounds is partly caused by the fact that sometimes no burst could be found and thus 0 ms was scored. Where the relative burst intensity (*relbint*) is concerned, a higher intensity was measured in the voiceless bursts. Slis & Cohen (1969) also studied burst duration and burst intensity. They found a mean difference of 15 ms between voiced and voiceless sounds, which is longer than the 9 ms that was found for TE speakers and the 5 ms for NL speakers in this study. The 10 ms as found by van Alphen & Smits (2004) come much closer to the findings of this study. They also found a greater intensity for the voiceless than for the voiced plosives. Slis & Cohen (1969) found that the amplitude of the voiceless burst was 6dB higher than the amplitude for the voiced burst. This is the case for TE speakers in this study, but not for the NL speakers. However, the burst intensity is expressed as ratio of the maximum intensity of the whole signal, which may explain differences. In addition, Slis & Cohen (1969) found that the intensity of the burst is lower than that of the vowel following, which is also the finding in this study.

Phonation onset time (*pot*) was longer in the voiceless sounds for both speaker groups. This finding shows that it is more difficult to start up a vowel after a voiceless sound than a voiced sound.

Both speaker groups also show a higher percentage of voiced frames (*percvoiced-frames)* in the voiced sounds. For NL speakers, a very high percentage of 43% was found for the voiceless plosives, which was unexpected. This high percentage was caused by two outliers who scored around 60%. This high amount of voicing, however, has not impaired their intelligibility.

A comparison between *percvoicedframes* with phonation time in the closure (*phonclosure)* seems logical. Phonation was considered to be either voiced or voiceless speech. In the following table, the percentage of phonation is compared with the percentage of voiced frames.

From Table 3.4 it can be seen that for NL speakers, in voiced sounds, the phonation was also always voiced. It can also be observed from Table 3.4 that the 20 dB cut-off point used to determine the relative phonation time does not mean that no phonation is present as the *percvoicedframes* has a higher percentage than the *relphonclosure* (and voiced frames always count as phonation).

When the quality of the voicing is considered, both speaker groups distinguish voiced and voiceless sounds by having a higher *hnrvoicedframes* for voiced sounds than for voiceless sounds. The same result is found for the *hnrcd*. This is logical as the voiced sounds will contain more voicing or at least more harmonic structure.

Table 3.4: Relative phonation time in the closure in percentage (r*elphonclosure*) and the percentage of voiced frames in the closure (*percvoicedframes)* for both speaker groups (NL and TE) in voiced and voiceless plosives.

	Speaker group	relphonclosure	percvoicedframes
Voiced	NL	100	98
	TE	72	77
Voiceless	NL	16.8	42.7
	TE	24	39.6

Vowel type

Both the preceding and the following vowel show an effect, with [a] being the longest vowel (*V1*: F(2,139)=68.45, p<.01; *V2*: F(2,139)=46.95, p<.01)). This is not unexpected because the open vowel [a] in Dutch is considered a long vowel, whereas the high close vowels [i] and [u] are considered short vowels.

A vowel effect was also found for *phonclosure* and *relphonclosure* with [a] showing the longest phonation offset: F(2,139)=10.88, p<.01 and F(2,139)=11.63, p<.01, respectively.

The *cd* showed longest durations for the vowel [i] (F(2,139)=3.93, p=.022). This was also found for the *cddurb* (F(2,139)=5.24, p<.01).

Durb also showed an effect for vowel type (F(2,139)=4.61, p=.012), with the vowel [a] showing the shortest burst. Where *relbint* is concerned, the vowel [u] showed the highest intensity (F(2,139)=10.64, p<.01).

The vowel [u] also had the longest *pot* (F(2,138)=8.13, p<.01). This can be explained by the fact that [u] is a vowel with low frequency energy and is susceptible to noise. These characteristics may elongate the noise in the *pot*.

For the *percvoicedframes* a vowel effect was present with [a] having the lowest and [u] having the highest *percvoicedframes* (F(2,138)=17.15, p<.01). The same effect for the vowel [a] was found for the *hnrvoicedframes* (F(2,118)=12.18, p<.01) and the *hnrcd* (F(2,118)=14.14, p<.01).

Place of articulation

An effect for place of articulation (labial plosives versus alveolar plosives) was found for *cd* (p<.01): labial plosives had the longest closure duration. When the total duration of the consonant (*durb*) is taken into account, the effect for place of articulation disappears.

An effect was also found for *durb* ($p<.01$), with the alveolar plosives showing the longest bursts. The place of articulation effect is in line with findings in the literature for Dutch initial plosives (Smits, 1995). Slis & Cohen (1969) did not find this effect for place of articulation.

No effect was found for place of articulation for *pot*. However, Van Alphen & Smits (2004) describe VOT of voiceless plosives as being the length of the burst, which includes everything up to the following phoneme, thereby measuring the same segment as this study did. They found an effect for place of articulation, which was not present in our NL speakers. Their mean VOT is about 5 ms longer than the *pot* for the NL speakers in this study.

3.3.2.3 Fricatives: results and discussion

For TE speakers, 116 out of a possible 132 fricatives were segmented (11 speakers x 4 fricatives x 3 vowels). The fact that 16 stimuli had to be discarded confirms earlier results that fricatives are difficult to produce for TE speakers, just as the fact that only 23 fricatives were perceived more than 80 percent correct (9 voiced and 14 voiceless). For NL speakers, all 60 fricatives were segmented (5 speakers x 4 fricatives x 3 vowels), of which 56 fricatives were perceived more than 80% correct (28 voiced and 28 voiceless).

For the TE speakers, the low number of stimuli included means that not all speakers contributed to the dataset and that even not all stimuli were included: [ufu] and [izi] were not perceived correctly often enough for inclusion. Only one speaker contributed to all the remaining stimuli and only five speakers in total were perceived more than 80 percent correct. Nevertheless, we believe that useful observations can be made from these limited data.

The following correlates were measured: preceding vowel length (*V1*), consonant duration (*cd*), following vowel length (*V2*), percentage of voiced frames in the closure (*percvoicedframes*), HNR of voiced frames in the consonant duration (*hnrvoicedframes*) and HNR of the whole consonant duration (*hnrcd*). The results are given in Table 3.5. Just as with the plosives, these are the raw measures, not corrected yet for any main effects that might be found. The results will again be discussed according to the four factors of the statistical model: TE versus NL speakers, voiced versus voiceless fricatives, type of vowel, and place of articulation.

Due to the low number of stimuli for TE speakers, one has to be careful when interpreting results: when no differences are found this might be because the sample was simply too small.

Table 3.5: Mean values of the acoustic correlates for TE and NL speakers for the voiced (V+) and unvoiced (V-) fricatives. Significance is given for the voiced-voiceless distinction (Sign. V+/V- in the last row of each block) and for the speaker groups (Sign. TE/NL in the last column).

Correlate	Voicing	TE mean (N=23)	SD	NL mean (N=56)	SD	Sign. TE/NL
durational correlates						
V1 (ms)	V+	211	62	207	60	NS
	V-	186	70	163	44	NS
Sign. V+/V-		*p<.01*		*p<.01*		
cd (ms)	V+	174	18	112	48	NS
	V-	198	43	181	54	NS
Sign. V+/V-		*p<.01*		*p<.01*		
V2 (ms)	V+	222	51	226	49	NS
	V-	239	63	218	56	NS
Sign. V+/V-		*NS*		*NS*		
calculated correlates						
percvoicedframes	V+	39	44	95	13	p<.01
(%)	V-	15	15	51	44	p<.01
Sign. V+/V-		*p<.05*		*p<.01*		
hnrvoicedframes	V+	1	2	15	7	p<.01
(dB)	V-	1	2	11	7	p<.01
Sign. V+/V-		*NS*		*p<.01*		
hnrcd (dB)	V+	2	3	15	6	p<.01
	V-	2	2	8	9	NS
Sign. V+/V-		*NS*		*p<.01*		

Table 3.5 shows that a difference between the two groups of speakers is only found for the pitch-related variables: percentage of voiced frames in the consonant (percvoicedframes), HNR of the voiced frames in the consonant (hnrvoicedframes) and HNR during the whole consonant duration (hnrcd).

TE versus NL speakers

For *percvoicedframes* TE speakers show a significantly lower percentage than the NL speakers, both for the voiced and voiceless fricatives. Just as with the plosives, an unexpected high percentage of 51% is found for the voiceless fricatives of NL speakers. This time, however, only two speakers show a percentage below 30%, while the rest shows voicing for over 50% of the consonant. One speaker even voices his whole voiceless fricative. Apparently, voicing in fricatives is a less important correlate for NL speakers, even though the values still distinguish between voiced and voiceless.

TE speakers do not seem to use *hnrvoicedframes* to differentiate between voiced and voiceless, whereas the NL speakers do, who show a higher HNR in the voiced fricatives. The speaker groups differ from each other significantly for both the voiced and voiceless fricatives. The very low values for TE speakers can be explained by the fact that TE speech already contains noise, which might be strengthened by the noise inherent to fricatives.

For *hnrcd*, the two speaker groups differ significantly from each other only for the voiced fricatives.

Voiced versus voiceless

For both speaker groups, the preceding vowel was significantly longer before a voiced fricative, with a difference in *V1* of 25 ms for TE and 44 ms for NL speakers. Slis & Cohen (1969) found a similar difference of 40 ms for their NL speakers. For the duration of the following vowel (*V2)*, no significant differences could be found.

For the consonant duration (*cd*), significant differences were found between voiced and voiceless fricatives for both speaker groups, with longest durations for the voiceless fricatives. When the differences found (24 ms for TE and 69 ms for NL speakers) are compared with values from literature, it can be seen that NL speakers differ from the speakers of Slis & Cohen (1969) with 19 ms (more) and that the TE speakers in this study show only half the difference as was found by Searl & Carpenter (2002). The 69 ms found for the NL speakers in this study may have been caused by the experimental set up and the overarticulation of some of the speakers.

Both speaker groups show a significantly higher percentage of voiced frames for the voiced than for the voiceless fricatives, though the absolute difference for the NL speakers is larger.

For *hnrvoicedframes* and *hnrcd*, TE speakers do not show a difference between voiced and voiceless, while the NL speakers do.

Type of vowel

Just as with the plosives, a vowel effect was found (*V1*: $F(2,63)=23.20$, $p<.01$; *V2*: $F(2,63)=39.86$, $p<.01$), with [a] being the longest vowel.

The same vowel effect was found for the *cd* with [a] having the shortest consonant duration ($F(2,63)=6.36$, $p<.01$). This can be explained by the length compensation, discussed in the section on plosives: as [a] is very long, the consonant is made shorter. For *hnrvoicedframes* and *hnrcd* a moderate vowel effect was present with [i] showing the lowest HNR ($F(2,56)=3.5$, $p<.05$).

Place of articulation

An effect for place of articulation (labial versus alveolar) could be found for only one variable: labial fricatives had the highest percentage of voiced frames ($p<.05$).

3.3.3 Conditional inference trees

In the previous sections on plosives and fricatives, it was shown that for plosives, all correlates differentiate significantly between voiced and voiceless. For fricatives, slightly different results are found: the duration of the following vowel (*V2*) does not distinguish between voiced and voiceless for both speaker groups and HNR of the voiced frames (*hnrvoicedframes*) and HNR of the closure (*hnrcd*) do not distinguish between voiced and voiceless for TE speakers. The other correlates do differentiate significantly. However, based on the data described above, it is not immediately clear which of the correlates is most important for conveying a correct voiced-voiceless distinction. Further analyses are necessary to find out which of the correlates is the prime cue. Even though an acoustic prime cue is not per se also the perceptual prime cue, it can be assumed that the acoustic prime cue found is also important in the perception of the voiced-voiceless distinction.

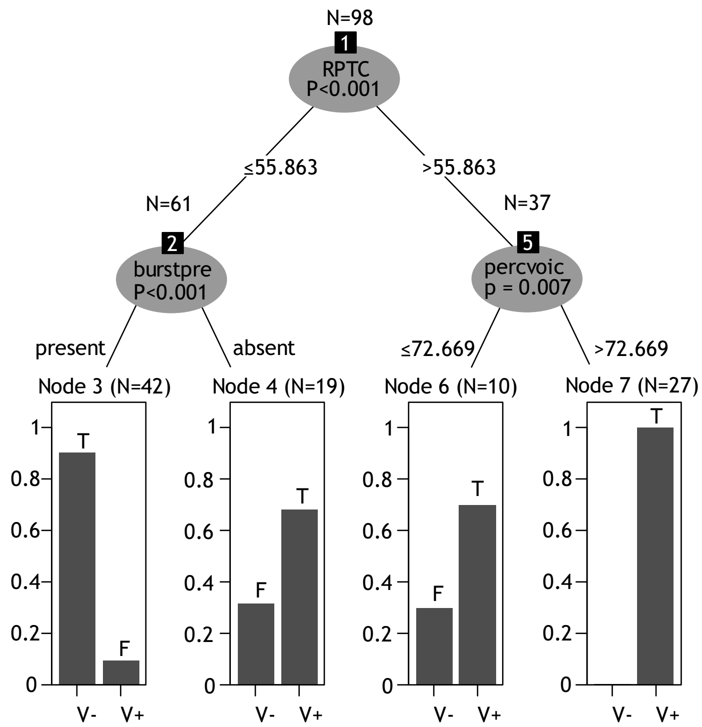


Figure 3.3a: The outcome for the 'conditional inference tree' for TE speakers, with all variables included. For each inner node, the Bonferroni-adjusted p-values are given, the actual number n of voiced and voiceless sounds is displayed for each terminal node. (RPTC=relphonclosure, burstpre=burst present, pervoic=percvoicedframes). T= true prediction; F=false prediction.

In this study, use has been made of a so called conditional inference tree. Conditional inference trees estimate a regression relationship by binary recursive partitioning in a conditional inference framework (for some more details see sect. 3.2.7 and Hothorn et al., 2006; 2006a). This method results in two trees for the plosives: one for TE speakers (Figure 3.3a) and one for NL speakers (Figure 3.3b) in section 3.3.3.1. In Figure 3.4, the tree for the fricatives is shown where no distinction is made between TE and NL speakers (section 3.3.3.2.).

3.3.3.1 Conditional inference tree for the plosives

It can be seen in Figure 3.3a that the relative phonation time in the closure (*relphonclosure)* is the most distinguishing correlate for TE speakers. If phonation is found in less than 55.9% of the closure duration, it is very likely that the stimulus is voiceless. If phonation is found in more than 55.9% of the closure, then a voiced stimulus is most likely. The cut-off point of 55.9% is related to the percentages for relphonclosure in Table 3.3b: 55.9% is about halfway between the 72% and 24% found for the voiced and voiceless plosives, respectively.

Additional correlates are needed for some of the stimuli to determine if they are really voiced or voiceless. For the voiceless stimuli this is the presence or absence of a burst. With the burst present, almost all stimuli in that node (N=42) are correctly labeled voiceless. Where the burst is absent, the plosive will most likely be voiced.

For the voiced stimuli, the additional correlate is *percvoicedframes*. When more than 72.7% of the closure is voiced, all stimuli in that category are labeled voiced. When less than 72.7% is voiced, 7 out of 10 stimuli in that node are still categorized as voiced, but 3 are wrongly labeled as voiceless. Again, the 72.7% is not surprising, considering the fact that the mean percentage of voiced frames in voiced plosives is 77%. Overall though, *relphonclosure* alone is sufficient to distinguish between voiced and voiceless.

Based on the conditional inference tree in Figure 3.3a, the accuracy of the decisions in the tree was calculated. The outcome of these calculations are shown in Table 3.6.

Table 3.6: The class predictions of the tree for the learning sample (and for new observations as well) when compared to the true membership in actual numbers.

	True	
Predicted	**Voiceless**	**Voiced**
Voiceless	38	4
Voiced	9	47

From Table 3.6 it can be seen that the accuracy is 80.9% (38 of 47) for the voiceless stimuli and 92.2% (47 of 51) for the voiced stimuli. The overall accuracy for this tree is 86.7% (85 of 98), with a 95% Confidence Interval of 78.4-92.7.

For NL speakers the following tree was generated:

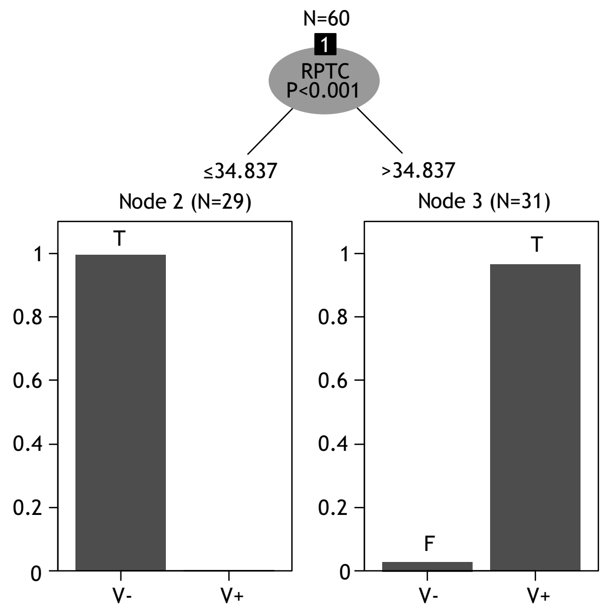


Figure 3.3b: The outcome for the 'conditional inference tree' for NL speakers, with all variables included. For the inner node, the Bonferroni-adjusted p-values is give, the actual number n of voiced (V+) and voiceless (V-) sounds is displayed for each terminal node (RPTC=relphonclosure). T=true prediction; F=false prediction.

For NL speakers, the same prime cue was found as for TE speakers. The division is even more clear-cut than for TE speakers. *Relphonclosure* is the only variable needed to distinguish voiced and voiceless plosives correctly. Only one stimulus from a total of 60 was wrongly identified. The fact that *relphonclosure* is the only correlate needed can probably be explained by the fact that the NL group is more homogeneous. The percentage *relphonclosure* on which the split is based is somewhat lower than for TE speakers: 34.8% instead of 55.9%.

The fact that *relphonclosure* is the main distinguishing correlate is interesting. It is a correlate based on the phonation offset, but in its current form it has not been discussed or even measured before in literature. This correlate, just as *poffvowel* and *relpoffvowel*, is based on intensity, rather than pitch. The cut-off point for phonation in the closure was set at the point where the intensity was 20 dB lower than the average intensity of the previous vowel. Where no cut-off point could be indicated, it was assumed that phonation occurred during the whole closure. As was discussed, this phonation could be either voiced or voiceless. Based on what was found in the tree analysis, it seems that the presence or absence of sound in the closure determines whether a phoneme is perceived as voiced or voiceless, respectively. This finding also implies that complicated pitch measurements are not necessary to classify a sound as voiced or voiceless. However, it needs to be kept in mind that the phonation time in the closure was very similar to the percentage of voiced frames in the closure (Table 3.4). Indeed, if *relphonclosure* is taken out of the analysis, *percvoicedframes* is then considered to be the most distinguishing correlate. Whether *relphonclosure* is still considered the main cue if the phonation is voiceless should be investigated in a further study.

In Table 3.7, just as for TE speakers, the class predictions of the tree for the learning sample (and for the new observations as well) are given when compared with true membership. This table shows that the accuracy for the voiceless plosives is 96.7%, whereas the prediction for the voiced plosives is 100%. The overall accuracy for this tree is 98.3%, with a 95% Confidence Interval of 91.1 - 100. For the NL speakers, the model is even more accurate than for the TE speakers.

Table 3.7: The class predictions of the tree for the learning sample (and for new observations as well) when compared to the true membership in actual numbers.

	True	
Predicted	**Voiceless**	**Voiced**
Voiceless	29	0
Voiced	1	30

3.3.3.2 Conditional inference tree for the fricatives

For the fricatives, the same type of conditional inference tree was generated as for the plosives. This time it was not possible to distinguish between TE and NL speakers, because there were too few stimuli that could be included for the TE speakers.

Figure 3.4 shows the results for both speaker groups combined. It can be seen that even though only one distinguishing correlate was found, results are not as clear-cut as for the plosives. If the consonant duration (*cd*) is shorter than 115 ms almost all stimuli in that group are considered to be voiced. If the duration is longer than 115 ms categorization is more difficult: of the 60 stimuli in this group, almost 70% is voiceless, and about 30% is wrongly categorized as voiced. When we compare the 115 ms with the durations in Table 3.5, it can be seen that the 115 ms matches the findings for the NL speakers more closely than for the TE speakers. The fact that *cd* is found as prime correlate is not surprising, as literature has also found this to be the most salient cue for the voiced-voiceless distinction in fricatives (Lisker & Abramson, 1964; Slis & Cohen, 1969).

Just as for the plosives, the accuracy of the tree predictions are calculated. In Table 3.8, the class predictions of the tree for the learning sample (and for the new observations as well) are given in comparison with true membership for the TE and NL speakers together. This table shows that the accuracy for the voiceless fricatives is 97.7%, but that the accuracy for the voiced fricatives is only 50%. The overall accuracy is 76.0%, with a 95% Confidence Interval of 65.0-84.9. This finding matches the perceptual finding that voiced fricatives are often perceived as voiceless.

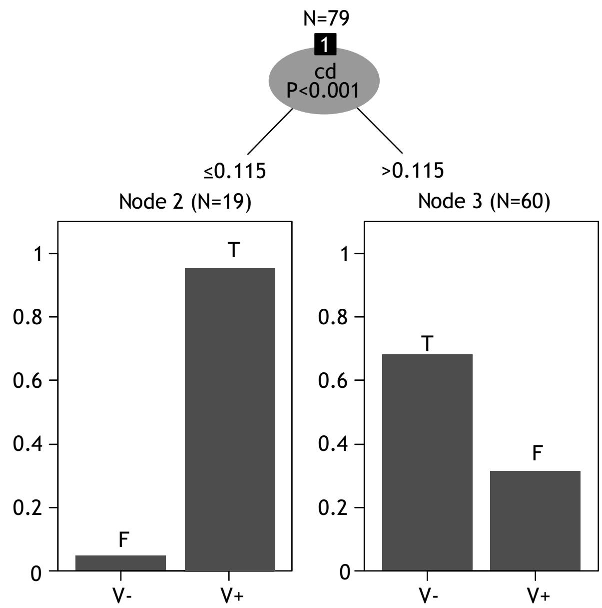


Figure 3.4: The outcome for the 'conditional inference tree' for TE and NL speakers, with all variables included. For the inner node, the Bonferroni-adjusted p-values are given. The actual number N of voiced (V+) and voiceless (V-) sounds is displayed for each terminal node.

Table 3.8: The class predictions of the tree for the learning sample (and for new observations as well) when compared to the true membership in actual numbers

	True	
Predicted	**Voiceless**	**Voiced**
Voiceless	42	18
Voiced	1	18

3.4 General discussion and conclusion

Perceptual results showed that TE phoneme intelligibility is significantly compromised. Multiple confusions were found, of which the most common confusion was between voiced and voiceless plosives and fricatives. This type of confusion is not unexpected when one considers the changed anatomy and physiology after a total laryngectomy. The neoglottis is not as pliable as the glottis and vibrations are harder to initiate or end voluntarily. Studying the voiced-voiceless distinction in more detail should provide information about the functioning of the neoglottis and about compensating strategies that may be employed by TE speakers. This information should be highly valuable for therapeutic purposes as well (see chapters 4-7). That is why we performed an acoustic study to investigate the voiced-voiceless distinction in cognate pairs of plosives and fricatives.

In the introduction, it was hypothesized that TE speakers would exaggerate (certain segmental) correlates to convey a voiced-voiceless contrast and that this would distinguish them from NL speakers. The results do not confirm this hypothesis, as for the segmental durations only phonation onset time shows a significant difference between TE and NL speakers, and only for the voiceless plosives. The explanation offered for this was the elongated burst and the fact that it takes TE speakers longer to start up phonation than NL speakers due to the different anatomy and physiology of the neoglottis and slightly altered vocal tract. However, the expectation that actual voicing (pitch) would be a problem was confirmed: TE speakers show a lower percentage of voiced frames in voiced plosives than NL speakers (the unexpectedly high percentage of voiced frames in voiceless sounds for NL speakers (42.7%) was caused by two outliers). The correlate HNR in the voiced frames is related to the quality of the voicing in the voiced frames. For this correlate, the speaker groups differ significantly as well: TE speakers show a lower HNR both for the voiced and voiceless plosives. Related to the voiced frames in the closure is the HNR in the closure. Voiced sounds have a better HNR, which was also found for both speaker groups, but the TE speakers perform worse than the NL speakers, especially in the voiced plosives. For the fricatives, similar results are found: TE speakers score significantly worse than NL speakers for percentage of voiced frames and HNR in the voiced frames, both for voiced and voiceless fricatives. For the HNR in the closure, only a difference for the voiced sounds was found. Again, NL speakers show an unexpectedly high percentage of voiced frames, which is caused by three of the five speakers. At the moment, there is no explanation as to why so much voicing is present in the voiceless fricatives.

The results mentioned above suggest that TE speakers have more difficulty employing actual voicing as a distinguishing correlate than NL speakers. Also the quality of the voicing is poorer than that of NL speakers. It does not mean, however, that TE speakers do not make use of these correlates to make a correct voiced/voiceless contrast. Based on the results TE speakers seem capable, to a greater or lesser extent, to employ voicing at appropriate times, at least for the plosives. In the general introduction (chapter 1) it was mentioned that studying the voiced-voiceless contrast in more detail might shed more light on the question if TE speech is myoelastic-aerodynamic or only aerodynamic. The fact that TE speakers are capable of using periodicity as a distinguishing correlate suggests the presence of a myoelastic component. However, we have only analyzed the correctly perceived stimuli. As was mentioned above, even though TE speakers use periodicity, they are less consistent than NL speakers. This implies that even if a myoelastic component is present, it is not present in all speakers and not always used consistently in speakers where it is present. These findings confirm earlier findings by Moon & Weinberg (1987), Mohri et al. (1994) and Deschler et al. (1999).

The difference between speaker groups was discussed above. When the differences between voiced and voiceless plosives are taken into account within the separate speaker groups, it is shown that TE speakers showed significant differences between voiced and voiceless plosives for all acoustic correlates and in that respect do not differ much from NL speakers. This was contrary to expectations. TE speakers were expected to have less correlates available to them to distinguish between voiced and voiceless sounds. The fricatives, however, show a different pattern: TE speakers only use two correlates (initial vowel duration and consonant duration) to distinguish between voiced and voiceless fricatives, whereas NL speakers use all six (preceding vowel, following vowel, consonant duration, percentage of voiced frames in the consonant, HNR of the voiced frames, HNR of the consonant).

A word of caution is necessary for the fricatives. As stated in paragraph 3.2.3, only very few fricatives were perceived more than 80% correct. The differences that are found are truly significant, but whenever no differences were established, this could easily have been caused by the limited number of stimuli available. The low intelligibility scores for fricatives do confirm that this category of consonants is hard to produce, not only for TE, but also for NL speakers, who showed lower scores as well.

It also needs to be kept in mind that overall very high standard deviations were found (see tables 3.3 and 3.5), both for the TE and NL speakers. Especially TE speakers are a heterogeneous group and the speakers do not all employ the same set of correlates to make the distinction between voiced and voiceless. This heterogeneity may also explain why sometimes differences were found between results from the literature and the results in this study. Also the fact that most other studies were performed on other languages, rather than Dutch, explains some of the discrepancies found, also for NL speakers. Finally, not all literature described effects of place of articulation or vowel type. This also made it difficult to compare results.

This study was first of all an acoustic study and it appeared to be difficult to link the acoustic results to the perceptual results. In the present set up, finding an acoustic correlate does not necessarily mean it is perceptually relevant, especially since it was found that almost all correlates distinguish between voiced and voiceless. In order to find the correlate that is most distinctive, 'conditional inference trees' were generated. This method found that *relative phonation time in the closure* was the main cue for both TE and NL speakers for the plosives. This result implies that the presence or absence of sound (periodic or aperiodic) is a sufficient cue to categorize a plosive as voiced or voiceless, respectively. As far as I am aware, no one has ever measured this correlate. Therefore, no studies are known in which the perceptual relevance of this cue are investigated. As mentioned in this paper, the percentage of phonation was highly similar to the percentage of voicing. Voicing has been found to be a main cue in literature. In the future, an investigation evaluating the impact of the relative phonation time on perception is desirable, whereby the amount of voicing should be taken into consideration as well.

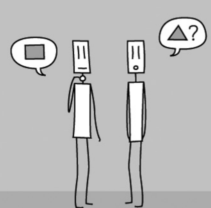
For fricatives, it was found that the *duration of the consonant* was the main cue. This is not a surprising finding as it has been discussed in literature before (e.g. Slis & Cohen, 1969).

It would also be interesting to compare very poorly perceived stimuli that are (almost) always confused with their voicing cognate with correctly perceived stimuli to see which acoustic correlates differ, but the present sample did not allow that comparison. There were too few phonemes that were misperceived 80-100% and when they were, the types of confusions were too diverse. In retrospect this study should have been set up differently, with fewer factors. Especially the vowels in this particular study showed effects regularly, which made interpretation of the results more complex.

In summary, it was established that TE speakers do not differ much from NL speakers. This was contrary to the expectations. In addition, TE speakers have many correlates available to them to produce a voiced-voiceless distinction. However, they are not very consistent in using them. The results found in this paper are relevant for therapeutic purposes, as it was so far not known how a correct distinction can be produced. By focusing on the burst, for example, speech training improved the production of voiceless plosives considerably (see chapters 4-7). More research is needed to establish how perceptually relevant the relative phonation time in the closure is and also to find out in what way poorly produced voiced and voiceless sounds differ from the correctly produced sounds.

4

Developing a therapy program to improve TE speech intelligibility



Abstract

After total laryngectomy, speakers are less intelligible than normal laryngeal speakers. This is partly caused by the fact that TE voice quality is generally poorer and also because surgery leads to changes in the vocal tract. Several studies, including our own perception experiment, have shown that there are typical problems at the phoneme and sentence level. However, these problems have not yet been addressed in an evidence-based therapy program that focuses on improving speech intelligibility. As speech intelligibility improves communication and hence quality of life, optimal intelligibility is important. In this chapter, a therapy program will be described, which is based on the results of our perception experiments. The aim was to develop a comprehensive training program based on published studies, including our own, on TE speech intelligibility and (evidence-based) TE speech therapy. The techniques and strategies that are incorporated in this program are explained and motivated and the structure and set-up of the sessions are described.

An adapted and Dutch version of this chapter has been accepted for the journal Stem-, Spraak-, Taalpathologie:

Van Rossum, M.A., Jongmans, P., Van As-Brooks, C.J. & Hilgers, F.J.M. Een therapie-programma voor het verbeteren van spraakverstaanbaarheid bij tracheoesofageale sprekers

4.1 Introduction

Several studies have found that intelligibility of tracheoesophageal (TE) speakers is poorer than that of normal laryngeal (NL) speakers, both for Dutch and for other languages (Doyle et al., 1988; Hammarberg et al., 1990; Nord et al., 1992; Miralles & Cervera, 1995; Roeleven & Polak, 1999; Oubrie, 1999; Boon-Kamma, 2001; Searl et al., 2001; Lundström & Hammarberg, 2004; Jongmans, 2006)

The TE speakers in our perception experiment showed an overall consonant intelligibility rate of 72% and a vowel intelligibility rate of 74% (chapter 2). This rate is much lower than the 93% consonant correct score for NL speakers as found by Pols (1983) for CVCVC nonsense words and the 84% vowel correct score as found by Koopmans-van Beinum (1980) for CVC words.

Considering the different anatomy and physiology of the neoglottis, as compared to the glottis, problems with voice quality and intelligibility can be expected (Van As et al., 1998). The instability of the neoglottic vibratory pattern causes a less clear harmonic structure in TE speech and undetectable or highly irregular periodicity patterns. Furthermore, TE speakers with little control over F0 did not convey accent as accurately as F0-users (Van Rossum et al, 2002). Problems with intonation and timing were also found (Roozen, 2005). McColl (2006) concluded that a laryngectomy also alters the form of the vocal tract, as during the operation the hyoid bone is removed, which changes the position of the base of the tongue. Due to the variable positions of the neoglottis, the length of the vocal tract is altered as well. These changes most probably impact the precision with which TE speakers produce certain phonemes, thus compromising articulation. As a consequence acoustic properties in TE speech are changed, which may lead to a decrease or deviancy in a number of acoustic-phonetic cues that listeners rely on during speech identification (see chapter 3).

To communicate, one needs to be intelligible. Optimizing intelligibility will improve communicative ability (e.g., conversing over the telephone), independence, and ultimately quality of life (Ackerstaff et al., 1994).

Despite the evidence that TE speakers have problems with a number of phonemes and features, as well as with intonation and timing, speech language pathologists (SLPs) do not routinely provide the patient with a rehabilitation program that covers both voice and speech quality. At present, the main goal is usually to establish an audible voice of acceptable quality. Also, prosthesis care (deservedly) receives much attention during therapy. This contrasts with the training of electrolarynx (EL) and esophageal (E) speech, during which articulation does receive more attention. Studies have compared the different speech modes and found TE intelligibility more favorable than EL or E speech (e.g. Williams & Watson, 1987; Pindzola et al., 1988; Debruyne et al., 1994

& Max et al., 1996). It might be that TE speech is perceived to be so much more intelligible than electrolarynx and esophageal speech, that it causes a shift in the SLP's focus of training from articulation training to prosthesis use/care and stoma occlusion. Additionally, busy hospital schedules and limited resources might prevent SLPs from investing extra time in intelligibility training. Many researchers and SLPs throughout the world, however, recognize that there is more to be gained by the patient through specific intelligibility training. To quote Searl (2004): "I simply want to remind myself and others not to settle for merely usable speech, but rather to work towards the best speech possible for each laryngectomee." In the same article he expresses surprise that about 10% loss of intelligibility in stroke patients is considered serious enough to warrant training whereas professionals are "content to let an intelligibility deficit of this magnitude (or greater!) go unaddressed in the laryngectomee patient."

Apart from the abovementioned reasons there is another explanation why intelligibility in TE speech might be neglected. For Dutch, no large-scale studies analyzing TE speech intelligibility had been performed so far. Thus, there was no experimental evidence that intelligibility might be compromised. Without this evidence, there would not be much incentive to develop an (evidence-based) therapy program. However, results in chapter 2 have shown that a rationale for developing an evidence-based therapy program is obvious.

The goal of our project is to develop a program that has the potential to improve intelligibility in TE speech. The typical TE speech problems catalogued in chapter 2 and elsewhere in the literature form the basis of this program. Optimally, a therapy program should consist of those evidence-based strategies and techniques that address the problems of the population in question. Section 4.2 describes the literature search that was executed to find relevant therapy strategies and techniques. Section 4.3 describes the content and structure of the therapy program.

4.2 Literature Search

The aim of the literature search was to find information on therapy or treatment to improve speech intelligibility in alaryngeal speech. As so far only very few papers were known to address improvement of speech or intelligibility in alaryngeal speakers, a systematic literature search was performed.

4.2.1 Method

Literature databases Pubmed, PiCarta, DocOnline and AHRQ, were searched. The key words included were: "laryngectomy" "intelligibility" "speech" "training" "therapy" “rehabilitation”, and “improvement”. In Figure 4.1, the search combinations are shown. To find potentially overlooked studies, the references of articles were hand searched and local experts were contacted. Studies in Dutch, English and German were included. Papers or books were considered useful if the patient category included alaryngeal speakers and some sort of therapy to improve speech / intelligibility was described.

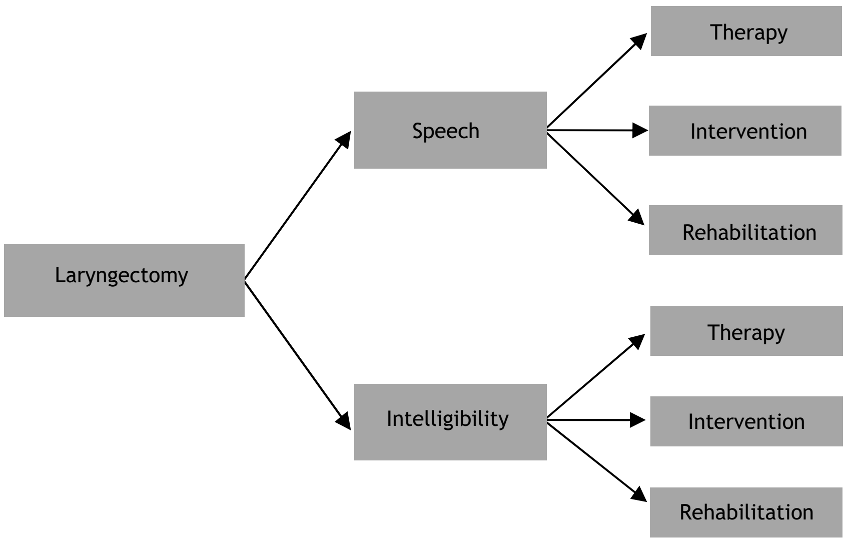


Figure 4.1: Combination of search terms used for finding literature on TE speech intervention studies.

4.2.2 Results

The results of the search are categorized in Table 4.1 according to type of publication (we chose “evidence-based” or “experience-based”, the latter referring to training manuals).

As can be seen in Table 4.1, seven evidence-based studies were found, but only two of them concerned TE speakers. Fitzpatrick et al. (1980) employed self-administered home practice for consonant training and found that both naïve listeners and the speakers themselves showed better discrimination scores after training. Mase-Goldman et al. (1988) trained back consonants and clusters in esophageal speech, using the same self-administered home practice Fitzpatrick et al. used. Comparison of pre and post test scores showed significant improvement. Zeine & Brandt (1988) investigated the effect of noise on the intensity of esophageal speech and electrolarynx speech. If a positive effect would be found, it could be used for rehabilitative purposes. Although they did find an increase of intensity for the esophageal speakers, it cannot be concluded that this increase is perceptually relevant. Christensen & Dwyer (1990) described a technique that increases oral pressure by giving the instruction 'pushing harder'. They found that this improved the articulation of voiceless consonants in electrolarynx speech and esophageal speech. Belloc et al. (1997) found that a general training of voiced and voiceless plosives and fricatives also improved the intelligibility of these phonemes. De Maddalena & Pfrang (1993) discussed a psychological approach to change the speaker's attitude towards his or her speech. This was part of a comprehensive rehabilitation program and rendered positive results, with improved communication behavior. Wong Chung et al. (1998) discussed their intensive, multidisciplinary post- laryngectomy training program. This program focused on voice production in the period immediately after the operation rather than on typical problems found in TE speech. Their study did show that intensive rehabilitation improved voice quality and also increased speech rate significantly. There are also some experience-based programs. Drost (1978) wrote a book on improving Dutch esophageal speech. Rood-Nieuwland (1982) described group lessons for esophageal speech. Bors et al. wrote an exercise book, mainly meant for use in their own hospital. Van der Meulen (1990-I+II) also wrote a reference work containing many exercises for esophageal speakers, which, in the same series, was followed by a book by Postma (1997) for TE speakers. For TE speech, Searl (2005) described the general problems found in TE speech and suggested therapy techniques. The efficacy of the experience-based programs has never been evaluated.

Table 4.1: Results of the literature search, author and date of publication, population (E=esophageal, EL=electrolarynx, TE=tracheoesophageal) with number of subjects (N) if stated, outcome and whether the publication was evidence-based or experience based.

Author and date	Population (N)	outcome	Type of publication
Fitzpatrick et al. (1980)	E (8)	improvement	evidence-based
Mase-Goldman et al. (1988)	E (7)	improvement	evidence-based
Zeine & Brandt (1988)	EL and E (18)	improvement	evidence-based
Christensen & Dwyer (1990)	EL and E (8)	improvement	evidence-based
De Maddalena & Pfrang (1993)	EL, E and **TE**	improvement	evidence-based
Belloc et al. (1997)	**TE**	improvement	evidence-based
Wong Chung et al. (1998)	EL and E (125)	improvement	evidence-based
Drost (1978)	E	unknown	experience-based
Rood-Nieuwland (1982)	E	unknown	experience-based
Bors et al. (unknown)	**TE**	unknown	experienced-based
vd Meulen (1990,I+II)	E	unknown	experienced-based
Postma (1997)	**TE**	unknown	experience-based
Searl (2005)	**TE**	unknown	experience-based

Many of the existing speech training techniques described in the literature above are incorporated in our training program. However, these techniques address only part of the speech problems typical in TE speech. The literature search, then, did not provide sufficient information to develop an all-round therapy program. As was mentioned in the introduction, phrasing and effective use of sentence accent, for example, also seem to be affected. Furthermore, techniques taught at the phoneme level should ultimately be carried over to the discourse level, if overall intelligibility of TE speech is to be improved. For this reason, aspects of prosody (accentuation and timing) are added to our therapy program, as well as an approach that is used in audiology, namely teaching clear speech. This "clear speech" approach stimulates carry-over to discourse level and is evidence-based (Picheny et al. 1986; Cutler & Butterfield, 1990; Uchanski et al. 1996; Krause & Braida 2002; Krause & Braida 2004).

The next section describes our therapy program.

4.3 Therapy Program

In this section, the content and structure of the therapy program is discussed. The entire Therapy Program, translated into English, can be found in appendix 4.1. In the following section, typical problems that are covered in the program will be described in more detail. Thereafter, the training techniques for specific phonemes/features will be discussed, followed by techniques for specific problems at the sentence level. Finally, the general techniques, auditory feedback and clear speech, will be discussed. Efficacy results will be presented in the next chapters.

4.3.1 Aspects included in the therapy program

In Table 4.2 the most common errors in TE speech are given.

Table 4.2: Summary of typical problems that affect intelligibility in TE speakers.

Level	Type of problem
Phoneme + word	Incorrect voicing, nasals, liquids, /x/ and /h/
Sentence	Inappropriate accent and pausing
Discourse	Omission of syllables and phonemes in running speech, especially at end of phrase/ sentence

The most dominant error we found was the voicing distinction (see chapter 2), which affects the plosives and fricatives, and also consonant clusters with a voicing transition. For plosives, confusions from voiceless to voiced were most often seen, whereas for fricatives confusions from voiced to voiceless were also very common. Besides voicing, there were also problems with nasals, the back fricatives /x/ and /h/ and the liquids /l/ and /r/. For this last group, the pattern of confusion was less clear: confusions seemed to occur across the board. From this information it seems reasonable that a therapy program aimed at improving intelligibility in TE speech should incorporate the above-mentioned features.

However, training consonant production alone is most likely not sufficient to improve intelligibility of spontaneous speech. Preliminary results of a pilot study performed on material of Roozen (2005) showed that TE speakers did not use accents optimally and tended to pause at inappropriate times (personal communication, van Rossum, 2007). These factors will also contribute to a diminished intelligibility. A pause provides listeners with extra time for perceptual processing, and as Nooteboom et al. (1990) say: "… speech pauses are first of all interpreted by listeners as cues to syntactic structure. When pause positions conflict with other cues to syntactic structure, listeners become confused and recognition is hampered". Sanderman & Collier (1997) gave a similar explanation for prosodic structures and concluded that listeners use the perceived prosodic information to compute the meaning of the input speech. Inappropriately phrased utterances take longer to process as reanalysis of earlier speech material is often necessary in order to comprehend the sentence. Therefore, in the therapy program attention is also given to sentence accent and phrasing (mostly proper timing). The design of this therapy program allows for a progression from phoneme to discourse level.

4.3.2 Specific: phoneme/feature training

The literature revealed that certain therapy techniques are expected to improve production of some of the problem features found in TE speakers. Most of these therapy techniques aim at strengthening perceptual cues that are still available to TE speakers or stimulating the use of secondary perceptual cues. In this section, specific techniques are matched with specific phoneme problems.

4.3.2.1 Voiced-voiceless distinction in plosives and fricatives

In normal laryngeal (NL) speech, voice onset time is one of the most important cues for distinguishing between a voiced and voiceless sound. This cue, however, is often difficult to produce for a TE speaker, who has less control over the onset and offset of vibration of the neoglottis. In chapter 3 it was described that TE speakers indeed differ significantly from NL speakers on the acoustic correlates based on pitch. Onset and offset of vibrations are more difficult to control in the neoglottis, the latter causing a long *Phonation Onset Time* (POT; similar to VOT) for TE speakers for voiceless plosives.

This is also found by Saito et al. (2000) who state that: "These characteristics of high intra-oral pressure and voluminous neoglottis are assumed to be responsible for the extension of VOT in TE speakers". This intra-oral pressure (a build-up of pressure behind the point of maximum constriction) is important in producing another important perceptual cue: a higher-intensity noise burst for /p t k/ than for their voiced cognates (e.g. Slis & Cohen, 1969). Christensen & Dwyer (1990) state that "Pushing harder on voiceless stops effectively increases intra-oral air pressure and provides a sharper burst and more aspiration noise capable of overriding an inappropriate voice onset time cue". They used this 'push-harder' technique for esophageal speakers and found that this approach rendered more intelligible speech. However, the aerodynamics of esophageal speakers is different from TE speakers, as esophageal speakers cannot use air from the lungs and have a relatively small esophageal air reservoir. One has to be cautious, then, when transferring techniques to TE speakers. Searl (2002) later investigated oral pressure in TE speakers and found that good TE speakers were capable of employing high intra-oral pressure. He suggests that the push-harder technique will also be effective for TE speakers who have problems producing a proper voiced-voiceless contrast.
Similarly, extending the duration of the frication noise will cue voiceless fricatives (Slis & Cohen, 1969). The techniques used for the plosives and fricatives in our therapy program teaches TE speakers to increase intra-oral pressure (thus producing a higher intensity noise burst) and to lengthen the duration of frication noise. Note that the emphasis in this program was on training of voiceless plosives and fricatives, not on the production of the voiced counterparts.

4.3.2.2 nasals and /l/

Techniques from a program aimed at improving esophageal speech (Drost, 1978) are incorporated in our program in order to improve the production of the /l/ sound and the nasals. According to this program, salience of /l/ may be improved by producing a slightly more retroflex quality. For the nasals, no specific 'tricks' exist to improve articulation, except using more force: if one presses the lips together with more force and for a longer period of time, /m/ will be more intelligible, whereas pressing the tongue against the alveolar ridge with more force and for a longer period of time will help improve the intelligibility for /n/.

4.3.2.3 /x/ and /h/

Drost (1978) also proposed techniques for the production of /x/ (as in Dutch 'la*ch*' or '*g*eef'). He said this phoneme might be more easily taught using either vowels or /k/. In the program /x/ is first practiced in final position, preceded by a front vowel in order to elicit a more fronted articulation of /x/. If this poses difficulties for the speaker, another technique is to start from the phoneme /k/ and then slowly release the constriction. A third technique for Dutch speakers is to use the plosive /ɣ/ as a starting point and then move the articulators forward during production. One or more of these techniques can be applied during therapy.

The phoneme is practiced in final position preceded by a front vowel, or in initial position if the technique with preceding /k/ is used.

Two techniques are included to train /h/: subjects are either taught to use a hardly audible /i/ or a very soft prolonged /ɣ/.

4.3.3 Specific: sentence level

Besides problems with specific phonemes and features, problems at sentence level also occur, with accent and phrasing causing most problems.

4.3.3.1 Accentuation and phrasing

An experiment on TE speech produced in silence and in noise revealed that TE speakers display within-phrase pausing quite frequently (Roozen, 2005). However, not only do TE speakers pause within phrases, pauses within words also occur. Pauses that are not located at prosodically motivated positions in a sentence are known to have a negative effect on speech recognition (e.g. Scharpff & Van Heuven, 1988 and Van Heuven & Scharpff, 1991) and poorly phrased utterances slow down a listener's speech processing (Sanderman & Collier, 1997). The Therapy Program therefore includes instructions and material to increase awareness, and practice proper phrasing. The speakers are also made aware, using minimal pair sentences differing only in place of accent, that correct placement of accent is important for a good understanding.

Apart from specific techniques at phoneme and sentence level, general techniques are also incorporated to facilitate learning new techniques and to allow carry-over to discourse level. These general techniques are described in the next section.

4.3.4 General techniques

4.3.4.1 Auditory feedback

Several studies in the literature describe the beneficial effects of auditory feedback on the intelligibility of the speaker. Auditory feedback here means that speech is recorded so that speakers can listen to it attentively later. Mase-Goldman et al. (1988), for example, describe this method for esophageal speakers where recordings were played back to the speakers, who scored and noted their errors. Auditory feedback was also found to be helpful by talkers participating in a study that was aimed at eliciting clear speech (Krause & Braida, 2002). Based on these findings, we have incorporated auditory feedback throughout the intervention, both at the phoneme and discourse levels. Recordings are used mainly to illustrate problems or to increase awareness of the speaker's own speech and are not a main part of the sessions.

4.3.4.2 Clear speech

Some speakers are easier to understand than others, even in identical listening conditions. Bond & Moore (1994) identified acoustic-phonetic aspects that differentiate less intelligible from more intelligible speech. This study found that the less intelligible speaker had shorter word and segment durations, a reduced vowel space (and therefore lack of clearly distinguished vowels), elimination of the noise bursts in plosives, lack of systematic pausing and rapidly decreasing vocal effort towards the end of a sentence. Furthermore, the more intelligible speaker made more effective use of pre-boundary lengthening and distinguished sentence-internal from sentence-final pausing. This knowledge is, however, only useful if one can improve these aspects, and thus the intelligibility, through training.

Lombard already discovered in 1909 that speakers adapt their speech in order to be understood in noise (Lane & Tranel, 1971). This and subsequent studies found that a speaker takes into account the need of the listener and will therefore compensate for changes in the signal-to-noise ratio. This behavior is described in the H&H (Hyper and Hypo speech) theory (e.g. Lindblom, 1990; Summers et al., 1988). If, however, a speaker is asked to read aloud in noise, but does not realize that his speech will serve a communicative purpose, the acoustic-phonetic adaptations, and therefore the effect on intelligibility, will be marginal (e.g. Lane & Tranel, 1971). Indeed, in a pilot experiment in which TE speakers (and one normal speaker) were asked to read a passage aloud in quiet and in added noise, but were not specifically instructed as to the communicative goal, the only significant result was the variation in intensity, which was greater in the condition with added noise (Roozen, 2005).

We have the strong expectation, based on personal observation, that TE speakers will not compensate sufficiently in certain noisy circumstances and that the H&H theory might not apply. TE speakers have an impaired speech apparatus, but if one considers their speech behavior (poor phrasing, deletions, especially in final position, poor use of accent and no final lengthening), it would seem that compensation is not automatic. In other words, these speakers do not spontaneously or instinctively adjust their post-laryngectomy speech to accommodate to the needs of the listener. In fact, as was discussed in chapter 3, TE speakers do not differ from NL speakers for the acoustic correlates measured. This might indeed suggest that TE speakers do not (over) compensate. We therefore believe they need extra help to learn to articulate clearly and improve overall speech intelligibility.

A number of studies have investigated production and perception of *deliberately* clear speech (Cutler & Butterfield, 1990; Picheny et al., 1986; Uchanski et al., 1996; Krause & Braida, 2002; Krause & Braida, 2004) as opposed to the subconscious adaptations in noisy environments as described by the H&H theory. These studies found intelligibility of deliberately clear speech to be superior to normal conversational speech. Added noise is not a prerequisite for the production of clear speech. Clear speech has certain phonetic-acoustic characteristics, such as increased number and duration of pausing, a slower articulation rate, higher F0 peaks in the intonation contour, more energy in the mid-to-high frequencies, more power in e.g. nasals and the noise bursts of plosives, and an expanded vowel space. This last point is of special interest, because Roeleven & Polak (1999) and Jongmans et al. (2005) found problems with vowel intelligibility in TE speech, which was also found in a study that revealed shifts in formant frequencies (Oubrie, 1999). Thus, clear speech potentially addresses a number of issues that are known to affect the intelligibility of TE speakers.

Moreover, Krause and Braida (2002) explained a clear speech training procedure that resulted in increased intelligibility. Their training procedure is, therefore, implemented in the present Therapy Program, and is further extended by including instructions that are generally used to improve the speech intelligibility of people who routinely communicate with the hard-of-hearing.

Since training of clear speech is at the sentence and discourse levels, it further facilitates carry-over of the aspects that are taught at the phoneme level, and also provides an opportunity to improve phrasing in TE speech.

Instructions that stimulate clear speech are to:

1. slow down articulation (but not unnaturally slow)
2. produce more powerful speech (but without shouting)
3. pause regularly and at appropriate places in the sentence
4. articulate each sound as precisely and clearly as possible
5. speak lively and melodiously, making effective use of accent

Clear speech is *not* monotonous, exaggerated, staccato, artificial or too loud.

4.4 Structure of the speech training program

The speech training program consists of nine one-hour sessions. An identical program is offered to all participants. Due to limited capacity, a maximum of ten subjects could participate in this study. It was decided to divide the subjects into two groups of five and offer them sessions twice a week during five weeks.

There are several advantages of a limited-session therapy program:

1. (re)funding by medical insurance is guaranteed within the Dutch system
2. knowing the number of sessions beforehand simplifies logistics for speech language pathologists and patients
3. an identical program simplifies statistical analyses and interpretation of results. By including questions on compliance and a tally sheet, on which the speakers marked how often they had practiced, information on difficult aspects or stages in the therapy program should come to light. This allows a better understanding of the results and possible ideas on improvement of the program
4. the structure of the program allows one to highlight or concentrate on those aspects that individual patients have difficulty with throughout the nine sessions of the therapy
5. part two (clear speech) allows facilitation and carry-over of the aspects in part one (specific phoneme/feature training)
6. if the intelligibility has been significantly improved at the end of the program, it can be said to be evidence-based and cost effective, making it a useful tool for Dutch speech language pathologists working with TE speakers. Most likely, similar results will be found if a similar speech training program will be used in TE speakers speaking other languages, although each language may have its specific issue that need to be addressed in a speech training program

Therapy sessions

In Figure 4.2, the structure of the program is made visible by a flowchart. Part one (lesson 1-6) consists of training of the individual articulatory features. The second part (lesson 7-9) focuses on training of clear speech and phrasing. This design allows for a progression from phoneme to discourse level.

The individual sessions in part one (phonemes) have the following structure: auditory discrimination (only for plosives and fricatives), explanation of the techniques, phonemes in isolation, phonemes in one-syllable structure, phonemes in multi-syllable structure and phonemes in sentences.

In part two of the program (lesson 7-9) first an explanation of what clear speech means is given. After that, the subjects listen to examples of clear speech on CD and are asked to repeat the sentences in the same way. Auditory feedback is used to demonstrate what went well and what could be improved. Thereafter, the subjects are asked to read aloud sentences and short texts focusing on clear speech and correct phrasing. In session nine, speakers practice monologues and question and answer dialogues. During the last three sessions, there is also room to pay extra attention to phonemes that are still difficult to improve.

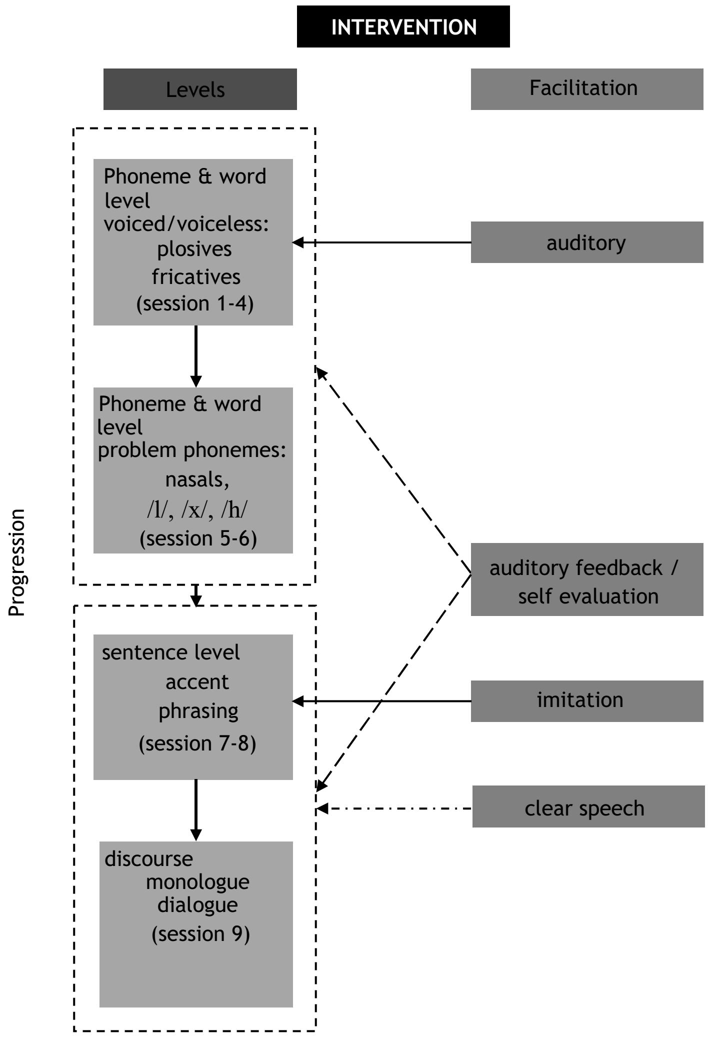


Figure 4.2: The structure of the program, with sessions on the left and the main general techniques on the right. The arrows indicate which technique is used in which part of the training.

Homework

Sessions in the clinic are mainly meant for instruction. The subject is expected to practice at home. She is advised to practice the exercises learnt in the last session for approximately ten minutes three times a day. Per session, practice sheets are available for the subject with exercises. The subject is asked to write down how many times and how long he practiced each week.

Evaluation forms on the progress and compliance of the subjects were included after every session for the SLP to fill out. Through these forms, we hoped to find out whether techniques were useful and easy to teach, whether the subject improved and whether the subject was compliant.

4.5 Conclusion

In chapter 2, it was shown that TE speech intelligibility is severely compromised. In this chapter, we have described the development of a speech training program that can potentially improve intelligibility of TE speakers. Because the literature review revealed very few papers that investigated evidence-based strategies to improve intelligibility, the Therapy Program described in this chapter also consists of a number of untested strategies and approaches. This poses us with two questions: does intervention improve TE speech intelligibility and is the current design of the program appropriate for training TE speech intelligibility.

To evaluate whether this specific therapy program was effective, a variety of measurements were performed, testing the (overall) intelligibility in nine TE speakers before and after therapy. Results of these tests are discussed in chapter 5 and 6. SLPs and the subjects themselves also gave a subjective impression of the structure and contents of the program. This evaluation is described in chapter 7.

Appendix 4.1 The Therapy Program (in English)

This is the program as it was handed out to the speech language pathologists. A Dutch version of the Therapy Program, including practice sheets, and a copy of the cd with clear speech exercises is available on request. Due to limited space, the practice sheets for homework are also not given here.

Therapy Program: Speech Intelligibility

Introduction

Aims

This study has two aims: first, to establish whether a structured intervention positively affects intelligibility; second, to determine if this specific program is useful.

Structure

This program consists of two parts. A: articulation training of specific phonemes (lesson 1 to 6) and B: teaching clear speech and optimal prosody (phrasing and use of accent; lesson 7 to 9). This program is short, but intensive. One lesson will be covered per therapy session. If, therefore, a speaker is unable to apply a specific technique which was, say, taught in lesson 4, the next session will still start with lesson 5. Part B of this program offers the opportunity to recapitulate and review problematic aspects which could not be covered adequately in previous lessons due to time constraints. The choice with regard to -- 1) sounds, 2) strategies (or techniques) and 3) number -- have been explained in "Improving intelligibility in Tracheoesophageal speakers".

Main aim of the therapy sessions is instruction and feedback, practicing should be done at home. Instruct the speaker to revise and practice the lessons covered during therapy sessions. The speaker should strive for three practices, of 10 minutes each, per day. Obviously, it is allowed to practice more often and longer. The speaker is asked to keep tally sheets (Appendix C). Appendix D contains more practice material to take home.

Short evaluations concerning progress and compliance have also been included (Appendix B). At the start and end of each lesson questions were asked. These questions were developed to gain insight into the usefulness of the strategies and techniques, the speaker's understanding and ability to apply the techniques that were taught in the previous therapy session, and compliance. This information is especially helpful if more than one SLP is involved, and one SLP has to take over from another SLP.

But this information will also help to (subjectively) evaluate the program. After this project has ended, you will be asked to evaluate the program by means of more extensive questionnaires. Your evaluation will allow us to further improve on this therapy program in the future.

Materials
You will be provided with a supply of practice, evaluation, and tally sheets. Recording equipment will also be provided.

General: please address stoma noise problems (evaluate occlusion / coordination). You might want to record the stoma noise while the speaker is speaking and play the recording back, so that he/she will be aware of its effect on the intelligibility and quality of speech. All participants' hearing will be screened beforehand, to identify hearing problems that might influence therapy.

Articulation training

There are three production levels (isolation, then progressing to monosyllabic words, etc). However, some speakers might find it easier to apply a technique at word level than in isolation. Be flexible; use your own judgment with each speaker.
Plosives (lesson 1 and 2)

1. Auditory discrimination plosives:

paas	-	baas	toek	-	doek
poos	-	boos	tak	-	dak
pes	-	bes	tek	-	dek
poot	-	boot	tik	-	dik
peet	-	beet	tas	-	das
pak	-	bak	toos	-	doos

2. Build-up of oral pressure:

Available air in oral cavity is 'held captive' through closure of lips and palate molle; the air is then put under pressure by decreasing the size of the oral cavity (movement of lips, cheeks, floor of mouth). This pressurized air is subsequently 'forced' out of the mouth. You can also instruct the speaker to "push down" on the voiceless sound (fricative or plosive).

Explain to the speaker that, although he now breathes through his tracheo-stoma, air is still present in his mouth. By puffing out each cheek in turn, for example, the speaker can be made aware of the presence of this air. The speaker has to learn that he is able to pressurize this oral air; and, after putting this air under pressure, that he is able to force this air out, either by using a plosive (/p/ or /t/) or a fricative (/f/ or /s/). Basically the air is allowed to escape, but very forcefully.

The /p/ sound is usually the easiest to start with. Ask the patient to prolong lip closure, using extra effort to keep lips together before releasing the air forcefully through the lips.

3. Production

Start with voiceless plosives (spend most of your energy and time on the voiceless sounds), then work on the voiced sounds.

Use auditory feedback to facilitate production: record the speaker's attempts, and together listen to the recording and judge what he has done.

3.1 In isolation :

P, T en K

3.2 cv vc:

p-a	aa-p	t-ee	ee-t	k-aa	aa-k
p-o	o-p	t-ui	ui-t	k-ui	ui-k
p-oe	oe-p	t-oe	oe-t	k-oe	oe-k

3.3 monosyllabic words:

p-ij-p	t-ij-d	k-ij-k
p-aa-p	t-aa-k	k-aa-p
p-ui-k	t-u-t	k-ui-p
p-aa-k	t-ij-k	k-ui-t
p-eu-k	t-eu-t	k-aa-k
p-au-k	t-ouw	k-ee-t
p-auw	k-ou-d	

poot	pot	teek	tip	keek	kik
peet	poet	took	tot	kook	kok
pak	puur	tak	tuut	kat	kuur
pet	pit	tep	toet	kep	koek
piet	tiek	kiek			

3.4 bisyllabic words:

pa-tat	patat	ka-pot	kapot
ta-pijt	tapijt	ko-ket	koket
ka-pok	kapok	ca-to	cato
pa-ket	pakket	pa-té	paté
ta-puit	tapuit	ko-pie	kopie
pi-ket	piket	ca-cao	cacao
pi-pet	pipet		

theetijd	theekop	tiktak
kaalkop	theepot	kooktijd
paktouw	theetuin	tiptop
kookpot	pappot	pieppop

praat	traan	kraak
prak	trap	krak
prauw	trein	krap
prijk	tref	krijt
prijs	trek	kreet
pret	trijp	krijs
preek	trein	krop
prut	trouw	kruk
prei	tros	kraan

b and d

boot	bief	dek	dop
beet	bad	dik	dip
bak	bol	das	baas
bok	biet	dof	bes
beef	doos		

brutaal	dromen
dorpeling	bedek
boeket	draaien
drager	bewijs
baard	dagdeel

Fricatives (lesson 3 and 4)

1. Auditory Discrimination s and z; f and v

Natives of Amsterdam might not be able to discriminate between these voiced and voiceless fricatives. Note which speakers were unable to discriminate; in this way we can take this into account in the statistical analyses. All speakers should attempt and practise production. Longer, sharper production of fricatives will affect intelligibility positively.

sip	-	zeep	fit	-	vis
sik	-	ziek	fout	-	vouw
saus	-	zout	fop	-	vos
sus	-	zus	fut	-	vul
sijs	-	zeis	fok	-	vod
suis	-	zuid	foei	-	voet

2. Production

See techniques described above, under plosive section. Articulation of fricatives should be sharp; the friction noise should be prolonged. First teach the /s/ and /f/ in final, then in initial position. Use auditory feedback to facilitate production: record the speaker's attempts, and together listen to the recording and judge what he has done.

2.1 S:

2.1.1 Isolation

2.1.2 Final en initial:

paas	tas	kaas	sap	saal
pees	thijs	kuis	sok	saai
pas	thuis	kees	sik	soel
poos	toos	keus	sop	sot
pus	kies	soep	seel	
poes	koos	sip	siel	

2.1.3 clusters:

spek	spook	staal	stof
spel	spiek	steel	stik
spit	spot	stijl	stok
spon	spin	stel	staak
spat	speen	stil	steek

2.1.4 medial:

passant	toestel	tijdsein	pastei
kassa	testen	pistool	kopstuk
tussen	kasteel	prutsen	potspel
tassen	kaaskop	koetsen	
kiespijn	koestal		
passief	pastoor		
passer	proesten		

2.2 F:

2.2.1 Isolation:

2.2.2 Final:

paf	kuif	toef
tof	poef	kef
tijf	pof	kijf
tof	puf	teef
bataaf		

2.2.3 Initial:

fout	fik	foto
feit	fok	fakkel
fuik	fijn	feest
faam	fideel	fel
foei	foedraal	fabriek

2.2.4 medial:

koffie	koffer	poffen	café	tuffen
plafond	tafel	buffet	keffer	suffie

z and v:

zie	zo	vee	vat
zeef	zout	val	van
zand	ziek	voer	vaas
zuiden	ziel	vaak	vader
zuivel	zak	veer	vis

Short sentences fricatives en plosives:

stap op kees	pak die stapel
pak de theekop	koffie en koek
pak de koffer	koffie en thee
touw kopen	tapijt en tafelkleed
kom kijken	de klok tikt
pen en papier pakken	piet praat
peper en tuinkruiden	conducteur en controleur
kaarten tekenen	koude koffie
kort na pasen	kasten takelen
typist en kantoor	trompettist en partituur

More difficult sounds (lesson 5 and 6)

Explain / illustrate different places of articulation, then demonstrate the sounds. Record the speaker's sound productions; listen to the recording and evaluate the sounds that were produced, giving feedback as to the quality of each production. Encourage the speaker to give his own feedback / comments on his production.

Lesson 5

1. Nasals:

There are no specific tricks, but it helps to articulate more forcefully: compress the lips and keep them compressed slightly longer for the /m/; similarly, tongue contact with alveolar ridge should be quite forceful and slightly prolonged when producing the /n/. First in final, then medial position.

kam	kan	zang
kim	ken	tang
tam	pien	zing
tom	kien	bang

pam	tien	stang
pim	teen	tong
kom	pan	beng

maat	nar
mik	nek
mat	nier
mol	nies
mild	nee
mop	nat
mier	nut

poema	kano	bengel
tomaat	kanaal	vangen
kamer	kaneel	longen
coma	tonen	zangen
tuimel	banaan	zingen
kameel	tanen	stengel

2. L:

Explanation: the /l/ sound becomes clearer by producing it more forcefully. The effect might be enhanced by producing a slightly more retroflex quality (tongue tip curled backward).

Demonstrate and ask speaker to imitate (in isolation), give feedback and repeat instruction if necessary

loop	leef	leek	lat	loos
lip	lief	leuk	lid	loes
lap	lied	leut	luis	lot
lok	lat	luim	luik	laaf

talen	poelen
kelen	delen
telen	balen
stelen	dalen
tellen	pellen

Lesson 6

3. G:

Explanation: the *g* or *ch* is formed by forcing air from the oropharynx through a dorso-palatal constriction. Combine production of /x/ with a front vowel; this can be achieved by starting in the final position:

wieg	weeg	wig	weg
lieg	leeg	lig	leg
veeg	big	keg	deeg
zich	deug	steeg	zeg
liegen	degen	begieten	teugel
liggen	buigen	begin	duigen

giet	gids	geel
giek	gis	geest
gier	gips	geep
giebel	gister	gum
gieter	gissen	geen

One may also attempt to produce the /x/ starting from the position of the /k/. Alternatively: Bring the /x/ forward by producing the typical /ɣ/ that is spoken in Limburg.

kiet	-	giet = ‘k giet (ik giet)
kiek	-	giek = ‘k giek (‘’)
kier	-	gier = ‘k gier (etc)
kil	-	gil = ‘k gil
kuit	-	guit = ‘k guit
keep	-	geep = ‘k geep
kein	-	gein = ‘k gein
keur	-	geur = ‘k geur
keel	-	geel = ‘k geel

giet giek gier gil guit geep gein geur geel *clusters*:

glas	graat
glip	groef
gluur	greep
glip	gras

Difficult combinations:

zaagje	haagje	dagje	kraagje
wilgen	ingooien	zwelgen	
ingeven	belgen	ingaan	
algemeen	ingang	walging	

schik	scheel	schoen
schip	scheef	schoon
schil	scheren	schots

schapen	schrapen
schikken	schrikken
schiften	schriften
schande	schrander
schijven	schrijven
schaal	schraal

4. H:

One may produce the /h/ by articulating and prolonging a nearly imperceptible /ie/, or a nearly imperceptible Limburg /ɣ/

uit	-	hh_uid	huid
aan	-	hh_aan	haan
ier	-	hh_ier	hier
aar	-	hh_aar	haar
een	-	hh_een	heen
oet	-	hh_hoed	hoed
at	-	hh_ad	had
ot	-	hh_ot	hot
al	-	hh_al	hal

a-hha	o-hho	ie-hhie	ahum

5. extra material clusters:

lamp	gesp	pulp	werp	kalm
pomp	wesp	hulp	dorp	walm
lomp	hesp	gulp	worp	halm
wolf	werf	fiets	fietst	
dolf	korf	poets	poetst	
delf	durf	beits	beitst	
zelf	derf	plaats	plaatst	

kans	hals	baars	haaks	kast
gans	bals	kaars	hoeks	vast
dans	vals	beers	waaks	bast
mans	bols	heers	baks	meest

borst	kant	belt	hard	ligt
worst	band	halt	baard	legt
korst	hand	valt	kaart	weegt
dorst	land	stalt	vaart	zegt

zacht	voorn	wolk	werk	dwerg
vacht	hoorn	dolk	vork	berg
lacht	karn	bulk	kurk	burg
nacht	doorn	balk	lurk	barg
dacht	deern	kalk	berk	borg

Review Plosives and Fricatives

extra material

stappen	spreektaal	stapeltje	slaapzak
stoten	stikstof	spotprijzen	strijkorkest
stuiten	stiksteek	speelkwartier	struikelblok
stakker	strijkstok	spoorstudent	specifiek
stinken	steekspel	spektakel	afrikaans
steken	stokpaard	ofschoon	argentijn
sokken	station	avontuur	effect
soepen	citroen	advent	alkoof
stoffen	soepen	saffraan	saffier
situatie	specerij	solopartij	suikerfabriek

B. Clear Speech

Teaching clear speech may be subdivided into a number of components: 1.) explanation (what does clear speech consist of, how does it differ from normal everyday speech); 2.) imitation of clear speech samples; 3.) instruction (attempt to pronounce each individual sound as clearly and precisely as possible); 4) feedback (point out speaker's typical errors) and self evaluation (speaker listens to, and identifies his own errors in his (recorded) speech).

Lesson 7

Explanation

Familiarize speaker with typical characteristics of clear speech:

Let the speaker listen to a recording of a piece of normal, conversational speech, and the same piece, but then deliberately and clearly articulated. Compare and identify the differences between the different types of speech. The material on the cd provided with this program consists of short pieces, each spoken conversationally and then clearly. The text of these pieces and further explanation is given in Appendix A.

- Words are spoken more slowly, but not unnaturally so
- Speech is more powerful, without being shouted
- More, and especially more appropriately positioned, pauses are inserted
- No sound is swallowed up. Each individual sound is well-formed; clearly and precisely articulated
- Speech is lively, melodic, with good and appropriate use of accents

Take care! Clear speech is NOT: monotone, staccato, loud and shrill, exaggerated, and does not sound artificial or weird.

Record the speaker's running speech and together evaluate typical errors. What typifies this speaker in terms of omissions, substitutions, indistinct sounds, etc? List the most obvious and most frequent problems.

Imitation

Ask the speaker to listen to, and then imitate sentences that are spoken clearly (clearly spoken short and longer sentences are also available on the cd as stimulus material for this purpose, see appendix A for the written version of these sentences). Record the speaker's imitations and give feedback with regard to clarity, and problem sounds. If necessary, review some of the techniques and strategies that were taught earlier in the program.

Lesson 8

Reading Aloud

Ask the speaker to bring interesting reading material from home (topics that interest the speaker). Repeat the characteristics of clear speech before starting. Especially the instruction to articulate each individual sound clearly and precisely will be helpful. Record the speaker's read-aloud pieces. Listen to the recorded speech and give feedback. Encourage the speaker to comment on his own speech and to identify the errors he still hears. Review the sounds that were especially problematic for this speaker.

To stimulate the use of sentence accent, highlight the important words in the written text (the text that the speaker brought from home, or sentences that lend themselves for this type of exercise). Ask the speaker to produce these highlighted words using extra speaking effort and a higher pitch. Demonstrate if necessary and ask the speaker to imitate.

PAY ATTENTION TO PHRASING!! Some speakers tend to pause at inappropriate places in a phrase, or even pause within words. When phrasing is a problem, explain the effect that wrongly positioned pauses have on the listener (wrongly positioned pauses affect speech intelligibility, because it tangles up the sentence structure). Mark the places in the written text where pauses would be appropriate. Ask the speaker to do the same. If necessary, phrasing may be practiced separately, using these types of phrases:

(1+2)....x3	Ax...(B-C)
1+...(2x3)	(AxB)...-C

Geen zout,...maar <u>peper</u>
We willen best helpen,... maar dan moeten jullie ons <u>eerst</u> betalen.
Had hij het over een <u>rode</u> mercedes,... of over een <u>groene</u>.
Ze hebben niet alleen de <u>minister</u> beledigd,...maar ook nog eens z'n <u>vrouw</u> voor schut gezet.

IK ZOU	(PIET EN TOOS), OF KEES UITNODIGEN.
IK ZOU	PIET EN (TOOS OF KEES) UITNODIGEN.
IK ZOU	PIET, OF (TOOS EN KEES) UITNODIGEN.
IK ZOU	(PIET OF TOOS) EN KEES UITNODIGEN.

IK ZOU (PATRICIA EN CORNELIUS), OF CATHARINA UITNODIGEN.
IK ZOU PATRICIA, EN (CORNELIUS OF CATHARINA) UITNODIGEN.
IK ZOU PATRICIA, OF (CORNELIUS EN CATHARINA) UITNODIGEN.
IK ZOU (PATRICIA OF CORNELIUS) EN CATHARINA UITNODIGEN.

IK ZOU (JOZEFIEN EN JOHANNES), OF WILLEMIJN UITNODIGEN.
IK ZOU JOZEFIEN, EN (JOHANNES OF WILLEMIJN) UITNODIGEN.
IK ZOU JOZEFIEN, OF (JOHANNES EN WILLEMIJN) UITNODIGEN.
IK ZOU (JOZEFIEN OF JOHANNES) EN WILLEMIJN UITNODIGEN.

IK ZOU PIET, EN (TOOS EN KEES), EN (PLEUN EN THIJS) UITNODIGEN.
IK ZOU (PIET EN TOOS), EN (KEES EN PLEUN), EN THIJS UITNODIGEN.
IK ZOU (PIET EN TOOS), EN KEES, EN (PLEUN EN THIJS) UITNODIGEN.

IK ZOU PATRICIA, EN (CORNELIUS EN CATHARINA) EN (CHRISTOFFEL EN PETRONELLA), UITNODIGEN.
IK ZOU (PATRICIA EN CORNELIUS), EN (CATHARINA EN CHRISTOFFEL), EN PETRONELLA UITNODIGEN.
IK ZOU (PATRICIA EN CORNELIUS), EN CATHARINA, EN (CHRISTOFFEL EN PETRONELLA) UITNODIGEN.

IK ZOU JOZEFIEN, EN (MARIUS EN WILLEMIJN), EN (JOHANNES EN ANNEMARIE) UITNODIGEN.
IK ZOU (JOZEFIEN EN MARIUS), EN (WILLEMIJN EN JOHANNES), EN ANNEMARIE UITNODIGEN.
IK ZOU (JOZEFIEN EN MARIUS), EN WILLEMIJN, EN (JOHANNES EN ANNEMARIE) UITNODIGEN.

Lesson 9

Monologue and Question-Answer Dialogue

Repeat the typical characteristics of clear speech. Especially the instruction to articulate each individual sound clearly and precisely will be helpful. Ask the speaker to talk about a topic of his choice (e.g., his hobby, favorite TV program), or, if no topics are forthcoming, suggest some easy, interesting topics that he can talk about. Record the speaker's monologue. Listen to the recorded speech and give feedback. Encourage the speaker to provide feedback on his own speech.

Review sounds that were problematic for this speaker; or review any of the previously taught techniques and strategies that might be helpful to improve his speech.
Ask open-ended questions, record this question-answer discourse and listen to the recording. Evaluate if the speaker is starting to apply what he has been taught during spontaneous speech. Pay attention to any problems that might still be affecting intelligibility negatively.

Appendix A: Stimulus material “clear speech” (on cd)

Discourse samples were recorded in “conversational” and in “clear” speech style. *Sentences* were recorded only as “clear” speech and are meant as imitation exercises.

“conversational” versus clear speech: discourse samples (on cd)

Explanation:

Every sound file (e.g. “dieven 1”) starts with sentences in “conversation style” which are then repeated in “clear speech”. The written text is therefore also given twice. First, the transcription of the conversational speech is given (including all omissions), second, the transcription of the clear speech version is given (the clearly pronounced sounds are highlighted in the written text). On first impression, clear speech sounds slower than conversational speech. Speaking rate does play a role, but more is involved. Pauses are positioned more grammatically (less randomly). Words, syllables, and sounds are no longer omitted (swallowed, deformed or distorted). Accented words are more noticeable. Fricatives, plosives, clusters, and vowels are pronounced more clearly. In contrast to clear speech, less acoustic information is available during normal “conversational” speech. The samples recorded on this cd might sound excessive, but analyses of “normal” speakers having a “normal” conversation will demonstrate this principle equally well. Omitted, distorted sounds and syllables, incomplete, unfinished words and sentences, are typical of conversational speech. Normally, this would not matter: under reasonable circumstances (barring noisy environments, hearing disorders and speech disorders), the average listener will have no problem understanding the average speaker. However, when the voice and speech quality has been affected, a listener will have more trouble understanding the speaker.

The written texts aid comparison of the two speech styles, as the texts highlight the contrasts between conversational and clear speech. The recordings on the cd are meant as a tool, it is also possible to give a live illustration of conversational and clear speech. It is important to record, and then listen carefully and critically to the speech of the participant (e.g., during conversation). The participant’s typical mistakes, omissions, speaking habits can then be noted. People are usually not aware of their own speech; how often they omit (part of) words or sentences, how ineffectively pauses and accents are used; how imprecise and sloppy their speech is. Awareness of typical mistakes helps to improve overall quality and intelligibility of speech.

Sound files:

Dieven 1:

Diev- -ebbe- dinsdag vo- de tweed- keer i- éé- wee(k) een ‘aanzienlij-’ gelb-dra(g) bui-gemaa(kt) in Ams-(e)-dam.

Diev**en** **-pauze-** hebb**en** dinsda**g** voor de tweede **k**eer in één wee**k** **-pauze-** een ‘aan-zienlijk’ geldbedra**g** bui**t**gemaa**kt** in Amsterdam.

Dieven 2:

-n twee e-ve-t-g ja-ge man w-d -p d- Nieuw-zij-s Vo-bu-gwa- beroof- v-n e- koffe-tje met on(g)-ve- ho-de-vijf-(t)win-e duize- euro.

Een **t**wee-en-veer**tig**-jarige man **-pauze-** wer**d** op de Nieuwe**z**ijds Voorbur**g**wal **-pauze-** beroo**fd** van een **k**offer**t**je **-pauze-** me**t** on**geveer** **h**onder**d**vijfen**t**winti**g** **d**ui-zen**d** euro.

Dieven 3:

Vo-ge wee- we-(d) ee(n) ‘aanzienlij-’ gel-b-drag (g)stol(e)- v-(n) twee Duitsers. **-pauze-** D-e p-lisie on(e)-zoe-t (o)f e- ve-ban- bestaa- tusse- (d)e diefstalle-.

Vori**g**e wee**k** **-pauze-** wer**d** een ‘aan**zien**lijk’ **g**eldbedrag ge**s**tol**en** van **t**wee Dui**t**sers. **-pauze-** De po**li**tie onderzoek**t** **-pauze-** of ee(n) ver**b**and bestaa**t** **-pauze-** **t**ussen **d**e diefstall**en**.

Dieven 4:

- ma(n) die dinsdag we- beroof(d), -ad -e- gel- bij zich voo- -n ‘zak-lijke transactie’, (me)-t -e Ams-dams- p-l-sie (w)oensdag.

De ma**n** **d**ie **d**insda**g** werd beroo**fd**, **-pauze-** **h**ad het gel**d** bij zi**ch** voor een ‘**zak**elij**k**e **t**ransac**t**ie’, **-pauze-** meld**t** (d)e Ams**t**e-damse **pol**i**t**ie woen**s**da**g**.

Dieven 5:

D- dad- ren- na- - g-reedstaan- blauw- auto, waarin -n (m)ededade- zat. **-pauze-** D-wa(g)e ree- d-rop (m)-t- (h)o(g)- snel-eid -n d- ri(ch)ting --n -et C-traa- S-atio-, meldt - p-lisie.

De **dader** **-pauze-** **r**en**de** naar -n **g**ere**e**d**st**aan**de** **bl**auwe au**t**o **-pauze-**, waarin -n **m**ede**d**ader **zat**. **-pauze-** **De** wa**gen** **r**eed **d**aarop me**t** **ho**g**e** **s**nel**heid** in **de** **richt**ing van het **C**en**t**raal **S**tation, **-pauze-** meld**t** de **pol**i**t**ie.

Dieven 6:

Vo-ge week we-de- (t)wee Dui-se man- in de buur- van -t Ams-e- s-atio(n) -et slachtof- van -n soo-(t)g-lijk- ov-al. **-pauze-** - ve-schil is dat -e Duits- zak-lied- -n afspraak -adde-, zeg- p-lisiewoor-voe-de- Rob van -e- Veen.

Vori-**ge** wee**k** **-pauze-** werde- **t**wee **D**uits**e** manne**n** in **d**e buur**t** van -e**t** Am**st**el **st**ation **-pauze-** **h**et sla**chtoff**er van een soo**rtg**elij**ke** overval. **-pauze-** **H**et ve**r**sch**i**l i**s** **d**at **d**e **D**uits**e** **z**ake-**l**ie**d**en een afspraak **h**a**dd**en, **-pauze-** ze**g**t p**o**l**i**t**i**ewoo**rdv**oerder Rob van der Veen.

Dieven 7:

De ma(n) die dinsda- wer- beroof(d), -ad gee- afspraa(k). **-pauze-** D- p-lisie on-e-zoe- -de signalement- en – g-brui-t- auto's met e-kaa- ov-reenkome-.

De man **d**ie **d**ins**dag** wer**d** beroo**fd**, **-pauze-** **h**ad **g**een a**fs**pra**a**k. **-pauze-** De **polit**ie onde**r**zoe**kt** **-pauze-** **of** **d**e signalemen**ten** en **d**e **gebr**uik**t**e auto's **-pauze-** me**t** el**k**aa**r** ove**r**een**k**om**en**.

Ganzen 1:

Ganze- Aa(n) e- g-oe(b)j- tamme ganz- o(b) -et erf (k)-n vee(l) pl-zie- worde- b-leefd, v-ral v-weg- -u- s(o)ciale g-drag.

Ganze**n** Aan een **gr**oe**p**je **t**amme **g**anz**en** o**p** -e**t** e**rf** –**pauze**- **k**an **v**eel **pl**ezie**r** wor**d**e- belee**fd**, –**pauze**- voora**l** v**anw**e**g**e hun so**c**iale **g**ed**r**a**g**.

Ganzen 2:

-t is bevo-be boeie- te zien hoe de lede- van e- g-oe(b) m-t e-ka- com-nicere-:-**pauze**- onde- mee® do- geluide t- make, - hals te strekke- -**pauze**- -n an-ere groepslede- m-t -e og- t- fixere-.

Het is **b**ijvoor**b**eeld **b**oeiend **t**e zien **-pauze-** **h**oe de **l**e**d**en van een groep me**t** el**k**aar communicer**en**: **-pauze**- onder meer **d**oor **g**el**u**ide**n** **t**e make**n**, **-pauze**- de **h**al**s** **t**e **str**ekke**n** –**pauze**- en an**d**ere **groeps**l**e**den me**t** de o**g**e**n** **t**e fixere**n**.

Ganzen 3:

-n an-r- uiting van -u- s-ciale g-drag is d- waakzaamheid. –**pauze**- -n e- groe- ganz(e)- z(ij)n alt(ij)d we 2 dier- te vinde- die vo- de veilighei(d) v-n de anderen doorlop- -omgev- in de gate- houwe- - met e- -eleboel lawaai -n indring- (p)robere- t- ve-jag-.

Een an**d**ere ui**t**ing van **h**un **so**ciale **gedrag** –**pauze**- **is** **d**e waakzaam**heid**. **-pauze**- In een **g**roep **g**anze**n** zij**n** al**tijd** wel **t**wee **d**iere**n** te vinde**n** –**pauze-** **d**ie **v**oor de **v**eili**g**hei**d** van de an**d**ere**n** –**pauze-** **d**oorlo**p**end **d**e om**g**eving in de **g**ate**n** **h**oude**n** –**pauze**- en me**t** een **h**eleboe**l** la*waai* (verlenging) een in**dr**inge**r** **pr**obere**n** **t**e verja**gen**.

Ganzen 4:

Hu- waa(k)z-m-eid gaa- zelf- zo ver, -**pauze**- dat in -e broe-tijd oo- de verzorg- als -n indring- wordt beschouwd. –**pauze**- V-ral -e gent, he- mannetj-, is – ag-sief.

H**un** waakzaamhei**d** –**pauze**- **g**aat zel**fs** zo ver, -**pauze**- da**t** in de **br**oedt**i**jd ook **d**e **v**erz**o**r**g**er als een in**dr**inge**r** wordt beschouw**d**. –**pauze**- Vooral de gen**t**, -**pauze**- he**t** manne**tje**, -pauze- i**s** a**g**re**ss**ie**f**.

Ganzen 5:

Ze snav-l is -n geduch(t) wape- en oo(k) -et slaa- m- (d)- vleuge-s ka flinke blauw- plek- veroorzak-.

Zij**n** snav**el** is een **ged**u**cht** wape**n** –**pauze**- en oo**k** het **slaan m**et de v**leu**g**els k**an **fl**in**k**e **bl**auwe **pl**ekke**n v**eroo**rza**k**en**.

Ganzen 6:

e ku- ma- bet-r uit -e buu- blij- va- broed-nde ganze-. –**pauze**- Ma- buiten -t broed- seizoen b-leeft u dr veel p(l)-zier aan.

U **kunt m**aar **b**e**t**e**r** uit **d**e **b**uurt **bl**i**jven** van **br**oede**nd**e **g**a**nzen**. –**pauze**- Maar **b**ui**ten** -et **br**oe**ds**ei**z**oen –**pauze**- **b**e**l**ee**ft** u er **v**ee**l** p**l**e**z**i**er** aan.

Clear speech: Sentences for imitation (on cd)

The sentences on the cd are also meant as a tool. You may model the sentences yourself, asking the participant to imitate your clear pronunciation.

The sounds /p, t, k, s/ are present in most sentences. The /f/ sound can be found in sentences 12, 13 and 20; nasals in sentences 3, 10, 12; the /l/ sound in sentences 19-21; the /x/ sound in sentences 10, 14, 16; the /h/ sound in sentences 11, 12 and 23.

1. Tegen de kant
2. Pa, kom kijken!
3. Dat kost kapitalen
4. Piet pakt peperkoek
5. Kitty koopt piepkuikens.
6. Bakt de kok kaaskoekjes?
7. Trek het touw niet kapot.
8. De bel van de voordeur is kapot
9. Ik zág dat ze perziken wilde pikken.
10. De tuinman heeft het gras gemaaid.
11. Het natte hout sist in het vuur.

12. Dat hotel heeft een slechte naam.
13. De nieuwe fiets is gestolen.
14. De aardappels liggen in de schuur.
15. De portier ging met vakantie.
16. De kachel is nog steeds niet aan.
17. De appels aan de boom zijn rijp.
18. De kleren waren niet gewassen.
19. De tuinen liggen er kaal en verlaten bij.
20. Lopen is niet gezonder dan fietsen.
21. Straks gaan we de ballonnen opblazen.
22. Als je niet oppast, loopt hij zo weg.
23. Soms heb ik het helemaal gehad met het huishouden.
24. Je moest eens weten hoe moeilijk die situatie was.

Appendix B: Evaluation, Handing over and Compliance

Name:

Date of Birth:

At the end of Lesson 1

Evaluation

1. Was the technique (push harder) useful?
   1. -yes
   2. -no, too difficult to explain
   3. -no, had no effect on this speaker
   4. -no, not necessary with this speaker (production of plosives is excellent)

Handing over (to another SLP)

1. Does the speaker (still) have trouble with stoma noise?
   1. -always 2. -often 3. -sometimes 4. -never

2. Could the speaker discriminate between voiced and voiceless plosives?
   1. -never 2. -sometimes 3. -often 4. -always

3. Did the speaker understand the instruction to "push harder"?
   1. -not at all 2. -somewhat 3. -reasonably 4. -well

4. Could the speaker apply the technique? (More than one answer is possible).
   1. -not at all
   2. -in isolation
   3. -monosyllabic words only: -**fin**, -**init**
   4.-bi words: -**fin**, -**in**, -**med**

5. Comments/suggestions?

At the start of Lesson 2

Compliance[1]

1. How many days did the speaker practice?

    1. -not at all 2. -every other day 3. -every day

2. How many times did the speakers practice per day?

    1.- not at all 2.- 1x 3.- 2-3x 4.- >3x

3. Did the speaker find the exercises:

    1.-difficult 2.-neither difficult, nor easy 3.-easy

4. Let the speaker demonstrate what he learnt during the previous session (ask the speaker to produce 10 words (or 10x in isolation) and mark each production as correct / incorrect). Give the score out of 10:______________________

At the end of Lesson 2

Handing over (to another SLP)

1. Does the speaker (still) have trouble with stoma noise?

    1. -always 2. -often 3. -sometimes 4. -never

2. Could the speaker discriminate between voiced and voiceless plosives?

    1. -never 2. -sometimes 3. -often 4.- always

3. Did the speaker understand the instruction to “push harder”?

    1. -not at all 2. -somewhat 3. -reasonably 4. -well

4. Could the speaker apply the technique? (More than one answer is possible).

    1. -not at all
    2. -in isolation
    3. -monosyllabic words only: **-fin**, **-init**
    4. -bi words: **-fin**, **-in**, **-med**

5. Comments/suggestions

1 Compliance sheets were the same for each session and should be used for each session, but will only be given once here as an example.

At the start of Lesson 3

Compliance sheet

At the end of Lesson 3

Evaluation

1. Was the technique (build-up of oral pressure: sharper, longer fricatives) useful?
    1. -yes
    2. -no, difficult technique to explain
    3. -no, had no effect with this speaker
    4. -no, this speaker could already produce clear fricative

Handing over (to another SLP)

1. Does the speaker (still) have trouble with stoma noise?

    1. -always 2. -often 3. -sometimes 4. never

2. Could the speaker discriminate between voiced/voiceless fricatives?

    1. -not at all 2. -sometimes 3. -often 4. -always

3. Could the speaker apply the technique? (More than one answer is possible).
    1. -not at all
    2. -in isolation
    3. -monosyllabic words only: **-fin**, **-init**
    4. -bi words: **-fin**, **-in**, **-med**

4. Comments/suggestions

At the start of Lesson 4

Compliance

At the end of Lesson 4

Handing over (to another SLP)

1. Does the speaker (still) have trouble with stoma noise?

    1. -always 2. -often 3. -sometimes 4. -never

2. Could the speaker discriminate between voiced/voiceless fricatives?
   1. -not at all 2. -sometimes 3. -often 4. -always

3 . Could the speaker apply the technique? (More than one answer is possible).
   1. -not at all
   2. -in isolation
   3. -monosyllabic words only: **-fin**, **-init**
   4. -bi words: **-fin**, **-in**, **-med**

4. Comments/suggestions

At the start of Lesson 5

Compliance

At the end of Lesson 5

Evaluation

1. Which technique did you use to teach the /l/?
   1.- forceful contact 2. -retroflex 3. -combination 4. -other:

Handing over (to another SLP)

1. Does the speaker articulate nasals clearly?
   1. -not at all
   2. -in isolation
   3. -monosyllabic words (fin, init)
   4. -polysyllabic words and phrases

2. Does the speaker articulate the /l/ clearly?
   1. -not at all
   2. -in isolation
   3. -monosyllabic words (fin, init)
   4. -polysyllabic words and phrases

3. Comments/suggestions

At the start of Lesson 6

Compliance

At the end of Lesson 6

Evaluation

1. Which technique did you use to teach the /g/?
   1. -starting from /k/ 2.-more to the front, as a soft /g/
   3. -combination 4. -other:

2. Which technique did you use to teach the /h/?
   1. -scarcely perceptible, prolonged /ie/ 2. -scarcely perceptible /g/
   3. -combination 4.-other:

Handing over (to another SLP)

1. Does the speaker articulate the /g/ clearly?
   1. -not at all
   2. -in isolation
   3. -monosyllabic words (fin, init)
   4. -polysyllabic words and phrases

2. Does the speaker articulate the /h/ clearly?
   1. -not at all
   2. -in isolation
   3. -monosyllabic words (fin, init)
   4. -polysyllabic words and phrases

3. Comments/suggestions

At the start of Lesson 7

Compliance

At the end of Lesson 7

Evaluation

1. Did you demonstrate the difference between conversational and clear speech?
    1. -using cd recordings
    2. -using own speech (demonstration)
    3. -combination

Handing over (to another SLP)

1. Which typical problems did you notice in this speaker's speech (omission of [end of] words, syllables, sounds; indistinct plosives, fricatives; inappropriate [breathing] pauses; etc; etc)

    ....

    ....

    ....

    ....

2. Does the speaker understand the notion of clear speech?

    1. -yes   2. -no

3. Which sounds were the most problematic during imitation of clear speech sentences?

    ....

    ....

    ....

    ....

4. Comments/suggestions

At the start of Lesson 8

Compliance

At the end of Lesson 8

Handing over (to another SLP)

1. Which sounds were the most problematic during reading (omission of [end of] words, syllables, sounds; indistinct plosives, fricatives; inappropriate [breathing] pauses; etc; etc)

    ....

    ....

2. How would you rate the speaker's phrasing?
   1. -poor 2. -moderate 3. -reasonable 4. -good

3. Does the speaker use sentence accent?
   1. -poor 2. -moderate 3. -reasonable 4. -good

4. Comments/suggestions

At the start of Lesson 9

Compliance with a different question 4:

1. Ask the speaker what he learnt during the previous lesson. Can he apply what was taught?
   1. -not at all 2. -sometimes 3. -often 4. -always

At the end of Lesson 9

Evaluation

1. Do you think the speaker's intelligibility has improved?
   1. -not at all 2. -a little bit 3. -somewhat 4. -a lot

2. Which parts of the program were the most difficult for this speaker?

3. Which strategies / techniques had effect (or not) with this speaker?

4. Was the available time sufficient to train everything?
   1. -completely insufficient
   2. -sufficient
   3. -more than sufficient

5. Comments/suggestions

Appendix C: Tally sheet

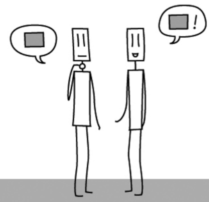
Name:

Date of Birth:

Date	Number of times practiced today?	Approximately how long did you practice (minutes)?	Comments (problems, difficulties)

5

Therapy evaluation: phoneme and sentence intelligibility tests



Abstract

In evidence-based medicine, it is customary to base therapy on techniques that have proven to be successful in other studies as well as on own expertise and experience. Since patients, and TE speakers in particular, generally form heterogeneous groups, it cannot be assumed that techniques that were successful in one group are equally successful in another group. It is therefore important to study the effects of a therapy program before it is incorporated in daily practice. Whether the therapy program discussed in this thesis was successful has been tested by means of pre and post tests. The following division between the tests was made: in this chapter the phoneme and sentence identification tests are discussed and in the next chapter the qualitative judgments of TE speech intelligibility (semantic scales and study-specific questionnaires). Results in this chapter show that scores have improved for the fricatives in initial position, and for the voiced-voiceless distinction. For medial position, all manners of articulation, except the approximant /l/ score better in the post test. The SUS sentences also show improvement. The nasals and vowels do not show significant changes, but these categories already showed very high scores in the pre test. It can be concluded that the therapy program seems to have a positive effect on intelligibility at the consonant and the sentence level.

5.1 Introduction

Over the last few years evidence-based rehabilitation (EBR) (just like evidence-based medicine (EBM) in the general medical field) has received more attention and has become increasingly important in the allied health field. The main reason for the increase in evidence-based practice is the fact that professionals have a responsibility towards their clients, towards people that refer clients to them (for example general practitioners) and towards the health insurance companies, especially when considering the social developments (Leemrijse et al., 2004). In oncology-related areas of research, new developments in medical treatments of oncology patients require an active, yet cost-effective approach of the speech language pathologist, especially since communicative disorders can be very invalidating and might severely diminish the quality of life.

There are two complementary approaches to evidence-based rehabilitation. Firstly, the literature should be critically reviewed to investigate which therapy would be beneficial for the group one is working with and secondly, the effects of a specific therapy program should be studied. MacKenzie et al. (2001) found a positive effect of speech therapy in a randomized clinical trial in 204 patients suffering from persistent hoarseness for at least two months. Speyer et al. (2004) published the results of a research project, which started in 1997 on the authority of the Board of Health Insurance Companies that investigated if speech therapy is beneficial at all. These authors restricted themselves to subjects with chronic voice disorders and overall found a small, but significant improvement in voice quality. Although these results show a positive effect of speech therapy, they can obviously not be generalized to other populations or therapy programs. Each therapy and/or population should be evaluated on their own merit.

In chapter 4, literature on speech therapy for laryngectomized individuals was reviewed and all choices in the therapy program, both for the contents of the therapy and for the therapy techniques used, were based on findings in the literature. Whenever possible, techniques were chosen that had already been found to be successful. Where this was not possible, the experience of the speech language pathologists writing the therapy program was used.

In this chapter and in chapter 6, the effect of the therapy program will be discussed. In effect studies like the present one, five levels of evidence can be distinguished (Moore et al., 1995). This study has a level III evidence: Evidence from well designed trials without randomization, with a single group performing pre and post tests. In order to asses the effect of therapy, we performed several pre and post tests. By comparing pre and post test scores, it can be established whether therapy has been successful. These tests consisted of three different listening experiments performed by different groups of listeners as well as study-specific questionnaires filled out by the subjects and their relatives. The idea behind this variation in tests and listeners is that TE speech intelligibility can be tested at different levels (from phoneme to discourse level) and that each listener group will evaluate speech intelligibility in a different way. Speech language pathologists, for example, having their professional frame of reference, will use different strategies when judging TE speech than naïve listeners will. By using several tasks it is hoped to create as complete a picture of intelligibility as possible implying different levels: from overall subjective impressions to transcription at phoneme level.

Our tests can be divided into two groups: on the one hand phoneme and sentence intelligibility tests and on the other hand qualitative judgments of intelligibility and the influence of the speaker's intelligibility on daily life (semantic scales and questionnaires). In this chapter, the phoneme and sentence intelligibility tests will be discussed, whereas in chapter 6 the 'qualitative' tests are described (the listening experiments with naïve listeners and the study-specific questionnaires for the speakers and their relatives). In this chapter then, two tests are included: the phoneme and sentence identification task performed by phonetically trained listeners and the phoneme identification task performed by speech language pathologists (SLPs). Both tests are devised to score the actual amount of errors.

Phonetically trained listeners were asked to listen to syllables containing phonemes that proved to be problematic (as shown in chapter 2), and to Semantically Unpredictable Sentences (SUS sentences). Their judgment is a specialistic one: it shows whether phonemes are articulated better (are perceived better) after training or not. The SLPs were asked to listen to a series of existing Dutch words and to judge whether they were produced correctly. Their judgment is important clinically, as SLPs are the ones that will offer speech therapy.

The results of the analyses will be used to answer the following questions:

1. Can TE speech intelligibility be improved by specific speech training
2. Is the set-up of the therapy program appropriate.

Answering the second question is only possible when the answer to the first question is affirmative. If no effect of speech training is found, this can either be caused by the inability to improve intelligibility in laryngectomized individuals at all, or by a suboptimal set-up of the therapy program. In any case, even if a positive effect of therapy is found, adjustments to the therapy program should be anticipated in order to achieve an optimal therapy approach and an optimal validation in future studies.

5.2 Subjects and methods

5.2.1 Subject selection

For the therapy evaluation study, it was important to have a homogeneous group of speakers. For that reason, selection criteria were set up (for the questionnaire used to select the subjects, see appendix 5.1). The main criteria stated that subjects should be male tracheoesophageal speakers using a Provox2 voice prosthesis, who had undergone a standard total laryngectomy without any additional resections or reconstructions and who had had their surgery at least six months prior to the investigations (questions 1-4). Also they should not have great difficulty producing voice as this usually indicates other physiological or anatomical problems (question 5). Seven more questions (6-12) were included to gain more insight in the problems these subjects have when speaking, both socially and physically, to determine if the set-up (contents) of the training would be beneficial for them.

Subject selection started by placing an advertisement in the magazine of the Dutch Laryngectomee Association, called 'de Tweede Stem' (the second voice). Besides general information on the therapy program itself, also some of the selection criteria were mentioned. In addition, we actively approached laryngectomees and asked them to participate. Based on the records of the Netherlands Cancer Institute a list was devised with possible participants. Based on the experience of the participating ENT specialist plus a colleague researcher and the speech language pathologists, a target group was devised. These people were phoned personally and received an explanation of the therapy program. After this introduction, they were asked if they would be interested in taking part in the study. Due to limited interest and/or reluctance of the patients to personally invest a considerable amount of time and energy in this additional rehabilitation program, it was necessary to be somewhat less strict in the selection criteria for this study. Eventually it was possible to put together a group of 9 subjects (there was no overlap with the group from chapter 2). As from the start of the project a maximum of 10 subjects in total could logistically be accommodated, a number of 9 participants was considered to be an acceptable number. The following table shows their characteristics.

Normally, it would be desirable to have a control group that is not offered speech therapy. However, due to the above mentioned difficult accrual of patients, it was not possible to establish such a group. In addition, one would also be obliged to offer the control group speech therapy at a later stage, and this was logistically impossible. In the present study the lack of a control group is not necessarily a problem, as with chronic voice and speech problems it can be assumed that no spontaneous recovery will take place during the therapy program. Hence, effects that are found after therapy may be seen as the direct effect of the therapy (Speyer et al., 2004).

Table 5.1: Subject characteristics, with age and time after Total Laryngectomy (TLE) in years and months as well as the timing of radiotherapy. Reconstruction refers to the closure of the pharynx after the removal of the larynx (- indicates primary closure of the pharyngeal mucosa; gastric pull-up means that the stomach is used to reconstruct the pharynx; PM flap indicates that the pharynx is partially reconstructed with a pectoralis major myocutaneous (muscle-skin) flap from the chest).

Subjects	Age	Time after TLE	Radiotherapy	Reconstruction
1	59;1	3	Post operative	-
2	68;6	11;10	Recurrence after radiation	-
3	63;2	0;7	Recurrence after radiation	-
4	54;11	2;10	Post operative	-
5	70	13;7	Recurrence after radiation	-
6	68;8	8	Recurrence after radiation	-
7	61;6	6;3	Post operative	Gastric pull-up
8	82;9	0;6	Recurrence after radiation	PM flap
9	55;4	1;8	No radiation	-
Average	64;9	5;6	-	

5.2.2 Recordings

Recordings were made in a sound-treated room with a Sennheiser MD421 Dynamic Microphone and an Edirol (Roland) R-1 portable 24 bit digital wave audio recorder. The maximum recording level was chosen for all speakers and then fixed.

5.2.3 Speech material

In this section, the speech material for the phoneme and sentence identification tests will be discussed; first the material for the listening experiment with phonetically trained listeners, and then the material for the listening experiment with SLPs. The same speech material was used for the pre and post test.

5.2.3.1 (Phoneme) identification experiment with phonetically trained listeners

The speech material the speakers had to read out loud included CV and VCV (nonsense) syllables for the consonants, meaningful CVC words for the vowels and Semantically Unpredictable Sentences (SUS) (Benoit, 1990) (see appendix 5.2). The nonsense syllables were used to avoid learning effects with the listeners and SUS sentences were used to avoid contextual cues that may help the listener. The CV and VCV syllables consisted of the consonants /p b t d k h x f v s z l n m/ in combination with the vowels /a i u/. These consonants had been trained in the therapy program. The CVC syllables consisted of existing words incorporating all Dutch vowels, including diphthongs. Where possible, minimal pairs were used differing only in the length of the vowel (for example *bom* versus *boom,* see also appendix 5.2). One of the reasons why the h-V-t structure discussed in chapter 2 was not used again is that /h/ is notoriously difficult for TE speakers to produce. Also, one of the most common problems with vowels concerned the feature length. By using minimal pairs, a limited set of vowels was established with the focus on length. All syllables and existing words were incorporated in a carrier phrase. Five different kinds of carrier phrases were used to avoid a 'summing up' intonation (Zeg het woord … (say the word …), nu volgt het woord … (now follows the word … =now the word … follows), noem het woord … (name the word …), type het woord … (type the word) … and schrijf het woord … (write the word …). Dummy sentences were placed at the beginning and end of the list and one dummy sentence was placed at the end of the page. The dummy phrases also avoided certain intonation problems common at the beginning and end of a list.

Each participant also read aloud a different subset of five SUS sentences. The program that generates SUS sentences can generate five different syntactical structures. Each subset consisted of one sentence of each syntactical category. For more details about the SUS sentences, see Chapter. 2.

5.3.2.2 Dyva: phoneme identification test with Speech Language Pathologists

The so called 'Dyva' wordlist (see appendix 5.3) was employed as speech material for the second experiment discussed in this chapter. Speakers were asked to read out loud the words on this list. The 'Dyva' is a standardized test devised originally for dysarthric speakers and this list includes existing words of one and two syllables (Dharmaperwira-Prins, 1998). The advantage of this wordlist is that it includes all Dutch vowels and consonants plus a range of consonant clusters. The fact that the test is standardized makes it easier to compare the outcomes of the subjects. Another reason why this test was used is that this is the only known standardized intelligibility test. This lack of tests may be filled in the future by the development of a new instrument for testing intelligibility, developed in Belgium by De Bodt et al. (2006). Unfortunately this test was not available yet at the time of testing.

5.2.4 The listening task

In this section, the listening tasks will be described for the two experiments.

5.2.4.1 Phoneme and sentence identification experiment with phonetically trained listeners

This experiment was performed by ten phonetically trained listeners. However, they were not specifically trained for this experiment. Listeners had a mean age of 45.4 (range 27-65). These highly experienced listeners had no prior experience with TE speech.

The carrier phrases used to embed the syllables with the phonemes were not included in the listening experiment, which means that the target syllables were segmented and extracted from the carrier phrases.

Listeners performed an online experiment in the same way as described in chapter 2 (2.2.4.)

As existing words were used for the vowels there was a risk of learning effects. The use of minimal pairs will minimize this risk somewhat as the listeners still have several options. This risk is further reduced by repeating some of the stimuli in the listening experiment and by stating clearly that stimuli may occur more than once. That way, listeners cannot decide based on what they have heard before and will still have to decide whether they heard a long or a short vowel.

5.2.4.2 Dyva: phoneme identification test with Speech Language Pathologists

Six Speech Language Pathologists (all female) participated because they are considered analytical listeners, who are not distracted by the voice quality of the speaker and hence may give a different judgment as compared to naïve experienced listeners (phoneticians). SLPs had to have at least 6 months experience with TE speakers. Mean age of the SLPs was 40 (range 26-59).

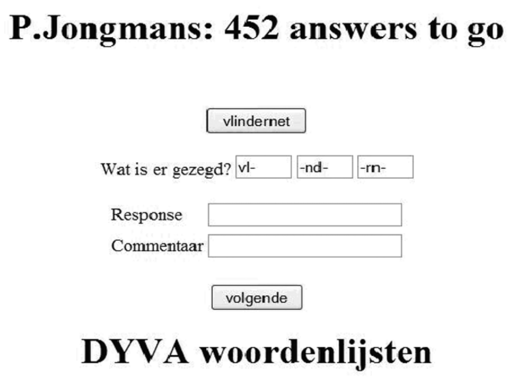


Figure 5.1: The response screen for the Dyva experiment, with the whole word visible in the ‘play’ button, followed by the three target sounds of this stimulus. Beneath, there is room for the whole response (how the whole word was heard) and comments. With the ‘volgende’ button, listeners could go to the next stimulus.

This listening experiment was again performed online. At the start of the test an instruction screen was shown, which again included sound files of all the individual speakers. Two practice items were included at the start of the test. During the test, the listeners saw on the screen which word was spoken at that moment (see Figure 5.1). First, the 'target' consonants and vowels were judged, which were shown separately on the screen: if they were produced correctly, the listener could go to the next stimulus; if they were not produced correctly, listeners deleted that particular target sound in the text box. Then, the listener typed in exactly what she perceived in normal spelling (the whole word). There was also room for additional comments. The test only counts the number of errors (how many target sounds were erased) and not the type of errors. However, the results of the phoneticians in the first experiment will provide information on the type of errors made, just as the whole response given by the SLPs. Just as with the test described above, speakers were randomized for each listener as were the pre and post test items and the individual stimuli. Again, response files were generated automatically.

5.2.5 Statistics

By letting all participants perform the same tests before and after the therapy program, a comparison of the test results will be able to establish whether or not the therapy has been successful. Based on Moore at al. (1995) this constitutes a level III evidence: "evidence from well designed trials without randomization, single group pre and post (…)".

The main goal in this chapter is to establish whether speakers perform better in the post test than in the pre test. First, the inter-listener reliability was calculated using Cronbach's alpha. To test for significance, the non-parametric Mc Nemar test for two related samples was used, with Bonferroni correction. This test is used for correlated samples of dichotomous nominal data and is similar to a t-test for related samples, except that for a Mc Nemar no assumptions exist about the distribution of the data (Kinnear & Gray, 1999).

5.2.6 Praat

Since automatic segmentation, as described in chapter 2, was difficult in this study because of the preceding carrier phrases, the CV, VCV and CVC stimuli were manually segmented. This was still a difficult task, especially for the stimuli starting with a fricative, as the plosive sound of the preceding word in the carrier phrase remained audible after segmentation. By using a smoothing script, the start of the stimuli was more gradual and hence the unwanted plosive perception disappeared. Smoothing occurred over a period of 30 ms before and after the stimulus using the flanks of a Gauss window.

As the recording level was frequently too low for the listening experiment, stimuli were amplified using a script. First, the absolute highest intensity of all speakers together was determined. After that, all stimuli were multiplied with the same factor, such that the highest peak was not clipped. Peaks that did not belong to the speech signal had been manually removed first. By using these methods the differences in intensity between speakers are kept intact.

5.3 Results

In this section, first the results of the phoneme and sentence identification experiment as performed by the phonetically trained listeners will be discussed, followed by a discussion of the results of the Dyva experiment as performed by the speech language pathologists (SLPs).

5.3.1 Phoneme and sentence identification experiment

First the results for the CV and VCV syllables will be discussed, followed by a discussion of the vowels and lastly the results of the Semantically Unpredictable Sentences (SUS) will be addressed.

5.3.1.1 Consonants

For the consonants the inter-listener reliability was .747 and .796 for initial and medial position in the pre test and .747 and .767 for initial and medial position in the post test, respectively. Even though these scores are lower than the inter-listener reliability found in chapter 2, reliability is sufficiently high for using averages.

Table 5.2 shows the overall percentage correct scores for all nine speakers individually, as well as the difference values and mean scores, both for initial and medial consonants in the pre and post test. The table shows a range of 56-83 percent correct for initial position in the pre test. The range is smaller (66-86) in the post test. There are large differences between the percentages of improvement of individual speakers. Speaker 8 shows the highest improvement for initial position (a difference of 14%), whereas speaker 7 shows a small deterioration. For medial position, the pre test shows a range of 63-82 percent correct, against a range of 67-83 percent correct in the post test. Again considerable differences between individual improvements are found: speaker 4 shows the highest improvement of 17%, whereas speaker 3 shows a decrease of 8%. This table also shows that speakers do not always score consistently:

a high improvement in initial position does not necessarily mean a high improvement in medial position, just as a decrease in score in initial position does not mean the speaker will show a decrease in medial position. It can also be seen from Table 5.2 that the mean overall improvement is 6.5%.

Table 5.2: Percentage correct scores for the individual speakers with the difference values and the mean scores per test and per position. The bold and italicized numbers indicate the speaker with the greatest deterioration in score; the underlined numbers indicate the highest improvement.

	1	2	3	4	5	6	7	8	9	Mean
Pre initial	62	75	71	61	67	56	***71***	<u>58</u>	83	67
Post initial	72	76	79	68	77	66	***69***	<u>72</u>	86	74
Difference value	10	1	8	7	10	10	***-2***	<u>14</u>	3	7
Pre medial	68	82	***78***	<u>69</u>	75	63	75	66	78	73
Post medial	80	78	***70***	<u>86</u>	84	67	83	79	82	79
Difference value	12	-4	***-8***	<u>17</u>	9	4	8	13	4	6
Mean difference value	11	1.5	***0***	12	9.5	7	3	<u>13.5</u>	3.5	6.5

The following table lists the mean percentage correct score for the pre and post test per manner of articulation for *initial* position. The included phonemes are /p b t d k h x f v s z l n m/.

Table 5.3: Mean percentage correct score per manner of articulation and the overall score for the pre and post tests for initial position, with the ranges between brackets and the significance in the last column.

	Initial position		
	Pre test (range)	**Post test (range)**	**Sign.**
Plosives	77 (66-89)	79 (61-92)	NS
Fricatives	50 (35-83)	62 (31-87)	p<.01
Nasals	88 (72-97)	87 (75-93)	NS
Approximant /l/	77 (50-95)	83 (60-95)	NS
Mean	67 (56-83)	74 (66-86)	p<.01

It can be seen in Table 5.3 that there is an overall significant improvement between the pre and post test. Closer inspection of Table 5.3 shows that this improvement is entirely caused by the fricatives that score much higher in the post test. The other categories only show marginal improvements and for the nasals no improvement could be established.

Even though the overall scores are important, it is just as important to consider the individual scores per manner of articulation. The large ranges indicate great variability between speakers. The following bar graphs show the pre and post test scores per manner of articulation in initial position for all nine speakers individually.

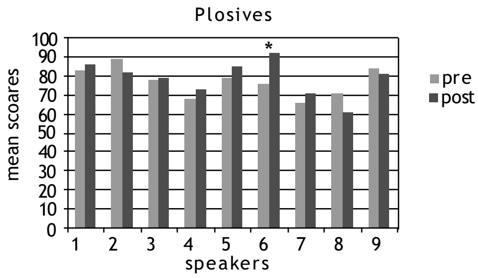


Figure 5.2a: Individual scores for plosives in initial position. * indicates significant difference.

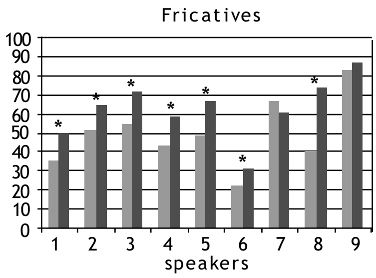


Figure 5.2b: Individual scores for fricatives in initial position. * indicates significant difference.

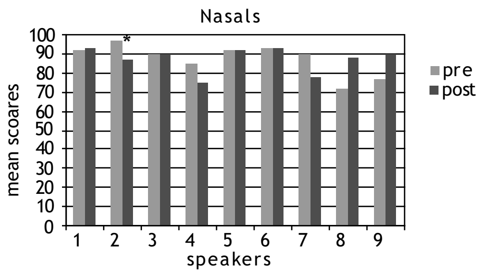


Figure 5.2c: Individual scores for nasal in initial position. * indicates significant difference.

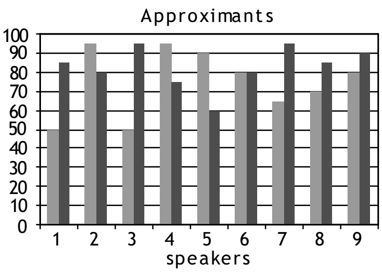


Figure 5.2d: Individual scores for approximant /l/ in initial position.

For the plosives, it can be seen that six speakers improved their score, whereas three speakers showed somewhat lower scores in the post test. Only one speaker (6), though, showed a significant improvement: from 76% to 92%.

For the fricatives, eight speakers show improvement and only one subject (7) scores lower in the post test (NS). All of the improvements are significant, except for speaker 9. Speaker 8 shows the greatest improvement: from 40% to 74%.

The nasals show improvement for three speakers, a lower score for three speakers (speaker 2 significantly lower) and the same score for another three speakers. Most of the differences found are very small, except for speaker 9 who shows an improvement from 77 to 90%. However, only the reduction for speaker 2 is significant.

The approximant /l/ shows a mixture of great improvements and great deterioration. Five speakers score higher in the post test, one speaker shows the same score and three speakers score worse. Except for speaker 6, all differences found are quite large. However, none of the differences are significant.

In the next table, the mean results for *medial* consonant position are given.

Table 5.4: Mean percentage correct score per manner of articulation and the overall score for the pre and post tests for medial position, with the ranges between brackets and the significance in the last column.

	Medial position		
	Pre test (range)	Post test (range)	Sign.
Plosives	81 (64-98)	86 (66-97)	p<.01
Fricatives	62 (49-79)	69 (59-83)	p<.01
Nasals	73 (40-92)	86 (72-97)	p<.01
Approximants	90 (57-100)	92 (70-100)	NS
Mean	73 (63-82)	79 (70-86)	p<.01

Table 5.4 for medial position shows even more promising results than Table 5.3. Again, there is a significant overall improvement. However, in medial position, three of the four categories show significantly higher scores. Only the approximant /l/ does not improve significantly. The nasals show the greatest improvement in percentages.

In Figures 5.3a-d again the individual scores are shown for all manners of articulation separately.

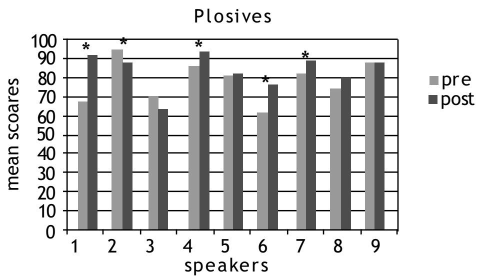


Figure 5.3a: Individual scores for plosives in medial position. * indicates significant difference.

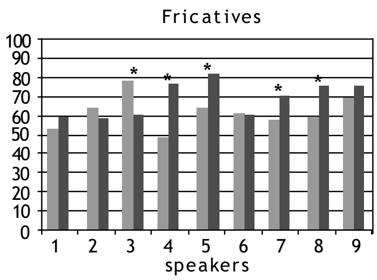


Figure 5.3b: Individual scores for fricatives in medial position. * indicates significant difference.

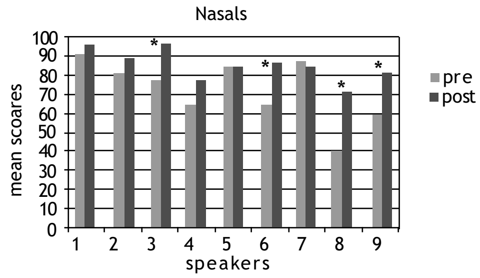


Figure 5.3c: Individual scores for nasals in medial position. * indicates significant difference.

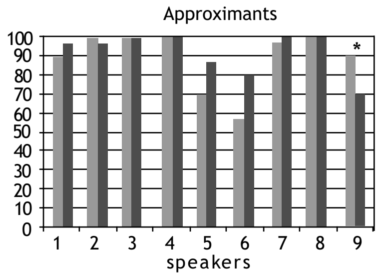


Figure 5.3d: Individual scores for the approximants /l/ in medial position. * indicates significant difference.

For the plosives, six speakers improved their score (four of whom significantly), two speakers scored lower (subject 2 significantly) and one speaker scored the same. Speaker 1 shows the greatest improvement: from 70% to 95%.

For the fricatives, six speakers showed improvement (four of whom significantly), and three speakers showed a lower score (one of whom significantly). Speaker 4 shows the greatest improvement (from 49% to 77%), whereas speaker 3 shows the largest decline (from 79% to 61%).

Seven subjects improved their scores on the nasals, one shows a lower score (not significant) and one the same score. Speaker 3, 6, 8 and 9 show a significant increase, with the improvement of speaker 8 being the largest: from 40% to 72%.

Only four speakers showed a higher score on the post test for the approximant /l/. However, the scores are already rather high for the pre test: no less than four speakers showed a 100% correct score. Speaker 6 shows the greatest improvement: from 57% to 80%. Speaker 9 scores significantly worse in the post test.

Table 5.5: Statistically significant increases (+) and decreases (-) in phoneme intelligibility per patient and per category.

	1	2	3	4	5	6	7	8	9
Initial plosives						+			
Initial fricatives	+	+	+	+	+	+		+	
Initial nasals		–							
Initial approximants									
Medial plosives	+	–		+		+	+		
Medial fricatives			–	+	+		+	+	
Medial nasals			+					+	+
Medial approximants									–

When initial position and medial position are compared with each other, it can be seen that more categories improve significantly in medial position. The mean scores, both for the pre test and the post test are also higher in medial position. It seems to be easier for TE speakers to produce consonants between vowels (and apparently it is also easier to apply the techniques learned for consonants in medial position). In chapter 2 it was also found that initial position is the most difficult position for TE speakers.

Table 5.5 summarizes the significant increases and decreases in score for the individual speakers per manner of articulation and per position.

5.3.1.2 Vowels

For the vowels, the inter-listener reliability was .78 for the pre test and .76 for the post test, respectively. One of the ten listeners was left out for this particular part of the experiment as this person failed to complete this vowel test.
Table 5.6 shows the individual overall scores for the pre and post test.

Table 5.6: Individual and mean scores for vowels in the pre and post test, plus the difference value. The bold italicized numbers indicate the speaker with the greatest deterioration; the underlined numbers indicate the highest improvement.

	1	2	3	4	5	6	7	8	9	Mean
Pre	97	97	97	***98***	89	92	97	<u>76</u>	90	93
Post	98	99	93	***91***	91	92	99	<u>86</u>	95	94
Difference value	1	2	-4	***-7***	2	0	2	<u>10</u>	5	1

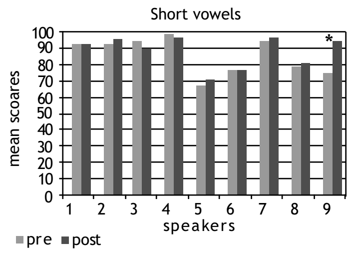
The first thing to notice from Table 5.6 is that the overall scores are all very high, except the pre test score of speaker 8. This is the person with most room for improvement, which he exploits rather well with an absolute improvement of 10%. Speakers 3 and 4 show a decrease in score, but their post test scores are still close to normal laryngeal scores (Koopmans-van Beinum, 1980). As with the consonants, it is also interesting to look at the scores for the types of vowels: long and short vowels and diphthongs. Table 5.7 shows these results.

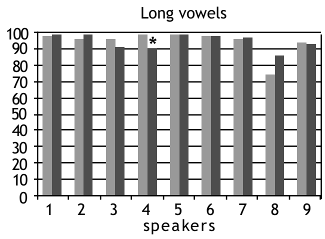
Table 5.7 shows that none of the categories improve significantly, which was to be expected considering the high overall pre test scores. The short vowels show the lowest score, which is in contrast with what was found in chapter 2, where the long vowels were the most difficult category. However, all the overall scores for the vowels found in this experiment are higher than the overall scores discussed in chapter two. The fact that the subject groups are different in the two experiments and the fact that different speech material was used, may partly explain the difference in findings.

Table 5.7: Mean scores per vowel type and overall scores with the ranges between brackets and the significance in the last column.

	Pre test (range)	Post test (range)	Sign. (range)
Short vowels	87 (68-100)	89 (72-98)	NS
Long vowels	95 (75-100)	96 (87-100)	NS
Diphthongs	96 (74-100)	96 (80-100)	NS
Mean	93 (76-98)	94 (86-99)	NS

As the individual improvements are important as well, the following figures show the individual scores per vowel category.





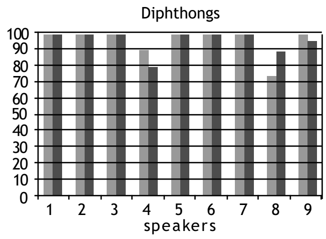


Figure 5.4a: Individual scores for short vowels. * indicates significant difference.

Figure 5.4b: Individual scores for long vowels. * indicates significant difference.

Figure 5.4c: Individual scores for diphthongs

The figures above show that similar scores for the pre and post tests are found for almost all speakers. Exceptions are the scores of speaker 9 for the short vowels, where a significant improvement is found from 76% to 96% correct. For the long vowels speaker 4 is the exception with a significantly lower score for the post test (from 100% to 91%). However, even the lower post test score for speaker 4 is still very high.

5.3.1.3 SUS sentences

The inter-listener reliability for the SUS intelligibility pre test is .88 and for the post test .85, respectively. Even though a perception task with Semantically Unpredictable Sentences (SUS) is difficult since every word in the sentence has to be scored correctly, the reliability of this test is higher than the reliability for the consonants and vowels.

Table 5.8 shows the individual percentage correct scores and the mean scores for the SUS sentences in the pre and post test.

It is apparent from Table 5.8 that there are large individual differences with pre test scores ranging from 20% to 80% and post test scores from 34% to 86%. It can also be seen that all speakers improve their scores, most of them quite considerably, with the greatest improvement for speaker 6 who also had most to gain. However, significance could only be established for speakers 4 and 7 (p<.05) and speakers 5 and 6 (p<.01). It should be emphasized that also the overall improvement is significant (p <.01).

Table 5.8: Individual scores for the SUS sentences and the mean score, plus the difference value. The underlined percentages indicate the greatest improvement. * indicates a significant difference.

	1	2	3	4	5	6	7	8	9	Mean
Pre	46	80	70	26	32	<u>20</u>	70	30	40	46
Post	64	86	78	44*	54*	<u>46*</u>	86*	34	44	60*
Difference value	18	6	8	18	22	<u>26</u>	16	4	4	14

5.3.2 Dyva phoneme identification test with Speech Language Pathologists

Inter-listener reliability for the Dyva test results is .72 for the pre test and .71 for the post test based on six speech language pathologists.

Even though in this test the target phonemes consist of consonants, consonant clusters and vowels, in the final scoring no distinction is made between these different categories. The test in its original set-up is not concerned with the *kind* of mistakes made, but only with the *number* of mistakes made. Based on the number of confusions, speakers belong to one of five categories (see also appendix 5.3) according to the protocol of this test: normal (0 mistakes), lightly impaired (maximum of 10 wrongly produced phonemes/clusters), moderately impaired (11-20 wrongly produced phonemes/ clusters), severely impaired (21-30 wrongly produced phonemes/clusters) and maximally impaired (over 30 wrongly produced phonemes/clusters). A maximum of 68 errors could be made. None of the speakers in this study were rated severely or maximally impaired. Table 5.9 shows which category the individual speakers belong to in the pre and post test (average of six listeners).

It can be observed in Table 5.9 that both for pre and post test the listeners are rated as lightly impaired (2). Hardly any improvement can be observed when categorizing the data like this and no significant effect is found when the original rules for scoring from the Dyva test are employed. As the Dyva test in the present study is used for a different category of 'pathological' speakers, and dysarthric patients are obviously not comparable to laryngectomized individuals, this study also looked at the percentage correct score of the individual speakers for the pre and post test. Table 5.10 shows those results.

Table 5.9: Individual and mean scores for the pre and post test based on the categories normal (1), lightly (2) and moderately (3) impaired and averaged over six listeners. None of the speakers belonged to category 4 (severely) or 5 (maximally).

Speakers	Pre test	Post test
1	2.2	2
2	1.8	1.5
3	1.8	2
4	2	2
5	2	2
6	2.2	2.2
7	2	2
8	2.2	2
9	2.2	2
Mean	2	1.7

Table 5.10: Individual score for the 'dyva' test, the difference value and the overall score in percentages. The bold and italicized percentages indicate the greatest deterioration, the underlined percentages indicate the greatest improvement. * indicates a significant difference.

	1	2	3	4	5	6	7	8	9	Mean
Pre	89	95	96	***91***	94	89	92	87	<u>84</u>	91
Post	93*	97	96	***90***	94	91	94	90	<u>95*</u>	93*
Difference value	4	2	0	***-1***	0	2	2	3	<u>11</u>	2

Table 5.10 shows that all speakers scored high in the pre test. This means that there is not much room for improvement. However, six out of nine speakers improve their score in the post test, with speaker 9 showing the greatest improvement. Even though the improvements are only small and only significant for speaker 1 ($p<.05$) and 9 ($p<.01$), the overall improvement is significant. This way of scoring shows a more positive result.

Even though this test was validated for a different type of speaker, the outcome of this validated test shows that overall the impairment in intelligibility at the phoneme level only seems to be moderate and that there is a trend towards a positive effect of the therapy program on TE speech intelligibility.

5.4 Discussion and conclusion

As was discussed in the introduction, it has become increasingly important to provide evidence as to why certain therapies are employed. Especially for total laryngectomy, the surgical techniques for voice and speech rehabilitation have changed so fundamentally in the last decades, that different/adjusted therapy strategies are in high demand. In view of these fundamental changes, with prosthetic voice rehabilitation presently being the method of choice for voice restoration in many parts of the world, it is surprising that there was no complete rehabilitation program available yet for prosthetic TE speakers. To fill that void, a therapy program was developed based on information obtained with a systematic literature search as was described in chapter 4. In the present chapter, part of the efficacy study has been described, meant to investigate if speech therapy, of which the content was based on specific TE speech problems, identified in the beginning of this research project (chapter 2 and 3), and on techniques presented in literature, can improve TE speech intelligibility. This question is one of the two main questions stated in the introduction. The second question, as to whether the set-up of the program was successful, will be discussed in chapter 7.

In this chapter, the phoneme and sentence intelligibility tests have been discussed. The tests concerned consonants, vowels and sentences and two listener groups were involved: phonetically trained listeners (who are, however, naïve listeners with respect to TE speech), and speech language pathologists (who are also trained listeners, but familiar with TE speech). When the mean scores found in this chapter are compared with those found in chapter 2, it can be seen that for the pre and post tests the scores are higher than in the first listening experiment. This is not unexpected, as there were now two groups of experienced listeners, whereas for the first experiment inexperienced and naïve listeners participated. This difference in score between experienced and inexperienced listeners has been found in literature as well (van As, 2001; Roeleven & Polak, 1999; Williams & Watson, 1985; Bridges, 1991; Hammarberg et al., 1992; Heaton et al, 1996 and Doyle et al., 1989). Furthermore, the two patient groups differ to some extent: in the rehabilitation part of the study there were 9 speakers, with a mean age of 64;9 (range 54;11-82;9) and average time after total laryngectomy (TLE) of 5;6 (range (0;6-13;7). In the first study (chapter 2) the 11 subjects had a mean age of 66;9 years (range 44-78) and average time after TLE of 9;4 (range 2;2-17;5). This means that the therapy group is slightly younger and had been operated more recently than the first group of subjects. Although differences are not very large, this may also have influenced results.

It was also found in this chapter that inter-listener reliability for consonant and vowel intelligibility, though sufficiently high, is much lower than the reliability scores in chapter 2 for consonants and vowels. The reliability score for the Dyva test is also low, but cannot be compared to any data from chapter 2. The inter-listener reliability scores for the SUS sentences are comparable. It has been discussed in literature that the reliability score depends on the type of rater used and the speech material.

Kreiman et al. (1990) found that naïve listeners tend to use similar strategies when judging similarity of voices, whereas clinicians differ greatly in the parameters they consider important when judging similarity of voices. Especially for pathological voices, disagreement between clinicians is found (Kreiman et al., 1992). This disagreement may partly be caused by the level of experience clinicians have with a particular pathology. Even though we chose raters that work with TE speakers on a daily basis, the level of experience did differ between raters. It can be assumed that for the highly experienced phonetically trained listeners similar disagreements can be found. These are then not based on the amount of experience with TE speech, but more on the different levels of expertise between listeners.

When the actual scores are considered, it was found that in initial position, only fricatives improved significantly. Fricatives are the most difficult category of consonants to produce for a TE speaker (see chapter 2) and most improvement could be obtained for these sounds. An improvement could even be seen for /x/ and /h/ separately, two phonemes that are notoriously difficult to improve and for which no evidence-based techniques existed. Fricatives were also trained extensively in the training program and the technique used had been described in previous studies and had been used in daily practice before. The fact that plosives have not improved may be due to the fact that initial plosives are always harder to hear as they start so abruptly. Additionally, in this case, the carrier phrase ended in a plosive. It is likely that it is more difficult to produce a plosive following another plosive than it is to produce a plosive after a vowel. In retrospect, other carrier phrases should have been used with more variation in the last phoneme of that phrase.

When the voiced-voiceless distinction is considered, it can be seen that perception, and thus the production, of this distinction has improved as well, but this score cannot be seen as separate from the fricative scores. The scores for nasals are already quite high in the pre test and close to normal laryngeal scores and hence leave little room for improvement. The approximant category consists of only one phoneme: /l/, which may have influenced results as the statistics are based on a limited number of realizations only. Moreover, no clear technique exists for improving articulation of /l/, which may also explain the lack of improvement.

In medial position, the situation is even more favorable, with all categories, except the approximant /l/, improving significantly. In medial position, the perception of plosives improves as well. This means that the 'push harder' technique that was described by Christensen & Dwyer (1990) and used in the speech training program of this study does improve the intelligibility of voiceless plosives. However, they tested this technique in initial position only where they found an effect, whereas in the present data no direct effect could be found for the plosives in initial position. A possible explanation for this difference between these data and theirs is that Christensen & Dwyer only tested one subject.

For vowels, no improvement could be established. The scores are already very high in the pre test leaving only limited room for further improvement, the so called 'ceiling effect'. The fact that the scores for all three vowel categories (long, short and diphthong) is higher than the vowel scores found in chapter 2, can be explained by the fact that real words were used instead of nonsense syllables. Even though it was tried to limit learning effects by using minimal pairs and by repeating some of the stimuli, the listeners will probably learn in which words the individual vowels occurred: if listeners learn that /ɑ/ and /aː/ only occur between /p/ and /s/, the only choice they have to make is whether they heard a long or a short vowel. The fact that the listeners are highly experienced will also have caused higher scores.

The scores for the SUS sentences on the other hand are lower in the pre test than the score found in chapter 2. As yet, no explanation is available as to why experienced listeners would score lower in this test than naïve listeners. However, improvement is found, with the post test score being similar to the score in chapter 2. Even though the score remains somewhat low, it is promising that sentence intelligibility has improved.

The Dyva phoneme intelligibility test showed mixed results, depending on how the test was scored. An overall improvement could only be found if all individual items were taken into account. However, when the five scoring categories from the test itself were used, the significance disappeared. Even though this test is standardized, it is in fact meant for dysarthric speakers, a completely different type of speakers. By using the original scoring categories, one basically compares TE speakers with dysarthric speakers. It can therefore be defended to compare individual items. Just as with the phonetically trained listeners, scores were already quite high prior to the onset of the training program and could hardly be further improved. The problem with this test is that the listener sees the word that is meant and then has to decide whether that word was pronounced correctly by the speaker. Once the word is known, it is hard to hear mistakes. Even though speech language pathologists are trained to listen analytically, it is a difficult task and this may have influenced results.

For all tests, age and time after TLE in relation to the scores found in the pre and post test and the actual improvement made (scores post test minus scores pre test) were investigated. For none of the tests, a relation could be established: neither age nor time after total laryngectomy seems to make a difference.

Conclusion

In the introduction two main questions were formulated:

1. Can TE speech intelligibility be improved by specific speech training?
2. Is the set-up of the therapy program appropriate?

The tests in this chapter have shown that after only five weeks of intensive training, improvement is found for many categories. Training consonant, vowel and sentence intelligibility seems to be beneficial for this group of speakers. However, a better articulation of phonemes alone is not sufficient for intelligible speech. In chapter 6 the other tests are discussed, focusing on spontaneous speech and personal perceptions of intelligibility. In chapter 7, an overall evaluation of the therapy program will be given, that will hopefully answer the second question.

Appendix 5.1 Selection criteria speakers

The selection criteria stated below were for internal use only as a guide for the researcher in the selection process.

1. What kind of operation did you undergo?
   Ask for reconstructions and other surgery in the head and neck area.
   Only include people with a *Total Laryngectomy without reconstructions*

2. Do you have problems producing voice or speech? ***Exclusion criterion!!***
   (Does it happen often that no audible sound can be produced or that you have to push really hard?) Explain that these types of problems should be solved by an ENT doctor and that our program focuses solely on speech intelligibility.

3. How long ago did you have your surgery? ***Inclusion: a minimum of 6 months ago***

4. What kind of voice prosthesis do you use? ***Include only Provox2 users, to avoid an extra variable***

5. Do you use an HME or a hands free valve? ***Speakers are asked only to speak with an HME and/or digital occlusion of the stoma during the study, to avoid an extra variable.***

6. On a scale of 1 to 10, how do you rate your own intelligibility?

7. Do you have a partner or close relative that you speak to often? And does this person have trouble understanding you?

8. Have acquaintances trouble understanding you?

9. Can strangers understand you?

10. Are you intelligible in company, at parties or in noisy environments, or do you avoid these situations?

11. Are you asked very often to repeat things; do people often ask ‘what did you say?’. If so, are these people strangers, or also acquaintances and relatives?

12. Can you repeat the following words (words are in Dutch)?

pijp	*tuit*	*koets*
thuis	*graad*	*saai*
gips	huid	hoed
patat	toestel	gister
plafond	schapen	fabriek
tapijt en tafelkleed	conducteur en controleur	

If it appears from questions 7-12 that the speaker has intelligibility problems and in particular has problems with the phonemes /p, t, f, s, x, h/, then it is likely he is a good candidate for this intervention study.

Appendix 5.2 Speech material phoneme identification task with phonetically trained listeners

Phonemes

Consonants in initial position (CV)

paa	die	goe	zaa	mie
pie	doe	faa	zie	moe
poe	kaa	fie	zoe	
baa	kie	foe	laa	
bie	koe	vaa	lie	
boe	haa	vie	loe	
taa	hie	voe	naa	
tie	hoe	saa	nie	
toe	gaa	sie	noe	
daa	gie	soe	maa	

Consonants in medial position (VCV)

aapaa	iedie	oegoe	aazaa	iemie
iepie	oedoe	aafaa	iezie	oemoe
oepoe	aakaa	iefie	oezoe	
aabaa	iekie	oefoe	aalaa	
iebie	oekoe	aavaa	ielie	
oeboe	aahaa	ievie	oeloe	
aataa	iehie	oevoe	aanaa	
ietie	oehoe	aasaa	ienie	
oetoe	aagaa	iesie	oenoe	
aadaa	iegie	oesoe	aamaa	

Vowels (CVC)

pas	beet	voet
paas	kip	touw
bom	kiep	neus
boom	fut	huis
bed	fuut	pijp

Nonsense sentences (SUS sentences)

subject 1:
Het land huilt langs de vieze naam
Het scherpe boek vormt de tuin
Wek de pijn en de stoel
Waar streelt het feit de smalle oom?
De zee prijst het dier dat smeekt

subject 2:
Het beeld stapt voor de dunne nacht
De lege krant vult de heer
Schenk het bed of de zoon
Wanneer wenst de staat het lange plan?
Het bos wast het geld dat praat

s***ubject 3:***
De taal komt na de volle lijn
De grote prijs slacht het been
Eis de hand of de vraag
Wanneer haat de rest de dikke vrouw?
De brief zoek de geest die gaat

subject 4:
De mens woont bij het lage woord
De diepe stem leent het volk
Blus het paar of de kerk
Waarom boeit de taak de trage maand?
Het blad schept de raad die kruipt

subject 5:
De lucht klimt naar de warme droom
De mooie school temt het jaar
Huur de wind en de broer
Waarom poetst de kaart de lauwe bank?
Het uur doodt de strijd die schrikt

subject 6:
De borst zit op de korte zak
De hete macht noemt het werk
Kweek het raam of de bloem
Waarom troost de tand de ruime weg?
De wijn scheert het paard dat zwijgt

subject 7:
Het huis loopt in het rare dorp
De vuile kracht neemt de fles
Zaai het hart of de kunst
Waarom groet de hond het gladde doel?
De dienst krijgt de eeuw die stijgt

subject 8:
De film rent uit de schone rug
De snelle grens meldt de stad
Keur het spel en de hoek
Waar waagt de arm de hoge muur?
De gang kookt de vorm die trilt

subject 9:
Het licht zwemt om het leuke vuur
Het koude glas ploegt de kop
Meng de deur en het ding
Waarom likt het lid de dure grond?
De week sloopt het schip dat zweeft

Appendix 5.3 Dyva: clinical test with Speech Language Pathologists

First, the original Dutch text is given. After that, the English translation of the scoring section is given.

Articulatie **SCORE:**

Woordenlijst naspreken

scoring

omcirkel fonemen en foneemclusters die bij eerste poging niet of niet goed werden gerealiseerd wanneer 2e poging wel succes had, zet dan een [2] achter de omcirkelde foneem/cluster

5=Normaal alles goed nagesproken

4=Licht gestoord maximaal 10 fonemen/clusters niet, niet goed gerealiseerd

3=Matig 11-20 niet, niet goed gerealiseerd

2=Ernstig 21-30 niet, niet goed gerealiseerd

1=Maximaal gestoord >30

scoring

Circle phoneme and phoneme clusters that were not pronounced correctly the first time. If a second attempt was successful, place a 2 after the circled phoneme or cluster.

5=Normal, everything pronounced correctly

4=Mildly impaired, incorrect pronunciation of a maximum of 10 phonemes/clusters

3=Moderately impaired, 11-20 phonemes/clusters not or not correctly pronounced

2=Severely impaired, 21-30 not, or not correctly pronounced

1=Maximally impaired, over 30 phonemes/clusters not, or not correctly pronounced

Dyva Woordenlijst

stimulus	respons	targetvocaal	targetcons	targetcons	opmerking
min		I	m-	-n	
nieuw		ie	n-	-w	
jaap		aa	j-	-p	
luik		ui	l-	-k	
poes		oe	p-	-s	
beul		eu	b-	-l	
koor		oo	k-	-r	
wijf		ij	w-	-f	
deeg		ee	d-	-g	
vet		e	v-	-t	
gum		u	g-	-m	
hopje		o	h-	-pj-	
zakje		a	z-	-kj-	
foutje		ou	f-	-tj-	
tuurt		uu	t-	-rt-	
ruist			r-	-st-	
plicht			pl-	-cht-	
smurf			sm-	-rf-	
plaatsen			pl-	-ts-	
zwartboek			zw-	-rtb-	
bruidspaar			br-	-dsp-	
schoensmeer			sch-	-nsm-	
grootscheeps			gr-	-tsch- -ps	
kleinbedrijf			kl- dr-	-nb-	
vlindernet			vl-	-nd- -rn-	

6

Therapy evaluation: semantic scaling and study-specific structured questionnaires

Abstract

When offering speech therapy it is necessary to investigate how effective this therapy is for a particular patient group. Effectiveness can be measured through pre- and post testing and checking whether any significant improvements can be found. To test the effect of the therapy program discussed in chapter 4, several pre and post tests were performed. In chapter 5 the phoneme and sentence intelligibility tasks were discussed. In this chapter the results of the study-specific structured questionnaires for speakers and close relatives, plus the validated questionnaire of the Voice Handicap Index (VHI-30) for the speakers, and a semantic scale judgment task for naïve listeners to judge speech intelligibility and voice quality of a read aloud text and a retold story are discussed.

Results show that TE speakers are more positive about their own intelligibility after completing the therapy program. In addition, the semantic scale judgments on intelligibility and articulation show positive effects. The other scales, however, show no improvement or even a deterioration. Overall it can be concluded that, although already effective in its present set-up, some extension of this 5 week rehabilitation program is necessary to ensure an optimal progression from phoneme level to spontaneous speech level.

6.1 Introduction

In chapter 5, two pre and post tests were described. These tests counted the actual number of errors made in the perception of phonemes and sentences. In this chapter, the three other tests will be discussed. These tests consist of study-specific structured questionnaires filled out by the speakers themselves as well as a close relative, plus the validated Voice Handicap Index (VHI-30), and a listening experiment performed by naïve and inexperienced listeners. The listening experiment will provide an overall impression of the intelligibility of read-aloud and semi-spontaneous speech by using semantic scale judgments instead of counting errors.

The study-specific structured questionnaire that TE speakers and their close relatives filled out contained questions on the impact of their intelligibility on daily life. In the end, it is most important if the speaker notices improvement in his speech and if the speaker is satisfied with the training. Although the Voice Handicap Index is meant to measure voice-quality-related quality of life rather than speech-quality-related quality of life, it was still included here as no validated speech-quality-related quality of life test was known.

In this chapter, the results of the tests mentioned above will be presented whereas in chapter 7, the therapy program as a whole (chapters 5 + 6) will be evaluated.

6.2 Subjects and methods

6.2.1 Subjects

Subjects were the same 9 male laryngectomized individuals as in chapter 5 (mean age 64;9, range 54;11-82;9).

6.2.2 Recordings

Recordings were made in a sound treated room using a Sennheiser MD421 Dynamic Microphone and an Edirol (Roland) R-1 portable 24 bit digital wave audio recorder. The maximum recording level was determined for each speaker and then fixed.

6.2.3 Pre and post tests

Study-specific structured questionnaires

The speakers and their relatives filled out the study-specific structured questionnaires before and after the therapy program (see appendix 6.1.). The questions used were in part taken from a laryngectomy quality of life questionnaire devised and in long-term use in the Netherlands Cancer Institute (Ackerstaff et al., 1994), and then complemented by specific questions formulated for the current study. The questions were all phrased as multiple-choice and required the expression of the opinion on how speech intelligibility of the subject affected daily life as in the following example:

Can your partner or relatives understand you?

a. poorly
b. moderately
c. fairly
d. well

The questions for the partners/relatives were phrased differently, but the information asked remained the same:

Can you understand your partner or relative?

a. poorly
b. moderately
c. fairly
d. well

The questionnaires were used to get an idea of how satisfied the speakers and their relatives are with the speech of the TE speaker and also if there is a difference in satisfaction rate between the TE speakers and their relatives.

The Dutch version of the Voice Handicap Index (VHI-30) was used (Hakkesteegt et al., 2006) because of the lack of a standardized intelligibility-related quality of life test. Though the VHI-30 is a validated, internationally accepted and standardized questionnaire, it has to be regarded with care as it tests voice-quality-related quality of life rather than speech-intelligibility-related quality of life. The VHI was introduced by Jacobson et al. (1997) and consists of 30 items, divided into three categories of ten items each. These categories consist of Functional (F), Emotional (E) and Physical (P) items. The scores range from 0-4 (0 indicating no problems and 4 indicating serious problems). A maximum score of 120 can be obtained, whereby 0 would be the best score and 120 the worst score, indicating more serious voice disorders. The reproducibility and internal consistency of the test are considered to be good. Various studies have since found that the VHI can be used to evaluate the effectiveness of various (voice) therapies. The test has been translated into Dutch (de Bodt et al., 2000) and the reproducibility of this Dutch version was investigated by Hakkesteegt et al. (2006) and was considered to be good.

As it is difficult for untrained people to distinguish between voice quality and speech intelligibility, the study-specific questionnaire was filled out *before* we recorded the speech material of the TE speaker and the VHI *after* the recordings (but in the same session). Before filling out the VHI form it was explained to the speaker what the difference between speech and voice is. By filling out the two tests at separate times, it was hoped that it would be possible to avoid confusions between voice and speech.

Semantic scaling experiment for naïve listeners

Speakers were asked to read out loud an adapted text from the newspaper about maltreated bears (see appendix 6.3) and were then told to retell the story. The story was adapted at a school for children with language disabilities, which mostly means that sentences have been made shorter. The reason why a simplified text was chosen is that all participants have to be able to read the text quite easily. A test text that was used first, turned out to be too difficult for TE speakers to read, because of long sentences and difficult grammatical structures.

Ten naïve listeners (mean age 52 years; range 40-68 years) were asked to evaluate speech intelligibility and voice-quality-related aspects of (semi) spontaneous and read-aloud speech. These types of speech come closest to normal connected speech. In this experiment, the naïve listeners rated how well TE speakers can be understood and how their voice sounds. For this purpose, semantic bipolar 7-point scales were used, similar to the listening experiment discussed in chapter 2. A more detailed discussion on the use of semantic scales can also be found in chapter 2. In Table 6.1, the semantic scales and their English translations can be found. One has to be aware that the English translations can have different connotations than the original Dutch terms.

In the present study more scales are used than in the first listening experiment described in chapter 2. The reason is that a judgment was wanted on specific aspects of TE speech, especially on those aspects that were trained during the therapy program. Voice quality is known to be affected as well and as this may influence the overall intelligibility, voice quality related scales were also included. The scales represent the following aspects: intelligibility, articulation, intonation, rhythm, loudness and voice quality. There is overlap in meaning for some of the scales. The difficulty with semantic scales is that the terms used may not mean the same to all listeners (Fagel et al., 1983). Results were expected to show that some terms give better results than others and that based on these results a subset of scales can be distinguished that represents the underlying perceptual dimension.

The scales were randomized, but after that the order was kept the same for all listeners. However, a clear distinction between the intelligibility and voice quality scales was maintained. By the time the listeners started with the voice quality scales it was pointed out to them that these scales concerned the quality of the voice and how they experienced the sound of the voice. This was done to avoid confusion between speech and voice. The last scale concerned an overall judgment. This scale had only three options: good, moderate and poor. It was included here as it had been proven to be a useful scale in the voice quality related perception experiment performed by Van As et al. (2003) and also in chapter 2.

Before the listeners started with the online experiment, an explanation was offered first on how to perform the test. The experiment started with a sound file with examples of all speakers, so that listeners could get used to the sound of the speakers and the variation between speakers. Speakers and pre and post test samples were randomized. Four items were repeated at the beginning and end of the experiment.

Table 6.1: The semantic scales in Dutch as used in the rating experiment and their English translations.

Dutch term	English term
Intelligibility scales	
1. moeilijk verstaanbaar - makkelijk verstaanbaar	difficult to understand - easy to understand
2. onduidelijke articulatie - duidelijke articulatie	unclear articulation - clear articulation
3. slordig - precies	sloppy - precise
4. slecht verstaanbaar - goed verstaanbaar	poor intelligibility - good intelligibility
5. klanken onduidelijk uitgesproken - klanken duidelijk uitgesproken	phonemes articulated unclearly - phonemes articulated clearly
6. eentonig - melodieus	monotonous - melodious
7. te langzaam spreektempo - goed spreektempo	speaking rate too slow - good speaking rate
8. niet vloeiend - vloeiend	non fluent - fluent
9. slecht spraakritme - goed spraakritme	poor speech rhythm - good speech rhythm
10. te snel spreektempo - goed spreektempo	speaking rate too fast - good speaking rate
11. traag - vlot	slow - fluent
12. saai - levendig	boring - lively
Voice quality scales	
13. afwijkend - normaal	deviant - normal
14. zacht - luid	soft - loud
15. niet natuurlijk - natuurlijk	not natural - natural
16. onaangenaam - aangenaam	unpleasant - pleasant
17. zwak - krachtig	weak - powerful
Overall judgment	
18. goed – matig - slecht	good – moderate - poor

6.2.4 Statistics

For the questionnaires, a Wilcoxon signed ranks test for related samples was used to investigate the difference between the pre and post test scores. A Wilcoxon test assumes neither normality, nor homogeneity of variance, but it does assume that the two samples are from populations with the same distribution shape. A Mann-Whitney test for independent samples was used to investigate the difference between the speakers and their relatives. This non-parametric test makes no assumptions on the distribution of the data and uses the median to compare groups rather than the mean, as is done in the t-test (Kinnear & Gray, 1999).

For the VHI-30, also a Wilcoxon signed ranks test for related samples was used to investigate the difference between the pre and post test scores. To investigate differences between the categories, an ANOVA was used with bonferroni correction.

For the semantic scales, first the reliability of each scale was calculated using Cronbach's alpha. Only scales with a coefficient higher than .70 were included for further analysis. A Principal Component (PC) analysis was performed to investigate the relations between the semantic scales. After the PC analysis a VARIMAX rotation was performed, by which the variance of the component loadings is maximized per factor. Only components with eigenvalues>1 were included. For a more detailed account of the statistics on semantic scales, see 2.2.5.

6.2.5 Praat

Samples of 30-35 s were chosen from the read-aloud text and the recounted story. Just as with the stimuli described in the previous chapter (5.2.5), the samples were amplified without affecting the differences between speakers.

6.3 Results

In this section, first the results of the different questionnaires will be discussed, followed by a discussion of the semantic scale judgments.

6.3.1 Questionnaires

6.3.1.1 Study-specific structured questionnaires for speakers and their relatives

The study-specific structured questionnaires (see appendix 6.1) consisted of eleven questions for the speakers and ten questions for the relatives with slight adaptations in the formulation. The questions were the same for both types of raters, except that the eleventh question applied only to TE speakers themselves. Each question had four options, with the first option being the most negative one and the last option the most positive. The numbers one to four were allocated as scores to these options.

In the following table the overall average scores are shown for the speakers and their relatives for the pre and post test scores. These scores are based on eight speakers and eight relatives, as one speaker failed to hand in his questionnaire (speaker 9, see chapter 5). A word of caution is necessary: the questionnaires are limited in length and there are only two raters per speaker: the speaker himself and one relative. The limited amount of data makes it difficult to perform statistics.

Table 6.2: Average scores for the speakers and their relatives on the questionnaire. The significance level indicates whether a difference exists between pre and post test (rightmost column) or between speakers and relatives (last row).

	Pre	Post	Significance
Speakers	2.88	3.18	$p<.01$
Relatives	3.21	3.35	NS
Significance	$p<.01$	NS	

From Table 6.2 it can be seen that, on average, speakers themselves scored significantly higher in the post test, meaning they were more positive about their intelligibility than at the start of the therapy. The judgment of the relatives did not improve significantly. Relatives generally appear to have a more favorable opinion of the speaker's intelligibility than the speakers themselves. This difference is significant in the pre test, but disappears in the post test.

It is also necessary to take individual scores into account. The following figures show the individual overall scores on the pre and post test for the speakers and their relatives.

The individual scores show that, even though the overall improvement for speakers was significant, only two speakers score significantly better: speaker 1 ($p<.05$) and speaker 2 ($p<.01$). However, six out of eight speakers do improve their score.
For the relatives in Figure 6.1b, it can be seen that scores are the same in the pre and post test for three relatives, worse in one case and that scores improve in four cases, though not significantly.

It is also interesting to look at the scores for each of the eleven questions that are shown in Table 6.3. It is however difficult to perform statistics, as there are only 8 speakers or relatives per question.

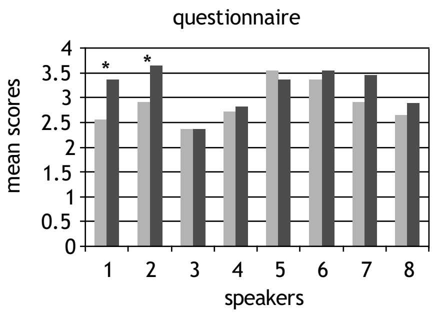


Figure 6.1a: Individual scores of speakers on the pre and post test. * indicates significant improvement.

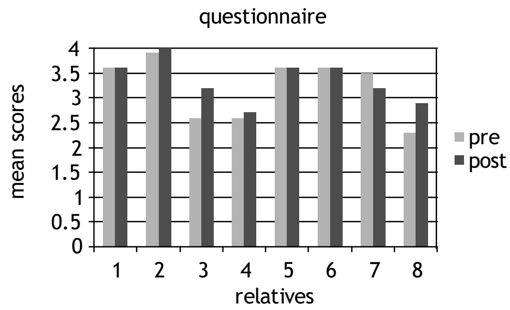


Figure 6.1b: Individual scores of relatives on the pre and post test.

Table 6.3: Scores per question for the speakers' and their relatives' pre and post test, plus difference value. NA means this question was not scored in the questionnaire of that relative. Scores are based on eight subjects.

Questions		1	2	3	4	5	6	7	8	9	10	11
speakers	Pre	3.4	3.3	3.1	1.7	3.1	2.9	2.2	2.9	2.0	3.2	3.2
	Post	3.6	3.4	3.1	2.2	3.3	2.9	3.2	3.2	2.6	3.3	3.6
	Diff. value	.2	.1	0	.5	.2	0	1	.3	.6	.1	.4
relatives	Pre	3.6	3.4	3.2	2.1	3.6	3.6	2.9	3.8	2.9	3.3	NA
	Post	3.7	3.4	3.3	2.8	3.7	3.4	3.2	3.4	2.9	3.1	NA
	Diff. value	.1	0	.1	.7	.1	-.2	.3	-.4	0	-.2	NA

It can be seen from Table 6.3 that the speakers show higher scores on all questions in the post test, but these improvements are mostly small. Three questions show a higher improvement: questions 4, 7 and 9. These were also the questions that initially scored lowest and hence had most room for improvement. Question 4 concerns the intelligibility at parties. It can be assumed that by training clear speech, intelligibility in especially these circumstances will improve. Question 7 concerns the speech rate. Again, this is an aspect that has been trained specifically during therapy. Question 9 concerns the pleasantness of the speech. This aspect has not been trained as such, but it is likely that when other aspects of intelligibility improve, the overall pleasantness will improve as well.

The relatives show a slightly different pattern: not all questions show improvement, but both increases and decreases in scores are small. Just as with the speakers, the largest improvements can be found for question 4 and 9, followed by question 7.

6.3.1.2 Voice Handicap Index (VHI-30)

Each subject filled out a VHI-30 questionnaire both at the start and at the end of the therapy program (see appendix 6.2). A lower score indicates a more positive rating. In Table 6.4 the overall scores per category are given.

Table 6.4: Mean and range of the total scores and the sub scores of the VHI forms, both for the pre and post test. The lower the score, the more positive the judgment. VHI-F: Functional category; VHI-E: Emotional category; VHI-P: Physical category. NS: statistically not significant. * indicates which subcategory scored significantly different from the other subcategories in each individual test.

	Pre test (range)	Post test (range)	Significance
VHI-total	38.9 (21-57)	33.4 (12-46)	NS
VHI-F	13.67 (8-22)	11.44 (3-17)	NS
VHI-E	8.44 (4-14)*	6.33 (1-16)*	NS
VHI-P	16.78 (7-24)	15.67 (4-26)	NS

When the scores in Table 6.4 are considered separately for the two tests, significant differences are found between the sub scores. Both in the pre test and the post test, the E-scores are significantly lower than the P-scores ($p<.05$ for pre-test and $p<.01$ for post-test), meaning that the physical discomfort is considered to be worse than the emotional discomfort. Although post test values are slightly lower (means better) than pre test scores, this difference is statistically not significant.

For this research it is important to know the individual scores on the VHI. According to Hakkesteegt et al. (2006) an overall difference of 14 points is significant. For the subscales F, E and P differences of 6, 7 and 8 points, respectively are significant.

Table 6.5: Differences between the pre and post test scores for the subcategories and the whole test, for all patients individually, where T stands for total, F for functional, E for emotional and P for physical. A + score means lower quality of life and a – score a better quality of life in the post test. The bold italicized numbers indicate a significant deterioration in score. The underlined numbers indicate a significant improvement.

	1	2	3	4	5	6	7	8	9
VHI-T	-2	-11	-5	-9	***+15***	0	-7	<u>-23</u>	-8
VHI-F	+2	-5	-2	+2	+5	+2	<u>-6</u>	<u>-11</u>	<u>-7</u>
VHI-E	-1	-3	-5	<u>-8</u>	***+8***	-1	-3	-3	-3
VHI-P	-3	-3	+2	-3	+2	-1	2	<u>-9</u>	+3

Only two subjects show a significant difference between the total scores of the first and second measurement. Contrary to expectations, one of these (speaker 5) shows a deterioration in quality of life. For the subcategories it is found that 4 subjects show a significant improvement for at least one category and 1 subject shows a significant deterioration on two categories.

6.3.2 Semantic scaling experiment for naïve listeners

Ten naïve and inexperienced listeners judged the subjects' speech by using semantic bipolar 7-point scales. Two types of speech material were used for this experiment: read-aloud text (see appendix 6.3) and a recount of the story in the text. Inter-rater reliability was .887 both for the retelling and reading aloud task in the pre test, and .872 and .884 for the retelling and reading aloud task in the post test, respectively. Table 6.6 shows the overall results for both types of speech material for all 18 scales together. The higher the mean score, the better speech intelligibility was judged to be.

Table 6.6: Overall score plus ranges for retelling and reading aloud in the pre and post test of the semantic scale judgment (NS: statistically not significant). The higher the score, the more positive the rating was.

	Pre (range)	Post (range)	Significance
Retelling	4.07 (2.97-4.98)	3.83 (2.64-4.29)	p<.01
Reading aloud	3.91 (2.98-5.07)	3.94 (3.03-4.97)	NS

Table 6.6 shows that the rating for the retelling task is significantly lower in the post test than in the pre test. The rating for the reading aloud task improves slightly, but not significantly.

As this study is more interested in the different aspects of TE speech intelligibility than the overall score, the separate scales were considered and it was investigated whether components (sets of scales) could be found for the two separate tasks in both the pre test and the post test. In order to do this, the inter-rater reliability was calculated for each separate scale using Cronbach's alpha. Only scales with an inter-rater reliability higher than .70 were included. Scales with a lower reliability were considered unreliable and excluded from further analysis. In Table 6.7, the mean scores per scale are shown for the pre test, together with the standard deviations and the correlations. In Table 6.8 the results of the post test are shown.

It can be seen for the pre test that for the retelling task, three scales score below .70: *sloppy – precise, monotonous – melodious*, and *speaking rate too fast – good speaking rate*. For the reading aloud task a different result was found. For this task the scales *non fluent – fluent*, *poor speech rhythm – good speech rhythm*, and *speaking rate too fast – good speaking rate* had to be excluded from further analysis.

In the post test it can be seen that for the retelling task the scales *poor speech rhythm – good speech rhythm* and *speaking rate too fast – good speaking rate* had to be excluded, whereas for the reading aloud task only the scale *speaking rate too fast – good speaking rate* was unreliable.

Table 6.7: Mean scores, ranges, standard deviations (SD) and Cronbach's alpha (α) for the individual scales for the retelling and reading aloud task in the pre test. The bold numbers in the α-column indicate high reliability.

Pre test	**Retelling**			**Reading aloud**		
	Mean (range)	**SD**	**α**	**Mean (range)**	**SD**	**α**
Intelligibility scales						
1. difficult to understand-easy to understand	3.93 (2.5-5.10)	1.75	**.86**	3.64 (2.6-5.2)	1.67	**.82**
2. unclear articulation-clear articulation	4.02 (2.8-5.1)	1.59	**.81**	4.03 (3-5.2)	1.34	**.82**
3. sloppy-precise	4.37 (3.5-5.1)	1.36	.56	4.48 (3.6-5.4)	1.24	**.78**
4. poor intelligibility-good intelligibility	4.04 (2.2-5.5)	1.78	**.89**	3.84 (2.5-5.5)	1.61	**.88**
5. phonemes articulated unclearly-phonemes articulated clearly	4.16 (2.5-5.4)	1.47	**.84**	3.89 (2.3-5.3)	1.5	**.88**
6. monotonous-melodious	3.94 (2.67-5.1)	1.31	.66	3.69 (2.5-4.4)	1.18	**.70**
7. speaking rate too slow-good speaking rate	4.92 (4.0-5.7)	1.37	**.85**	4.78 (4-5.8)	1.23	**.83**
8. non fluent-fluent	4.34 (2.9-5.8)	1.52	**.87**	4.38 (3.4-5.6)	1.49	.55
9. poor speech rhythm-good speech rhythm	4.59 (3.7-5.8)	1.38	**.83**	4.42 (3.6-5.5)	1.39	.69
10. speaking rate too fast-good speaking rate	4.99 (3.6-5.7)	1.23	.52	4.71 (4.3-5.5)	1.21	.29
11. slow-fluent	4.8 (3.6-5.7)	1.23	**.82**	4.64 (3.6-5.7)	1.23	**.86**
12. boring-lively	4.12 (2.7-5.2)	1.36	**.81**	3.93 (2.8-5.3)	1.16	**.88**
Voice quality scales						
13. deviant-normal	3.09 (1.5-4.7)	1.67	**.89**	2.81 (1.4-4.7)	1.36	**.85**
14. soft-loud	4.57 (3.5-6.0)	1.12	**.89**	4.34 (3.0-5.8)	1.38	**.92**
15. not natural-natural	3.32 (1.6-5.1)	1.65	**.88**	3.08 (1.5-5.1)	1.46	**.93**
16. unpleasant-pleasant	3.72 (1.8-5.1)	1.54	**.92**	3.48 (1.7-5.3)	1.43	**.92**
17. weak-powerful	4.42 (3.8-5.44)	1.26	**.83**	4.3 (3.2-5.8)	1.4	**.85**
Overall judgment						
18. good-moderate-poor	1.86 (1.2-3)	.76	**.92**	2.02 (1-3)	.76	**.91**

Table 6.8: Mean scores, ranges, standard deviations and Cronbach's alpha (α) for the individual scales for the retelling and reading aloud task in the post test. The bold numbers in the α-column indicate high reliability.

Post test	Retelling			Reading aloud		
	Mean (range)	SD	α	Mean (range)	SD	α
Intelligibility scales						
1. difficult to understand-easy to understand	3.8 (1.6-4.9)	1.78	**.85**	3.9 (2.4-5.5)	1.83	**.88**
2. unclear articulation-clear articulation	4.13 (202-5.0)	1.55	**.83**	4.3 (2.9-5.7)	1.5	**.88**
3. sloppy-precise	4.43 (3.0-5.1)	1.33	**.77**	4.74 (3.7-6.0)	1.31	**.85**
4. poor intelligibility-good intelligibility	4.19 (1.7-5.2)	1.74	**.89**	4.23 (2.7-5.7)	1.71	**.88**
5. phonemes articulated unclearly-phonemes articulated clearly	4.22 (1.9-4.9)	1.5	**.88**	4.26 (2.9-5.6)	1.52	**.73**
6. monotonous-melodious	3.72 (2.6-4.9)	1.32	**.85**	3.88 (2.9-4.7)	1.3	**.81**
7. speaking rate too slow-good speaking rate	4.38 (3.6-5.0)	1.38	**.72**	4.7 (3.6-5.4)	1.29	**.78**
8. non fluent-fluent	3.88 (2.9-4.5)	1.41	**.72**	4.44 (3.0-5.2)	1.49	**.70**
9. poor speech rhythm-good speech rhythm	4.15 (3.5-4.7)	1.29	.66	4.53 (3.2-5.4)	1.42	**.76**
10. speaking rate too fast-good speaking rate	4.77 (4.2-5.2)	1.1	.54	4.99 (4.4-5.5)	1.05	-.33
11. slow-fluent	4.22 (3.4-5.1)	1.27	**.73**	4.31 (3.3-4.9)	1.23	**.78**
12. boring-lively	3.86 (2.4-5.5)	1.36	**.88**	3.83 (2.8-5.0)	1.34	**.78**
Voice quality scales						
13. deviant-normal	2.8 (1.3-4.2)	1.5	**.86**	2.83 (1.4-4.7)	1.54	**.86**
14. soft-loud	3.99 (2.7-5.2)	1.1	**.83**	3.66 (2.8-4.6)	1.04	**.86**
15. not natural-natural	3.1 (1.5-4.3)	1.4	**.83**	3.1 (1.6-4.6)	1.47	**.91**
16. unpleasant-pleasant	3.61 (1.8-4.5)	1.29	**.86**	3.6 (2.0-5.2)	1.42	**.90**
17. weak-powerful	4.01 (2.5-5.1)	1.13	**.81**	3.71 (2.9-4.9)	1.23	**.78**
Overall judgment						
18. good-moderate-poor	1.87 (1.5-2.9)	.65	**.90**	1.88 (1.0-2.8)	.68	**.93**

A Principal Component (PC) analysis was performed on the remaining scales of the pre and post tests. The results for the retelling task and the reading aloud task in the pre test can be found in Tables 6.9 and 6.10, respectively, whereas the results for the retelling task and the reading aloud task in the post test can be seen in Tables 6.11 and 6.12 respectively.

Table 6.9: The scales with the factor loadings per component and the communalities (common factor variance) for the retelling task in the pre test. Factor loadings higher than .50 are taken into consideration and are in bold. The percentage of variance explained is given between brackets in the column headings.

Retelling pre test	Comp. 1 (54.84%)	Comp. 2 (12.8%)	Comp. 3 (10.65%)	Communalities
Intelligibility scales				
1. difficult to understand-easy to understand	**.68**	.47	.06	.68
2. unclear articulation-clear articulation	**.84**	.30	-.02	.79
4. poor intelligibility-good intelligibility	**.75**	.50	.02	.80
5. phonemes articulated unclearly-phonemes articulated clearly	**.63**	.55	.05	.70
7. speaking rate too slow-good speaking rate	**.83**	.03	.30	.79
8. non fluent-fluent	**.78**	.24	.30	.76
9. poor speech rhythm-good speech rhythm	**.85**	.19	.16	.79
11. slow-fluent	**.74**	.10	.45	.77
12. boring-lively	.38	.40	**.59**	.65
Voice quality scales				
13. deviant-normal	.12	**.92**	.19	.89
14. soft-loud	.08	.06	**.85**	.74
15. not natural-natural	.19	**.90**	.26	.92
16. unpleasant-pleasant	.28	**.90**	.06	.89
17. weak-powerful	.17	.23	**.84**	.78
Overall judgment				
18. good-moderate-poor	-.38	**-.79**	-.21	.81

Table 6.10: The scales with the factor loadings per component and the communalities (common factor variance) for the reading aloud task in the pre test. Factor loadings higher than .50 are taken into consideration and are in bold. The percentage of variance explained is given between brackets in the column headings.

Reading aloud pre test	Comp. 1 (46.64%)	Comp. 2 (12.64%)	Comp. 3 (10.08%)	Comp. 4 (7.2%)	Communalities
Intelligibility scales					
1. difficult to understand-easy to understand	**.81**	.25	.20	.09	.76
2. unclear articulation-clear articulation	**.80**	.13	.36	.16	.81
3. Sloppy-precise	**.81**	.10	.03	.14	.69
4. poor intelligibility-good intelligibility	**.85**	.33	.12	.15	.85
5. phonemes articulated unclearly-phonemes articulated clearly	**.77**	.27	.21	.28	.80
6. monotonous-melodious	.15	.37	.39	.30	.40
7. speaking rate too slow-good speaking rate	.27	-.03	**.87**	.02	.83
11. Slow-fluent	.20	.10	**-.83**	.21	.79
12. boring-lively	.06	.41	**.66**	.17	.64
Voice quality scales					
13. deviant-normal	.15	**.85**	-.02	.18	.78
14. soft-loud	.24	.15	.17	**.86**	.86
15. not natural-natural	.22	**.89**	.12	.13	.87
16. unpleasant-pleasant	.19	**.88**	.21	-.07	.87
17. weak-powerful	.21	.20	.16	**.88**	.88
Overall judgment					
18. good-moderate-poor	-.34	**-.76**	-.17	-.15	.75

Table 6.11: The scales with the factor loadings per component and the communalities (common factor variance) for the retelling task in the post test. Factor loadings higher than .50 are taken into consideration and are in bold. The percentage of variance explained is given between brackets in the column headings.

Retelling post test	**Comp. 1 (48.8%)**	**Comp. 2 (13.4%)**	**Comp. 3 (11.2%)**	**Comp. 4 (7.1%)**	**Communalities**
Intelligibility scales					
1. difficult to understand-easy to understand	**.75**	.31	.26	.01	.74
2. unclear articulation-clear articulation	**.91**	.19	.10	.10	.88
3. sloppy-precise	**.85**	.17	.06	.25	.81
4. poor intelligibility-good intelligibility	**.82**	.26	.35	.09	.87
5. phonemes articulated unclearly-phonemes articulated clearly	**.86**	.26	.15	.22	.88
6. monotonous-melodious	.14	**.69**	.37	.36	.77
7. speaking rate too slow-good speaking rate	.30	**.86**	-.01	-.07	.83
8. non fluent-fluent	.32	**.82**	.23	.03	.83
11. slow-fluent	.15	**.72**	-.05	.40	.71
12. boring-lively	.28	**.72**	.14	.40	.77
Voice quality scales					
13. deviant-normal	.08	.06	**.91**	.09	.85
14. soft-loud	.14	.29	.17	**.78**	.74
15. not natural-natural	.06	.11	**.89**	.20	.84
16. unpleasant-pleasant	.38	.15	**.80**	.06	.81
17. weak-powerful	.11	.10	.23	**.85**	.84
Overall judgment					
18. good-moderate-poor	-.47	-.12	**-.65**	-.27	.73

Table 6.12: The scales with the factor loadings per component and the communalities for the reading aloud task in the post test. Factor loadings higher than .50 are taken into consideration and are in bold type. The percentage of variance explained is given between brackets in the column headings.

Reading aloud post test	Comp. 1 (51.8%)	Comp. 2 (12.9%)	Comp. 3 (8.0%)	Communalities
Intelligibility scales				
1. difficult to understand-easy to understand	**.81**	.32	.179	.79
2. unclear articulation-clear articulation	**.85**	.23	.11	.79
3. Sloppy-precise	**.66**	.24	-.17	.52
4. poor intelligibility-good intelligibility	**.82**	.41	.19	.88
5. phonemes articulated unclearly-phonemes articulated clearly	**.76**	.27	.18	.69
6. monotonous-melodious	.35	.44	.49	.55
7. speaking rate too slow-good speaking rate	**.78**	-0.6	.40	.77
8. non fluent-fluent	**.74**	-.01	.43	.74
9. poor speech rhythm-good speech rhythm	**.78**	.09	.42	.80
11. Slow-fluent	**.59**	.16	.58	.72
12. boring-lively	.43	.33	**.59**	.65
Voice quality scales				
13. deviant-normal	.10	**.86**	.19	.78
14. soft-loud	.14	.26	**.80**	.73
15. not natural-natural	.09	**.85**	.33	.84
16. unpleasant-pleasant	.27	**.85**	.15	.81
17. weak-powerful	.08	.28	**.72**	.61
Overall judgment				
18. good-moderate-poor	-.44	**-.68**	-.26	.72

Table 6.9 shows the components found for the retelling task in the pre test. The first component consists partly of articulation/intelligibility scales and partly of rate and rhythm scales. This component explains 54.84 % of the variance. The second component is a little smaller. It consists of voice quality scales that concern the overall impression of pleasantness or naturalness of the voice, and the overall judgment scale, which apparently is perceived by the listener as an overall judgment of voice quality rather than speech intelligibility. This component explains 12.8% of the variance. The third component is the smallest, with only three scales: one scale on how boring or lively the speech is and two on loudness of the speech. This scale explains 10.65% of the variance.

Table 6.10 shows the components found for the reading aloud task in the pre test. For the reading aloud task, one extra component is found. The first component now consists entirely of articulation/intelligibility scales and explains 46.64% of the variance. The second component, just as with the retelling task, consists of voice quality scales concerned with the general impression of the sound of the voice and again the overall judgment scale. It explains 12.64% of the variance. The third component again includes the scale *boring-lively*, but this time combined with the scales on rate. The variance explained is 10.08%. The loudness scales are represented by component four, which explains 7.2% of the variance.

The scale *monotonous – melodious* does not have sufficiently high factor loadings and cannot be included in any of the components.

From Table 6.11 it can be seen that four components can be distinguished for the retelling task in the post test. That is one component more than in the pre test. The first component consists solely of the articulation/intelligibility scales and explains 48.8% of the variance. The second component concerns the scales on pitch, rate and rhythm. These scales belonged to component one in the pre test. In the post test the listeners seem to be able to distinguish articulation from scales concerned with pitch, rhythm and rate. The second component explains 13.4% of the variance. The third component consists of the voice quality scales concerned with the sound of the naturalness or pleasantness of the voice and the overall judgment and is the same as component 2 in the pre test. It explains 11.2% of the variance. The last component consists of scales on loudness and explains 7.1% of the variance. It is very similar to component 3 in the pre test, except that the scale *boring-lively* is now part of component two.

For the reading aloud task in the post test (Table 6.12) three components can be distinguished, which in this case is one component less than in the pre test. The first component again consists of the articulation/intelligibility scales, but also of the rate and rhythm scales. This component is in fact equal to component 1 of the retelling task in the pre test. 51.8% of the variance is explained by this component. The second component consists of the voice quality scales (except loudness) and the general judgment scale, thereby being equal to component 2 in the pre test. It explains 12.9% of the variance. The third component includes the scales on loudness and the scale boring-lively and is equal to component 3 in the retelling task in the pre test. The third component explains 8% of the variance.

Just as in the pre test, the scale *monotonous – melodious* does not have sufficiently high factor loadings to be included in one of the components. It might be that this scale is not applicable to a reading task, as reading intonation differs greatly from intonation used in spontaneous speech.

The Principal Component analyses show that for each task and for both the pre and post test, the scales on articulation and intelligibility, just as the scales on rate, rhythm and pitch are the most important. In paragraph 6.2.3 it was stated that the scales were meant to represent the following categories: intelligibility, articulation, intonation, rhythm, loudness, voice quality and an overall judgment. When these categories are compared to the components found, it can be seen, however, that this division does not match the components found in this chapter. Listeners do not distinguish, for example, between articulation and intelligibility. The categories rhythm and intonation are also taken together and the general judgment is always a judgment on voice quality rather than intelligibility. Fewer categories are used by the listener than were anticipated beforehand. Unfortunately, the components found for the tasks, and the scales belonging to a certain component, differ somewhat for the pre and post test and for the two tasks. This makes it harder to make comparisons between pre and post tests. It was hoped that with a reduction of scales, the overlapping scales of the pre and post tests could be included in the reduced components, which would then make comparison of pre and post test scores much easier. In addition, it would mean that in future experiments fewer scales would be needed to evaluate TE speech. The criteria used for reducing the set of scales were first described by Fagel et al. (1983). Three main criteria are stated in that article: 1) each of the components should be represented by at least two scales, 2) scales with high communalities in the component solution are preferred over scales with a low communality and 3) scales with a high reliability are preferred over scales with a lower reliability. When applying these criteria to the data in this study, unfortunately no components could be established that were the same for the pre and post test. In addition, the amount of variance decreased for the reduced sets. A reduction of scales was therefore considered not to be useful in this particular study.

To make comparison possible, it was decided to take the components from the pre test and to compare the scores of the scales belonging to these components with the scores in the post test. The name given to the components is meant to represent the common theme of the component. For that reason, it has been decided to choose the term ‘intelligibility’ for both the intelligibility and the articulation scales. The following tables give the scores per component for the pre and post test. The scales that are included in the category are given in numbers between brackets (see also Table 6.9 and 6.10).

Table 6.13a: Scores for the different components for the pre and post test for the retelling task, based on the pre test components. The number of the scales included in the categories are given between brackets. The higher the score, the more positive the rating was.

Retelling			
Category	**pre test**	**post test**	**sign.**
Intelligibility/rate/rhythm (1, 2, 4, 5, 7-9, 11)	4.35	4.12	p<.01
Voice quality (13, 15, 16)	2.99	2.82	p<.05
Liveliness/loudness (12, 14, 17)	4.36	3.94	p<.01

Table 6.13b: Scores for the different components for the pre and post test for the reading aloud task, based on the pre test components. The number of the scales included in the categories are given between brackets. The higher the score, the more positive the rating was.

Reading aloud			
Category	**pre test**	**post test**	**sign.**
Intelligibility (1, 2, 3, 4, 5)	3.98	4.28	p<.01
Voice quality (13, 15, 16)	2.85	2.85	NS
Liveliness/rate (7, 11, 12)	4.46	4.29	p<.05
Loudness (14, 17)	4.32	3.68	p<.01

From Table 6.13a above it can be seen that intelligibility deteriorates for the retelling task. This component also consists of the scales on rhythm and rate. In the reading aloud task, however, intelligibility improves according to the scores. The voice quality scales deteriorate slightly in the retelling task and stay the same in the reading aloud task. It has to be kept in mind, however, that there were no exercises in the program to improve voice quality. Component 3 of the retelling task includes the scales on loudness and liveliness. Even though lower scores are seen for the post test, it does not necessarily mean the speaker has performed worse. Just as for the scales on liveliness and rate (component 3 for the reading aloud task) the clear speech of the training program advocates a slower speaking rate and a comfortable loudness level without straining the voice, rather than very loud speech. The scores in this test suggest that speakers actually employed this technique correctly, but in a somewhat exaggerated manner. This may have caused their speech to sound more boring or unnatural.

6.4 Discussion and conclusion

The tests discussed in this chapter provide valuable information on how the speakers perceive their own intelligibility (questionnaire and VHI), how a close relative perceives their intelligibility (questionnaire) and how several aspects of intelligibility are judged to be by naïve listeners (semantic scaling experiment).

The study-specific structured questionnaires showed overall improvement for the speakers. At closer inspection, this result was mainly based on the improvement of two speakers. However, it is difficult with a short questionnaire consisting of 4-point answers (1-4) to obtain significant results. Nevertheless, subjects apparently were more critical about their own intelligibility than their relatives were. After completing the therapy program, however, they felt more comfortable with their speech and this difference of opinion with their relatives disappeared. The fact that the relatives did not improve their scores is likely to be caused by the fact that mean scores were already quite high. Improvement would have meant near perfect scores, which even after intensive training would be very unlikely.

The validated Voice Handicap Index (VHI) was also included in this study, mostly because as far as the author is aware no standardized intelligibility related quality of life questionnaire exists. The VHI is normally used to measure the effect of voice quality on quality of life. Even though intelligibility was investigated in this study, it was thought it might be interesting to use VHI since the intelligibility might influence the answer on many of the questions. The lower (=better) scores found for the subcategory *emotion* match the personal experience of the researcher and the subjects: this specific group of subjects is very aware of the functional and physical implications of a low intelligibility and motivated to improve it, but generally does not encounter any specific emotional problems related to their intelligibility. At the moment, it is not possible to explain why one subject (nr. 5) scored so much higher (meaning worse quality of life) on the post-test. The subject told the researcher the therapy had made him more aware of his speech and articulation. It might mean he has also become more aware of his limitations, but this is conjecture. The only subject (nr. 8) that showed an overall significant improvement was a highly motivated subject that practiced diligently for at least an hour a day. He was very positive about his improvements even though it could not be shown to him at the time that an actual improvement had taken place. Still, it is interesting that his positive personal experience has been reflected in an improved VHI score. Even though results of the VHI have been discussed here, the results need to be interpreted with great care. In retrospect, it is not verifiable whether the subjects had their voice quality or speech intelligibility in mind when filling out the form irrespective of the instructions. It is not believed that the VHI-30 is a suitable test in intelligibility studies and hopefully a speech intelligibility related quality of life test will appear soon.

The results of the semantic scale judgments were somewhat disappointing: the score for the reading aloud speech improved, but not significantly, whereas the scores for retelling the story decreased significantly in the post test. At this point, one can only speculate as to why these results were found. Relatively little time was devoted to connected (spontaneous) speech in the therapy program. In only three sessions, subjects were instructed on clear speech and then asked to practice on sentences, short reading passages and short monologues and dialogues. It may be the case that more time is needed for the progression from phoneme/word level to spontaneous speech. The fact that reading aloud was practiced during therapy probably explains the fact that a (non significant) improvement was found for the reading aloud task, but not for the retelling of the story (which was not practiced). In addition, the use of clear speech may not be an automatic process yet. Another reason for the low scores could be that retelling a story is a difficult task. Even though the subject was told the information did not need to be correct, subjects did try and remember what they had been reading, causing a slow speech rate and many hesitations. And since patients after the therapy were trying harder, the retelling task became more difficult and is even less comparable with running spontaneous speech. In retrospect, it would have been better to let the subjects talk spontaneously about matters that occur in their daily life.

As was already mentioned in the introduction of chapter 5, the fact that in one or the other test no positive result is found obviously might be caused by the fact that there is indeed no actual improvement, but this might also be caused by the fact that the underlying test is not suitable or accurate enough to detect meaningful differences. A rating task with semantic scales is always a difficult undertaking to perform and to analyze. Especially for naïve raters it is difficult as they do not always know what exactly the terms mean on either side of the scale. The scales might not be sensitive enough to detect small changes in speech. The type of test used for naïve raters may need to be reconsidered. However, it is important to include naïve raters as they best represent the people TE speakers encounter in daily life. It is then also logical that naïve raters should be used to evaluate connected speech. Even though semantic scales may not be the perfect test, as far as the researcher knows, it is the only way to obtain an overall impression of someone's speech intelligibility.

Conclusion

Overall, the tests discussed in this chapter show some improvement although smaller and less unambiguous than the tests described in chapter 5. Where semi-spontaneous speech is concerned, even a decrease in intelligibility is suggested. There are several explanations which could explain these results. Subjects become more aware of the limitations when performing a difficult task and longer training is needed to translate all the improvements on the phoneme and word level to the discourse level. However, our data also suggest that the VHI is most likely not adequate for measuring speech intelligibility. The semantic scales might not be adequate in measuring possible changes over time due to the therapy program. This leaves the results for the study-specific structured questionnaires as the most useful outcomes of this part of the study. And, despite the obvious limitations of such questionnaires, the subjects themselves did consider their intelligibility to be improved, which seems a relevant and important result. Apparently, therapy has boosted their confidence and has made them more satisfied with their speech. In chapter 7 the therapy program and the pre and post tests will be evaluated and recommendations for the future will be made. Chapter 8 will contain the final (overall) discussion and conclusion of this thesis.

Appendix 6.1 Study-specific structured questionnaires

For the speakers (first the Dutch version will be given, followed by an English translation)

Dutch

1. Kan uw partner/uw familie u goed verstaan?
   1. slecht
   2. matig
   3. redelijk
   4. goed

2. Kunnen vreemden u verstaan?
   1. slecht
   2. matig
   3. redelijk
   4. goed

3. Bent u aan de telefoon te verstaan?
   1. slecht
   2. matig
   3. redelijk
   4. goed

4. Bent u te verstaan in gezelschap/op een feestje?
   1. slecht
   2. matig
   3. redelijk
   4. goed

5. Vindt u dat u duidelijk spreekt?
   1. helemaal niet
   2. een beetje
   3. redelijk
   4. heel duidelijk

6. Vindt u dat u vloeiend spreekt?
   1. helemaal niet
   2. een beetje
   3. redelijk
   4. heel vloeiend

7. Hoe is het tempo van uw manier van spreken?
   1. heel laag
   2. redelijk
   3. gewoon
   4. vlot

8. Hoe is het ritme van uw manier van spreken?
   1. slecht
   2. matig
   3. redelijk
   4. goed

9. Vindt u uw spraak aangenaam klinken?
   1. helemaal niet
   2. een beetje
   3. redelijk
   4. aangenaam

10. Vindt u spreken vermoeiend?
    1. heel erg
    2. redelijk
    3. een beetje
    4. helemaal niet

11. Hoe vindt u dat u te verstaan bent?
    1. slecht
    2. matig
    3. redelijk
    4. goed

English:

1. How well can your partner/relatives understand you?
    1. poorly
    2. moderately
    3. fairly
    4. well

2. Can strangers understand you?
    1. poorly
    2. moderately
    3. fairly
    4. well

3. How is your intelligibility on the phone?
    1. poor
    2. moderate
    3. fair
    4. good

4. Are you understood in company/ at parties?
    1. poorly
    2. moderately
    3. fairly
    4. well

5. Do you think you speak clearly?
    1. not at all
    2. a little
    3. fairly
    4. very clearly

6. Do you think you speak fluently?
    1. not at all
    2. a little
    3. fairly
    4. fluent

7. How fast do you speak?
    1. slowly
    2. moderately fast
    3. normal
    4. fast

8. What do you think of your speech rhythm?
    1. poorly
    2. moderately
    3. fairly
    4. good

9. Do you think your speech sounds pleasant?
   1. not at all
   2. a little
   3. fairly
   4. pleasant

10. Do you think speaking is tiring?
    1. very
    2. fairly
    3. a little
    4. not at all

11. How is your intelligibility?
    1. poor
    2. moderate
    3. fair
    4. good

For the relatives

Dutch:

1. Hoe vindt u dat uw partner/ familielid te verstaan is?
   1. slecht
   2. matig
   3. redelijk
   4. goed

2. Kunnen vreemden uw partner/ familielid verstaan?
   1. slecht
   2. matig
   3. redelijk
   4. goed

3. Kunt u uw partner/familielid aan de telefoon verstaan?
   1. slecht
   2. matig
   3. redelijk
   4. goed

4. Kunt u uw partner/familielid in gezelschap/op een feestje verstaan?
   1. slecht
   2. matig
   3. redelijk
   4. goed

5. Vindt u dat uw partner/familielid duidelijk spreekt?
   1. helemaal niet
   2. een beetje
   3. redelijk
   4. heel duidelijk

6. Vindt u dat uw partner/familielid vloeiend spreekt?
   1. helemaal niet
   2. een beetje
   3. redelijk
   4. heel vloeiend

7. Hoe is het tempo van de manier van spreken van uw partner/familielid?
    1. heel laag
    2. redelijk
    3. gewoon
    4. vlot

8. Hoe is het ritme van de manier van spreken van uw partner/familielid?
    1. slecht
    2. matig
    3. redelijk
    4. goed

9. Vindt u de spraak van uw partner/ familielid aangenaam klinken?
    1. helemaal niet
    2. een beetje
    3. redelijk
    4. aangenaam

10. Vindt uw partner/familielid zijn manier van spreken vermoeiend?
    1. heel erg
    2. redelijk
    3. een beetje
    4. helemaal niet

English:

1. How well can you understand your partner/relative?
    1. poorly
    2. moderately
    3. fairly
    4. well

2. Can strangers understand your partner/relative?
    1. poorly
    2. moderately
    3. fairly
    4. well

3. How intelligible is your partner/ relative on the phone?
    1. poorly
    2. moderately
    3. fairly
    4. well

4. Do you understand your partner/ relative in company/ at parties?
    1. poorly
    2. moderately
    3. fairly
    4. well

5. Do you think your partner/relative speaks clearly?
    1. not at all
    2. a little
    3. fairly
    4. clearly

6. Do you think your partner/relative speaks fluently?
    1. not at all
    2. a little
    3. fairly
    4. fluent

7. How fast does your partner/relative speak?
   1. slowly
   2. moderately fast
   3. normal
   4. fast

8. What do you think of the speech rhythm of your partner/relative?
   1. poorly
   2. moderately
   3. fairly
   4. good

9. Do you think your partner's/relative's speech sounds pleasant?
   1. not at all
   2. a little
   3. fairly
   4. pleasant

10. Does your partner/relative think speaking is tiring?
   1. very
   2. fairly
   3. a little
   4. not at all

Appendix 6.2 Voice Handicap Index

Only the Dutch translation is given here (De Bodt et al., 2000; Hakkesteegt et al., 2006). For the original, English version, see Jacobson et al., 1997.

Instruction:
These are statements that many people use to describe their voice quality and the effect their voice has on daily life. For each statement, indicate how often you experience this statement.

		0=nooit, 1=bijna nooit, 2=soms, 3=bijna altijd, 4=altijd	0	1	2	3	4
1.	F	Mensen kunnen me slecht horen vanwege mijn stem					
2.	P	Ik raak buiten adem tijden het spreken					
3.	F	In een lawaaierige omgeving hebben mensen moeite me te verstaan					
4.	P	Mijn stemgeluid varieert gedurende de dag					
5.	F	Als ik roep in huis, heeft mijn familie moeite mij te verstaan					
6.	F	Ik telefoneer minder dan ik zou willen					
7.	E	Wanneer ik met anderen spreek, raak ik gespannen vanwege mijn stem					
8.	F	Ik ben geneigd een gezelschap te vermijden vanwege mijn stem					
9.	E	Het lijkt of mensen geïrriteerd raken door mijn stem					
10.	P	Mensen vragen:”Wat is er aan de hand met je stem?”					
11.	F	Ik spreek minder met vrienden, kennissen en buren vanwege mijn stem					
12.	F	Wanneer ik iemand spreek, vragen ze me vaak te herhalen wat ik gezegd heb					
13.	P	Mijn stem klinkt krakerig en droog					
14.	P	Ik heb het gevoel dat ik moet persen / forceren om stem te geven					
15.	E	Ik heb het gevoel dat andere mensen mijn stemprobleem niet begrijpen					
16.	F	Mijn stemproblemen beperken mijn persoonlijke en sociale leven					
17.	P	Het is nooit te voorspellen hoe goed/helder mijn stem zal zijn					
18.	P	Ik probeer van alles om mijn stem anders te laten klinken					
19.	F	Ik voel me tijdens een gesprek vaak buitengesloten vanwege mijn stem					
20.	P	Spreken kost heel veel inspanning					
21.	P	Ik heb ‘s avonds meer last van mijn stem					

22.	F	Ik verlies (een deel van) mijn inkomen door mijn stem
23.	E	Ik raak van streek door mijn stem
24.	E	Ik ben minder spontaan vanwege mijn stemproblemen
25.	E	Ik voel me gehandicapt door mijn stem
26.	P	Mijn stem begeeft het tijdens het spreken
27.	E	Ik erger me wanneer mensen me vragen iets te herhalen
28.	E	Ik voel me opgelaten als mensen me vragen iets te herhalen
29.	E	Door mijn stem voel ik me minderwaardig
30.	E	Ik schaam me voor m'n stem

F	**Functional**
P	**Physical**
E	**Emotional**
T	**Total**

Appendix 6.3 Reading aloud text

(adapted text from De Volkskrant, 4-10-1997)

In 1993 nam de Turkse politie in Istanbul veertien dansende beren in beslag. De beren waren door zigeuners gehouden, die de dieren leerden dansen op gloeiende platen of kolen. De arme beren hadden een ring door neusgaten of bovenlip, waaraan een ketting bevestigd was. Ze hadden geen enkele bewegingsvrijheid; 's nachts stonden ze aan een boom geketend.

Na hun bevrijding waren de dieren van de mishandeling verlost, maar de redders stonden voor de lastige vraag hoe het nu verder moest. Konden zij de beren een prettige toekomst bieden of waren de beesten zo beschadigd dat ze geen normaal leven meer konden leiden? In dat geval zou het beter zijn ze rustig te laten inslapen.

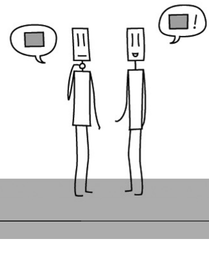
De beren kregen een kans: de meeste van hen werden opgevangen in een park in Turkije. De drie die er het slechtst aan toe waren –ze waren blind– kwamen in 1994 naar Nederland. Ze kregen onderdak in het Berenbos bij Ouwehands Dierenpark in Rhenen, dat speciaal voor de opvang van misbruikte bruine beren was aangelegd.

Naast de drie Turkse beren wonen er in Rhenen ook twee 'oorlogsslachtoffers' uit het voormalige Joegoslavië, twee beren uit het Russische Staatscircus en zes beren uit dierentuinen, waar ze zaten opgesloten in te kleine ruimtes.

Het was de vraag of de opvang goed zou uitpakken; vooral de komst van de blinde beren was spannend, want hun akelige verleden beloofde grote problemen. Het is gelukkig reuze meegevallen. "De beren gedragen zich net zo als in het wild", zegt de bioloog van het Berenbos. "Dat is een teken dat ze zich prettig voelen. Een nog sterkere aanwijzing is dat ze vaak spelen. Tot onze verrassing spelen de blinde mannetjes met de andere mee. Enige hinder van hun handicap hebben ze wel, want ze wagen zich maar in een deel van het bos en ze worden door de anderen niet helemaal voor vol aangezien. Dat blijkt uit de ruzietjes over de verdeling van het voedsel. Maar toch kunnen we zeggen dat de blinde beren, net als de andere, hun draai gevonden hebben in het Berenbos. Ze zijn weer beer en dansen doen ze nooit meer!"

7

Overall evaluation of the therapy program



Abstract

In chapter 4, the development of the training program meant to improve TE speech intelligibility was discussed. Chapters 5 and 6 described the therapy effect measurements. In this chapter, the entire therapy program is evaluated (except the VHI, which appeared not to be useful for intelligibility research), whereby the personal experiences of the speakers and the participating speech language pathologists (SLPs) are also taken into account. Main findings are that the four pre tests do not correlate highly, which indicates that there is not one perfect test to measure improvement in intelligibility. The speakers themselves appear to be quite satisfied with their progress. The SLPs have some criticisms, but on the whole are quite satisfied with the therapy program. Based on the outcomes of all tests and the evaluation of the therapy program, recommendations for the future are given.

7.1 Introduction

In chapter 4, the development of the rehabilitation program for this study was discussed. In order to test the effectiveness of this program in a group of nine TE speakers, five different pre and post tests were used, results of which were discussed in chapter 5 and chapter 6. Since, as concluded in chapter 6, the VHI is not an appropriate questionnaire for intelligibility research, it will not be further taken into account in this chapter. This leaves four pre and post tests for comparison in this chapter: three listening experiments set up to evaluate the speech tasks of participants and a set of study-specific structured questionnaires. In chapter 5, two main questions were formulated:

1. Can TE speech intelligibility be improved by specific speech training?
2. Is the set-up of the therapy program appropriate?

The first question has partly been answered in chapters 5 and 6: it could be seen that TE speakers improved their intelligibility of consonants and SUS sentences, but that no improvement was seen in read aloud speech and that the intelligibility during the retelling of the story was rated even lower in the post test than in the pre test.

It was also stated in chapter 5 that the second research question could complicate the interpretation of the results: if no improvement is found, is this then caused by the set-up of the therapy program, by the techniques used, or by the simple fact that TE speakers cannot be trained. In order to be able to answer this second research question, in the present chapter the whole therapy program will be evaluated. This means that the set-up of the program, the use of pre and post tests and the results will be evaluated with the knowledge that has been gained throughout the process. First, the results of the phoneme and sentence intelligibility tests, the questionnaires, and the semantic scaling experiment will be compared in order to investigate whether some of the tests measure the same aspects or whether all tests measure different aspects of speech intelligibility. Thereafter the personal experiences of the participants and the speech therapists will be discussed. The chapter will end with recommendations for the future.

7.2 Subjects and methods

7.2.1 Subjects

Speakers were 9 male laryngectomized individuals (mean age 64;9, range 54;11-82;9). All speakers had also received radiotherapy. Two speakers had had additional reconstructions: one PM flap and one gastric pull up (see also chapter 5). All subjects completed the study.

7.2.2 Statistics

In this study, several pre and post tests were chosen that together investigated various levels of speech. Different types of raters were chosen for each test as each type of rater usually judges speech differently. However, evaluating effects of speech training by using four instruments is a time-consuming process and therefore it should be investigated whether all the tests indeed measure different aspects of speech, or if some of the tests measure the same aspects. Correlations between tests can show whether scores on one test can be predicted by scores on another test. However, the interpretation of the results is difficult as there are only nine subjects in this study. In addition, no base-line study was performed due to time constraints, meaning that the tests used were not fully validated. Theoretically, this may mean that speakers are inconsistent when performing a test twice. The therapy itself will also have influenced the behavior of the speakers during the post test. Therefore, the calculation of correlations has been limited to the pre tests. The non-parametric Spearman's rank correlations were used that show whether the rank ordering of subjects is the same in the separate tests. However, if this is the case, based on only nine subjects it cannot be concluded that some of the tests can be discarded. Results can only be used as indicators of certain trends. In the following section, the correlations between the different pre tests are discussed.

7.3 Correlations between the pre tests

The correlations are subdivided into three sections due to the large amount of correlations. In the first section, the phoneme and sentence identification task performed by the phonetically trained listeners is correlated with the phoneme identification task performed by the Speech Language Pathologists (SLPs). In the next section, the questionnaire questions are correlated with the tests from the first section. Finally, in the last section, the phoneme and sentence identification task, the phoneme identification task and the questionnaires are correlates with the 18 separate scales of the intelligibility rating task as performed by naïve listeners.

7.3.1 Correlations between the tests performed by phonetically trained listeners and SLPs

In chapter 5 two tests were discussed: a phoneme (initial, medial and final consonants, as well as vowels) and SUS sentence identification experiment performed by phonetically trained listeners, with no prior knowledge of TE speech, and a phoneme identification experiment (Dyva test) performed by SLPs, also experienced listeners, but acquainted well with TE speech. Even though the two groups of listeners differ, the speech tasks for the speakers were very similar and therefore correlations between the results of the two experiments might be expected.

In this section, the Spearman's rank correlations between the pre tests described in chapter 5 will be discussed. In the following table, the relevant results are shown. The correlations are based on the overall score per speaker.

Table 7.1: Correlations between the overall scores per speaker of the different phoneme and sentence identification tests in the pre test (only relevant part of the table is given); empty cells: not significant. * indicates significance at .05 level, ** indicates significance at .01 level.

	medial consonants	SUS sentences
dyva	.667*	
initial consonants	.867**	.778**
medial consonants	NA	.762*

We found that the medial consonants correlate with all the other tests except the vowels, and that the initial consonants also correlate with the SUS sentences. The fact that medial consonants correlate highly with all the other tests can be explained by the fact that these same consonants were also present in the other tests except, of course, where the vowels were tested.

The fact that vowels show low correlation with the other tests can be explained by the fact that vowels scored high in the pre test (see Table 5.5), which means there was little variation in scores. Secondly, it is hard to compare vowels with consonants as they are produced in a very different way.

Even though correlations are found between some of the tests, this does not mean that some tests can be discarded, as was discussed in the section on statistics. The correlations mainly indicate that the rank order of speakers is the same for these tests.

7.3.2 Correlations between the questionnaires and the phoneme and sentence identification tasks

In this section, the correlations between the two study-specific structured questionnaires (one for the speaker and one for the close relative, see appendix 6.1), and the phoneme and sentence identification tasks as performed by phonetically trained listeners and SLPs (chapter 5) are investigated.

First, the 10 answers by the speakers themselves on the questionnaire (4-point scales) were compared with the 10 answers by their close relative. No sufficiently high correlation could be found. In chapter 6 it was described that the close relative scored quite high, both in the pre and post test (see Table 6.3). This means that not much variation was present, which makes it hard to find high correlations. It is conceivable that the family members did not want to appear too negative about their relative's speech. The results might have been different if the family member would have filled in this questionnaire without the laryngectomized individual present. Otherwise, it might be concluded that the structured questionnaire used in this study does not provide opinions that reflect reality.

For the correlations between the questionnaires, the Dyva and the phoneme and sentence identification task, overall scores per speaker were used. For the questionnaires this means that the average over ten questions was used, rather than the scores on the separate questions. It was found that only the Dyva test, with SLPs as raters, correlated highly with the questionnaire filled out by the speakers themselves (.683, $p<.05$). The ratings from the SLPs seem to reflect best the speaker's opinion. The fact that the phoneme and SUS sentence identification experiments did not correlate with the study-specific questionnaires may indicate that the actual phoneme intelligibility of the speakers does not play a significant role when speakers and family are asked for their opinion on the influence of the overall intelligibility of the speaker on daily life. It is conceivable that questionnaires test different aspects of speech than phoneme and sentence identification tests.

7.3.3 Correlations between the phoneme and sentence identification tests, the semantic scales and the questionnaires

In this section, the phoneme (Dyva, initial and medial consonant, vowel) and SUS sentence identification tests, and the study-specific questionnaires for both speakers and family will be compared with all 18 semantic scale judgments (see Table 6.7). Again, overall scores per speaker are used, also for the questionnaires. Correlations between the scales themselves will not be investigated as the Principal Component analysis in chapter 6 already provided information on which scales belong together.

Results are given separately for the scale judgments on the retelling task (Table 7.2a) and the scale judgments on the reading aloud task (Table 7.2b). Due to limited space, not the whole correlation matrices are given, but only the relevant parts.

Table 7.2a: Significant correlations (p-values) between the phoneme identification tests and questionnaire, and specific semantic scale judgments on the retelling task in the pre test. Empty cells: not significant. The numbers in the row on top indicate the numbers of the semantic scales. * indicates significance at .05 level, ** indicates significance at .01 level.

Retelling task (scales)						
Tests	**1**	**3**	**4**	**5**	**9**	**10**
initial consonants		.895**	.731*			
medial consonants	.689*	.870**		.778*	.706*	.711*
questionnaire speakers			.831**			

Table 7.2b: Significant correlations (p-values) between phoneme identification tests and specific semantic scale judgments on the reading aloud task in the pre test. Only the significant correlations are shown. Empty cells: not significant. The numbers in the row on top indicate the numbers of the semantic scales. * indicates significance at .05 level, ** indicates significance at .01 level.

Reading aloud (scales)				
Tests	**3**	**4**	**5**	**9**
dyva	.828**	.703*	.817**	
Initial consonants				.683*

Very few significant correlations were found. The scales that do show correlations are either the intelligibility scales (1-5) or the rate/rhythm scales (7-11). The voice quality component (scales 13-17) does not correlate at all with the phoneme and sentence identification tests. This result suggests that intelligibility is not necessarily influenced by voice quality.

The fact that the rate/rhythm scales (7-11) do correlate with the other tests is somewhat surprising, as variation in rate and rhythm is hardly possible when reading out word lists. However, it was also found in chapter 6 that rate and rhythm scales often belonged to the same component as the articulation scales. Apparently listeners feel that these scales belong together.

What is also surprising is that the SUS sentence intelligibility test does not correlate with any of the scales. In chapter 2, however, this intelligibility correlated highly with all scales and it was suggested that this was caused by the fact that sentences are closest to connected (spontaneous) speech. It was concluded that SUS sentences seemed to be a good measure for overall intelligibility. An explanation for the present results might be due to the fact that the group of raters differed and also of course the speaker group. It also shows that care is needed when interpreting results of only nine participants. The conclusion from chapter 2 seems to have been too optimistic: it is not possible or advisable (yet) to use only one test for intelligibility, especially since the speakers form such a heterogeneous group. Furthermore, different listeners will use different cues when judging TE speech, all providing valuable information.

7.4 Personal experiences with the therapy program

The day subjects came in for their post test, they were asked their personal opinion about the therapy program. This served a two way purpose: it broke the ice, which made them less nervous at the beginning of the post test and it provided useful information on the therapy program itself.

From the interviews, it appeared that broadly two groups of TE speakers could be distinguished: one group that participated to help the researcher and to help other people rather than helping themselves. These subjects were often good speakers already. A second group hoped their own speech would improve by participating in this study.

Everyone reacted positively to the training. All of them said they had been made aware of their speech, of how articulation works and what their own speech sounds like. Also psychologically, this had been beneficial for some of the subjects. At the start of the therapy, some subjects were very apprehensive about listening to their own voice. At the end of the therapy, this apprehension had disappeared.

A few speakers claimed their speech had become more intelligible. One subject said his friends and family had made positive remarks on his improved intelligibility. About half of the group said that even though this was a positive experience for them, they probably would not practice anymore, whereas the other half was enthusiastic and planned on practicing on their own and improving their speech further.

Even though these personal communications were valuable, a study-specific structured questionnaire was also handed out in the last session for evaluating the therapy program (see appendix 7.1).

The questionnaire consisted of 16 questions concerning the information they received beforehand, their expectations of the therapy, the form and contents of the training and their practicing behavior. All questions were in multiple-choice format; some questions had three possible answers (9) and some four (7).

In the following table the answers are given per question and per speaker.

Table 7.4: Scores per question (summary of questions given) by the individual subjects and averages per question, empty cells are missing values. In the last column also the maximum score per question, either 3 or 4, is indicated.

Subjects Questions	1	2	3	4	5	6	7	8	9	Average (/maximum) score
1. Satisfaction with information beforehand	3	3	4	4	4	4	4	4	4	3.8 (/4)
2. Satisfaction info/explanation during sessions	4	4	4	4	4	4	4	4	4	4 (/4)
3. Clear expectations	4	2	4	4	3	4	4	4	3	3.6 (/4)
4. Clearness exercises	4	4	4	4	3	4	4	3	3	3.7 (/4)
5. Difficulty exercises	2	3	3	1	2	1	1	1	1	1.8 (/4)
6. Variation exercises	2	2	2	2	2	2	2	2	2	2 (/3)
7. Sufficiency practice material sessions	2	2	2	2	2	2	2	2	2	2 (/3)
8. Sufficiency number of sessions	2	2	2	2	2	2	2	2	2	2 (/3)
9. Sufficiency duration of sessions	2	2	2	2	2	2	2	2	2	2 (/3)
10. Satisfaction amount of homework	3	2	2	2	3	2	2	2	2	2.2 (/3)
11. Satisfaction frequency homework	2	2	2	2	2	2	4	2	2	2.2 (/3)
12. Satisfaction type of homework	3	4	4	4	3	3	1	4	3	3.2 (/4)
13. Intention to practice in future	3	1	3	3	3	3		3	3	2.8 (/3)
14. If so, how much	1		2	1	2	2	2	1	1	1.5 (/3)
15. Benefit of repeating sessions	3	3	2	1	2	2	3	3	1	2.2 (/3)
16. Willingness 2nd post test after 6 months	3	3	3	3	3	3	3	3	3	3 (/3)

The first point to notice is that on most questions answers are very similar. On questions 2, 6, 7, 8, 9 and 16 the answers are even all the same. Below, results will be discussed per group of questions.

The first three questions dealt with the information and explanation received beforehand and during the therapy program. Subjects were reasonably (score 3) to very satisfied (score 4) with the information they received beforehand and all subjects were very satisfied with the information and the explanations during the therapy sessions. For all but one subject it was reasonably to very clear as to what to expect from the therapy. Only for one subject was it a little clear (score 2).

Questions 4 to 7 dealt with the exercises during the therapy program. Subjects thought the exercises were reasonably to very clear and believed they were not at all, or only a little hard. Two subjects thought they were reasonably hard. Everyone believed there was enough variation in the kind of exercises during a session and that the amount of practice material was sufficient.

Where the number and duration of therapy sessions was concerned (questions 8 and 9), all subjects thought both the number and length of the sessions was sufficient (score 2).

Questions 10 to 12 concerned the homework. Seven subjects thought the amount of homework was sufficient, while two subjects thought it was too much. Eight subjects thought that the frequency of practicing (2 times a short period a day) was just right. Furthermore, subjects were reasonably to very satisfied with the kind of homework they received, except one subject who was not satisfied at all.

The last few questions (13-17) concerned the intention of the subject to continue practicing in the future. Seven subjects intended to practice after the therapy program had ended, one subject had no intention and one did not answer this question. The subjects that do have the intention to practice intend to do so on a daily or weekly basis. Not everyone believed to gain from a repeated session; some were not sure if that would be beneficial and the rest believed it would be good to repeat the instructions. All subjects were willing to return after six months for another speech evaluation.

Overall, it can be concluded that the subjects were quite satisfied with the whole set-up of the therapy program. As almost all of the answers are positive, it can be questioned how honest the subjects have been in answering these questions. Although the speech therapist was not present, the researcher was in the room at the moment of filling out the questionnaire, and thus the answers were not anonymous and this may have caused the homogeneous and mostly positive answers. However, during the therapy and the post test there has been ample opportunity for comments. Also the fact that no one dropped out half way and that subjects remained quite motivated, leads to the belief that the therapy program in fact was rated positively, albeit maybe less so than the answers to the questionnaire would suggest.

7.5 The Speech Language Pathologists' evaluation of the program

The two speech language pathologists (SLPs), who actually gave the therapy to the subjects, were also given a study-specific structured questionnaire (see appendix 7.2) at the end of the therapy program with 14 questions on its content and structure. The aim of the questionnaire was to find out how experienced SLPs deal with the program and how useful and easy to use SLPs think this program is. Their comments can be used in the possible adjustment of the program in the future. It should be emphasized that these SLPs deliberately had not been involved in the development of the program, which allows them to be neutral towards the whole program and probably makes their responses more reliable than if they would have been involved. In the following table the scores per question are shown for each SLP.

Table 7.4: Scores per question by the two individual SLPs. In the bottom row also the maximum score is given per question.

Questions														
SLP's	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
1	3	4	4	2	3	2	3	2	2	2	4	2	2	3
2	3	3	4	3	3	3	2	1	2	1	3	2	2	2
Max score	4	4	4	4	4	4	4	3	3	3	4	3	3	4

Questions will again be discussed per group (theme).

The first question concerned the expectations of the SLPs. Both SLPs had a reasonable idea of what to expect from the program, but one of them commented it would have been good to have a practice session first. In the future, users should be recommended to do a trial session first to see exactly what a session entails.

Questions 2 and 3 dealt with the set-up of the therapy program as a whole and the set-up within the sessions. Where the set-up of the program was concerned, one SLP said it was 'good', one said it was 'reasonable', but missed transfer to spontaneous speech. Both SLPs agree that the set-up within the session was '(very) good'.

In question 4 and 5 the content and the level of the therapy program was addressed. One SLP believed that the contents were 'a little' related to the problems TE speakers tend to have, while the other SLP believed the contents were 'reasonably' related to the problems of TE speakers. On the question as to whether the level of the therapy matches that of the participants both answered 'reasonable'.

Questions 6 and 7 were concerned with the techniques that were used within the program. One SLP said the techniques were only 'a little' useful, whereas the other SLP said their usefulness was 'reasonable'. On the question whether techniques were easy to explain to the participants, again one answered 'a little', whereas the other answered 'reasonable'.

Questions 8 to 11 were concerned with the exercises of the program. On the question if there was enough variation of exercises during the sessions, one SLP said there was 'enough' variation, but the other SLP felt she also had to use some of her own exercises to complement the ones in the program. She used her own exercises, because they were more familiar to her and easier to explain. It also provided for more variation in the exercises. The topic of the exercises used by the SLP, however, was the same as the topic of the exercises in the training program. Both SLPs believed that the amount of exercises was ‘sufficient’. Where the practice material to use at home was concerned, one answered that there were ‘enough’ exercises, but the other SLP answered there were ‘too few’ exercises. She commented that she mostly meant there was not enough variation in the exercises.

Question 11 asked whether the speech material and the exercises were a useful contribution to the session. One answered ‘useful’, while the other answered ‘reasonable’.

On the questions whether the amount of sessions and the duration of sessions was (long) enough, both SLPs thought the length of the sessions (1 hour) was ‘sufficient’ as well as the two sessions a week.

The last question asked whether the participants had learned something, at least in the short run. One SLP answered that they had learned a ‘reasonable’ amount, whereas the other said they had learned only ‘a little’.

On the whole, both SLPs were reasonably satisfied with the program, except where certain details were concerned. From personal communication it appeared that not all material was considered useful or easy to use and that own material was used to fill gaps. It was also said that some speakers did not have problems at all with some of the topics and that it was demotivating for them and the SLP to go through these lessons.

Based on these opinions, the exercises will need to be adjusted somewhat. It is also apparent that a more thorough preparation with the SLP is helpful to make it even clearer what is expected of them, in order to streamline this new concept with their own traditional way of working. However, even though the program is structured and based on the (scarcely) available evidence of their usefulness, there should be room to deviate from that structure. The idea is that people practice most that what they have problems with. If, for example, the first two lessons are too easy, one should be allowed to spend these lessons on topics scheduled for later in the program that do cause problems.

The fact that one SLP missed transfer to spontaneous speech is backed up by the data in chapter 6. In that chapter it was found that retelling in the post test scored worse than in the pre test and that reading aloud showed no improvement. The possible explanation offered was that more time is needed to use the techniques learned automatically. Speakers are asked to forget all the techniques they learned or taught themselves and start all over again with the techniques trained in the therapy program. The acquisition of clear speech seems to require more time than the three lessons offered. It seems logical to spend more time on spontaneous speech in future programs. However, the nine sessions were chosen deliberately to ensure acceptation by Dutch health insurance companies, who tend to approve speech therapy in blocks of nine sessions. However, when founded with the proper arguments, additional blocks of 9 sessions mostly will be granted. Therefore, spending more time on spontaneous speech should not mean spending less time on training phonemes, which clearly have been shown to 'profit' from this dedicated training, and thus more training sessions are called for.

7.6 Discussion and conclusion

In chapter 4, the theoretical background of the therapy program was discussed. In chapters 5 and 6 the effect measurements were described and results were discussed. In the current chapter, the whole therapy program has been evaluated based on the results found and the personal experiences of the participants, the speech language pathologists (SLPs) and the researcher. Two main questions were formulated in chapter 5:

1. Can TE speech intelligibility be improved by specific speech training?
2. Is the set-up of the therapy program appropriate?

The first question can be answered affirmatively, at least where phonemes and sentences are concerned. An improvement was also found in the way the speaker thought about his own intelligibility. However, it has to be kept in mind that not all tests showed progress, i.e. no improvement could be found for the reading aloud and retelling task.

This leads to the second question about the set-up of the program, which is more complicated to answer. With 'set-up', both the structure of the program and whether the pre and post tests were appropriate is meant. Where the structure of the program is concerned, adjustments will have to be made, mostly based on the opinion of the SLPs and the experience of the author. To improve spontaneous speech as well, more time spent on training spontaneous speech intelligibility is required. The best option would probably be to increase the number of sessions. For this, the health insurance authorities have to be convinced and the results obtained with the present program provide good arguments for this increase in the number of sessions. Where the phonemes are concerned, it seems that the right techniques were chosen and that the time spent on training them was sufficient.

The various pre and post tests and the various groups of listeners were deliberately chosen in order to be able to investigate intelligibility at as many speech levels as possible. Trained raters such as phoneticians and SLPs are the best choice when an expert judgment on the articulation and sound of a phoneme is wanted. It is also important to know the judgment of the speakers themselves and of their relatives about their intelligibility. Untrained, naïve listeners show best how a TE speaker is perceived in daily life, but may be less useful for the evaluation of therapy effects. Additionally, it has to be kept in mind that pleasing as it may be to have a satisfied TE speaker, who indicates to have gained better insight into and control over the intelligibility of his speech, this is not enough to 'prove' the validity of a therapeutic approach. The perceived improved satisfaction could still be a 'placebo effect' and the result of the extra attention the speaker has received throughout the therapy program. The aforementioned 'objective results' of the different tests is critical for being able to possibly call the therapy program evidence-based.

There are some limitations in the tests that were used, though. The VHI was considered to be unsuitable for the purposes of this study. This as such is not surprising, as this questionnaire was devised to test quality of life related to voice quality and not to speech intelligibility. It was seen in this study that voice quality and speech intelligibility are not necessarily related. Unfortunately, as far as the author is aware, there is no validated test that investigates quality of life in relation to speech intelligibility. It was for that reason that study specific questionnaires were devised, based on standard questions used in earlier quality of life research in the hospital and complemented with questions devised specifically for this study. Although these questionnaires are relatively short, and not validated, they are well structured and based on the vast experience of the researchers involved in this study (Ackerstaff et al., 1994), and therefore a good 'second-best' to validated questionnaires. They may render useful information for the researcher, are statistically analyzable and may form the bases for future validated questionnaires on speech intelligibility.

It has already been discussed in chapter 6 that using semantic scales for naïve listeners may not be the optimal choice, as it is a difficult task. It is not always clear to the listener what the terms on either side of the scale mean. However, it is difficult to obtain judgments on spontaneous speech without using semantic scales. Also the inter-rater reliability was sufficiently high, which seems to justify the use of scales. What was seen, however, is that naïve listeners are not always capable of distinguishing between categories of scales. For example, the intelligibility scales were sometimes considered to belong to the same component as the rhythm and rate scales. They did, however, distinguish between voice quality and speech intelligibility. Overall, however, it is more likely that the speakers had problems implementing the techniques they had learned during the therapy and that the naïve listeners actually performed rather well.

In the current chapter, correlations between the different tests were investigated. Not many significant correlations could be found, though. This can be explained by the fact that all tests measure quite different aspects and that all groups of listeners were different. Also the fact that some tests showed improvement while others did not makes it unlikely for tests to correlate highly. If anything, these results show that there is not one single test that can test (improvement of) intelligibility. More, rather than fewer tests should be included in the future. In addition, it was stated in section 7.2.2 on statistics that results have to be interpreted very carefully since only nine subjects were involved.

Future studies

This therapy evaluation study was the first step towards an evidence-based training program for TE speakers. Although already quite useful as it now is, more work is needed to optimize this program. Several improvements can be suggested for the future:

1. In this program, phoneme exercises focused mainly on teaching compensatory strategies rather than training better control of the neoglottis. Some speakers have shown that they are able to control the neoglottis, though not consistently. If it is possible to improve control, for example by using visual feedback with digital high-speed recordings, this would most likely have a positive effect on intelligibility.
2. Where the voiced-voiceless distinction was concerned, main focus was on training voiceless sounds. In the future, time should also be devoted to training voiced sounds.
3. Even though auditory feedback was used, it was not incorporated in a structural way. Also visual and aerodynamic feedback should be considered.
4. There should be more time for practicing clear speech and self evaluation should receive more attention.
5. Only a small group of subjects was used, and this group was reasonably homogeneous. To obtain an even better insight in the effect of speech training, a larger group (>20) is necessary. Within this group, speakers with extended surgery and reconstructions should be included as these people in particular show many problems with voice quality and speech intelligibility.
6. It should be investigated whether time after the total laryngectomy influences results of speech therapy. In this study, no effect for time after laryngectomy was found, but as it was only a small group, this should be studied in more detail.
7. Acoustic analyses should be performed on the clear speech before and after therapy to see if the improvements aimed at with this technique can be found.

When the program has been adjusted, a second study (if possible, a multi-center study) should be initiated to investigate the effects of the improved program. The main aim of course is that this program will be validated further and made available for general use in hospitals for all TE speakers.

Appendix 7.1 Evaluation of the therapy program by TE speakers

Dutch version

Kruis aan wat het beste van toepassing is:

1. Bent u tevreden over de voorlichting en informatie die u kreeg voor aanvang van de therapie?
    1. helemaal niet
    2. een beetje
    3. redelijk
    4. heel tevreden

2. Bent u tevreden over de informatie en uitleg tijdens de therapie?
    1. helemaal niet
    2. een beetje
    3. redelijk
    4. heel tevreden

3. Was het duidelijk wat u van de therapie kon verwachten?
    1. helemaal niet
    2. een beetje
    3. redelijk
    4. heel duidelijk

4. Vond u de oefeningen tijdens de sessies duidelijk?
    1. helemaal niet
    2. een beetje
    3. redelijk
    4. heel duidelijk

5. Vond u de oefeningen tijdens de sessies moeilijk?
    1. helemaal niet
    2. een beetje
    3. redelijk
    4. heel erg

6. Vond u dat er voldoende verschillende oefeningen tijdens de sessies waren?
    1. te weinig
    2. voldoende
    3. te veel

7. Vond u dat er voldoende oefen-materiaal tijdens de sessies was?
    1. te weinig
    2. voldoende
    3. te veel

8. Vond u dat er voldoende therapie-sessies waren?
    1. te weinig
    2. voldoende
    3. te veel

9. Vond u de duur van de therapie-sessies voldoende?
    1. te kort
    2. voldoende
    3. te lang

10. Wat vond u van de hoeveelheid huiswerk?
    1. te weinig
    2. voldoende
    3. te veel

11. Wat vindt u van de frequentie van het huiswerk (2 keer een korte periode per dag)?
    1. te laag
    2. precies goed
    3. te hoog

12. Bent u tevreden met het soort oefeningen dat u als huiswerk meekreeg?
    1. helemaal niet
    2. een beetje
    3. redelijk
    4. heel tevreden

13. Bent u van plan te blijven oefenen?
    1. nee
    2. weet ik nog niet
    3. ja

14. Zo ja, hoe vaak gaat u dan oefenen?
    1. dagelijks
    2. wekelijks
    3. maandelijks

15. Denkt u baat te hebben bij een (aantal) herhalingssessie(s)?
    1. nee
    2. weet ik nog niet
    3. ja

16. Bent u bereid na 6 maanden eenmalig terug te komen om te kijken hoe het op dat moment met uw spraak gaat?
    1. nee
    2. weet ik nog niet
    3. ja

English version

Circle the answer most applicable to you:

1. Are you satisfied with the advice and information you received before the start of the therapy?
    1. not at all
    2. a little
    3. fairly
    4. satisfied

2. Are you satisfied with the information and explanation you received during the therapy?
    1. not at all
    2. a little
    3. fairly
    4. satisfied

3. Was it clear what to expect from the therapy?
    1. not at all
    2. a little
    3. fairly
    4. clear

4. Did you think the exercises during the sessions were clear?
    1. not at all
    2. a little
    3. fairly
    4. clear

5. Did you think the exercises during the sessions were hard?
    1. not at all
    2. a little
    3. fairly
    4. hard

6. Did you think there was enough variation in the exercises during the sessions?
    1. too little
    2. enough
    3. too much

7. Did you think there was enough practice material during the sessions?
    1. too little
    2. enough
    3. too much

8. Did you think there were enough therapy sessions?
    1. too few
    2. enough
    3. too many

9. Did you think the therapy sessions were long enough?
    1. too short
    2. sufficiently long
    3. too long

10. What did you think of the amount of homework?
    1. too little
    2. enough
    3. too much

11. What did you think of the frequency of the homework (2 times a short period a day)?
    1. too low
    2. just right
    3. too high

12. Were you satisfied with the kind of exercises you received for home-work?
    1. not at all
    2. a little
    3. fairly
    4. satisfied

13. Are you planning to continue practicing?
    1. no
    2. not sure yet
    3. yes

14, If yes, how often are you planning to practice?
   1. daily
   2. weekly
   3. monthly

15. Do you think a (couple of) repetition session(s) is/are useful?
   1. no
   2. not sure yet
   3. yes

16. Are you prepared to come back once after six months to re-evaluate your speech?
   1. no
   2. not sure yet
   3. yes

Appendix 7.2 Evaluation of the therapy program by the participating Speech Language Pathologists

Dutch version

Kruis aan wat het beste van toepassing is:

1. Was van te voren duidelijk hoe het programma gebruikt moest worden?
    1. helemaal niet
    2. een beetje
    3. redelijk
    4. heel duidelijk

2. Was de opbouw van het programma in zijn geheel goed?
    1. helemaal niet
    2. een beetje
    3. redelijk
    4. goed

3. Was de opbouw binnen een sessie goed?
    1. helemaal niet
    2. een beetje
    3. redelijk
    4. heel goed

4. Sloot de inhoud van het programma aan bij de problemen die de sprekers hebben?
    1. helemaal niet
    2. een beetje
    3. redelijk
    4. helemaal

5. Sloot het niveau van het programma aan bij dat van de deelnemers?
    1. helemaal niet
    2. een beetje
    3. redelijk
    4. heel goed

6. Waren de technieken zoals ze in het programma besproken werden bruikbaar?
    1. helemaal niet
    2. een beetje
    3. redelijk
    4. heel bruikbaar

7. Waren de technieken over het algemeen makkelijk uit te leggen aan de patiënt?
    1. helemaal niet
    2. een beetje
    3. redelijk
    4. makkelijk

8. Waren er voldoende verschillende oefeningen tijdens de sessies?
    1. te weinig
    2. voldoende
    3. te veel

9. Was er voldoende oefenmateriaal tijdens de sessies?
   1. te weinig
   2. voldoende
   3. te veel

10. Waren er voldoende oefeningen om thuis te doen?
    1. te weinig
    2. voldoende
    3. te veel

11. Waren het stimulusmateriaal en de oefeningen een nuttige aanvulling op wat in de sessie besproken was?
    1. helemaal niet
    2. een beetje
    3. redelijk
    4. nuttig

12. Was de duur van een sessie voldoende?
    1. te kort
    2. voldoende
    3. te lang

13. Waren de twee sessies in de week voldoende?
    1. te weinig
    2. voldoende
    3. te veel

14. Hebben de patiënten er (op korte termijn) iets van opgestoken?
    1. helemaal niet
    2. een beetje
    3. redelijk wat
    4. voldoende

English version

Circle the answer most applicable to you:

1. Was it clear beforehand how the program should be used?
   1. not at all
   2. a little
   3. fairly
   4. clear

2. Was the set-up of the therapy program as a whole good?
   1. not at all
   2. a little
   3. fairly
   4. good

3. Was the set-up within the session good?
   1. not at all
   2. a little
   3. fairly
   4. good

4. Did the contents of the therapy program match the problems TE speakers generally have?
   1. not at all
   2. a little
   3. fairly
   4. well

5. Did the level of the therapy match the level of the participants?
   1. not at all
   2. a little
   3. fairly
   4. well

6. Were the techniques as discussed in the program useful?
   1. not at all
   2. a little
   3. fairly
   4. useful

7. Were the techniques generally easy to explain to the participants?
   1. not at all
   2. a little
   3. fairly
   4. easy

8. Was there enough variation in practice material during the sessions?
   1. too little
   2. enough
   3. too much

9. Were there enough exercises during the session?
   1. too few
   2. enough
   3. too many

10. Were there enough exercises to do at home?
    1. too few
    2. enough
    3. too many

11. Were the stimulus material and the exercises a useful contribution to the topic discussed in that session?
    1. not at all
    2. a little
    3. fairly
    4. useful

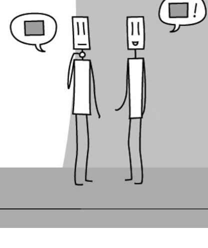
12. Was the duration of the sessions long enough?
    1. too short
    2. long enough
    3. too long

13. Were two sessions a week enough?
    1. too few
    2. enough
    3. too many

14. Do you believe the participants have learned something, at least in the short term?
    1. not at all
    2. a little
    3. fair amount
    4. much

8

General discussion and conclusion



Abstract

In this final chapter, the results of the assessment and therapy program will be summarized and discussed in relation to the aims formulated in chapter 1. Those two main aims were: 1) to assess the intelligibility of tracheoesophageal (TE) speech, and 2) to see if specific speech training can help improve the intelligibility of TE speech. The assessment consisted of perceptual and acoustic analyses. For the speech training, a therapy program was developed and effects were measured with pre and post tests. Limitations will be described, followed by directions for future research. The chapter ends with a general conclusion.

8.1 Introduction

With a total laryngectomy, one of the most important components for speech production is removed: the larynx with the vocal folds. This removal of the voice source has a profound effect on the voice quality of the laryngectomized individual. It is therefore not surprising that most studies investigating alaryngeal speech have focused on voice quality. However, based on the changes in the anatomy and the physiology after a total laryngectomy, and based on the literature on TE speech intelligibility, problems with TE speech intelligibility are also expected to occur. As diminished intelligibility will have a negative effect on communication, we argued in the introduction (chapter 1) that intelligibility should be investigated in order to establish what the typical problems are TE speakers have. We also argued that when a problem is identified, a possible solution should be offered. For that reason, a rehabilitation therapy study was performed to see if TE speech intelligibility can be improved with speech training focusing specifically on the typical problems TE speakers have.

This thesis project thus consisted of two main parts: an assessment study and a therapy study. The assessment study consisted of perceptual analyses that provide insight in the most common intelligibility problems, and acoustic analyses that focused on the most commonly found problem: the inability to produce a correct voiced-voiceless distinction. The therapy part consisted of the development of a speech training program and the evaluation of the effects of this program. In this final chapter, the results from the previous chapters will be summarized and discussed in relation to the main aims formulated in this thesis. Also, the limitations of the studies will be described and directions for future research will be given.

8.2 The assessment of TE speech intelligibility

First the perceptual analyses will be summarized and discussed, followed by a discussion of the acoustic analyses

8.2.1 Perceptual analyses

Perceptual analyses were chosen because they provide insight in how the TE speaker is perceived and how intelligible TE speakers are. As we were particularly interested in how well TE speakers are perceived in their daily communicative environment, naïve, inexperienced and untrained listeners were asked to participate in the listening experiment. Because the experiment was time consuming, it was decided to put the experiment online. This allowed listeners to perform the experiment whenever it suited them and from the comfort of their own home. An online experiment brings with it certain risks: the quality of the sound card or the noise in the environment cannot be controlled. Even though uniform high quality headphones were provided, people might have chosen to use their own. However, despite these risks, inter-listener reliability was sufficiently high to conclude that the online experiment was reliable.

In chapter 2, the set-up and the results of the perception experiment have been discussed. Eleven TE speakers participated in the experiment. We only included patients with a standard laryngectomy, i.e. no patients with some form of pharyngeal reconstruction were included and all patients used a Provox2 voice prosthesis. These selection criteria were developed in order to create a group as homogeneous as possible. The reason for excluding patients with reconstructions is that, potentially, different reconstructions can have different effects on speech intelligibility.

The perception experiment showed that the phoneme scores were much lower for the TE speakers than the scores found for normal laryngeal (NL) speakers in comparable experiments. It was found that the overall consonant correct score was only 72%, which is 21% lower than the score found by Pols (1983). Initial position was hardest to perceive correctly, while final position was easiest. The fricatives as a group were the most difficult category in initial and medial position, while the nasals were most difficult in final position. The most common error was the confusion between voiced and voiceless counterparts in plosives and fricatives.

The consonant clusters show results similar to the single consonants: an overall cluster correct score of 78% was found and initial position was more difficult than the other positions. The consonant clusters all contained a transition from voiced to voiceless or the other way round. Just as with the consonants, it appeared that this transition was difficult to produce correctly.

The overall vowel correct score was 74%. Again, this is a much lower score (10% lower) than the score described for NL speakers (Koopmans-van Beinum, 1980). Long vowels were the most difficult category, with the most common error the diphthongization of the vowel. The feature height was also a problem, both for short and long vowels.

The sentences scored rather low: only 57% correct. However, Semantically Unpredictable Sentences (SUS) are difficult to perceive, as there is little help of context. There was also much variation between the speakers, which suggests that this test differentiates between the poorer and the good speakers.

The scores for the spontaneous speech are more difficult to interpret. Semantic scales were used to obtain more general judgments on particular aspects of speech. The Principal Component analysis found only one component, which makes a VARIMAX rotation impossible and results harder to interpret, especially since almost all scales showed high factor loadings. Both the voice quality scales and the speech intelligibility scales made up this one component, which seems to indicate that naïve listeners do not distinguish between voice quality and speech intelligibility. However, we also included one overall judgment scale, consisting of the three options *good– moderate –poor*, and this scale distinguished well between the values on the other scales. A discriminant analysis showed that the scale *easy to understand – difficult to understand* best predicts class membership (good-moderate-poor).

The results found in chapter 2 confirm our expectations as formulated in chapter 1 (introduction). In that chapter it was stated that problems with the voiced-voiceless distinction were to be expected, as well as with the glottal phoneme /h/. Also problems with vowels, pitch and phrasing were expected to occur. With the perception experiment, the first aim of this thesis, i.e. to establish what the typical intelligibility problems are TE speakers have, has been reached.

The assessment of TE speech intelligibility is very time consuming, one of the main drawbacks of a perception experiment. As we are particularly interested in the intelligibility of spontaneous speech, correlations were calculated between the different phoneme and sentence tests of the experiment and the semantic scales. It was found that SUS sentences correlate highly with all scales (except the unreliable scales *low – high* and *deep – shrill*). This result seems to suggest that SUS sentences are a good predictor for TE speech intelligibility. As SUS sentences are less time consuming than a semantic scaling experiment, this is considered a positive result. However, it should be kept in mind that with only eleven speakers, it cannot be concluded that the latter test can be discarded.

8.2.2 Acoustic analyses

Where perceptual analyses indicate *what* goes right or wrong, acoustic analyses can provide insight as to *why* something is perceived the way it is. In the acoustic study, the focus was only on the most common confusion found in the perception experiment: the voiced-voiceless distinction. The aim was to find out which acoustic correlates TE speakers use when they produce the contrast correctly and if they differ from NL speakers in that respect. The plosives /p b t d/ and the fricatives /f v s z/ were included. Fourteen acoustic correlates were investigated for the plosives and a subset of six for the fricatives. The assumption was that TE speakers would use compensatory strategies and would exaggerate the durational correlates. The results are somewhat surprising: all correlates distinguish between voiced and voiceless for both speaker groups. In addition, hardly any differences are found between TE and NL speakers. Even voicing itself (periodicity) is employed by TE speakers, albeit in a less consistent way than the NL speakers do. In order to find out which of these correlates was the most important cue, conditional inference trees were used. These show that *relative phonation time* *during the closure* of the plosives is the prime cue. This is an interesting result, as this is the first time ever this variable has been used. This variable measures the percentage of time a sound, periodic or aperiodic, is present during the closure. This result suggests that complicated pitch measurements might not be necessary to categorize a sound as voiced or voiceless and that intensity measurements alone are enough. For the fricatives, it was found that the consonant duration is the most important cue. This is not unexpected, as this is previously described in the literature.

Even though the acoustic aspects of the voiced-voiceless distinction have been studied extensively in this thesis, there are some limitations that need to be discussed. We have deliberately only included stimuli that were perceived more than 80% correct. This allowed us to investigate the difference between TE speech and NL speech. In addition, if we know what good speakers do, this information can be used in speech training. However, it did mean that only a limited number of stimuli could be included which was especially a problem with the fricatives. Having only few stimuli makes statistical analyses harder. It would have been interesting to investigate the difference between poorly perceived stimuli and well-perceived stimuli, but with the current data set, this was not possible, as was discussed in chapter 3. In addition, it was stated in chapter 3 that we hoped that acoustic analyses would provide information about the functioning of the neoglottis. With the current results, it is not possible to make any firm statements about the amount of control TE speakers have over their neoglottis. The fact that pitch is used, suggests that at least some of the speakers have control over their neoglottis, but probably in a less consistent way. Other studies have found similar results, as was discussed in chapter 3. At this point, we do not know, however, what happens in the poorly produced stimuli. More research is needed to study the functioning of the neoglottis and to see if TE speech is myoelastic-aerodynamic, or only aerodynamic.

Despite the fact that *relative phonation time in the closure* emerged as the prime distinguishing cue, it does not necessarily mean that this is also the prime cue used in perception. The perceptual relevance of this cue will need to be investigated by means of a perception experiment in which the value of the cue is systematically manipulated.

8.3 The speech training program for TE speakers

The results from the intelligibility assessment confirmed our expectations that TE speech intelligibility is moderately to severely diminished. It was decided that a speech training program should be developed that focuses specifically on the problems found in the perception experiment.

In chapter 4, the development of the evidence-based training program is described. A first step towards an evidence-based training program is a systematic literature search. This search revealed that hardly any studies have been performed on improving TE speech intelligibility. The techniques that were described in literature and were expected to be successful were incorporated in our program. Other techniques, such as the use of clear speech, were taken from studies on different populations, in this case the hard of hearing. The final program consisted of two main levels: phoneme level and discourse level. In order to improve overall intelligibility, a progression from phoneme to discourse level is necessary. The first six lessons were dedicated to training problematic phonemes, whereas the remaining three lessons were devoted to training intonation and phrasing as well as monologues and dialogues.

In chapters 5 and 6 the pre and post tests used to evaluate the effect of the therapy were discussed and results of these tests were presented. After only nine lessons in five weeks of training, and dedicated homework by the participants, the consonants, and with that the voiced-voiceless distinction, improved significantly. The speakers themselves were more confident and more satisfied with their speech after the therapy. Unfortunately, reading aloud and retelling did not improve. Possible explanations for that were given in chapter 6: especially the retelling task is challenging as it places a high load on memory. Also, in therapy it is often seen that people perform worse, before they become better speakers. Overall, it can be concluded that TE speakers benefit from such a structured therapy program.

In chapter 7 the therapy as a whole was evaluated. The speakers and the participating SLPs had filled out an evaluation form. Based on the answers of those evaluations, the experience of the researcher and the evaluation of the pre and post tests, recommendations for the future were made. It was discussed that the Voice Handicap Index was not suitable to test intelligibility-related quality of life. Also the use of semantic scales to measure spontaneous speech will need to be reconsidered.

In chapter 5, two main questions were formulated: can TE speech intelligibility be improved by specific speech training, and is the set-up of the training program appropriate. The results show that speech intelligibility can be improved. Even though not all aspects improved, results are definitely promising. Where the second question is concerned, we believe that in principle the set-up that was chosen is correct, but that (minor) adjustments are necessary to further optimize the program.

8.4 Limitations

Both in the analytic part and the therapy part of the study, a limited number of subjects were included. Only eleven TE speakers participated in the listening experiment described in chapter 2. The reason for this was that the experiment was originally considered a pilot study. The idea was that the experiment would be repeated with a much larger group. However, this first study already yielded many useful results and based on the fact that results from this study matched results from other studies, both from abroad and the Netherlands, it was assumed that an experiment with more speakers would not yield different data.

In the therapy study, only nine TE speakers participated. There were several reasons for this. First of all, it appeared to be challenging to find laryngectomized individuals interested to participate. The program itself was quite demanding and many patients from the Netherlands Cancer Institute live too far away to travel to the hospital twice a week for five weeks. Furthermore, logistically, it was only possible to include a maximum of ten speakers. This last point was one of the main reasons why no control group was added either: we would have been obliged to offer speech therapy to the control group after the study and this was logistically not possible due to the present lack of infrastructure and funding to provide such time consuming therapy to more laryngectomized individuals on a more regular basis.

It should be noted that with only eleven speakers in the first and nine speakers in the second study, statistical results should be regarded with care. It could be the case with these small groups that no significance can be established, while in fact with a larger sample significance might be found. It was also discussed earlier that no firm conclusions can be made based on the correlations found. With a larger sample, statistical analyses would have had more power. However, most of our results were similar to results from other studies and hence we felt comfortable enough to make statements about our findings.

Both in the analytic and in the therapy part of this thesis, only perception experiments have been used to evaluate intelligibility. Perception experiments are still considered the gold standard as they provide information on what goes right or wrong. Unfortunately, no standardized (validated) intelligibility tests exist for TE speakers. We therefore chose several commonly used intelligibility tests. The reason for using different tests is that we wanted to test various levels of speech. We were not only interested in phoneme intelligibility, but also and mostly in overall (spontaneous) speech intelligibility. However, perception experiments have one major disadvantage: they are very time consuming, both for the listeners and the researcher, especially when several different experiments are used for evaluation. We had hoped that at the end of the study one test or a subset of tests would have emerged as being the best test(s) for intelligibility. For that reason, correlations between the tests were calculated in chapter 2 and 7. Correlations can be used to establish whether a score on one test can predict the score on another test. If the correlation is sufficiently high, it would be safe to discard one of the tests. Unfortunately, with only eleven subjects in the analytic part and nine subjects in the therapy part of this thesis, one has to be very careful in interpreting results. Even when high correlations are found, based on the low number of subjects, it still cannot be concluded that tests can be discarded.

Another problem with non-validated tests is that we cannot predict how speakers will behave. Theoretically, it is possible that a speaker shows very different behavior on the same test but at a different time. This possibility makes it difficult to compare pre and post tests as the difference in answers between pre and post tests could either be caused by the fact that the test is not validated or by the fact that a therapeutic intervention has taken place. In retrospect it would have been better if we had performed a baseline study, implying a test-retest without therapeutic intervention, for all of the separate tests we used. A baseline study would have shown how consistent the tests are and if changes can be attributed only to the therapy. The time constraints in this on itself already elaborate study, however, made this not possible. However, although the tests are not validated, they are commonly used in intelligibility research, especially the phoneme intelligibility tests. In addition, the results that were found in this study for the phoneme intelligibility tests match results from earlier studies. It can therefore be assumed that these tests are appropriate and show consistent behavior. Semantic scaling experiments are more difficult: we found different results for the group in chapter 2 than for the group in the therapy study. Semantic scaling experiments are also considered to be quite difficult to perform, especially for naïve listeners. We believe that at the moment they are the best test available for spontaneous speech, but a less complex and shorter test would be welcome.

This discussion clearly shows the need for a validated intelligibility test that can also be used to assess overall intelligibility. In Belgium, De Bodt et al. (2006, the SPACE project) have developed the Nederlandse Spraak Verstaanbaarheid Onderzoek (NSVO; Dutch Speech Intelligibility Research), a tool to measure intelligibility for all kinds of speech disorders. They focus mainly on phoneme intelligibility. The test is multiple-choice and the rater is the researcher. The great advantage is that it only takes 10 minutes per speaker to perform the test. The disadvantage at the moment is that the intra-rater reliability and the inter-rater reliability are low. Also, the poorer the intelligibility of the speaker, the less agreement can be found between raters.

Another option, less time consuming and more objective, is the use of automatic speech recognition for the evaluation of speech intelligibility. Schuster et al. (2006) are investigating the use of this instrument particularly for TE speakers. They conclude that automatic speech recognition serves as a good means to objectify and quantify global speech outcome of laryngectomees. Similar research is being performed in Belgium, by Martens (personal communication and based on presentation at "Spraakverstaanbaarheid meten en verbeteren"("Measuring and improving speech intelligibility")). Both studies find that correlations between subjective and objective raters are sufficiently high, but that more research will need to be done is this area in order to obtain correlations that approximate correlations between subjective raters and to obtain standard values for TE speech. Nevertheless, it is a promising tool, as no large groups of listeners are needed anymore to evaluate speech.

8.5 Suggestions for further research

In both the assessment and the therapy parts of our study only patients with a standard total laryngectomy have been included. In the therapy part, there were two deviations from that norm: one patient with a reconstruction of the neoglottis with a gastric transposition and one with a Pectoralis Major Musculocutaneous flap. The reason that only standard TLE's were (meant to be) chosen was because we wanted the speaker group to be as homogeneous as possible. In future studies it would also be quite interesting to include (more) speakers with reconstructions, especially since it can be assumed that their intelligibility is less good than for laryngectomized individuals with a standard TLE. However, before they can be included in the therapy study, it needs to be clear in what way the reconstructions impede intelligibility: if it is physically impossible to produce certain articulatory movements, there is no point in training them. This point is connected to the following point:

In this study, perception experiments have mostly been employed to evaluate TE speech intelligibility, except for acoustics in chapter 3 and the questionnaires in chapter 6. However, as was mentioned in the introduction, speech is multidimensional and many aspects need to be taken into account when investigating TE speech intelligibility. Other aspects than articulation are, for example, voice quality, prosody and aerodynamics. Perception experiments only tell us what goes wrong and not why or where something goes wrong. Other methods are available to study different aspects of speech. In the introduction, it was mentioned that in the original project proposal videofluoroscopy and digital high-speed imaging were envisaged. Videofluoroscopy visualizes the vocal tract by means of x-ray images and can be used to study vowels. When reference markers are used, objective measurements can be performed on every vowel production. This may provide insight as to how the anatomy/morphology of the neoglottis and vocal tract is linked to the production and perception of the vowels. Interesting though the results will be, for therapy purposes it is not likely to be a useful addition: even though videofluoroscopy may provide information on the cause of the problems, it cannot be repeated, due to strict health regulations. Hence, changes in the vocal tract anatomy cannot be monitored. In addition, it would be quite naïve to expect that the speakers could be able or trained to make minor adjustments in his/her vocal tract while producing vowels.

Another method that was mentioned was digital high-speed imaging and connected to that videokymography. These methods allow visualization of the neoglottis and can provide useful and interesting information on the vibratory characteristics of the neoglottis. In a small-scale pilot study (Jongmans et al., 2006) we asked the speakers to produce ‘patakabadaga’ in order to find out whether vibrations for the voiced sounds were alternated with no vibrations for the voiceless sounds. It would show whether speakers are capable of producing vibrations voluntarily and up to what level the vibratory patterns appear to be irregular and uncontrolled. It has been discussed that acoustic analyses alone cannot answer this question. The information about the neoglottis could thus be linked to intelligibility as well. In addition, information on the functioning of the neoglottis is also highly important for therapy purposes. When restrictions in anatomy and physiology are known, this knowledge can be taken into account when developing and offering articulation techniques.

Acoustic analyses were only used on a small-scale basis in this research, actually only to investigate the voiced-voiceless distinction. Much more could be investigated with acoustic analyses. Prosodic features, length and location of pauses, and vowel quality are examples of interesting research that would complement perceptual analyses, especially since clear speech aimed at improving these features.

It would be useful if a multidimensional protocol was developed, to be used in the clinic that would quickly measure a speaker’s capability and intelligibility. Van As et al. (2006) suggested the use of acoustic signal typing as a quick evaluation method. They found that these signal typings correlated highly with the measured voice quality. The European Society for Laryngology developed the ELS protocol for the assessment of voice disorders, resulting in the Dysphonia Severity Index (DSI, see Wuyts et al., 2000). Though such assessment tools seem to be available for voice quality, no assessment methods yet exist for speech quality and TE speech in particular. If such a tool would exist, it would be possible to assess the therapeutic needs of the speaker and afterwards the progress the speaker has made.

Where the therapy study is concerned, the aim was to develop an evidence-based training program. According to Moore et al. (2005), five levels of evidence can be distinguished. In our study we have level III evidence, meaning we had a well-designed trial, without randomization, with only one group of subjects performing a pre and a post test. The limitations of this study and suggestions for future research have been discussed in chapter 7. After further optimizing the program by using the information obtained during the therapy study, we should strive to obtain level II evidence, which means strong evidence from at least one properly designed randomized controlled trial of appropriate size. A multi-centre study should then be considered.

Long-term effects of the training program should also be studied in the future. Within the available time for this study, this was not possible and therefore remains something to be handed over to the next project.

8.6 Conclusion and future directive

In this study, we have provided an overview of the most common intelligibility problems of TE speech. Our expectation that TE speech quality is diminished compared to normal laryngeal speech was confirmed. This strengthened our belief that speech therapy for TE speakers deserves attention. It can be concluded that as a result of our study we now have a promising rehabilitation tool to improve TE speech intelligibility: the fact that phoneme intelligibility improved after only five weeks of training is a positive result that shows that TE speakers can be trained. In addition, all participants were positive about the program and the effect it had on them. Even when no objective improvement could be found, these laryngectomized individuals became more aware of their speech and felt more confident. It is our expectation, based on the results, that overall intelligibility can be improved as well. Based on the evaluation of the therapy program, this program can be optimized further with the aim of achieving the expected effects. And, obviously, the revised program should be evaluated again to see if the aims have been achieved. The main aim of a new evaluation study would be to further validate this training program and to make it available for general use in hospitals for all TE speakers.

English Summary

In the Netherlands, around 600 people are diagnosed with laryngeal cancer annually. When radiotherapy or laser surgery is not effective in curing the illness, the patient still has a good chance of surviving with a total laryngectomy. About 150 laryngectomies are performed each year. With a total laryngectomy, the entire larynx is removed, including the vocal folds. The windpipe (trachea) is disconnected from the digestive tract (esophagus). The trachea is bent forward and sutured in the skin at the base of the neck. This tracheal opening is called the tracheostoma.

There are three possible ways of voice rehabilitation after a total laryngectomy: the use of an electro-larynx, the use of esophageal speech and the use of tracheoesophageal (TE) speech. In this thesis, all emphasis is on TE speakers. To make TE speech possible, a fistula is created in the wall between the trachea and the esophagus. A voice prosthesis is then inserted, which acts as a one-way valve which allows air to pass from the lungs into the esophagus, but prevents food and liquids from entering the lungs. In order to be able to speak the speaker closes off his or her stoma, and air is then redirected through the voice prosthesis into the esophagus. In the esophagus, the air can cause vibrations of the pharyngoesophageal mucosa. These vibrations are used for speech production.

In **chapter 1** it is described that a total laryngectomy does not only have a profound effect on the *voice* quality of the speaker, but that *speech* quality and thus intelligibility is also affected. Several studies have shown that TE speech intelligibility is compromised. However, most of these studies involved languages other than Dutch. Only small scale studies had been performed for Dutch so far. The first aim of this thesis then is to study Dutch TE speech in more detail (both perceptually and acoustically) to gain insight in the most typical problems for these speakers. If intelligibility is indeed compromised, the next step would be to offer therapy to improve intelligibility. However, no structured speech rehabilitation programs were available for improving TE speech intelligibility. The second aim of this thesis is therefore to develop an evidence-based rehabilitation program. In the rehabilitation part of the study we hope to answer two main questions:

1. Can TE speech intelligibility be improved by means of a therapy program based on and addressing the typical problems found?
2. Is the set-up of the therapy program, and are the pre and post tests used to measure effects, appropriate?

In **chapter 2**, a large perception experiment is discussed which was developed in order to test TE speech intelligibility at various speech levels. Results are meant as a starting point for the development of the therapy program aimed at improving TE speech intelligibility.

11 TE speakers participated. Ten untrained and naïve listeners were asked to perform various intelligibility experiments, as we are most interested in how well a TE speaker is understood in his daily environment.

The speech material consisted of Dutch native consonants, consonant clusters, all native Dutch vowels, sentences and spontaneous speech. Nonsense syllables and semantically unpredictable sentences were used to minimize the effect of learning and context. For the phonemes and the sentences, listeners had to write down what they perceived. For spontaneous speech, various rating scales were used to evaluate speech intelligibility.

Main findings are high inter-listener reliability scores for all subtests, indicating that naïve listeners can reliably judge TE speech intelligibility. A difference between hands-free speakers, using an automatic speaking valve, and speakers who digitally close of their stoma is found for almost all subtests, with the hands-free speakers exhibiting higher intelligibility scores. Large individual differences are found between speakers, but none of the overall results are influenced by this. All tests show markedly lower scores than scores for normal laryngeal speakers as described in literature.

Many confusions are found for the consonants and vowels, with the most common confusions between voiced and voiceless fricatives and plosives. Sentence scores are rather low (57% correct). The judgments on spontaneous speech again show large variation between speakers, but overall speakers scored moderately. Correlations between the phoneme and sentence identification tasks and the rating scales showed that Semantically Unpredictable Sentences (SUS) correlated highly with all but the pitch related scales. This seems to indicate that SUS assessment would be a quick and easy way to evaluate TE speech intelligibility.

In **chapter 3** the difficulty of producing a correct voice distinction is studied in more detail by performing *acoustic* analyses on voiced and voiceless plosives and voiced and voiceless fricatives. The difficulty to produce this contrast correctly is attributed to the working of the neoglottis. As the (in)ability to produce the voicing contrast correctly can teach us more about the functioning of the neoglottis, acoustic analyses have been performed on this contrast. Main aim of this chapter is to investigate which acoustic correlates are used to produce a correct voiced-voiceless distinction and if TE speakers differ from normal laryngeal (NL) speakers in the use of correlates.

The plosives /p b t d f v s z/ in medial position were included for analysis. The recordings were the same as were employed in chapter 2. A control group of 5 NL speakers was included for comparison. A set of 14 acoustic correlates were investigated for the plosives and a subset of 6 for the fricatives. Only stimuli that were perceived more than 80% correct were included as the first interest of this study goes to a better understanding of the importance of specific acoustic features once the distinction has been produced correctly.

It was expected TE speakers would use fewer correlates to produce a correct distinction than the NL speakers, but results show that *all* correlates distinguish between voiced and voiceless plosives, for both speaker groups. For the fricatives, three out of six correlates distinguish between voiced and voiceless for the TE speakers and five out of six for the NL speakers. Contrary to our expectations, hardly any differences between the speaker groups are found, except for the correlates related to periodicity. As periodicity is produced at (neo)glottal level, this is not unexpected. It should be noted that periodicity was used by TE speakers to convey the contrast, but not as consistently as the NL speakers used it.

Even though all cues are found to differ between voiced and voiceless, it is important to study which cue best predicts class membership, in other word, which cue is the prime cue. Conditional inference trees were used to investigate this. For the plosives, it is found that the *percentage of phonation during the closure* is the main distinguishing cue. For the fricatives it is the consonant duration. Perception experiments in which the acoustic cues are manipulated will have to show if the main acoustic cue is also the main cue for perception.

Studies on TE speech intelligibility all found that TE speech is less intelligible than NL speech, including our own study as described in chapters 2 and 3. Mistakes are made both for manner of articulation and place of articulation and the most common confusions found are between voiced and voiceless sounds. Despite the low intelligibility rates and articulation problems, it is not common practice yet to offer speech therapy in a structured manner. It is to be expected however, that speech training focusing on the most common and persistent problem features will help improve overall speech intelligibility of TE speakers. However, a systematic literature search, described in **chapter 4**, rendered no evidence-based training programs. For other languages than Dutch, some individual features were trained and some improvement was found. For Dutch, only experience-based programs or techniques were found. For that reason, we developed a therapy program, based on the problem features that were found in our perception experiment. A program with fixed contents, consisting of nine one-hour sessions was developed. In the Netherlands this set up more or less ‘guarantees’ refunds by medical insurance and it makes planning for speech pathologists easier. Having the same program for all subjects also means that preparation time is limited. The program consists of two parts: the first part deals with individual phonemes/features and the second part with accentuation and phrasing in sentences and spontaneous speech. This set up allows for a progression from phoneme to discourse level. Our main objective of this study is to provide speech language pathologists (SLPs) with an evidence-based training program that is known to improve TE speech intelligibility. In order to see if the therapy has been successful, a number of pre and post tests were performed, the results of which are described in chapters 5 and 6.

The pre and post tests can be divided into phoneme and sentence identification tasks (described in chapter 5), and the more qualitative tests that are described in chapter 6 (questionnaires and a semantic scaling experiment). One test in **chapter 5** is performed by phonetically trained, but naive listeners, and one is performed by SLPs familiar with TE speech. In both tests listeners were asked to write down the phoneme (or sentence) they perceived. Results show that consonants improve significantly with the phonetically trained listeners, just as the semantically unpredictable sentences (SUS) do. Vowel scores are already high and do not improve. Overall, the SLPs give the speakers a higher rating on the post test. It can be concluded from this chapter that in only 5 weeks of training, improvement in phoneme intelligibility is obtainable.

In **chapter 6,** the other pre and post tests, used to evaluate the effect of the therapy discussed in chapter 4, are discussed. These tests consisted of structured study-specific questionnaires for both the speaker and one of his relatives, and a listening experiment in which TE speech was evaluated by naïve listeners. The questionnaires consisted of multiple choice questions on the impact TE speech intelligibility has on daily life. The questions were the same for speakers and their relatives, though the focus for the speaker was on himself and the focus for the relative on the speaker (How well are you understood on the phone *versus* how well can you understand your partner on the phone). Speakers also filled out a Voice Handicap Index (VHI).

For the listening experiment, the speakers read aloud a story and were then asked to recount the story. Both types of speech were evaluated by naïve listeners by means of 17 semantic scales, representing aspects of speech such as intelligibility, rate and rhythm, voice quality and loudness.

Results show that speakers are more satisfied with their speech after rehabilitation training. Before training they were significantly less satisfied about their speech than their relatives were, but this difference disappeared in the post test. The VHI did not render useful results. This may have been caused by the fact that the VHI is primarily meant to evaluate voice, rather than speech. Also the semantic scaling experiment shows less positive results. Scores on the reading aloud task do not improve and the scores on the retelling task decrease significantly. When looking at the separate aspects of speech, a positive result is found for the articulation scales, but not for the rest. It seems that 3 one-hour sessions in the therapy program are too short to carry over the improvements on the phoneme and sentence level to continuous speech. Results suggest that speakers did try to use the techniques learned during training, but in a too much exaggerated way and not yet automatically.

In **chapter 7,** the therapy program as a whole is evaluated. Main aim of this chapter is to answer the question if the set-up of the therapy program has been appropriate for the purposes of this study. In this case, the set-up includes the structure and contents of the training program as well as the pre and post tests used to measure effects. The results from chapters 5 and 6 as well as the personal experiences of the researcher, the participants and the speech language pathologists, are taken into account in this evaluation. Correlations between the different pre tests are investigated to see if one or just a few tests would be sufficient. It appears that only very few significant correlations exist. This is probably caused by the fact that the tests themselves are very different as well as the raters. Also, some tests showed improvements whereas others did not and this will influence the correlations. The conclusion is that several different tests should be used, as there is not one single test that can measure (improvement of) intelligibility in a satisfactory way. Except for the VHI, the use of all tests seems justified in this study.

Where the content and structure of the program is concerned, the participants and SLPs were interviewed by means of structured questionnaires. Both groups are quite satisfied with the program, but the SLPs had some remarks on the variety of the exercises and the transfer to spontaneous speech. The results from chapter 6 confirm the opinion of the SLPs and thus the conclusion is warranted that in the future, spontaneous speech should deserve much more attention. Based on all these findings, recommendations are made for the future.

In the final **chapter 8**, the separate experiments and their results are summarized and briefly discussed. The perception experiment discussed in chapter 2 has shown that TE speech intelligibility is (severely) diminished, especially when compared to normal laryngeal speech. This confirms our expectations and at the same time backs up our decision to develop a speech training program for TE speakers. The effect of the speech training program was measured with pre and post tests. The overall results are positive: after only five weeks, consonants intelligibility improves and speakers themselves are satisfied.

Of course, this thesis also has its limitations, such as a limited number of subjects and the lack of a control group. These and other limitations of this thesis are discussed, followed by suggestions for further research. However, the overall conclusion is that even though the set-up of the therapy and the use of certain pre and post tests may need to be altered slightly, dedicated speech rehabilitation for TE speakers is worthwhile. It is our belief that after an optimization of the program, SLPs will have a useful instrument to help improve (overall) TE speech intelligibility.

Nederlandse samenvatting

In Nederland krijgen jaarlijks zo'n 600 mensen de diagnose keelkanker. Als radiotherapie of laserchirurgie de ziekte niet meer kunnen genezen, heeft de patiënt toch nog een goede overlevingskans met een totale laryngectomie. Jaarlijks worden zo'n 150 laryngectomieën uitgevoerd. Bij een laryngectomie wordt de hele larynx, inclusief de stembanden, weggehaald. De luchtpijp (trachea) wordt gescheiden van de slokdarm (esofagus) en wordt vervolgens naar voren gebogen en vastgemaakt in de basis van de nek. De opening die hier ontstaat heet een tracheostoma.

Na een laryngectomie zijn er drie manieren van spraakrevalidatie mogelijk: het gebruik van een electrolarynx, het gebruik van esofageale (slokdarm) spraak en het gebruik van tracheoesofageale (TE) spraak. In dit proefschrift zal alleen TE spraak aan de orde komen. Om TE spraak mogelijk te maken, wordt een fistula (gaatje) gemaakt in de wand tussen de trachea en de esofagus. In deze fistula wordt een stemprothese geplaatst, die fungeert als eenrichtingsklepje waarbij lucht vanuit de longen wel naar de slokdarm kan, maar eten en drinken niet van de slokdarm in de luchtpijp kan komen. Om te kunnen spreken, sluit een spreker zijn stoma af. Vervolgens wordt de lucht van de longen via de stemprothese naar de slokdarm geleid. In de slokdarm kan deze lucht trillingen veroorzaken van het faryngoesofageale slijmvlies. Deze trillingen worden dan gebruikt voor de productie van spraak.

In **hoofdstuk 1** wordt beschreven dat een totale laryngectomie niet alleen een grote invloed heeft op de stemkwaliteit van de spreker, maar ook op zijn spraakkwaliteit, oftewel zijn verstaanbaarheid. Verschillende studies tonen aan dat TE spraakverstaanbaarheid is aangetast. Echter, de meeste van deze studies betreffen andere talen dan het Nederlands. Slechts enkele kleinere studies hebben naar Nederlandse TE spraak gekeken. Het eerste doel van dit proefschrift is dan ook om Nederlandse TE spraak in meer detail te onderzoeken (zowel perceptief als akoestisch) om inzicht te krijgen in de meest voorkomende problemen bij TE sprekers. Als verstaanbaarheid inderdaad blijkt te zijn aangetast, zou de volgende stap het aanbieden van spraaktraining zijn om de verstaanbaarheid te verhogen. Echter, er zijn geen gestructureerde spraak revalidatieprogramma's gevonden om TE spraakverstaanbaarheid te verbeteren. Daarom is het volgende doel van dit proefschrift het ontwikkelen van een evidence-based revalidatieprogramma. In dit revalidatiedeel van de studie hopen wij de volgende twee vragen te beantwoorden:

1. Kan TE spraakverstaanbaarheid worden verbeterd met een therapie- programma dat is gebaseerd op de typische problemen van TE sprekers?
2. Is de opzet van het therapieprogramma en zijn de pre- en posttesten die zijn gebruikt om effect te meten geschikt?

In **Hoofdstuk 2** wordt een groot perceptie-experiment besproken dat was opgezet om TE spraakverstaanbaarheid op verschillende spraakniveaus te testen. De resultaten zijn bedoeld als startpunt voor de ontwikkeling van het spraaktrainingsprogramma om TE spraakverstaanbaarheid te verbeteren.

Elf sprekers deden mee aan dit experiment. Tien ongetrainde en naïeve luisteraars zijn gevraagd mee te doen aan de verschillende verstaanbaarheidstesten, omdat wij vooral zijn geïnteresseerd in hoe een TE spreker in zijn dagelijkse omgeving wordt verstaan.

Het spraakmateriaal bestond uit syllaben met Nederlandse medeklinkers, medeklinkerclusters, alle Nederlandse klinkers, semantisch onvoorspelbare zinnen en spontane spraak. De syllaben en semantisch onvoorspelbare zinnen zijn gebruikt om leereffecten en gebruik van context zo klein mogelijk te maken. Bij de fonemen en zinnen moesten luisteraars intypen wat ze hoorden. Voor spontane spraak zijn verschillende semantische schalen gebruikt om de verstaanbaarheid te beoordelen.

Resultaten laten een hoge inter-beoordelaar betrouwbaarheid zien voor alle subtesten. Dit betekent dat naïeve luisteraars betrouwbaar zijn als zij TE spraakverstaanbaarheid beoordelen. Er is ook een verschil gevonden tussen sprekers met een automatische spreekklep en sprekers die hun stoma afsluiten met de vinger, waarbij de sprekers met een automatische spreekklep betere verstaanbaarheidsscores laten zien. Er zijn grote individuele verschillen zichtbaar tussen de sprekers, maar de gemiddelde resultaten worden hier niet door beïnvloed. Alle testen laten lagere scores zien dan vergelijkbare testen bij normale laryngeale (NL) sprekers die worden beschreven in de literatuur.

Er zijn veel verwarringen gevonden bij de medeklinkers en de klinkers, waarbij de meest voorkomende de verwarring tussen stemhebbende en stemloze plosieven en fricatieven is. De score voor de zinnen is behoorlijk laag (57%). De oordelen over de spontane spraak laten weer een grote variatie zien tussen de individuele sprekers, maar over het algemeen genomen scoorden de sprekers redelijk. Correlaties tussen de foneem- en zinsverstaanbaarheidstesten en de schaaloordelen laten zien dat de semantisch onvoorspelbare zinnen hoog correleren met alle schalen, behalve met de schalen die met toonhoogte te maken hebben. Dit lijkt aan te geven dat het gebruik van deze zinnen een snelle en makkelijke manier is om TE spraakverstaanbaarheid te evalueren.

In **hoofdstuk 3** wordt het probleem om een correct onderscheid te maken tussen stemhebbende en stemloze plosieven en fricatieven in meer detail onderzocht door middel van akoestische analyses. De moeilijkheid om dit contrast correct te produceren wordt toegeschreven aan de werking van de neoglottis. We hebben akoestische analyses uitgevoerd, omdat het wel of niet correct kunnen produceren van het stemcontrast ons meer kan leren over hoe de neoglottis werkt. Het hoofddoel van dit hoofdstuk is uit te vinden welke akoestische correlaten worden gebruikt om een correct stemcontrast te produceren en of TE sprekers verschillen van NL sprekers in hoe zij het contrast maken.

De plosieven /p b t d f v s z/ in mediale positie zijn geanalyseerd. De opnames die zijn gebruikt zijn dezelfde als in hoofdstuk 2. Er is een controlegroep van 5 NL sprekers ter vergelijking. Een set van 14 akoestische correlaten is onderzocht voor de plosieven en een subset van 6 voor de fricatieven. Alleen de stimuli die meer dan 80% correct zijn waargenomen zijn geïncludeerd, omdat de eerste aandacht uit gaat naar een beter begrip van het belang van specifieke akoestische kenmerken bij een correct geproduceerd contrast.

De verwachting was dat TE sprekers minder correlaten zouden gebruiken om een correct contrast te maken tussen stemhebbend en stemloos dan de NL sprekers, maar de resultaten tonen aan dat alle correlaten onderscheid maken tussen stemhebbende en stemloze plosieven, bij beide sprekergroepen. Bij de fricatieven maken drie van de zes correlaten onderscheid tussen stemhebbend en stemloos bij de TE sprekers, en vijf van de zes bij NL sprekers. In tegenstelling tot wat wij verwachtten is nauwelijks verschil gevonden tussen de sprekergroepen, behalve voor de correlaten die met periodiciteit te maken hebben. Aangezien periodiciteit wordt geproduceerd op (neo) glottaal niveau, is dit niet onverwacht. Opgemerkt dient te worden dat TE sprekers wel periodiciteit gebruiken om het contrast te maken, maar dat zij dit niet zo consequent doen als de NL sprekers.

Alhoewel alle cues verschil maken tussen stemhebbend en stemloos, is het toch belangrijk te onderzoeken welke van die cues het best voorspelt tot welke klasse (stemhebbend of stemloos) een foneem behoort, oftewel, wat is de meest belangrijke cue. Conditionele regressiebomen zijn gebruikt om dit te onderzoeken. Voor de plosieven is gevonden dat het *percentage fonatie tijdens de sluiting* het best onderscheid maakt tussen stemhebbend en stemloos. Bij de fricatieven bleek medeklinkerduur het meest belangrijk. Perceptie-experimenten waarbij de verschillende cues worden gemanipuleerd, zullen moeten uitwijzen of de belangrijkste akoestische cue ook perceptief het belangrijkst is.

Alle studies over TE spraakverstaanbaarheid tonen dat TE spraak minder verstaanbaar is dan NL spraak, net als onze eigen studie zoals beschreven in hoofdstuk 2 en 3. Fouten worden zowel voor manier als voor plaats van articulatie gevonden en de meest voorkomende verwarring is tussen stemhebbende en stemloze plosieven en fricatieven. Ondanks de lagere verstaanbaarheid en de articulatieproblemen, wordt spraaktherapie niet standaard en gestructureerd aangeboden. Het is echter te verwachten dat met spraaktherapie die zich specifiek richt op de gevonden problemen de algehele spraakverstaanbaarheid zal verbeteren. Echter, een systematisch literatuuronderzoek, beschreven in **hoofdstuk 4**, leverde geen evidence-based trainingsprogramma's op. In het buitenland zijn studies geweest waarin op bepaalde individuele kenmerken is getraind waarbij verbetering werd gevonden. In Nederland zijn alleen maar programma's of technieken gevonden die gebaseerd zijn op ervaring en niet op bewijs. Om die reden ontwikkelden wij een trainingsprogramma dat is gebaseerd op de specifieke problemen van de TE sprekers die in het perceptie-experiment zijn gevonden. Het resultaat is een programma met een vaste inhoud, verdeeld over negen lessen van een uur. Deze opzet garandeert in Nederland min of meer een vergoeding van de zorgverzekeraar voor de deelnemer. Bovendien is een vast programma makkelijker in te plannen voor de logopedist. Door hetzelfde programma aan te bieden aan alle patiënten, blijft de voorbereidingstijd voor de logopedist beperkt. Het programma bestaat uit twee delen: het eerste deel betreft het trainen van individuele fonemen en kenmerken; het tweede deel betreft het trainen van accentuering en frasering in zinnen, en het trainen van spontane spraak. Door deze indeling is een progressie mogelijk van foneemniveau naar discourse niveau. Ons hoofddoel van deze studie is om logopedisten een trainingsprogramma te kunnen geven waarvan het nut is vastgesteld. Om te zien of het programma daadwerkelijk succesvol is geweest, zijn een aantal pre- en posttesten uitgevoerd, waarvan de resultaten zijn beschreven in hoofdstuk 5 en 6.

De pre- en posttesten kunnen worden verdeeld in foneem- en zinsidentificatietesten (beschreven in hoofdstuk 5) en de meer kwalitatieve testen, besproken in hoofdstuk 6 (vragenlijsten en semantische schalen experiment). Een van de testen beschreven in **hoofdstuk 5** is uitgevoerd door fonetisch getrainde, maar naïeve luisteraars en de andere test is uitgevoerd door logopedisten die ervaring hebben met TE spraak. In beide testen is aan de luisteraars gevraagd om in te typen welk foneem of welke zin zij hadden gehoord. Resultaten laten zien dat de medeklinkers significant verbeteren in de test uitgevoerd door de fonetische luisteraars, net als de semantisch onvoorspelbare zinnen. De klinkers laten in de pretest al een zeer hoge score zien en verbeteren niet. In het algemeen geven ook de logopedisten een betere score aan de sprekers in de posttest. De conclusie die naar aanleiding van dit hoofdstuk getrokken kan worden, is dat na slechts vijf weken training een verbetering in foneemverstaanbaarheid haalbaar is.

In **hoofdstuk 6** worden de andere pre- en posttesten die zijn gebruikt om het effect van therapie te meten, besproken. Deze testen bestaan uit gestructureerde studiespecifieke vragenlijsten voor zowel de sprekers als een direct familielid, en een luisterexperiment waarin naïeve luisteraars TE spraak beoordelen. De vragenlijsten bestaan uit meerkeuzevragen over de invloed die de verstaanbaarheid van de TE spreker heeft op het dagelijkse leven. De sprekers en de familieleden kregen dezelfde vragen, alhoewel voor de sprekers de focus lag op hemzelf en voor de logopedisten lag op de spreker. (Hoe goed wordt u verstaan aan de telefoon versus Hoe goed kunt u uw partner verstaan aan de telefoon). De sprekers vulden ook nog een Voice Handicap Index (VHI) vragenlijst in.

Voor het luisterexperiment lazen de sprekers een verhaal voor waarna zij het verhaal navertelden. De spraak van beide opdrachten is beoordeeld door naïeve luisteraars door middel van 17 semantische schalen die aspecten zoals verstaanbaarheid, tempo en ritme, stemkwaliteit en luidheid representeerden.

Resultaten tonen dat de sprekers meer tevreden zijn over hun spraak na het volgen van het therapieprogramma. Voor aanvang van de training waren de sprekers significant meer ontevreden over hun spraak dan hun familieleden waren, maar dit verschil verdween weer in de post test. De VHI blijkt niet erg nuttig voor onze studie wat wellicht wordt veroorzaakt door het feit dat deze test primair bedoeld is om het effect van stemkwaliteit en niet van verstaanbaarheid op het dagelijks leven te meten. Ook het semantische schalenexperiment laat niet erg positieve resultaten zien. De scores voor het voorlezen verbeteren niet en de scores voor het navertellen verslechteren zelfs. Als wij vervolgens kijken naar de verschillende aspecten van spraak, dan zien wij dat de articulatieschalen wel verbeteren, maar de rest niet. Deze resultaten lijken er op te wijzen dat 3 sessies van een uur in het trainingsprogramma niet voldoende zijn om een progressie van foneemverstaanbaarheid naar spontane spraakverstaanbaarheid te waarborgen. De resultaten suggereren verder dat de sprekers wel geprobeerd hebben de geleerde technieken in de praktijk te brengen, maar dat zij dit op een te overdreven manier hebben gedaan en dat het nog geen automatisme is.

In **hoofdstuk 7** wordt het therapieprogramma in zijn geheel geëvalueerd. Het hoofddoel van dit hoofdstuk is het beantwoorden van de vraag of de opbouw van het programma geschikt is geweest voor het doel van onze studie. In dit geval betreft opbouw zowel de structuur en de inhoud van het programma als de pre- en posttesten die zijn gebruikt voor de effectmeting. De resultaten van hoofdstuk 5 en 6, samen met de persoonlijke ervaringen van de onderzoeker, de sprekers en de logopedisten, zijn betrokken bij de evaluatie. Om te kijken of er een of slechts een paar testen voldoende zouden zijn om de spraakverstaanbaarheid te meten, zijn correlaties tussen alle testen berekend. Het blijkt dat slechts weinig testen significant met elkaar correleren. Dit wordt waarschijnlijk veroorzaakt door het feit dat zowel de testen zelf als de beoordelaars erg van elkaar verschillen. Ook laten sommige testen wel verbetering zien en andere niet wat de correlaties zal beïnvloeden. De conclusie is dat meerdere verschillende testen gebruikt dienen te worden, omdat er niet een specifieke test bestaat die de (verbetering van) spraakverstaanbaarheid naar tevredenheid kan meten. Behalve VHI lijkt het gebruik van alle andere testen in deze studie gerechtvaardigd.

De deelnemers en de logopedisten is door middel van gestructureerde vragenlijsten gevraagd wat hun mening is over de structuur en inhoud van het programma. Beide groepen blijken redelijk tevreden over het programma, maar de logopedisten hebben wel wat opmerkingen over de variatie van de oefeningen en de transfer naar spontane spraak. De resultaten van hoofdstuk 6 bevestigen de mening van de logopedisten en daarom concluderen wij dat in de toekomst spontane spraak meer aandacht verdient. Op basis van al deze bevindingen worden aanbevelingen gedaan voor de toekomst.

In het laatste **hoofdstuk 8** worden de verschillende experimenten en hun resultaten samengevat en kort bediscussieerd. Het perceptie-experiment uit hoofdstuk 2 laat zien dat TE spraakverstaanbaarheid (ernstig) verminderd is, zeker als TE spraak vergeleken wordt met normale laryngeale spraak. Dit bevestigt ons vermoeden en onderbouwt tegelijkertijd ons besluit om een spraaktrainingsprogramma voor TE sprekers te ontwikkelen. Het effect van het trainingsprogramma is gemeten met verschillende pre- en posttesten. Over het algemeen zijn de resultaten positief: na slechts vijf weken training is de verstaanbaarheid van medeklinkers verbeterd en zijn de sprekers meer tevreden over hun eigen spraak.

Natuurlijk heeft dit proefschrift ook beperkingen, zoals het kleine aantal deelnemers en het gebrek aan een controlegroep. Deze en andere beperkingen worden bediscussieerd, gevolgd door suggesties voor verder onderzoek. Echter, de algemene conclusie is dat hoewel de opzet van de therapie en het gebruik van bepaalde pre- en posttesten wellicht moeten worden aangepast, intensieve spraakrevalidatie zinvol is voor TE sprekers. Wij geloven dat wij logopedisten, na het verder optimaliseren van het programma, een bruikbaar instrument in handen kunnen geven dat de (algehele) spraakverstaanbaarheid van TE sprekers kan verbeteren.

Bibliography

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

Ackerstaff, A.H., Hilgers, F.J.M., Aaronson, N.K. & Balm, A.J.M. (1994). Communication, functional disorders and lifestyle changes after total laryngectomy. *Clinical Otolaryngology, 19*, 295-300.

Alphen, P.M. van & Smits, R. (2004). Acoustical and perceptual analysis of the voicing distinction in Dutch initial plosives: the role of prevoicing. *Journal of Phonetics, 32*, 455-491.

As, C.J. van, Hilgers, F.J.M., Verdonck-de Leeuw, I.M. & Koopmans-van Beinum, F.J. (1998). Acoustical analysis and perceptual evaluation of tracheoesophageal prosthetic voice. *Journal of Voice*, *12* (2), 239-248.

As, C.J. van, Tigges, M., Wittenberg, T., Op de Coul, B.M., Eysholdt, U. & Hilgers, F.J.M. (1999). High-speed digital imaging of the neoglottic vibration after total laryngectomy. *Archives Otolaryngology Head Neck Surgery, 125,* 891-897.

As, C.J. van, Op de Coul, B.M., Van den Hoogen, F.J., Koopmans-van Beinum, F.J. & Hilgers, F.J.M. (2001). Quantitative videofluoroscopy: a new evaluation tool for tracheoesophageal voice production. *Archives Otolaryngology Head Neck Surgery, 127* (2),161-169.

As, C.J. van (2001). *Tracheoesophageal speech. A multidimensional assessment of voice quality*. PhD thesis. Amsterdam: Universiteit van Amsterdam.

As, C.J. van, Koopmans-van Beinum, F.J., Pols, L.C.W. & Hilgers, F.J.M. (2003). Perceptual evaluation of tracheoesophageal speech by naïve and experienced judges through the use of semantic differential scales. *Journal of Speech, Language, Hearing Research, 46*, 947-959.

As-Brooks, C.J. van, Koopmans-van Beinum, F.J., Pols, L.C.W. & Hilgers, F.J.M. (2006). Acoustic signal typing for evaluation of voice quality in tracheoesophageal speech. *Journal of Voice, 20* (3), 355-368.

Bassich, C.J. & Ludlow, C.L. (1986). The use of perceptual methods by new clinicians for assessing voice quality. *Journal of Speech and Hearing Disorders, 51*, 125-133.

Belloc, J.B., Vaissière, J. & Brasnu, D. (1997). Identification of consonants in tracheo-esophageal speech (phonatory implant). *Larynx-97*, 109-114.

Benoit, C. (1990). An intelligibility test using semantically unpredictable sentences: towards the quantification of linguistic complexity. *Speech Communication, 9*, 293-304.

Benoit, C., Grice, M. & Hazan, V. (1996). The SUS test: a method for the assessment of text-to-speech synthesis intelligibility using Semantically Unpredictable Sentences. *Speech Communication, 18*, 381-392.

Berg, J.W. van den (1958). Myoelastic-Aerodynamic theory of voice production. *Journal of Speech and Hearing Research, 1*, 227-244.

Bergem, D.R. van, Pols, L.C.W. & Koopmans-van Beinum, F.J. (1988). Perceptual normalization of the vowels of a man and a child in various contexts. *Speech Communication, 7*, 1-20.

Bertino, G., Bellomo, A., Miani, C., Ferrero, F. & Staffieri, A. (1996). Spectrographic differences between tracheo-esophageal and esophageal voice. *Folia Phoniatrica Logopedica, 48,* 255-261.

Bocca, E. & Calearo, C. (1963). Central hearing processes. In J. F. Jerger (Ed.), *Modern developments in audiology* (pp. 337-370). New York: Academic Press.

Bodt, M. de, Jacobson, B., Musschoot, S., Zaman, S., Heylen, L., Mertens, F., Van de Heyning, P.H. & Wuyts, F.L. (2000). De Voice Handicap Index, een instrument voor het kwantificeren van psychosociale consequenties van stemstoornissen (The Voice Handicap Index, an instrument for quantifying the psychosocial consequences of voice disorders). *Logopedie*, *13*, 29-33.

Bodt, M. de, Guns, C. & Van Nuffelen, G. (2006). *NSVO: Nederlandstalig SpraakVerstaanbaarheidsOnderzoek (Dutch Speech Intelligibility Study)*. Herentals: Vlaamse Vereniging voor Logopedisten.

Boersma, P. & Weenink, D. (1996). Praat, a system for doing phonetics by computer. *IFA report*, No. 132. Institute of Phonetic Sciences, Universiteit van Amsterdam, Amsterdam.

Bond, Z.S. & Moore, T.J. (1994). A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Communication, 14*, 325-337.

Boon-Kamma, B. (2001). *Verstaanbaarheid na totale laryngectomie. De verstaanbaarheid van sprekers met een Provox® 2 stemprothese (Intelligibility after total laryngectomy. The intelligibility of speakers with a Provox® 2 voice prosthesis).* MA thesis, Amsterdam: Universiteit van Amsterdam.

Bors, E.F.M., Dijkstra, J., de Vries, R.J. & Wicherlink, W.H. *Oefengang voor de button-oesophagusstem (Therapy program for prosthetic speech)*. Publicatie Centrum voor Stem-, Spraak- en Taalstoornissen, KNO-kliniek AZG.

Bridges, A. (1991). Acceptability ratings and intelligibility scores of alaryngeal speakers by three listener groups. *British Journal of Disorders of Communication, 26*, 325-335.

Cervera, T., Miralles, J.L. & Gonzalez-Alvarez, J. (2001). Acoustical analysis of Spanish vowels produced by laryngectomized subjects. *Journal of Speech, Language, and Hearing Research, 44*, 988-996.

Christensen, J.M. & Dwyer, P.E. (1990). Improving alaryngeal speech intelligibility. *Journal of Communication Disorders, 23***,** 445-451.

Christensen, J.M. & Weinberg, B. (1976). Vowel duration characteristics of esophageal speech. *Journal of Speech and Hearing Research, 19* (4), 678-89.

Cullinan, W.L., Brown, C.S. & Blalock, P.D. (1986). Ratings of intelligibility of esophageal and tracheoesophageal speech. *Journal of Communication Disorders, 19*, 185-195.

Cutler, A. & Butterfield, S. (1990). Durational cues to word boundaries in clear speech. *Speech Communication, 9*, 485-495.

Debruyne, F., Delaere, P., Wouters, J. & Uwents, P. (1994). Acoustic analysis of tracheo-esophageal versus esophageal speech. *Journal of Laryngology and Otology, 108*, 325-328.

Deschler, D.G., Doherty, E.D., Reed, C.G. & Singer, M.I. (1999). Effects of sound pressure levels on fundamental frequency in tracheoesophageal speakers. *Otolaryngology-Head and Neck Surgery, 121*, 23-26.

Dharmaperwira-Prins, R.I.I. (1998). *Dysartrie en Verbale Apraxie (Dysarthria and verbal apraxia)*. Lisse : Swets bv.

Doyle, P.C., Danhauer, J. & Reed, C. (1988). Listeners' perceptions of consonants produced by esophageal and tracheoesophageal talkers. *Journal of Speech and Hearing Disorders, 53*, 400-407.

Doyle, P.C., Swift, R.E. & Haaf, R.G. (1989). Effects of listener sophistication on judgments of tracheoesophageal talker intelligibility. *Journal of Communication Disorders, 22*, 105-113.

Drost, H.A. (1978). *Slokdarmspraak (Esophageal speech)*. Pioniersreeks, Vermande.

Ernestus, M. (2000). *Voice assimilation and segment reduction in casual Dutch*. PhD thesis. Utrecht: Holland Institute of Generative Linguistics.

Fagel, W.P.F., Van Herpt, L.W.A. & Boves, L. (1983). Analysis of the perceptual qualities of Dutch speakers' voice and pronunciation. *Speech Communication, 2*, 315-326.

Fant, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton.

Fitzpatrick, P.A., Gould, D. & Nichols, A.C. (1980). Self-administered intelligibility practice for oesophageal speakers. *Journal of Communication Disorders*, *13*, 341-346.

Hakkesteegt, M.M., Wieringa, M.H., Gerritsma, E.J. & Feenstra, L. (2006). Reproducibility of the Dutch version of the Voice Handicap Index. *Folia Phoniatrica Logopedica, 58* (2), 132-8.

Hammarberg, B., Lundström, E. & Nord, L. (1990). Consonant intelligibility in esophageal and tracheoesophageal speech. A progress report. *Phoniatric and Logopedic Progress Report*, No. 7. Department of Logopedics and Phoniatrics, Huddinge University Hospital & Karolinska Institute, Stockholm.

Hazan, V., Benoit, C. & Jekosch, U. (1989). Assessment of the intelligibility of synthetic speech using semantically unpredictable sentences. *Esprit-SAM project 1541 Extension Phase Final Report,* 61-82.

Heaton, J.M., Sanderson, D., Dunsmore, I.R. & Parker, A.J. (1996). Speech assessment of patients using three types of indwelling tracheo-esophageal voice prostheses. *The Journal of Laryngology and Otology*, *110*, 343-347.

Heuven, V.J.J.P. van & Scharpff, P.J. (1991). Acceptability of several pausing strategies in low quality speech synthesis: interaction with intelligibility. *Proceedings ICPhS XII*, 458-461.

Hilgers, F.J.M. & Schouwenburg, P.F. (1990). A new low-resistance, self retaining prosthesis (Provox®) for voice rehabilitation after total laryngectomy. *Laryngoscope 100*, 1202-1207.

Hilgers, F.J.M., Ackerstaff, A.H., Balm, A.J.M. & Gregor, R.T. (1996). A new heat and moisture exchanger with speech valve (Provox® Stomafilter). *Clinical Otolaryngology, 21*, 414-418.

Hilgers, F.J.M., Ackerstaff, A.H., Balm, A.J.M., Tan, I.B., Aaronson, N.K. & Persson, J.O. (1997). Development and clinical evaluation of a second-generation voice prosthesis (Provox® 2), designed for anterograde and retrograde insertion. *Acta Otolaryngologica, 117*, 889-896.

Hilgers, F.J.M., Ackerstaff, A.H., Van As, C.J., Balm, A.J.M., Van den Brekel, M.W.M. & Tan, I.B. (2003). Development and clinical assessment of a Heat and Moisture Exchanger with a multi-magnet automatic tracheostoma valve (Provox FreeHands HME) for vocal and pulmonary rehabilitation after total laryngectomy. *Acta Otolaryngology (Stockholm), 123*, 91-99.

Hoemeke, K.A. & Diehl, R.L. (1994). Perception of vowel height: The role of F1-F0 distance. *Journal of the Acoustical Society of America, 96* (2), 661-674.

Hothorn, T., Hornik, K., Van de Wiel, M.A. & Zeileis, A. (2006). A Lego system for conditional inference. The *American Statistician*, *60* (3), 257–263.

Hothorn, T., Hornik, K., Van de Wiel, M.A. & Zeileis, A. (2006a). Unbiased recursive partitioning: a conditional inference framework. *Journal of Computational and Graphical Statistics*, *15* (3), 651–674.

Huang, W. (2007). *English as a lingua franca: Mutual intelligibility of Chinese, Dutch and American speakers of English*. PhD thesis. Utrecht: LOT.

Isman, K.A. & O'Brien, C.J. (1992). Videofluoroscopy of the pharyngoesophageal segment during tracheoesophageal and esophageal speech. *Head and Neck, 14*, 352-358.

Jacobson, B.H., Johnson, A., Grywalski, C., Silbergleit, A., Jacobson, G., Benninger, M.S. & Newman, C.W. (1997). The Voice Handicap Index (VHI): development and validation. *American Journal of Speech Language Pathologie, 6*, 66-70.

Jongmans, P., Van As, C.J., Hilgers, F.J.M. & Pols, L.C.W. (2003). An introduction to the assessment of the intelligibility of tracheoesophageal speech. *IFA Proceedings 25*, 185-196.

Jongmans, P., Hilgers, F.J.M., Pols, L.C.W. & van As-Brooks, C.J. (2005). The intelligibility of tracheoesophageal speech: first results. *Proceedings Interspeech 2005*, 1749-1752.

Jongmans, P., Hilgers, F.J.M., Pols, L.C.W. & Van As-Brooks, C.J. (2006). The intelligibility of tracheoesophageal speech, with an emphasis on the voiced-voiceless distinction. *Logopedics. Phoniatrics. Vocology, 31*, 172-181.

Jongmans, P., Wempe, A.G., Hilgers, F.J.M., Pols, L.C.W. & Van As-Brooks, C.J. (2007). Acoustic correlates of the voiced-voiceless distinction in Dutch normal and tracheoesophageal speakers. *Proceedings ICPhS XVI*, 1997-2000.

Kent, R.D. & Read, C. (2002). *Acoustic analysis of speech*. Albany: Singular Thomson Learning.

Kinnear, P.R. & Gray, C.D. (1999). *SPSS for Windows made simple*. Hove: Psychology Press Ltd, Publishers.

Klein, W., Plomp, R. & Pols, L.C.W. (1970). Vowel spectra, vowel spaces, and vowel identification. *The Journal of the Acoustical Society of America, 48*, 4 (2), 999-1009.

Koopmans-van Beinum, F.J. (1969). Nog meer fonetische zekerheden (Even more phonetic certainties). *De Nieuwe Taalgids, 62*, 245-250.

Koopmans-van Beinum, F.J. (1980). *Vowel contrast reduction: an acoustic and perceptual study of Dutch vowels in various speech conditions*. PhD thesis. Amsterdam: Universiteit van Amsterdam.

Krause, J.C. & Braida, L.D. (2002). Investigating alternative forms of clear speech: the effects of speaking rate and speaking mode on intelligibility. *Journal of the Acoustical Society of America, 112*, 2165-2172.

Krause, J.C. & Braida, L.D. (2004). Acoustic properties of naturally produced clear speech at normal speaking rates. *Journal of the Acoustical Society of America, 115* (1), 362-378.

Kreiman, J.C., Gerrat, B.R. & Precoda, K. (1990). Listener experience and perception of voice quality. *Journal of Speech and Hearing Research, 33*, 103-115.

Kreimann, J., Gerratt, B.R., Precoda, K. & Berke, G.S. (1992). Individual differences in voice quality perceptions. *Journal of Speech and Hearing Research, 35*, 512-520.

Kreiman, J., Gerrat, B.R., Kempster, G.B., Erman, A. & Berke, G.S. (1993). Perceptual evaluation of voice quality: a review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36*, 21-40.

Lane, H. & Tranel, B. (1971). The Lombard sign and the role of hearing in speech. *Journal of Speech and Hearing Research, 14*, 677-709.

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge: Cambridge University Press.

Leemrijse, C., Steultjens, E., Dorgelo, M. & Van den Ende, E. (2004). *Prioritering van onderzoek naar de effectiviteit en doelmatigheid van ergotherapie en logopedie (Prioritizing of research in the effectivity and efficiency of occupational therapy and speech therapy)*. Nivel 2004.

Lieberman, P. (1977). *Speech physiology and acoustic phonetics: an introduction*. New York: Macmillan.

Lieberman, P. & Blumstein, S.E. (1988). *Speech physiology, speech perception, and acoustic phonetics*. Cambridge: Cambridge University Press.

Lindau, M., Jacobson, L. & Ladefoged, P. The feature advanced tongue root. *UCLA Phonetics Laboratory, Working Papers in Phonetics, 22,* 76-94.

Lindblom, B. (1990). Explaining phonetic variation: a sketch of the H&H theory. In Hardcastle, W.J. & Marchal, A. (Eds), *Speech production and speech modelling*. (pp.403-439) Dordrecht: Kluwer Academic publishers.

Lisker, L. & Abramson, A.S. (1964). A cross-language study of voicing in initial stops: acoustical measurements. *Word, 20*, 384-422.

Lundström, E. & Hammarberg, B. (2004). High-speed imaging of the voicing source in laryngectomees during production of voiced-voiceless distinctions for stop consonants. *Logopedics. Phoniatrics. Vocology. 29*, 31-40.

MacKenzie, K., Millar, A., Wilson, J.A., Sellars, C. & Deary, I.J. (2001). Is voice therapy an effective treatment for dysphonia? A randomised controlled trial. *British Medical Journal, 323* (7314), 658-660.

Maddalena, H. de & Pfrang, H. (1993). Improvement of communication behavior of laryngectomized and voice-rehabilitated patients by a psychological training program. *Hals-, Nase-, Ohrheilkunde, 41* (6), 289-95.

Mase-Goldman, D., Allen, E.J. & Nichols, A.C. (1988). Oesophageal intelligibility training: back consonants and clusters. *Journal of Communication Disorders, 21,* 437-445.

Max, L., Steurs, W., De Bruyn, W. (1996). Vocal capacities in esophageal and tracheoesophageal speakers. *Laryngoscope*, *106*, 93-96.

McColl, D.A. (2006). Intelligibility of tracheoesophageal speech in noise. *Journal of Voice, 20* (4), 605-615.

Meulen, Sj. van der (1990). *Spraakrevalidatie na laryngectomy, oefenprogramma I (Speech rehabilitation after laryngectomy, exercise program I)*. Harcourt Assessment B.V.

Meulen, Sj. van der (1990). *Spraakrevalidatie na laryngectomy, oefenprogramma II (Speech rehabilitation after laryngectomy, exercise program II)*. Swets & Zeitlinger Publishers.

Miralles, J. & Cervera, T. (1995). Voice intelligibility in patients who have undergone laryngectomies. *Journal of Speech and Hearing Research*, *38*, 564-571.

Mohri, M., Yoshifuji, M., Kinishi, M. & Amatsu, M. (1994). Neoglottic activity in tracheoesophageal phonation. *Auris Nasus Larynx (Tokyo), 21*, 53-58.

Moon, J.B. & Weinberg, B. (1987). Aerodynamic and myoelastic contributions to tracheoesophageal voice production. *Journal of Speech and Hearing Research, 30*, 387-395.

Moore, A., McQuay, A. & Gray, J.A.M. (1995). Evidence-based everything. *Bandolier, 12*, 1.

Most, T., Tobin, Y. & Mimran, R.C. (2000). Acoustic and perceptual characteristics of esophageal and tracheoesophageal speech production. *Journal of Communication Disorders, 33* (2), 165-180.

Nieboer, G.L.J., De Graaf, Tj. & Schutte, H.K. (1988). Esophageal voice quality judgements by means of the semantic differential. *Journal of Phonetics*, *16*, 417-436.

Nooteboom, S.G., Scharff, P. & Van Heuven, V.J. (1990). Effects of several pause strategies on the recognizability of words in synthetic speech. *Proceedings ICSLP, Kobe, Japan, 1*, 385-387.

Nord, L., Hammarberg, B. & Lundström, E. (1992). Voiced-voiceless distinction in alaryngeal speech-acoustic and articulatory observations. *Phoniatric and Logopedic Progress Report*, No. 2-3. Department of Logopedics and Phoniatrics, Huddinge University Hospital & Karolinska Institute, Stockholm, 19-22.

Op de Coul, B.M.R., Van As, C.J., Van den Hoogen, F.J.A., Marres, H.A.M., Joosten, F.B.M., Manni, J.J. & Hilgers, F.J.M. (2003). Evaluation of the effects of primary myotomy in total laryngectomy of the neoglottis with the use of quantitative videofluoroscopy. *Archives Otolaryngology Head and Neck Surgery, 129,* 1000-1005.

Osgood, C.E., Suci, G.J. & Tannebaum, P.H. (1957). *The measurement of meaning*. Urbana: University of Illinois Press.

Oubrie, A. (1999). *Formantwaarden en verstaanbaarheid van Nederlandse klinkers bij tracheo-esophageale sprekers (Formant values and intelligibility of Dutch vowels of tracheoesophageal speakers)*. MA thesis, Nijmegen: Universiteit van Nijmegen.

Picheny, M.A., Durlach, N.I. & Braida, L.D. (1986). Speaking clearly for the hard of hearing II: Acoustic characteristics of clear and conversational speech. *Journal of Speech and Hearing Research, 29*, 434-446.

Pindzola, R.H. & Cain, B.H. (1988). Acceptability ratings of tracheoesophageal speech. *Laryngoscope, 98*, 394-397.

Pols, L.C.W. (1983). Three-mode principal components analysis of confusion matrices, based on the identification of Dutch consonants, under various conditions of noise and reverberation. *Speech Communication, 2*, 275-293.

Pols, L.C.W. (2002). Spreken en verstaan – spraakklanken (Speech and understanding – speech sounds). In Appel, R., Baker, A., Hengeveld, K., Kuiken, F. and Muysken, P. (eds.), *Taal en Taalwetenschap*. (pp. 245-258). Blackwell Publishers.

Postma, G. (1997). *Spraakrevalidatie na laryngectomie/3. Oefenen met de ventielstemprothese (Speech rehabilitation after laryngectomy, exercise program III. Practising with the voice prosthesis)*. Harcourt Assessment B.V.

Rietveld, T. & Van Hout, R. (1993). *Statistical techniques for the study of language and language behaviour.* Berlin: Mouton de Gruyter.

Robbins, J. (1984). Acoustic differentiation of laryngeal, esophageal, and tracheoesophageal speech. *Journal of Speech and Hearing Research, 27*, 577-585.

Robbins, J., Fischer, H.B., Blom, E.D. & Singer, M.I. (1984a). A comparative acoustic study of normal, esophageal, and tracheoesophageal speech production. *Journal of Speech and Hearing Disorders, 49*, 202-210.

Robbins, J., Fischer, H.B., Blom, E.D. & Singer, M.I. (1984b). Selected acoustic features of tracheoesophageal, esophageal, and laryngeal speech. *Archives of Otolaryngology, 110*, 670-672.

Robbins, J., Christensen, J. & Kempster, G. (1986). Characteristics of speech production after tracheoesophageal puncture: voice onset time and vowel duration. *Journal of Speech and Hearing Research*, *29*, 499-504.

Roeleven, S. & Polak, M. F. (1999). *De Verstaanbaarheid van consonanten in de Nederlandse trachea-esophageale spraak (The intelligibility of consonants in Dutch tracheoesophageal speakers)*. BA thesis, Rotterdam: Rotterdam College of Logopedics.

Rood-Nieuwland, A. (1982). Groepslessen in slokdarmspraak (Group exercises in esophageal speech). *Logopedie en Foniatrie, 54*, 294-97.

Roozen, M. (2005). *The intelligibility of tracheoesophageal speech in spontaneous speech situations*. Report number 145, MA thesis, Amsterdam: Universiteit van Amsterdam.

Rossum, M.A. van, de Krom, G., Nooteboom, S.G. & Quené, H. (2002). 'Pitch' accent in alaryngeal speech. *Journal of Speech, Language and Hearing Research, 45* (6), 1106-1118.

Rossum, M.A. van (2005). *Prosody in alaryngeal speech*. PhD thesis. Utrecht: LOT.

Sackett, D.L. & Rosenberg, W.M.C. (1996). Evidence-based medicine: what it is and what it isn't. *British Medical Journal, 312* (7023), 71-72.

Saito, M., Kinishi, M. & Amatsu, M. (2000). Acoustic analyses clarify voiced-voiceless distinction in tracheoesophageal speech. *Acta Otolaryngology*, *120,* 771-777.

Sanderman, A.A. & Collier, R. (1997). Prosodic phrasing and comprehension. *Language and Speech, 40*, 391-409.

Scharpff, P.J. & Van Heuven, V.J.J.P. (1988). Effects of pause insertion on the intelligibility of low quality speech. In Holmes, J.N. (ed.), *Proceedings of the 7th FASE/Speech-88 Symposium* (pp.261-268). Edinburgh: The Institute of Acoustics.

Schuster, M., Haderlein, T., Nöth, E., Lohscheller, E., Eysholdt, U. & Rosanowski, F. (2006). Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating. *European Archives Otorhinolaryngology, 263*, 188-193.

Searl, J.P., Carpenter, M.A. & Banta, C. (2001). Intelligibility of stops and fricatives in tracheoesophageal speech. *Journal of Communication Disorders*, *34*, 305-321.

Searl, J.P. (2002). Magnitude and variability of oral pressure in tracheoesophageal speech. *Folia Phoniatrica Logopaedica*, *54* (6), 312-328.

Searl, J.P. & Carpenter, M.A. (2002). Acoustic cues to the voicing feature in tracheoesophageal speech. *Journal of Speech, Language and Hearing Research, 45*, 282-294.

Searl, J.P. & Ousley, T. (2004). Phonation offset in tracheoesophageal speech. *Journal of Communication Disorders, 37*, 371-387.

Searl, J.P. (2004). Minding our P's, Q's, and S's in alaryngeal training: focusing on articulation – adequacy to excellence. *VoicePoints on WebWhispers.org*.

Searl, J.P. (2005). Remembering articulation in tracheoesophageal speech rehabilitation. *VoicePoints on WebWhispers.org*.

Singer, M.I. & Blom, E.D. (1980). An endoscopic technique for restoration of voice after laryngectomy. *Annals of Otology, Rhinology and Laryngology, 89*, 529-533.

Sisty, N.L. & Weinberg, B. (1972). Formant frequency characteristics of esophageal speech. *Journal of Speech and Hearing Research, 15* (2), 439-448.

Slis, I.H. & Cohen, A. (1969). On complex regulating voiced-voiceless distinction. *Language and Speech, 12*, 80-102.

Slis, I.H. (1970). Articulatory measurements on voiced, voiceless and nasal consonants: A test of a model. *Phonetica, 21*, 193-210.

Smits, R. (1995). *Detailed versus gross spectro-temporal cues for the perception of stop consonants*. Unpublished manuscript, Institution for Perception Research, Eindhoven.

Speyer, R. (2003). *Effect of voice therapy: measurement and evaluation*. PhD thesis. Utrecht: Universiteit van Utrecht.

Speyer, R., Wieneke, G.H. & Dejonckere, P.H. (2004). Documentation of progress in voice therapy : perceptual, acoustic, and laryngostroboscopic findings pretherapy and posttherapy. *Journal of Voice, 18* (3), 325-340.

Strasser, H. & Weber, C. (1999). On the asymptotic theory of permutation statistics. *Mathematical Methods of Statistics, 8*, 220–250.

Summers, W., Pisoni, D.B., Bernacki, R.H., Pedlow, R.I. & Stokes, M.A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *Journal of the Acoustical Society of America, 84,* 917-928.

Švec, J.G. & Schutte, H.K. (1996). Videokymography: High-speed line scanning of vocal fold vibration. *Journal of Voice, 10*, 201-205.

Titze, I.R. (1994). *Principles of Voice Production*. Englewood Cliffs, NJ: Prentice Hall.

Traunmüller, H. (1981). Perceptual dimension of openness in vowels. *Journal of the Acoustical Society of America, 69*, 1465-1475.

Uchanski, R.M., Choi, S.S., Braida, L.D., Reed, C.M. & Durlach, N.I. (1996). Speaking clearly for the hard of hearing IV: Further studies on the role of speaking rate. *Journal of Speech and Hearing Research, 39*, 494-509.

Verdonck-de Leeuw, I.M. (1998). *Voice characteristics following radiotherapy: The development of a protocol*. PhD thesis. Amsterdam: IFOTT, Universiteit van Amsterdam.

Visser, O., Siesling, S. & van Dijck, J.A.A.M. (2003). *Incidence of cancer in the Netherlands 1999/2000*. Utrecht: Vereniging van Integrale Kankercentra.

Williams, S. & Watson, J. (1987). Speaking proficiency variations according to method of alaryngeal voicing. *Laryngoscope, 97*, 737-739.

Wittenberg, T., Moser, M., Tigges, M. & Eysholdt, U. (1995). Recording, processing and analysis of digital high speed sequences in glottography. *Machine, Vision and Applications, 8,* 399-404.

Wong Chung, R.P., de Bruin, Aalders & Mahieu. (1998). An integrated intensive post-laryngectomy rehabilitation programme. *Clinical Otolaryngology, 23* (2), 186-187.

Wuyts, F.L., De Bodt, M.S., Molenberghs, G., Remacle, M., Heylen, L., Millet, B., Van Lierde, K. & Raes, J. (2000). The Dysphonia Severity Index, an objective measure of vocal quality based on a multiparameter approach. *Journal of Speech, Language, and Hearing Research, 43*, 796-809.

Zeine, L. & Brandt, J.F. (1988). The Lombard effect on alaryngeal speech. *Journal of Communication Disorders, 21*, 373-83.

Curriculum Vitae

Petra Jongmans was born on October 12, 1975 in Hendrik Ido Ambacht. After primary school (Kruiswiel) in that village, she attended secondary school (grammar school) in Dordrecht (Dordts Christelijk Lyceum, later called Insula College) which she completed in 1994.

That same year, she started her study in English Language and Literature at Leiden University. She spent a year at the University of Wales, Bangor, where she studied linguistics and German and taught Dutch. She graduated in 2000, with an MA thesis on Bangor English. In 1998 she started her second study, applied linguistics, at the VU University Amsterdam. She graduated in 2002, with an MA thesis on a program to train the hearing of adult cochlear implant users.

Between 2003 and 2007 she was employed as a Ph.D student at the Institute of Phonetic Sciences of the University of Amsterdam and the Netherlands Cancer Institute, Amsterdam. Her project was sponsored by the "Stichting Breuning ten Cate". She investigated the intelligibility of tracheoesophegeal speakers, resulting in her Ph.D thesis "The intelligibility of tracheoesophageal speech: an analytic and rehabilitation study". While being employed as a Ph.D student, she was also a member of the board of the PhD Researchers' Association and she chaired the board of the "Werkverband Amsterdamse Psycholinguisten".

Since June 2007 she has been employed at the Nationale Hoorstichting (National Hearing Foundation) in Leiden.

For more information, please refer to:

www.fon.hum.uva.nl
www.nki.nl
http://www.medewerker.uva.nl/phd
www.hetwap.nl
www.hoorstichting.nl

Dankwoord

Na 4,5 jaar zwoegen zou ik bijna zeggen dat dit dankwoord het lastigst is om te schrijven. Je wilt graag allerlei mensen bedanken voor hun hulp, onvoorwaardelijke steun en/of vriendschap, maar om alleen maar ‘dank je’ te zeggen is zo kaal. Iets anders verzinnen is zo simpel nog niet. Je wilt grappig zijn, maar geloof me, in die laatste fase van je proefschrift is je gevoel voor humor tot ver beneden het vriespunt gedaald. Soms wil je een persoon laten weten hoeveel je om hem/haar geeft, maar alles klinkt ineens als een stuk tekst uit een slechte soapserie. Soms wil je iemand de hemel in prijzen voor wat hij/zij allemaal voor je heeft gedaan, maar vind je de juiste woorden niet. Maar dan, een dag voor je boekje naar de drukker moet, ben je ineens blij dat je überhaupt iets op papier hebt en hoop je alleen maar dat je niemand vergeet die er echt tussen had moeten staan. Als dat wel gebeurd is, wijt het dan maar aan stress geïnduceerde (tijdelijke) dementie. Dit moet m maar zijn:

Prof. Dr. V.J.J.P. van Heuven, beste Vincent, bedankt dat je ooit tegen me zei dat ik maar “gewoon” moest solliciteren op een promotieplaats en dank je voor al je hulp daarbij!

De sprekers van mijn eerste luisterexperiment, ik vind het fantastisch dat u zich zo belangeloos heeft ingezet en eindeloos voor mij klanken en zinnen heeft voorgelezen. Zonder u en de deelnemers aan het interventieprogramma zou er geen proefschrift zijn geweest!

De deelnemers aan het trainingsprogramma, ik heb veel bewondering voor uw inzet bij het trainingsprogramma. 5 weken 2 keer per week vrijwillig naar het ziekenhuis komen en dan zulke goede prestaties leveren: heel erg bedankt!

Luisteraars, uren hebben jullie moeten luisteren naar allerlei rare woorden en nog raardere zinnen. Dank jullie wel voor al het geduld en alle bereidwilligheid. Ik kon niet zonder jullie.

Marion (van Zuilen), dank je voor alle dingen die je al die jaren voor me geregeld hebt!

Annemieke en alle andere medewerkers van de afdeling Hoofd-Hals oncologie, jullie maakten deel uit van een stimulerende werkomgeving. Annemieke, fijn dat ik met vragen bij je terecht kon.

Harm van Tinteren, statistiek doen voor een onwetende taalkundige is weer eens iets anders, maar ik vind dat we ons er dapper doorheen geslagen hebben. Bedankt!

Karel, wat heb ik gelachen in Duitsland. Je bent zeer goed gezelschap geweest al die jaren!

Beste Herman, dank je wel voor alle interesse in mijn proefschrift en vooral ook voor alle ruimte die je me gaf! Mede daardoor ligt dit boekje er nu.

Michel, van oppaskind naar grafisch wonder ☺ Door jou heeft iedereen nu een mooie aanwinst in de boekenkast.

Marre, het is altijd leuk als mensen geïnteresseerd zijn in je onderwerp. Ik vond het heel leuk om je scriptie te begeleiden en je hebt geweldig werk verricht.

Lieve Diana, heel erg bedankt voor 4 jaar exota, Bijenkorf en Kalverstraat en dat je zo'n leuke kamergenoot was. En dan wil je ook nog m'n paranymf zijn! Ik mis je nog elke dag in Leiden! Wanneer gaan we winkelen?

Raffie, zonder jou in mijn kamer is werk zoveel saaier (en fatsoenlijker!) geworden. Het waren mooie tijden, zeker samen met al jouw "andere" vrouwen Wieneke, Irene en Diana. Dames, van de gratie is nog maar weinig over zo, wat moet de fonetiek nu zonder ons 4 samen…? (en ik zonder jullie?)

Nel dank je dat je zo'n goed voorbeeld was voor me! En natuurlijk voor onze vriendschap!

Dirk, jij was mijn Gert, ik jouw Hermien. Wat hebben we het leuk gehad!

Ton, bedankt voor alle onmisbare input en al het geanalyseer, en dan niet alleen van akoestische kenmerken. Voor een leuke discussie moet je bij jou zijn, en je weet: God bestaat en van zendmasten wordt je ziek!

Rob, altijd maar klaar staan voor je collega's. Je hebt mijn leven en het leven van mijn luisteraars een heel stuk prettiger gemaakt. Dank je!

Lieve IFA-ers, wat hebben we een mooie tijd gehad op de Herengracht! Wat een werkplek en wat een gezellige collega's!

Andere ACLC-ers, de verhuizing naar het BH werd dan wel alom gevreesd, ik zelf vond het eigenlijk achteraf heel gezellig. Ineens had ik er heel veel leuke collega's bij.

Collega's promovendi overleg, in het bijzonder Jarinde en Dennis, het was fijn om nog iets naast het promoveren te doen en het is nog fijner als je dat met zulke gezellige mensen kunt doen!

Lieve wappers, we laten toch elke keer weer zien hoe mooi de taalkunde is. We zijn een goed team! (en heel gezellig).

Mama, Jij hebt me altijd het goede voorbeeld gegeven. Van jou heb ik mijn leergierigheid en doorzettingsvermogen. Dit boek is het resultaat daarvan. Ik hou van je.

Marcel, lief broertje, wat fijn dat je altijd voor mij klaar staat, ook nu weer als paranymf. Met jou naast me op het podium kan het niet meer fout gaan! Je bent geweldig!

Lieve Yvonne, ik kan me werkelijk geen betere schoonzus bedenken. Een dikke 10!

Lieve Lon, Joos en Guid, wat hebben we toch al veel meegemaakt. En het is altijd zoveel leuker, fijner, gezelliger, troostender of makkelijker als we het met elkaar kunnen delen. Jullie zijn geweldig!

Lieve Amanda, jij druk met alweer twee kleintjes, ik met promoveren, maar gelukkig maken we nog steeds tijd voor elkaar en het zal nu alleen maar beter worden.

Dear, sweet Helen, always so busy, but still time to correct my English. I have never had such humoristic comments ☺ You're the greatest!

Lieve Sue, heel veel dank voor al het nakijkwerk! Nu heb ik weer tijd voor wijntjes in Gouda.

En natuurlijk al mijn andere lieve vrienden die voor zoveel leuke afleiding hebben gezorgd tijdens deze bij tijd en wijle zware bevalling. Wat fijn om jullie om me heen te hebben.

Lieve Rianne, ik wil je graag bedanken voor al je inzet tijdens mijn promotie, maar ik wil je nog meer bedanken voor alle gezelligheid in het NKI, onze treinreizen en borrels met saté!

Beste Pauline, ook jij enorm bedankt voor alle hulp en inzet en de gezellige lunches!

Lieve Maya, wat had ik toch zonder je gemoeten in dit proefschrift. Wat heb je veel voor me gedaan! Ik zal dat nooit vergeten. Ik ben erg blij dat we nu weer samen kunnen lunchen in Leiden!

Dr. C.J. van As-Brooks, beste Corina, Amerika was geen belemmering om betrokken te blijven. Onze telefoongesprekken lieten me altijd weer beter voelen. Fijn dat ik altijd mijn hart kon luchten bij je en natuurlijk bedankt voor alle waardevolle input.

Prof. Dr. Ir. L.C.W. Pols en Prof. Dr. F.J.M. Hilgers. Beste Louis, ik weet nog dat ik me geïntimideerd voelde tijdens mijn sollicitatie, maar dat gevoel was niet terecht! Ik beschouw onze samenwerking als uiterst waardevol, leerzaam en vooral heel plezierig! Beste Frans, jouw eeuwige optimisme bracht mijn eeuwige pessimisme aardig in balans. Geweldig hoe jij aan alles een positieve draai weet te geven en dat was vaak nodig om er weer tegen aan te kunnen. Bovendien denk ik dat ik de snelst nakijkende promotoren heb gehad van Nederland. Ik had me geen betere promotoren kunnen wensen en het was een eer jullie als promotor te hebben!

en als allerlaatste mijn allerliefste Gerard: had ik je al verteld dat….