SPEECHREADING SUPPLEMENTED WITH AUDITORY INFORMATION

M. Breeuwer

SPEECHREADING SUPPLEMENTED WITH AUDITORY INFORMATION

Het in dit proefschrift beschreven onderzoek werd uitgevoerd bij de groep Experimentele Audiologie van de vakgroep Keel-, Neus-, Oorheelkunde en Audiologie van de Faculteit der Geneeskunde van de Vrije Universiteit te Amsterdam.

isbn 90-6256-207-8 CIP

VRIJE UNIVERSITEIT TE AMSTERDAM

SPEECHREADING SUPPLEMENTED WITH AUDITORY INFORMATION

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van
doctor in de wiskunde en natuurwetenschappen
aan de Vrije Universiteit te Amsterdam,
op gezag van de rector magnificus dr. P.J.D. Drenth,
hoogleraar in de faculteit der sociale wetenschappen,
in het openbaar te verdedigen
op donderdag 12 september 1985 te 15.30 uur
in het hoofdgebouw der universiteit,
De Boelelaan 1105

door

MARCEL BREEUWER
geboren te Haarlem

VU Uitgeverij
Amsterdam 1985

Promotor: Prof. dr. ir. R. Plomp
Referent : Prof. dr. G.F. Smoorenburg

aan mijn vader

VOORWOORD

Het in dit proefschrift beschreven onderzoek werd gedurende de periode juli 1982 tot en met juni 1985 verricht bij de groep Experimentele Audiologie van de vakgroep Keel-, Neus-, Oorheelkunde en Audiologie van de Faculteit der Geneeskunde van de Vrije Universiteit te Amsterdam. Graag wil ik allen bedanken, die mij behulpzaam zijn geweest bij de uitvoering van dit onderzoek en bij de totstandkoming van dit proefschrift.

De promotor, prof.dr.ir. R. Plomp, dank ik voor zijn inspirerende wetenschappelijke begeleiding en voor de wijze waarop hij het manuscript van dit proefschrift van kritische opmerkingen heeft voorzien. Mijn collega's, Joost Festen, Jan de Laat, Peter Pabon, Cor de Boer, Janette van Dijkhuizen en Peter Anema dank ik voor de prettige werksfeer en voor de adviezen bij wetenschappelijke en technische problemen.

De referent, prof.dr. G.F. Smoorenburg, wil ik bedanken voor het op kritische wijze doornemen van het manuscript van het proefschrift.

Prof.dr.ir. L.C.W. Pols, dr. P.J.J. Lamoré, dr. D.J. Povel en drs. J.M. van der Stelt dank ik voor de deelname aan de begeleidingsgroep van het onderzoeksproject.

Tenslotte dank ik Corien van den Wall Bake voor het op videoband inspreken van het in de experimenten gebruikte spraakmateriaal, het Audiovisueel Centrum van de Vrije Universiteit voor de prettige samenwerking op audiovisueel gebied, Frank Oonk voor zijn adviezen betreffende de Engelse taal en mijn familie en vrienden voor de door hen getoonde interesse in de voortgang van mijn onderzoek.

Marcel Breeuwer

This research was supported by the Foundation for Linguistic Research, which is funded by the Netherlands Organization for the Advancement of Pure Research, ZWO.

CONTENTS

CHAPTER 1 - INTRODUCTION

1.1 The problem

Audiometric surveys, based on large groups of subjects, have shown that approximately 3 to 4% of the human population suffers from an inoperable hearing impairment, which causes them to have difficulties in understanding speech (see Plomp, 1978). These persons have average hearing losses at 500, 1000 and 2000 Hz (AHL) of approximately 35 dB or more. For a large range of hearing losses (AHL less than about 90 dB), amplification of the speech sound by a conventional hearing aid can significantly improve the perception of speech in quiet surroundings. In noisy surroundings hearing aids are of no or little help. For persons with a severe hearing loss (AHL>90 dB, approximately 0.1% of the Dutch population) amplification of the speech signal by a conventional hearing aid offers no or only very limited benefit; persons that do not benefit at all are often called totally or profoundly deaf. The latter group has to rely mainly on speechreading, i.e. they have to perceive spoken language by looking at the articulatory movements, the facial expression, and the gestures of the speaker.

Several factors limit the information that can be obtained by speechreading alone. First, different speech sounds are pronounced with identical lip and jaw movements, which makes them visually indiscriminable (Voiers, 1973; Lowell, 1974). For example, the consonants /p/ and /b/ are produced with the same articulatory movements; therefore words such as "pale" and "bale" can not be distinguished by speechreading alone. Second, several sounds do not produce clearly visible lip or jaw movements at all; e.g., Berger (1972) reported that the consonant /h/ as initial consonant before a vowel can not be perceived by speechreading alone. Third, through speechreading alone hardly any information can be obtained about the prosody (intonation and stress patterns) of the speech signal. The perception of prosodic features is of importance, for example, for deciding whether a sentence was spoken as a statement or as a question, and for perceiving which of the words or syllables in a sentence are emphasized. Fourth, speakers often do not articulate clearly, and often they have to be seen under poor conditions (e.g., under poor illumination).

Because of the above-mentioned factors, speechreading alone is not sufficient for full comprehension of speech. Therefore, it has to be supplemented with extra information about the speech signal. Up to the present a considerable amount of work has gone into research on methods for improving speech perception of deaf persons. Some researchers developed methods in which speechreading is supplemented with manually given information about the speech signal. Cornett (1967), for example, developed a method in which supplementary information about hard-to-distinguish phonemes is provided by different handshapes and handpositions near the mouth. However, the use of such systems is restricted to persons that are familiar with the manual signs; i.e. it can be useful for communication between a deaf person and his or her normal-hearing close relatives, or for communication between deaf persons. Most other researchers aim at developing a speech-perception aid, which extracts the necessary supplementary information automatically from the acoustic speech signal and presents this information to the deaf person through an alternative sense, i.e. through another sense than hearing (for a short review on speech-perception aids see Section 1.4). Three different alternative senses have been used: Upton (1968), for example, attempted to present supplementary information visually, other researchers used tactile stimulation (for reviews see Kirman, 1973; Sherrick, 1984) or electrical stimulation (cochlear implants, for reviews see White, 1982; Millar et al., 1984). Some of the researchers in this area try to achieve speech transmission without the help of speechreading, others aim explicitly to supplement speechreading. On the whole, results with speech-perception aids that do not aim at supplementing speechreading are rather disappointing. Therefore, it seems logical first to develop an efficient speechreading aid.

For a successful development of an automatic speechreading aid two problems have to be solved. First, it has to be clarified which information about the speech signal is the most appropriate for supplementing speechreading. The speech signal itself is a very complex signal and, since the information transmission capacity of the alternative sense is limited, it is not appropriate for direct presentation to the alternative sense. Therefore, processing of this signal is necessary in order to extract the most adequate information for transmission. Second, psychophysical studies should clarify how the alternative senses can be

stimulated most optimally. Very often, these two problems were treated as being one and the same. However, the confounding of these two problems makes it difficult to find the causes of the often observed disappointing results. They may lie in information extraction, in information transduction, or in both. Therefore, we have decided to focus our attention only on the first problem, namely on investigating which information about the speech signal is necessary to effectively supplement speechreading.

The acoustic speech signal can be processed in several ways. First, processing can be either segmental or non-segmental. In segmental processing techniques the speech signal is segmented into smaller units (mostly phonemes) and the information is extracted per unit. Of course, automatic recognition of phonemes would be appropriate; however, the problem of accurate, speaker independent, automatic speech recognition has not yet been solved. Therefore, several researchers use related techniques. Benedetto et al. (1982) and Cornett et al. (1977), for example, attempt to automatically classify each phoneme to a certain group. These groups are chosen in such a way that each phoneme can be recognized by supplementing speechreading with information about to which group it belongs. Phonemes that can not be recognized easily by speechreading alone are arranged in different groups. However, with the current technical means the classification results are far from perfect.

In non-segmental processing techniques no attempt is made to divide the speech signal into smaller units. Two approaches can then be followed. First, processing techniques can be used of which the result is transmitted directly to an alternative sense. For example, a specific frequency band can be filtered from the speech signal and can be presented directly to a single vibrator on the skin (e.g., see DeFillippo, 1984). Second, slowly varying speech parameters can be extracted from the speech signal, such as the fundamental frequency or the overall sound-pressure level, which are then presented to an alternative sense by modulating them on a carrier. For example, the above-mentioned parameters could be presented to a single vibrator on the skin by encoding the fundamental frequency as frequency of vibration and the overall sound-pressure level as amplitude of vibration. The extraction of slowly varying speech parameters has the advantage that the problem of developing an efficient speechreading aid can be split up in two parts: first, it can be investigated which speech parameters contain information that is not or only partly available through speechreading alone and second, it can be investigated on what carrier these parameters should be modulated before presentation to an alternative sense.

Our research is focused on the first part of the problem, i.e. extracting those speech parameters from the speech signal that contain sufficient linguistic information to achieve speech perception in combination with speechreading. We have restricted ourselves to the use of not more than two parameters.

1.2 Methods of research

The speech parameters that we consider to be candidates for supplementing speechreading (see Section 1.3) were evaluated experimentally, by presenting them auditorily to normal-hearing listeners in combination with the visible articulatory movements of the speaker. Then the difference between speech intelligibility under the conditions of speechreading-only and speechreading while listening to the auditorily presented parameters is a measure of the efficiency of the parameters concerned. Thus, in our scheme the alternative sense of the deaf person is simulated by the auditory modality of the normal-hearing listener.

This research method has several advantages. First, it allows for a systematic evaluation of what information is required to supplement speechreading effectively. Second, using the auditory modality of normal-hearing subjects means that we use the optimal way of presenting the supplementary speech information; we can be quite sure that if a parameter proves to be no efficient supplement to speechreading when presented auditorily, it also will not be efficient when presented to an alternative sense. Third, the role of training is reduced to a minimum. A successful stimulation of an alternative sense requires a substantial amount of training before different stimuli can be discriminated. However, the auditory modality is used to receive acoustic stimuli; therefore, the amount of training that will be required to adjust the subjects to the auditorily presented speech parameters can be reduced to a minimum by modulating these parameters on acoustic carriers that are familiar to the subjects. For example, in one of the experiments (see Chapters 1.3 and 4) speechreading was supplemented with information about the fundamental

frequency of the speech signal, by encoding the fundamental frequency as the frequency of a periodic pulse sequence that was subsequently band-pass filtered between 160 and 400 Hz. The resulting signal contains frequency variations that are similar to what is perceived as the intonation of the original speech signal. Thus, the acoustic stimuli are synthesized in such a way that they contain modulations that do not conflict with the modulations present in the original speech signal.

However, our method also has disadvantages. Parameters that prove to be efficient supplements to speechreading when presented auditorily indicate only the lower limit of the information that is required to supplement speechreading through an alternative sense. It may well be that information is lost during the transmission of the speech parameters through the alternative sense; then, extra information has to be added in order to compensate for this loss. It also may be that the best-scoring auditorily presented parameter is not the best for presentation to an alternative sense, e.g., because other parameters fit better to the characteristics of the alternative sense.

1.3 The selected speech parameters

In order to be a candidate for supplementing speechreading a speech parameter has to contain information that can not or only partly be perceived by speechreading alone. As was mentioned in Section 1.1, by speechreading alone limited information can be obtained about individual speech segments (phonemes), and hardly any information can be obtained about the prosody of the speech signal. Thus, both segmental and suprasegmental information are lacking.

The speech signal is a constantly changing, non-stationary signal with frequency components between approximately 80 and 10,000 Hz. Segmental information is present primarily in the gross shape of the amplitude spectrum (which displays the strength of the individual frequency components) and in the changes in time of this shape. The speech signal is highly redundant, not all the frequency components have to be perceived to indentify a particular sound. For example, in telephone systems only the frequencies between 400 and 3400 Hz are transmitted. The development of Vocoder systems, i.e. systems in which the speech signal is represented by the variations of the amplitude envelopes in a limited number of narrow frequency bands, has shown that approximately 10 to 15 bands are sufficient to reconstruct intelligible speech (e.g., see Flanagan, 1972). By speechreading alone some information can be obtained about individual sounds; therefore, the number of frequency bands required to achieve speech perception by speechreading in combination with information about the amplitude variations in these bands will probably be less than 10. Chapter 2 describes an experiment in which was investigated whether information from only one or two frequency bands is sufficient to supplement speechreading. The sound-pressure levels in one or two frequency bands with center frequencies of 500, 1600, or 3160 Hz, and with one-third or one octave bandwidth, respectively, were used as supplements. One or two frequency bands with the above-mentioned center frequencies and bandwidths were filtered from the speech signal, the envelopes of the outputs of these bands were detected, and these envelopes were used for modulating the amplitude of pure tones with frequencies equal to the center frequencies of the filter bands. The resulting signals were presented auditorily in combination with speechreading to normal-hearing subjects.

For voiced sounds the amplitude spectrum contains a number of peaks (mostly four or five), i.e. regions in which concentration of energy occurs. These peaks are called formants, and can be characterized by their frequency, their amplitude, and their bandwidth. It is well known that the frequencies of these formants cue part of the linguistic information in speech. Vowels are almost completely characterized by the frequencies of the lowest two formants (F1 and F2) (e.g., see Pols et al., 1973). Furthermore, F1 and F2 are also important for the perception of consonants. Thus, these frequencies contain important segmental information and can therefore be considered as candidates for supplementing speechreading. Chapter 3 reports the results of an experiment in which speechreading was supplemented with F1 and F2, which were presented to normal-hearing subjects either as pure tones or as a complex speech-like signal (see Section 3.2.3).

Chapter 4 describes an experiment in which speechreading was supplemented with parameters that contain suprasegmental, prosodic information. Three important ácoustic parameters related to the prosodic features quantity, tone, and stress are the duration of a sound, the fundamental frequency, and the overall sound-pressure level (Lehiste, 1970; see also Chapter 4.1). The following four supplements were used: (a) information about the duration of voiced speech segments, (b) information about the overall sound-pressure level, (c) information about the fundamental frequency, and (d) information about both the overall sound-pressure level and the fundamental frequency. Supplement (a) only gives limited temporal information; a tone with constant frequency was provided as long as the speech signal was voiced. The temporal structure of supplements (c) and (d) resembles that of supplement (a); a signal was given only when the speech signal was voiced. Supplement (c) consisted of a pulse sequence, band-pass filtered between 160 and 400 Hz, with frequency equal to the fundamental frequency of the speech signal; supplement (d) resembled (c), except that the amplitude of supplement (d) followed the amplitude of the speech signal. Supplement (b) gives information about the amplitude of both voiced and unvoiced signals (a constant-frequency tone of which the amplitude followed the amplitude of the speech signal).

Apart from giving segmental information, the parameters investigated in Chapter 2 (the sound-pressure levels in one or two frequency bands) and in Chapter 3 (the frequencies of the first and second formants) also give suprasegmental information. Both types of parameters support temporal information, and the sound-pressure levels, especially that in the 500-Hz band, will provide information about the overall sound-pressure level of the speech signal. However, both types of parameters do not give any information about the fundamental frequency of the speech signal.

In the experiments of Chapter 2, 3, and 4 normal-hearing subjects without experience in speechreading participated, and short everyday sentences were used for investigating the efficiency of the earlier-mentioned parameters. In the experiment reported in Chapter 5 the best-scoring supplements of the previous chapters were compared. In order to gain insight into the influence of experience in speechreading both inexperienced and experienced speechreaders participated (all with normal hearing). Furthermore, in addition to the perception of sentences the discrimination of phonemes (both consonants and vowels) was measured. The latter was done in order to gain insight into which speech sounds and which speech features can be perceived by the different supplements.

Finally, in Chapter 6 the question of what further research is needed to develop an efficient speechreading aid is discussed.

1.4 Review of literature on speech-perception aids

In this section a review is given of the literature on speech-perception aids for deaf persons. The review is not exhaustive, it is mainly ment to provide the reader a broad insight into what information has been used for presentation to an alternative sense. The following differentiations will be made. First, it will be mentioned whether an aid uses segmental or non-segmental processing of the speech signal. Second, differentiation will be made between aids that extract certain (slowly varying) speech parameters from the speech signal (henceforth this will be called parametric processing), and aids that do not use this technique (including aids with no processing at all). Third, three different stimulation techniques will be considered: visual stimulation, tactual stimulation, and electrical stimulation via a cochlear implant. Finally, it will be mentioned whether the aid was used in combination with speechreading or was tested only without speechreading.

Limited information will be provided about specific results with these aids. Very often they were only tested with closed-set speech material, and the discrimination scores on those types of material are not appropriate for the prediction of perception scores on open-set materials. Furthermore, differences in speech materials, methods, and subjects among the different studies make the comparison of the results difficult. Except for the cochlear implants most of the aids have only been used in a laboratory situation, and on the whole the results have been rather disappointing. Only recently some very interesting results with cochlear implants have been reported on open-set speech materials (Clark et al., 1981; Hochmair-Desoyer et al., 1983).

1.4.1 Aids with visual stimulation

The spectrogram was the first visual display that was tested as a speech-perception aid. The spectrogram displays the short-time energy spectrum as a function of time, i.e. it displays the distribution of the energy in the speech sound over frequency and time (e.g., see Cole et al., 1980). Potter et al. (1947) reported the first experiment in which the reading of spectrograms was investigated. They found that a deaf subject could acquire a vocabulary of about 800 words after over 200 hours of training. However, more recent studies have shown that real-time interpretation of spectrograms is an extremely difficult task, which probably can be learned only by a few persons after extensive training.

Most of the more recent aids with visual stimulation use segmental processing techniques and are designed to supplement speechreading. In 1968 Upton introduced his wearable eyeglass speechreading aid, a device that provides information about some hard-to-speechread phonetic features by means of an array of light-emitting diodes mounted on the frame of ordinary eyeglasses (Upton, 1968). The lights are reflected into the eye of the user in such a way that they seem to be superimposed around the lips of the speaker. One version, for example, gives information about voicing, plosion, friction, and place of articulation (for vowels). After a large amount of training the perception of connected discourse improved about 19% compared to speechreading alone (Harvard PAL S-1 sentences, 2 subjects, see Gengel, 1977).

Cornett (1967) developed Cued Speech, a method in which supplementary information about phonemes is given manually (see also Section 1.1). At the moment attempts are made to automatize the extraction of the supplementary cues, both by Cornett himself (1977) and by a group of researchers at the IBM Scientific Centre in Paris (Benedetto et al., 1982). They attempt to classify phonemes that can not be distinguished easily by speechreading alone into different groups. The groups are chosen in such a way that each phoneme can be discriminated by speechreading with information about to which group it belongs. However, the classification results are far from perfect.

Martony (1974) investigated a speechreading aid that provides information about the phonetic features voicing and plosion. A set of lights near and under the mouth of the speaker was used as visual indicators. For 11 normal-hearing subjects the discrimination of consonants improved from 28.7% for speechreading alone to 59.1% for speechreading with the aid.

An example of a visual speechreading aid that uses (non-segmental) parametric processing is an aid developed by Traunmüller (see Martony, 1974). In this aid the spectral gravity is extracted from the speech signal and is indicated on 10 light-emitting diodes arranged in a half circle under the mouth of the speaker. Consonant discrimination increased slightly from 29.7% for speechreading alone to 35% for speechreading with the aid (7 normal-hearing subjects).

1.4.2 Aids with tactual stimulation

Almost all tactile speech-perception aids use the non-segmental processing technique. An example of an aid with segmental processing is the speechreading aid used by Martony (1974). This aid gives information about the features voicing and plosion, and is actually similar to Martony's visual aid described in Section 1.4.1, but now the supplementary information is transmitted by a set of vibrators on the fingertips. Comparable results were obtained.

Apart from the differentations mentioned in the beginning of Section 1.4 tactile aids can be classified according to the number of vibrators that are used: (a) systems with a single vibrator, (b) systems with a number of single vibrators and (c) systems with a two-dimensional matrix of vibrators.

1.4.2.1 Tactile aids with one vibrator

Several investigators have used aids with one vibrator without parametric processing or with no processing at all. In 1926 Gault reported an experiment in which the unprocessed speech signal was fed to a hand-held vibrator. He demonstrated that homophenous words, i.e. words that can not be distinguished by speechreading alone, can be identified by their vibration pattern (Gault, 1926). Danhauer and Appel (1976) and Plant (1982) used similar devices. For 24 normal-hearing subjects, who were only shortly trained, Danhauer and Appel found no significant difference in consonant discrimination between speechreading-only and speechreading with the tactile aid. Plant did found improvements on different types of speech material when speechreading was supplemented with the single vibrator (4 totally deaf subjects, who received a short training). Finally, DeFillippo (1984) first lowpass filtered the speech signal before feeding it to a single vibrator on the palm of the hand. After 17 to 21 hours of training two normal-hearing subjects scored 45 wpm (words per minute) for speechreading with the aid and 38.5 wpm for speechreading-only on a connected discourse tracking task (CDT, see DeFillippo and Scott, 1978).

Single vibrators have also been used for transmitting the results of parametric processing techniques. Plant (1983) extracted the fundamental frequency (F0) and the amplitude (A) (i.e. the overall sound-pressure level) from the speech signal; F0 was encoded as vibration frequency, whereas A was encoded as vibration amplitude on a hand-held vibrator. He demonstrated that in the vibration-only condition a high level of information was transmitted about stress, word syllable number and word type (discrimination of monosyllables, trochees, spondees, and trisyllables). Rothenberg et al. (1977, 1979) also used the fundamental frequency encoded as vibration frequency of a single vibrator on the forearm. They demonstrated that a one or two-octave reduction in F0 enabled a group of five deaf subjects to discriminate between different intonation patterns (vibration-only condition). Finally, Traunmüller (1975) extracted the spectral gravity from the speech signal and encoded it as vibration frequency on one vibrator held against the cheek bone. The amplitude of the speech signal controlled the amplitude of vibration. He found a significant improvement in the recognition of (known) words when speechreading was supplemented with the aid (one normal-hearing subject).

1.4.2.2 Tactile aids with multiple vibrators

In this section tactile aids with more than one vibrator (except two-dimensional matrices of vibrators) are discussed. Most of the recent aids use parametric processing; an example of an aid with non-parametric processing is the device used by Kringlebotn (1968). In that device the speech signal is passed through a zero-crossing detector, which transforms the speech signal into a sequence of pulses, with one pulse for each positive zero crossing. This pulse sequence is then fed to a cascade of four frequency dividers (division factor 2); the outputs of the zero-crossing detector and the dividers are fed to five vibrators, one for each of the fingers of one hand. He demonstrated a nearly 100% correct recognition of (known) homophenous word pairs (by vibration only) and reported that the aid might be useful as a speechreading aid.

All the aids mentioned in the following of this section use parametric processing. DeFillippo (1984) performed experiments in which several speech parameters were used. For example, in one of these experiments speechreading was supplemented with information about the amplitudes (i.e. the sound-pressure levels) of the high-pass filtered speech signal (above 8 kHz), in the frequency band from 1.8 to 3 kHz, and in the frequency band from 250 to 900 Hz, and supplemented with information about the frequency of the first formant (F1). The amplitudes were encoded as vibration amplitude of three vibrators, whereas F1 was encoded as vibration frequency of one of the vibrators (the other two vibrators had a constant vibration frequency). Two normal-hearing subjects scored 44.5 wpm (words per minute) for speechreading-only and 62.5 wpm for speechreading with the aid (average of five hours of connected discourse tracking on both conditions).

Several researchers band-pass filtered a large number of narrow frequency bands from the speech signal and used the amplitude envelopes of the output of these bands, or the energy in these bands, for modulating the vibration amplitudes of a set of vibrators (they thus presented information about the sound-pressure levels in these bands). Oller et al. (1980), for example, divided the frequency range from 80 Hz to 10 kHz in 24 frequency bands (with coinciding -3 dB points), and used the energy in these bands to control 24 vibrators placed on the skin. They found that hard-to-speechread word pairs contrasting in manner of articulation could be identified above chance in the vibration-only condition. Pickett (1963) used the amplitudes of 10 frequency bands from 210 to 7700 Hz on 10 vibrators. He found an improvement of 8.7% and 25.1% (two different talkers) in the identification of (Swedish) words when speechreading was supplemented with the tactile aid (7 deaf subjects). Finally, Saunders et al. (1976) used a similar device with 22 frequency bands and 22 vibrators. They found an improvement of 21.5% in the identification of (known) words embedded in a fixed sentence frame (five normal-hearing subjects).

1.4.2.3 Tactile aids with a two-dimensional matrix of vibrators

All the aids mentioned in this section use parametric processing. Kirman (1974) used a tactile display consisting of 15-by-15 vibrators for presenting the frequencies of the first and second formants from the voiced parts of the speech signal. Frequency was represented by the 15 rows of the matrix, time by the 15 columns. The formant frequencies were entered into the display at the first column and shifted upwards one column every 10 ms, as new formant frequencies were entered. The display is similar to a spectrogram, in which only the frequencies of the first and second formants are represented. Six normal-hearing subjects learned to recognize two pronunciations of 15 words to an average accuracy of 83% and 70% (vibration-only condition). Some phonemes, such as /l/ and /r/, appeared to be consistently confused.

Several researchers used two-dimensional tactile matrices for displaying the amplitude or energy (thus information about the sound-pressure level) in different frequency bands of the speech signal. For example, Clements et al. (1982) used a 24-by-6 matrix for presenting the amplitudes in 18 frequency bands. The 24 rows were used to encode frequency (some frequency bands were assigned to two rows), the six columns were used to encode amplitude. They found a 77.5% correct discrimination of naturally spoken vowels (two normal-hearing subjects, tactile stimulation alone). Green et al. (1983) also used a 24-by-6 display, but filtered 24 frequency bands from the speech signal. They also found a significant vowel discrimination (six normal-hearing subjects, tactile stimulation alone). Sparks et al. (1978) developed the MESA (Multipoint Electrotactile Speechreading Aid). They used 36 frequency bands in the range from 85 Hz to 10.5 kHz, and encoded the amplitudes in these bands on a 36-by-8 tactile matrix. Three normal-hearing subjects were tested under the conditions of speechreading alone and speechreading with the aid. The test used was a connected-discourse tracking task. After extensive training, speechreading alone and speechreading with the aid scored practically equally (Sparks et al., 1979).

1.4.3 Cochlear implants

This section provides a short review of the cochlear implants that were developed by a number of leading research groups. First, implants with simple non-parametric processing of the speech signal will be discussed.

The research group at the House Ear Institute in Los Angeles implanted a single electrode into the cochlea of a large group of totally deaf subjects. They use a very simple speech-processing technique; the speech signal is low-pass filtered at 4000 Hz and transmitted to the electrode by amplitude-modulation on a 16-kHz carrier (Berliner and House, 1982; House, 1982; Danley and Fretz, 1982). Audiological test results of 135 implant users showed that word discrimination, word stress discrimination, and the discrimination of environmental sounds is better with the implant than with a hearing aid. However, without speechreading no significant speech perception was obtained (Thielemeir et al., 1982).

At the University of Utah (Salt Lake City) an implant was developed that presents the outputs of four band-pass filters, in the frequency region of the first, second, and third formants, and in the frequency region above the third formant, to four intra-cochlear electrodes. A high level of vowel discrimination and a significant level of open-set word recognition were reported (one subject) (Eddington, 1980; see also Millar et al., 1984). A similar approach was followed by a group at the University of California at San Francisco; however, without speechreading they obtained poor results on open-set speech material (see Millar et al., 1984).

Hochmair-Desoyer et al. (1983) developed a cochlear implant with four intra-cochlear electrodes of which the best functioning is used for stimulation. They use gain compression followed by frequency equalization in order to map the speech signal from 200 to 4000 Hz onto an equal-loudness contour at a comfortable loudness level. Significant scores on an open-set sentence perception task were reported (without speechreading). The individual scores appeared to vary largely: some subjects reached nearly 100% correct responses, while others could not use the implant at all (12 subjects).

The cochlear implants that are discussed in the remainder of this section all use parametric processing. At London a group of researchers developed a cochlear implant with one extra-cochlear electrode. Their strategy was to design an implant that provides information that can not be obtained by speechreading alone. Therefore, they used the fundamental frequency of the speech signal and encoded it as frequency of a pulse train on the electrode (Fourcin and Rosen, 1979; Moore et al., 1984). For one patient, for example, the perception of connected discourse improved from 22 wpm (words per minute) for speechreading-only to 35.7 wpm for speechreading with the implant (connected-discourse tracking task).

A research group at Paris designed an implant with 8 to 12 electrodes. The energy levels in 8 to 12 frequency bands from the speech signal were encoded as rate of pulsation on the electrodes. They reported that the implant gave a significant increase in speech intelligibility compared to speechreading alone (Pialoux et al., 1979; see also Millar et al., 1984).

At Stanford, California, an implant was developed with three electrodes placed into the cochlear-nerve fibers. The frequencies of the first and second formants were encoded as the frequencies of pulse trains on two electrodes, whereas a high-rate pulse train stimulated the third electrode for unvoiced sounds. Without speechreading no recognition of open-set speech material was obtained (see Millar et al., 1984).

Finally, at Melbourne a cochlear implant with 22 intra-cochlear electrodes was developed. The frequency of the second formant determined which of the 22 electrodes was stimulated. The fundamental frequency was encoded as the frequency of a pulse train on the chosen electrode and the amplitude of the speech signal was encoded as the amplitude of that pulse train. Significant open-set sentence perception scores (76% on the CID sentence test) were reported for speechreading with the implant (one subject) (Clark et al., 1979, 1981; Tong et al., 1981).

1.4.4 Conclusion

From the previous sections it can be concluded that several types of speech parameters have been used for presentation to an alternative sense.

The first type consists of the sound-pressure levels in a number of frequency bands filtered from the speech signal; from 3 (DeFillippo, 1984) to 36 (Sparks et al., 1978, 1979) frequency bands were used. These sound-pressure levels were presented visually (the spectrogram), tactually or via a cochlear implant. In the study reported in this dissertation this type of information was also investigated; however, only one or two frequency bands were used (Chapter 2).

The second type consists of the frequencies of the first and/or the second formants. This type of information was presented both tactually (Kirman, 1974) or via a cochlear implant (see Millar et al., 1984). The combination of these frequencies was also investigated in the present study (Chapter 3).

The third type consists of parameters that provide prosodic information: the fundamental frequency (Rothenberg, 1977, 1979; Fourcin and Rosen, 1979) or the combination of the fundamental frequency and the overall sound-pressure level (Plant, 1983). These two parameters were also investigated in the present study. Additionally, two other prosodic parameters were investigated: the overall sound-pressure level alone, and information about the duration of sequences of voiced sounds (Chapter 4).

The spectral gravity, that was used by Traunmüller (1975; see Section 1.4.2.1), was not investigated in the present study.

The results obtained with the speech-perception aids described in the previous sections are far from optimal. Many researchers reported improvements in the identification of closed-set speech material; however, except for some cochlear implant patients, only a limited speech intelligibility on open-set speech material has been reported.

1.5 Design of this dissertation

The remainder of this dissertation consists of five chapters. Chapters 2 and 3 are based on articles that have been published in the Journal of the Acoustical Society of America, with R. Plomp as coauthor (Breeuwer and Plomp, 1984, 1985). Chapters 4, 5, and 6 have been submitted to this journal for publication as one article. Since these articles are presented as chapters in this dissertation some overlap will exist among them.

CHAPTER 2 - EXPERIMENT I: SPEECHREADING SUPPLEMENTED WITH FREQUENCY-SELECTIVE SOUND-PRESSURE INFORMATION

ABSTRACT

The benefit of supplementing speechreading with frequency-selective sound-pressure information was studied by auditorily presenting this information to normal-hearing listeners. The sound-pressure levels in one or two frequency bands of the speech signal with center frequencies of 500, 1600, and 3160 Hz, respectively, and with one or one-third octave bandwidth were used to amplitude-modulate pure-tone carriers with frequencies equal to the center frequencies of the filter bands. Short sentences were presented to eighteen normal-hearing listeners under the conditions of speechreading-only and speechreading combined with the sound-pressure information. The mean number of correctly perceived syllables increased from 22.8% for speechreading-only to 65.7% when sound-pressure information was supplied in a single one-octave band at 500 Hz, and to 86.7% with two one-octave bands at 500 and 3160 Hz, respectively. The latter signal scored only 26.7% correct syllables without accompanying visual information.

2.1 Introduction

For the profoundly deaf person, the most important avenue to perceiving spoken language is through speechreading. Speechreading, however, is a poor substitute for auditory speech perception, since the distinguishable lip and jaw movements are often ambiguous. Different vowels and consonants can produce similar visual patterns, making them visually indiscriminable (Voiers, 1973; Lowell, 1971).

Up to the present time, considerable effort has gone into the design of speech-perception aids using other senses than hearing. Some of the researchers in this area try to achieve speech transmission without visual cues, others aim explicitly to supplement speechreading (for a review see Martony, 1974). Upton (see Gengel, 1976), for example, used visual indicators to supplement speechreading with a set of distinct features, consisting of voicing, friction, plosion, and front versus back. Other investigators have used tactile (for a review see Kirman, 1973) or electrical stimulation (cochlear implants; White, 1982). For successful transmission of speech information through an alternative sense two problems have to be solved. First, since the speech signal itself is too complex for direct presentation to the alternative sense, processing of this signal is necessary in order to extract the most adequate information for transmitting. Second, psychophysical study should clarify how the alternative sense can be stimulated most optimally. Very often, these two problems are treated as being one and the same. However, the confounding of these two issues poses the difficulty that if disappointing results arise it is not possible to pinpoint their causes. They may lie in information extraction or in transduction, or in both. Therefore, we have decided to focus our attention only on the first problem, namely on investigating which parameters of the speech signal contain the most information to successfully supplement speechreading.

Parameters that are considered to be candidates are evaluated experimentally, by presenting them auditorily to normal-hearing listeners in combination with the visible articulatory movements of the speaker. The difference between speech intelligibility under the conditions of speechreading-only and speechreading combined with the auditory presentation is a measure of the importance of the parameter concerned. Thus in our scheme the alternative sense of the profoundly deaf person is simulated by the auditory modality of normal-hearing listeners. This means that we use the optimal way of presenting the supplementary speech information, postponing the problem of how to stimulate an alternative sense most adequately. Additionally, this choice has the advantage that the role of training is reduced to a minimum. Once the most effective parameters are known an attempt can be made to present them through other senses than hearing. In the transmission of these parameters through an alternative sense information will be lost. Thus the parameters needed to reach an acceptable level of speech intelligibility when presented auditorily together with speechreading indicate the minimum information required to perceive speech through speechreading with stimulation of an alternative sense.

In the present study the speech parameters tested are the sound-pressure levels in one or two frequency bands filtered from the speech signal. These sound-pressure levels, fluctuating continuously over

time, are used to modulate the amplitude of pure-tone carriers presented to the listener.

2.2 Methods

2.2.1 Speech material

The speech material consisted of twenty lists of thirteen short, meaningful Dutch sentences (eight or nine syllables) typical of everyday conversation (Plomp and Mimpen, 1979), 260 different sentences in all. Eighteen of these lists were used for the actual experiment, the remaining two for pre-experimental training. A female speaker was recorded in color on video tape while pronouncing these lists. She was a speech therapist with clear pronunciation from a speechreading point of view. The lighting gave a shadow-free illumination of the face. Only the head and part of the shoulders were visible.

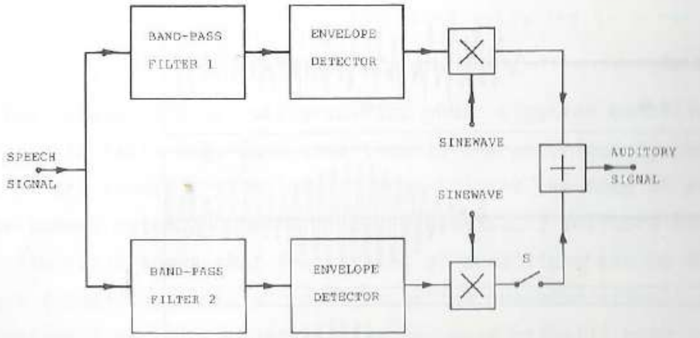

Fig. 2.1. Apparatus used for the derivation of the auditory stimuli.

2.2.2 Experimental conditions

The auditory stimuli used in the present study were derived as follows (Fig. 2.1). Speech was passed through one or two band-pass filters, and the outputs were full-wave rectified and smoothed by a low-pass filter (cutoff frequency 20 Hz). The resulting signals represent the amplitude envelope of the outputs of the band-pass filters. These envelopes were used to amplitude-modulate pure-tone carriers. In the case of two bands the resulting signals were added. The frequency bands used had center frequencies of 500, 1600, and 3160 Hz, respectively; their bandwidth was either one or one-third octave (Brüel & Kjaer Spectrum Shaper Type 5612). The 500-Hz band was chosen because it covers the middle part of the frequency region of the first formant. The 1600-Hz band covers the middle part of the second-formant region, and the 3160-Hz band lies in the region of the third and fourth formants. Figure 2.2 gives an example of two of these auditory stimuli derived from the speech signal in Fig. 2.2-a: Fig. 2.2-b represents a 500-Hz tone amplitude-modulated by the envelope of the output of the one-octave 500-Hz band, and Fig. 2.2-c represents the sum of a 500-Hz and a 3160-Hz tone amplitude-modulated by the envelopes of the outputs of the one-octave bands round 500 and 3160 Hz, respectively.

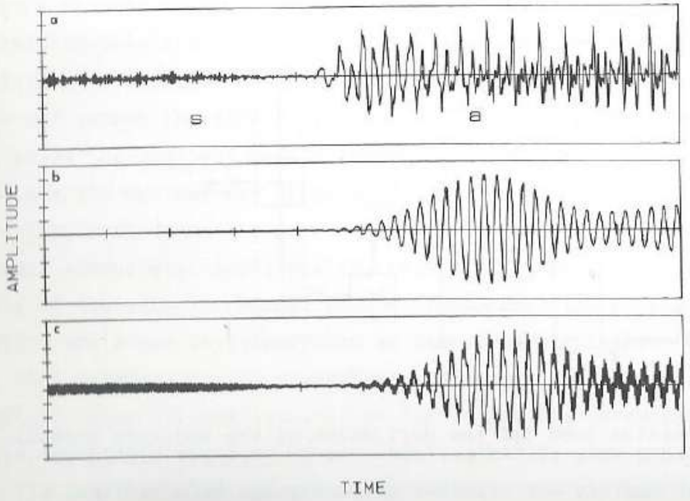

Fig. 2.2. Example of acoustic stimuli; (a) unfiltered speech signal of the utterance /sa/, (b) 500-Hz tone amplitude-modulated by the envelope of a frequency band with center frequency of 500 Hz and 1-oct bandwidth, (c) the sum of a 500- and 3160-Hz tone amplitude-modulated by the envelope of octave bands at 500 and 3160 Hz, respectively. Time scale: 10 ms per division, arbitrary amplitude scale.

condition	acoustic stimulus			mode of presentation	mean	standard deviation
	band 1 (Hz)	band 2 (Hz)	width (octaves)			
1	500	-	one-third	audiovisual	55.0	20.8
2	1600	-	one-third	audiovisual	44.1	18.9
3	3160	-	one-third	audiovisual	43.9	27.3
4	500	-	one	audiovisual	65.7	18.4
5	1600	-	one	audiovisual	46.3	21.0
6	3160	-	one	audiovisual	41.4	24.1
7	500	1600	one-third	audiovisual	72.5	20.7
8	500	3160	one-third	audiovisual	80.8	12.8
9	1600	3160	one-third	audiovisual	64.1	20.4
10	500	1600	one	audiovisual	81.7	17.7
11	500	3160	one	audiovisual	86.7	11.9
12	1600	3160	one	audiovisual	66.4	20.7
13	1600	500	one	audiovisual	43.7	22.3
14	500	1600	one	auditory only	22.9	12.7
15	500	3160	one	auditory only	26.7	16.3
16	1600	3160	one	auditory only	9.1	5.2
17	no acoustic stimulus			visual only	22.3	18.2
18	no acoustic stimulus			visual only	23.3	17.7

Table 2.1. The 18 conditions under which the speech material was presented. For each condition the mean and the standard deviation of the percentage correctly perceived syllables is given.

The speech material was presented under eighteen conditions, tabulated in Table 2.1. Each condition is characterized by the frequency bands of the acoustic stimulus (if present) and the mode of presentation of the speech material, being either audiovisual, auditory only or visual only. The table shows that four groups of conditions can be distinguished. Group 1 (conditions 1 to 6) consists of the one-band stimuli, group 2 (conditions 7 to 13) consists of the two-band stimuli, both groups presented audiovisually. In all cases except condition 13 the carrier frequencies are equal to the center frequencies of the filter bands. In condition 13 the envelope of the 500-Hz band is used to amplitude-modulate a 1600-Hz tone and vice versa. Comparison between the speech-intelligibility scores for the conditions 10 and 13, then, may reveal the effect of a less optimal adaption between the information presented and the process of perception.

The acoustic stimuli of the conditions in group 3 (14 to 16) equal those of conditions 10 to 12, but now without visual information. A pilot experiment showed that the one-band stimuli did not lead to significant intelligibility scores when presented auditorily only; therefore, these conditions were not included.

Finally, group 4 (conditions 17 and 18) consists of two cases in which no acoustic stimulus was presented. In these conditions the subjects had to rely upon speechreading alone.

2.2.3 Subjects

Eighteen normal-hearing subjects, nine male and nine female, 17 to 29 years of age, participated in the experiment. All subjects except one had pure-tone hearing levels less than 17 dB for their better ear, measured at frequencies of 125, 250, 500, 1000, 2000, 4000 and 8000 Hz, and had normal or normally corrected vision (Landolt C-test, visual acuity of 20/20 or better). One subject had a hearing level of 22 dB at 8000 Hz. None of the subjects had ever participated earlier in a speechreading experiment. They were paid for their participation.

2.2.4 Procedure

The subject was seated in a soundproof room at a distance of approximately 2 m from the 50 cm color video monitor. All acoustic stimuli were presented binaurally (Beyer DT-48 headphones) at a comfortable loudness level. The subject received one list (13 sentences) per condition. The order of presentation of the lists was fixed. In order to eliminate the influence of learning as much as possible, the conditions were presented according to a counterbalanced design.

The presentation of the signals was controlled by a digital computer. Each sentence was preceded by a short audible warning signal and was followed by a 9-s pause during which the subject had to reproduce verbally as much of the sentence as possible (syllables, words, part of or the whole sentence). The correct syllables were notated by the experimenter, seated in the same room, on prepared response sheets. Additionally, the responses were recorded on tape, so that in case of uncertainty the responses could be replayed afterwards.

Prior to the experiment the subjects were briefly trained with different sentences in all audiovisual conditions with one octave bandwidth. Total length of the training was approximately 20 min; total length of the experiment was two and a half hours.

2.3 Results and discussion

For each subject the percentage of correctly reproduced syllables per condition was counted. This measure was chosen because sometimes subjects scored only parts of words correctly. The percentage correct syllables and the percentage correct words, however, were found to follow very similar patterns. Mostly, the latter was only slightly lower (2-4%). The last two columns of Table 2.1 give the mean and the standard deviation of these percentage syllable scores for the group of eighteen subjects. Figures 2.3 and 2.4 provide a graphic comparison of the mean values for the one-band and two-band audiovisual conditions. In both figures the mean score for speechreading-only is plotted as a black bar. The white bars indicate the conditions with one-third octave bandwidth, the shaded bars those with one octave bandwidth. The figures indicate that supplementing speechreading with auditorily presented sound-pressure information leads to a substantial increase in speech intelligibility in all cases. When this information is presented auditorily only, the highest score is only 26.7% (condition 15). The mean speechreading-only score is 22.8% ; it should be mentioned, however, that three of the eighteen subjects had a speechreading-only score above 45% (46.0% , 54.9% , and 56.1%, respectively). These subjects seem to have a natural ability for speechreading.

For an individual comparison between two conditions the intelligibility scores of each subject on these conditions were considered as paired observations. A t-test on differences between paired observations was used to investigate if for the two conditions these differences were significant. This method is actually similar to a two-way analysis of variance on a subset of two conditions. The 500-Hz band is a more effective supplement to speechreading than the 1600-Hz and 3160-Hz bands (t(17)=5.5, and t(17)=6.6 , p<0.01). The bandwidth is a significant parameter only for the 500-Hz band (t(17)=2.4, p<0.026). These results may be explained as follows.

First, the 500-Hz band probably gives the best information about voicing; for unvoiced sounds the energy in this band is very low (cf. Fig. 2.2-b). On the contrary, the 1600-Hz and 3160-Hz bands respond to both

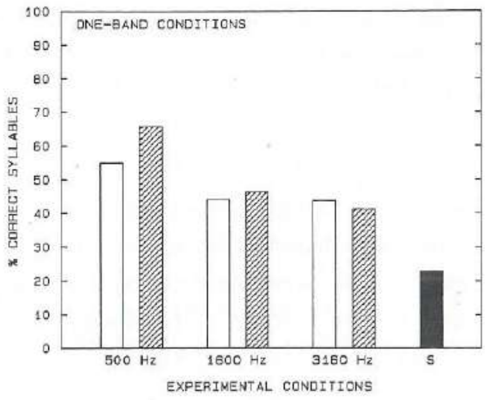

Fig. 2.3. Mean percentage correctly reproduced syllables for the audiovisual one-band conditions. The white bars indicate the conditions with 1/3-oct bandwidth, the shaded bars the conditions with 1-oct bandwidth. The mean speechreading-only score is indicated as a black bar.

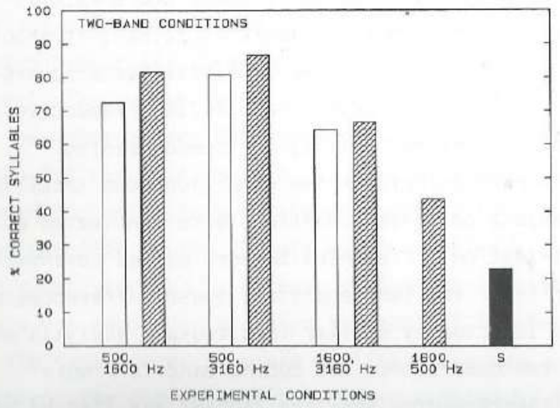

Fig. 2.4. Mean percentage correctly reproduced syllables for the audiovisual two-band conditions. The white bars indicate the conditions with 1/3-oct bandwidth, the shaded bars the conditions with 1-oct bandwidth. The mean speechreading-only score is indicated as a black bar.

voiced and unvoiced sounds and thus present more ambiguous information to the subjects. For the speechreader, voicing is of importance, since through speechreading alone no information about voicing can be perceived (Berger, 1972).

Second, the 500-Hz band with one octave bandwidth probably presents more information about the prosody of the speech signal than the other bands. This is important, since through speechreading alone hardly any information can be obtained about prosody (Risberg and Lubker, 1978). Prosodic features are tone, stress, and quantity (see Section 4.1). The perceptual correlates of these features are called pitch (intonation), stress, and duration. The perception of pitch depends on the fundamental frequency of the speech signal. Through the acoustic signals of the present study no perception of pitch is possible. But stress, which depends on fundamental frequency, on duration, and on sound pressure, and duration will be perceived at least partly. For a number of sentences the correlations between the sound-pressure levels in the 500-Hz, 1600-Hz, and 3160-Hz bands and the overall sound-pressure level were determined. These sound-pressure levels were sampled at 100 Hz, and correlations over time were computed within each sentence. The mean correlations are tabulated in Table 2.2. As can be seen, the 500-Hz band with one octave bandwidth reaches a higher correlation than the other bands, indicating that this band gives more precise information about the overall sound-pressure level. The 500-Hz band, therefore, probably provides more precise information about the stress pattern in a sentence. A close agreement can be seen between these correlations and the perceptual results (Table 2.2, Fig. 2.3).

frequency band (Hz)	bandwidth (octaves)	
	one	one-third
500	0.84	0.66
1600	0.67	0.65
3160	0.66	0.66

Table 2.2. Correlations between the sound-pressure-levels in the 500-, 1600-, and 3160-Hz bands and the overall sound-pressure level.

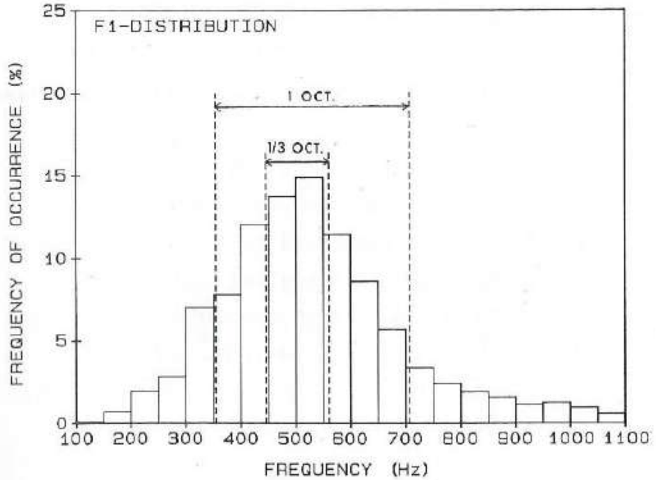

Fig. 2.5. Distribution of the frequency of the first formant together with the cutoff frequencies of the 500-Hz bands.

Both show that only for the 500-Hz band the bandwidth is of significance and that the results for the 1600-Hz and 3160-Hz band do not differ significantly. However, the correlation for the one-third octave 500-Hz band is not higher than that of the 1600-Hz and 3160-Hz bands, but the percentage correctly perceived syllables is higher. The higher score may be caused by the more precise information about voicing in the 500-Hz band.

Figure 2.5 gives the distribution of the frequency of the first formant together with the cutoff frequencies of the 500-Hz bands. These formant frequencies were determined using the method of Linear Predictive Coding (Atal and Hanauer, 1971). For 130 sentences F1 was determined every 10 ms, using overlapping speech segments of 20 ms. The one-octave band covers most of the frequency region of the first formant. The sound-pressure level in this band will mainly depend on the energy of the first formant. For most voiced sounds the energy of the first formant is closely related to the overall sound-pressure level. This explains the high correlation for the broad 500-Hz band. The one-third octave band covers only a small part of the first formant region. The sound-pressure level in this band depends on both the frequency and the energy of the formant. A

high-energy first formant situated outside the one-third octave band can result in the same sound-pressure level as a low-energy formant situated inside this band. This may explain the decrease in correlation when the bandwidth is lowered to one-third octave.

Analysis of variance showed that the audiovisual two-band conditions (conditions 7 to 12) scored significantly higher than the audiovisual one-band conditions (F(1,17)=105, p<0.00001). The combination of the 500-Hz and 3160-Hz bands (one octave bandwidth) appears to be the most effective supplement to speechreading (86.7% correct syllables); however, this condition differs only slightly from the combination of the 500-Hz and 1600-Hz bands (condition 10) (t-test on paired observations, t(17)=1.7, p<0.10) For the combination of the 1600-Hz and 3160-Hz bands the bandwidth is of no significance. For the combinations of the 500-Hz and 1600-Hz bands and the 500-Hz and 3160-Hz bands it is significant (t(17)=2.8, p<0.01, and t(17)=1.9, p<0.075). Thus only for those two-band conditions in which the 500-Hz band is included the bandwidth is of importance. This is consistent with the results for the one-band conditions (Fig. 2.3).

Supplementing speechreading with both the sound-pressure levels in the 500-Hz and 3160-Hz bands probably provides the subjects with more exact information about voicing and prosody than the 500-Hz band alone. Unvoiced sounds are conveyed more clearly by the presence of the 3160-Hz tone. Thus the durations of different speech sounds can be perceived more precisely. In Fig. 2.2-c the difference between the unvoiced /s/ and the voiced /a/ is very obvious. The 1600-Hz band will respond both to voiced and unvoiced sounds, and thus provides the subjects with more ambiguous information about voicing. This may explain the somewhat lower score for the combination of the 500-Hz and 1600-Hz bands.

Comparison between condition 10 and 13 (Table 2.1, Fig. 2.4) indicates that a reversal of the carrier frequencies led to a substantial reduction in intelligibility (from 81.7% to 43.7%). In the perception of the acoustic stimuli two processes play a role. The first process, taking place in the ear, is the reception of the acoustic stimuli and the coding into neural activities. The second process, located more centrally, performs the processing and the interpretation of the received information. We may assume that peripheral hearing at 500 Hz is not specifically adapted to the temporal fluctuations of the sound-pressure level at that frequency, so central processes should account for the reduction in intelligibility when the carrier frequencies are reversed. Extra training on condition 13 (reversed carriers) may overcome this difficulty. Blesser (1972), for example, demonstrated that spectrally rotated speech can be understood after a sufficient amount of training.

A similar problem arises when supplementary information is presented through other senses than hearing. For example, several investigators have evaluated speech perception by tactually presented sound-pressure information. Oller et al. (1980) divided the frequency range from 80 to 10,000 Hz into 24 overlapping frequency bands. The sound pressure in these bands was used to control 24 vibrators placed on the skin. They reported that hard-to-speechread word pairs contrasting in manner of articulation were identified significantly above chance level. Sparks et al. (1978) developed the MESA (Multipoint Electrotactile Speechreading Aid). They used the sound-pressure levels from 36 frequency bands in the range from 85 to 10,500 Hz to control a matrix of 36-by-8 vibrators placed on the abdomen. The 36 rows were used to encode frequency, the 8 columns to encode sound-pressure level. They reported that the MESA "enables receivers to achieve excellent recognition of vowels in CVC context and the consonantal features voicing and nasality" (Sparks et al., 1978, page 246). For the perception of connected discourse, however, speechreading alone and speechreading supplemented with tactile information scored practically equally (Sparks, 1979). The present study shows that acoustically presented sound-pressure information from only two appropriately chosen frequency bands can effectively supplement speechreading, even in the case of the perception of suprasegmental speech units such as sentences. Thus, the rather disappointing results with tactually displayed sound-pressure information can not be due to a shortcoming of the presented information. From a study of Lamoré et al. (1980) it is known that the skin is able to receive low-frequency modulations up to about 12 Hz. Most of the frequency components of the sound-pressure levels lie below 12 Hz, thus the skin must be able to receive these modulations at least partly. Possibly, much more training is required to get familiar with tactually presented sound-pressure information. It can be questioned, however, whether appropriate central processing of this information can take place.

Several investigators have evaluated the benefit of supplementing speechreading with speech parameters other than those of the present study. Erber (1972) supplemented speechreading with the sound-pressure level of the speech, low-pass filtered at 2 kHz. This sound-pressure level was used to modulate the amplitude of a one-octave noise band centered at 500 Hz. He found only a 6 to 8% gain over speechreading-only in the perception of known words. Risberg and Lubker (1978) supplemented speechreading with information about the sound-pressure level of the speech signal in the frequency region from 100 to 2700 Hz, or with both this sound-pressure information and information about the fundamental frequency. Normal-hearing subjects were asked to answer unknown questions under the conditions of speechreading alone (37.9% correct), speechreading plus sound-pressure information (42.4% correct), and speechreading plus sound-pressure and fundamental-frequency information (78.5% correct). Rosen et al. (1979, 1981) found a substantial gain in intelligibility of connected discourse when speechreading was aided by information only about the fundamental frequency.

We performed an experiment with auditory signals comparable to those of Rosen et al. (1979, 1981) and Risberg and Lubker (1978) and found the speech-intelligibility scores not to be as high as in the present study (see Chapter 4). For example, when speechreading was supplemented with information about both the fundamental frequency and the overall sound-pressure level, ten normal hearing subjects reached a mean score of 64.0% correct syllables. The speech material used in that pilot experiment was equal to that of the present study, although the experimental procedure differed somewhat; the subjects received no pre-experimental training, and each condition was presented twice instead. For some subjects, the lack of training may be partly responsible for the lower scores. The fact that five of the ten subjects reached a mean score of 83.6% on this information supports this view. In a subsequent study we plan to compare the parameters used by Risberg and Lubker, and by Rosen et al. with the best-scoring parameters of the present study.

2.4 Conclusions

From this study it can be concluded that:

1) Auditorily presented information about the sound-pressure levels in two appropriately chosen frequency bands appears to be a very effective supplement to speechreading of short sentences.

2) Of the frequency bands considered, the combination of two one-octave bands with center frequencies of 500 and 3160 Hz is the most effective supplement (86.7%). When only one band is used, the one-octave band at 500 Hz proves to be most effective (65.7%). Speechreading-only scored 22.8%.

CHAPTER 3 - EXPERIMENT II: SPEECHREADING SUPPLEMENTED WITH FORMANT-FREQUENCY INFORMATION FROM VOICED SPEECH

ABSTRACT

The benefit of supplementing speechreading with information about the frequencies of the first and second formants from the voiced sections of the speech signal was studied by presenting short sentences to eighteen normal-hearing listeners under the following three conditions: (a) speechreading combined with listening to the formant-frequency information, (b) speechreading-only, and (c) formant-frequency information only. The formant frequencies were presented either as pure tones or as a complex speech-like signal, obtained by filtering a periodic pulse sequence of 250 Hz by a cascade of four second-order band-pass filters (with constant bandwidth); the center frequencies of two of these filters followed the frequencies of the first and second formants, whereas the frequencies of the others remained constant. The percentage of correctly identified syllables increased from 22.8 in the case of speechreading-only to 82.0 in the case of speechreading with listening to the complex speech-like signal. Listening to the formant information only scored 33.2% correct. In the present experiment the best-scoring supplement of the previous experiment (Chapter 2), namely information about the sound-pressure levels in two one-octave filter bands at 500 and 3160 Hz, was also used as supplement. This sound-pressure information appeared to be a more effective supplement to speechreading (90.9% correct syllables) than the formant-frequency information.

3.1 Introduction

For the perception of speech the profoundly deaf rely mainly on speechreading. Unfortunately the information available through speechreading is very limited, and it is desirable to supplement speechreading with additional information about the speech signal. Our research is aimed at finding the best possible supplementary information which is extractable directly from the acoustic speech signal. We focus on isolating those parameters from the speech signal that carry enough information to successfully supplement speechreading. These parameters are presented auditorily to normal-hearing listeners by modulating an acoustic carrier. Their speech intelligibility scores under the condition of speechreading combined with this auditory information is measured and compared with their scores under the conditions of speechreading-only or listening to the auditory information only. Once the most successful parameters are known, an attempt can be made to present these parameters through other senses (tactual or visual), or via a cochlear implant to the profoundly deaf.

In a previous investigation (Experiment I, Chapter 2) we evaluated the possibility of aiding speechreading with information about the sound-pressure levels in one or two frequency bands with center frequencies of 500, 1600, or 3160 Hz, respectively, and with one or one-third octave bandwidth. Normal-hearing listeners with no experience in speechreading scored 86.7% correct syllables (in short sentences) when speechreading was supplemented with information about the sound-pressure levels in two one-octave bands with center frequencies of 500 and 3160 Hz. Speechreading-only scored 22.8%. In the present study we evaluate the benefit of aiding speechreading with information only about the frequencies of the first and second formants from the voiced parts of the speech signal. Eighteen normal-hearing listeners were requested to reproduce unknown sentences through speechreading combined with this formant-frequency information. Thus, no information was given about formant bandwidth or amplitude.

Many results indicated that formant frequencies convey part of the linguistic information in speech (e.g., Cooper et al., 1952; Blumstein et al., 1982; Remez et al., 1981), particularly information about manner and place of articulation. In contrast, through speechreading hardly any information about manner and only limited information about place of articulation can be obtained. This has led to our hypothesis that formant-frequency patterns might be a useful supplement to speechreading.

3.2 Methods

3.2.1 Formant extraction

3.2.1.1 The principle

The frequencies of the first and second formants were extracted using the method of Linear Predictive Coding (LPC) (Atal and Hanauer, 1971; Markel and Gray, 1976). In this method the n-th sample of the speech signal is considered to be composed of a weighted sum of the past M samples and some input u_n:

$$s_n = \sum_{k=1}^{M} a_k \cdot s_{n-k} + u_n = \hat{s}_n + u_n$$

where

s_n = the n-th sample of the speech signal

$\hat{s}_n$ = a prediction of the n-th sample of the speech signal by the weighted sum of the past M samples.

a_k = the k-th predictor coefficient

u_n = the n-th sample of the input signal

Thus the speech signal s_n is considered as the output of a filter with input u_n and with transfer function:

$$H(z) = \frac{S(z)}{U(z)} = \frac{1}{1 - \sum_{k=1}^{M} a_k \cdot z^{-k}}$$

where

$S(z)$ = the z transform of the sampled speech signal s_n
$U(z)$ = the z transform of the sampled input signal u_n

The difference $u_n = s_n - \hat{s}_n$ can be considered as a prediction error, the predictor coefficients a_k $(k=1,2,\ldots,M)$ can be found by minimizing the mean-squared prediction error over all speech samples in the segment to be analysed.

The transform $H(z)$ can be written as the product of $M/2$ (assuming M is even) second-order filter sections:

$$H(z) = \frac{1}{\prod_{k=1}^{M/2} (\,1+p_k \cdot z^{-1}+q_k \cdot z^{-2})}$$

Once the predictor coefficients a_k $(k=1,2,\ldots,M)$ are known the filter coefficients p_k and q_k $(k=1,2,\ldots,M/2)$ can be calculated (see Vogten, 1983). The k-th filter section can either have two complex conjugate poles or two real poles. In the case that two complex conjugate poles exist this section can be associated with a resonance peak in the frequency spectrum of the speech signal, with frequency F_k and bandwidth B_k:

$$F_k = \frac{1}{2\pi T} \arccos\left(- \frac{p_k}{\sqrt{2q_k}}\right)$$

$$B_k = - \frac{1}{\pi T} \ln(\sqrt{q_k})$$

where T is the sampling interval (in seconds).

3.2.1.2 The algorithm

The calculation of the predictor coefficients and the associated resonance frequencies is a laborious task which could not be done in real time with our equipment. Therefore the following method was developed. The videotaped speech material was sampled at 10 kHz by using a trigger pulse preceding each sentence as reference, and was stored on magnetic disk. The frequencies of the first and the second formants (F1 and F2) were determined with a digital computer every 10 ms, using overlapping speech segments of 20 ms. For each 20-ms segment the mean was extracted, a frequency preemphasis of 6 dB/oct was applied in order to account for lip radiation characteristics (+6 dB/oct) and glottal waveform (-12 dB/oct) (i.e. to equalize the spectrum), and a Hamming window was used to reduce spectral distortion. Then the predictor coefficients a_k $(k=1,2,\ldots,8)$ (i.e. M was chosen equal to eight), the filter coefficients p_k and q_k $(k=1,\ldots,4)$, and the resonance frequencies and bandwidths F_k and B_k $(k=1,\ldots,4)$ were calculated.

The resulting frequencies F_k $(k=1,\ldots,4)$ can be considered as estimates of the formant frequencies. However, estimation errors occured: sometimes formant frequencies were missing (i.e. one of the filter sections had real poles) or artificial frequencies were detected. The patterns of F1 and F2 were determined from F_k by first choosing F1 and F2 to obey $F1<1100$ Hz and $800<F2<2700$ Hz. If more than two resonance frequencies were situated in one of these regions, the frequency with the smallest bandwidth was chosen, if no resonance frequency was present in one of the regions the formant frequency was temporarily set to zero. In those cases that no more than two adjacent formant frequencies (20 ms) were missing or grossly out of line (frequency jumps of more than 300 Hz) the formant frequency was calculated by interpolation of the preceding and following values. If more than two values were missing the absent formant frequency was assigned a value of zero. After these corrections it sometimes occurred that F1 was detected as F2 or vice versa; in those rare cases, and in the case that more than two adjacent formant frequencies were missing, corrections were made by hand with the help of a digital computer and a graphic display.

3.2.2 Speech material

The speech material consisted of lists of short unknown, meaningful, Dutch sentences (eight or nine syllables) typical of everyday conversation (Plomp and Mimpen, 1979). Six of these lists were used for the actual experiment, three for pre-experimental training, 117 different sentences in all. The speaker was a female speech therapist with a clear pronunciation from a speechreading point of view (clearly visible lip and jaw movements, and for some sounds such as /l/ or /r/ also visible tongue movements). Only her head and part of her shoulders were recorded in color on videotape, under shadow-free illumination.

3.2.3 Experimental conditions

The speech material was presented under six conditions, tabulated in Table 3.1. Each condition is characterized by the nature of the acoustic stimulus (if present) and the mode of presentation of the speech material (audiovisual, auditory only, or visual only).

In condition 1 the acoustic stimulus consisted of the sum of two sinewaves with constant amplitudes, and with frequencies varying according to F1 and F2. This signal was made by updating two computer-controlled generators (Rockland model 5100) every 10 ms with the values of F1 and F2. The pure tones had equal amplitudes. No signal was generated for unvoiced sounds. We judged this signal to have a rather nonspeech (i.e. tonal) character. Since we thought it likely for this stimulus to require a long period of adjustment we have searched for another way of presenting F1 and F2 auditorily. In order to get a more speech-like signal the following method was used (condition 2). Whenever the speech signal was voiced, a periodic pulse sequence with frequency of 250 Hz was generated, which was filtered by a cascade of four second-order band-pass filters. Thus, no signal was generated for unvoiced sounds. The bandwidths of these filters were constant, namely B1=50 Hz, B2=100 Hz, and B3=B4=200 Hz. The center frequencies of the first two filters followed F1 and F2, the center frequencies of the third and fourth filters were fixed at 3500 and 4500 Hz, respectively. In this way a speech-like signal was synthesized with four formants (of which two were constant) and with a constant fundamental frequency of 250 Hz. The signal sounded somewhat like a very monotonous female speaker. This scheme is, in fact, similar to a cascade formant synthesizer (e.g., Klatt, 1980) in which only the frequencies of the first and second formants are varied and no information about unvoiced speech sounds is given.

condition	acoustic stimulus	mode of presentation	% correctly reproduced syllables	
			mean	standard deviation
1	F1,F2 as sinewaves	audiovisual	78.3	14.5
2	F1,F2 as complex signal	audiovisual	78.7	12.4
3	F1,F2 as complex signal + noise	audiovisual	82.0	9.7
4	sound-pressure information	audiovisual	90.9	8.0
5	F1,F2 as complex signal	auditory only	33.2	10.8
6	no acoustic stimulus	visual only	22.8	13.8

Table 3.1. Experimental conditions under which the speech material was presented, together with the mean and standard deviation of the percentage of correctly reproduced syllables.

In condition 3 the acoustic stimulus resembled that of condition 2 except that background noise with a speech-like spectrum was added. The level of the complex formant signal in this noise was 25 dB above threshold. This background noise was added in order to create a signal in which the absence of information about unvoiced sounds was masked. This signal sounded more continuous than the signal in condition 2.

The acoustic stimulus of condition 4 was equal to that of the best-scoring condition of our previous experiment (Experiment I, Chapter 2); i.e., it provided information about the sound-pressure levels in two one-octave filter bands with center frequencies of 500 and 3160 Hz. This signal was derived by filtering the speech signal with two one-octave filter bands at 500 and 3160 Hz and using the envelope of the output of these bands to modulate the amplitude of two pure-tone carriers with frequencies of 500 and 3160 Hz. After modulation the resulting signals were added. Conditions 1 to 4 were presented audiovisually. Figure 3.1 gives an example of the acoustic stimuli of conditions 1,2 and 4 for the utterance /sa/.

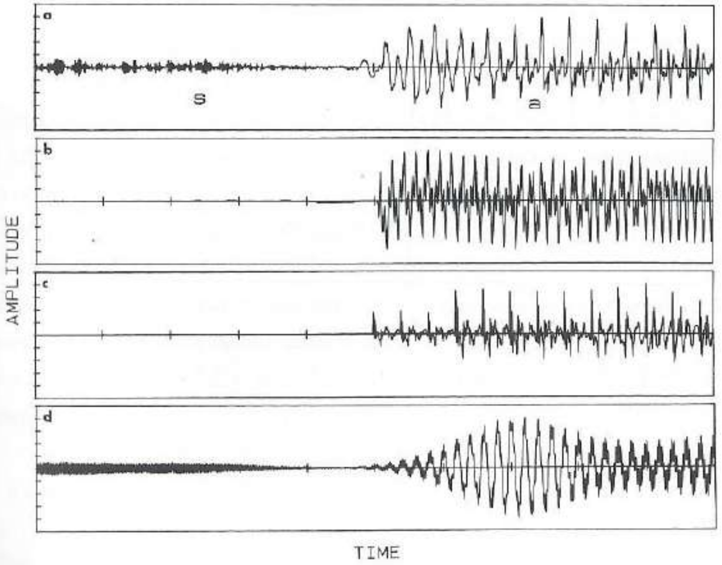

Fig. 3.1. Example of acoustic stimuli; (a) unfiltered speech signal of the utterance /sa/, (b) sum of two sinewaves with frequencies varying according to F1 and F2, and with constant amplitudes (condition 1), (c) complex formant signal of condition 2, (d) the sum of a 500-Hz and 3160-Hz tone amplitude-modulated by the envelopes of octave bands at 500 and 3160 Hz, respectively (condition 4). Time scale: 10 ms per division, arbitrary amplitude scale.

In condition 5 the complex signal of condition 2 was presented auditorily only. This condition was included to get an indication of the amount of information obtainable by listening to the formant-frequency information only. Finally, in condition 6 no acoustic stimulus was presented; the subjects had to rely on speechreading alone.

Since the formant frequencies could not be calculated in real time, the signals of conditions 1 to 3 had to be dubbed in on the videotape synchronously with the lip movements of the female speaker. This was done by using a trigger pulse preceding each sentence as reference. The same trigger pulse was used for sampling the speech signal in the formant-frequency analysis (see Section 3.2.1.2).

3.2.4 Subjects

Eighteen normal-hearing listeners, twelve male and six female, 17 to 30 years of age, with no experience in speechreading, participated in the experiment. They were paid for their participation. All subjects except two had pure-tone hearing levels of less than 17 dB for their better ear, measured at frequencies of 125, 250, 500, 1000, 2000, 4000, and 8000 Hz, and had normal or normally corrected vision (Landolt C-test, visual acuity of 20/20 or better). Two subjects had hearing levels of 25 dB (8000 Hz) and 21 dB (500 Hz), respectively.

3.2.5 Procedure

The experimental procedure was similar to that in our previous study. The subject was seated in a soundproof room 2 m from the 50 cm color video monitor. All acoustic stimuli were presented binaurally (Beyer DT-48 headphones) at a comfortable loudness level. The subject received one list (13 sentences) per condition. The order of presentation of the lists was fixed. In order to eliminate the influence of learning as much as possible, the conditions were presented according to a counterbalanced design. Each sentence was preceded by a short audible warning signal and followed by a 9-s pause during which the subject had to reproduce verbally as much of the sentence as possible. The experimenter, seated in the same room, took down the correct syllables on prepared response sheets. Additionally, the responses were recorded on tape, so that in case of uncertainty the responses could be replayed afterwards. The subject was told that the sentences consisted of meaningful words. Subjects did not often respond nonsense words. Using verbal report appeared to be an easy and accurate way of scoring the responses. Prior to the experiment the subject received some training in all six conditions (with different sentences). Total duration of the training was 30 min; total duration of the experiment 90 min .

3.3 Results

The last two columns of Table 3.1 show the mean and the standard deviation of the percentage of correctly reproduced syllables. As this table indicates, supplementing speechreading with the formant information brings about a substantial increase in the number of correctly perceived syllables.

Analysis of variance, applied to all six conditions, showed that 83.2% of the variance is explained by the conditions, while 6.3% is explained by the subjects. The subjects appear to form a homogeneous group. Analysis of variance on a subset, consisting of conditions 1 to 3 (formant information presented audiovisually), showed that these conditions do not differ significantly ($F(2,34)=0.91$, $p<0.58$). Apparently, the subjects were able to extract the same amount of information from the signal consisting of only two sinusoids as from the complex speech-like signals. Analysis of variance on conditions 1 to 4 (all audiovisual conditions) indicated that these conditions differ significantly ($F(3,54)=9.06$, $p<0.0002$). Individual comparison of the mean scores of conditions 1 to 4 showed that the sound-pressure information is a more effective supplement to speechreading that the formant-frequency information (t-test on paired observations, $t(17)>4.1$, $p<0.001$).

Formant-frequency information presented without visual information (condition 5) only scored 33.2% correct syllables. The speechreading-only score was 22.8%. This score exactly equals the speechreading-only score in our previous study. One subject appeared to be a born speechreader (61.8% correct syllables).

3.4 Discussion

Kirman (1974) presented the patterns of F1 and F2 tactually on a 15-by-15 tactual matrix to six subjects (without speechreading; see Section 1.4.2.3). He hypothesized that these patterns would probably not be sufficient for the perception of all speech sounds, since several sounds were persistently confused (e.g., /l/ and /r/). The present study supports these observations; when these patterns were presented without visual information (condition 5), a score of only 33.2% correct syllables was obtained. However, this study also indicates that these formant patterns do carry sufficient information to supplement speechreading effectively. A score of 82.0% correct syllables (condition 3) is very likely to be sufficient for meaningful conversation.

Contrary to our expectations, formant-frequency information presented (audiovisually) as pure tones did not lead to significantly lower scores than formant-frequency information presented as complex speech-like signals. The first time we listened to the pure-tone signal we felt that it might be difficult to integrate the two frequency-modulated tones into perception of speech. However, the subjects in the present study appeared to have no problem in doing so. They could adjust equally well to both types of signals. Adding noise to the complex signal of condition 2 did not significantly influence speech intelligibility either. Apparently, the subjects were not affected by the discontinuities created by omitting unvoiced speech sounds.

The sound-pressure information appears to be a more effective supplement than the formant-frequency information. To explain this result, it is necessary to investigate systematically the perception of more elementary speech material such as phonemes in monosyllabic nonsense words. Some speculative remarks, however, can be made. First, the sound-pressure information probably provides the subjects with more prosodic information than the formant-frequency information. This may be of importance, since by speechreading alone hardly any prosodic information can be obtained (Risberg, 1974). In the signal of condition 4 the temporal structure of the speech signal is more accurately preserved; information is given both about voiced and unvoiced sounds. Many results indicate the importance of the duration of different speech sounds as a cue for the perception of linguistic information (for a review, see Klatt, 1976). Furthermore, the signal of condition 4 gives information about the overall sound-pressure level of the speech signal. This may be of help in perceiving the stress pattern of the sentences. The formant-frequency information, however, does not give any information about the overall sound-pressure level. Second, the sound-pressure information probably contains more information about unvoiced sounds. In conditions 1 to 3 unvoiced sounds must be perceived solely by the formant-frequency transitions in adjacent voiced sounds; no signal is provided during the unvoiced portions of the speech signal. In condition 4, an unvoiced sound may be perceived by the intensity-ratio of the 500-Hz and the 3160-Hz tones.

3.5 Conclusion

From this experiment it can be concluded that for normal-hearing subjects with no experience in speechreading auditorily presented information about the frequencies of the first and second formants is a very effective supplement to speechreading of short sentences. However, information about the sound-pressure levels in octave bands at 500 and 3160 Hz appears to be a more effective supplement.

CHAPTER 4 - EXPERIMENT III: SPEECHREADING SUPPLEMENTED WITH PROSODIC INFORMATION

ABSTRACT

The benefit of supplementing speechreading with information about the prosody of the speech signal was studied by presenting short sentences to ten normal-hearing listeners under the conditions of speechreading-only and speechreading supplemented with: (a) information about both the overall sound-pressure level and the fundamental frequency of the speech signal, (b) information only about the fundamental frequency, (c) information only about the overall sound-pressure level, and (d) information about the duration of the voiced parts of the speech signal. The mean number of correctly perceived syllables increased from 16.7% in the case of speechreading-only to 64.0% in the case of speechreading with the fundamental-frequency and overall sound-pressure information. This study shows that supplementing speechreading with information about the prosody of the speech signal leads to a significant increase in speech intelligibility.

4.1 Introduction

Several factors limit the information obtainable through speechreading. First, different speech sounds can produce equal lip and jaw movements and are therefore visually indiscriminable (Voiers, 1973; Lowell, 1974). Second, several sounds do not produce visually perceivable lip movements, e.g. /h/ as initial consonant before a vowel (Berger, 1972). Third, through speechreading alone hardly any information about the prosody of the speech signal can be obtained. Risberg and Lubker (1978) showed that through speechreading alone it is very difficult to perceive prosodic cues such as vowel length, intonation pattern, juncture on different places in a word sequence, and emphasis on one word in a short sentence. Because of this lack of prosodic information the speechreader meets problems in parsing the utterances into smaller units such as words, syllables, and phonemes. The difficulties he has in choosing the correct phonemes from ambiguous lip movements are enlarged by the low accuracy with which the occurrence of these phonemes can be detected. Supplementing speechreading with prosodic information could simplify this task.

In the description of speech, a distinction is made between phonetic and prosodic features. A phonetic feature, such as plosion, friction, and nasality, is a distinctive property of a phonetic segment. By contrast, the domain of a prosodic feature extends over sequences of sounds that are longer than a segment (syllables, words, and sentences). Prosodic features are quantity, tone and stress. Quantity refers to the relative length of speech sounds (phonemes, syllables, words); the acoustic correlate of quantity is the duration of these speech sounds. Tone refers to the relative pitch, and stress to the relative force with which speech sounds are uttered. The acoustic correlate of tone is the fundamental frequency. Stress is not directly related to one single acoustic parameter; amplitude (i.e. the overall sound-pressure level) and duration play a role, but fundamental frequency appears to be the most important parameter for the perception of stress (Lehiste, 1970). Thus, three important acoustic parameters related to prosodic features are duration, fundamental frequency, and amplitude.

Rosen et al. (1978) showed that confusions between different phonemes were reduced when speechreading was supplemented with information about the duration of sequences of voiced sounds or with information about the fundamental frequency. They presented this information auditorily to normal hearing listeners in the form of a pulse sequence (band-pass filtered between 4 kHz and 20 kHz), either with constant frequency or with frequency equal to the fundamental frequency of the speech signal. Discrimination of consonants in a vowel-consonant-vowel context increased from 43.9% correct (speechreading alone) to 71.9% (constant frequency) and 73.6% (fundamental frequency). Risberg and Lubker (1978) studied the perception of different kinds of speech material presented to normal-hearing listeners under the conditions of speechreading alone, speechreading with information about the amplitude of the speech signal, and speechreading with information about both fundamental frequency and amplitude. They reported an increase in the number of correct responses to unknown questions from 37.9% (speechreading alone) to 42.4% (speechreading with amplitude information) and 78.5% (speechreading with fundamental-frequency and amplitude information). They further found an improvement in the discrimination of vowel length, intonation pattern, syllable stress, and emphasis on one word in a sentence.

In the experiment presented in this chapter speechreading was supplemented with acoustic stimuli similar to those of Rosen et al. (1978) and Risberg and Lubker (1978). The experiment was performed for two reasons: (a) to compare directly the relative effectiveness of the type of information used by Rosen et al. with that used by Risberg and Lubker, and (b) to investigate to what extent for the Dutch language these types of information can be used as supplements to speechreading.

4.2 Methods

4.2.1 Speech material

The speech material was equal to that of our previous studies, i.e. lists of thirteen short, meaningful Dutch sentences (eigth or nine syllables) typical of everyday conversation (Plomp and Mimpen, 1979). These sentences were recorded in color on videotape by a female speaker (speech therapist) with clear pronunciation. All sentences were completely unknown to the subjects. The lighting gave a shadow-free illumination of the face. Only her head and part of her shoulders were visible.

4.2.2 Subjects

Ten normal-hearing subjects, three male and seven female, aged 15 to 32 years, with normal or normally corrected vision (Landolt C-test, visual acuity of 20/20 or better), participated in the experiment. They had no experience in speechreading tests, and were paid for their participation. All except one had pure-tone hearing levels less than 20 dB for their better ear, measured at frequencies of 125, 250, 500, 1000, 2000, 4000, and 8000 Hz. One subject had a hearing level of 27.8 dB at 8000 Hz.

4.2.3 Experimental conditions

The sentences were presented to the subjects under the following five conditions:

(1) Speechreading with (auditorily presented) information about the fundamental frequency and the overall sound-pressure level of the speech

signal. The fundamental frequency was extracted as follows. First, the speech signal was band-pass filtered (Krohn-Hite model 3343 R, slope 48 dB/oct) between 160 and 320 Hz (the range of the fundamental frequency of the female speaker). Subsequently, the positive zero crossings were detected and converted into pulses (i.e. one pulse for each zero crossing). If the energy in the frequency band from 160 to 320 Hz was above a certain (experimentally determined) threshold the sound was considered to be voiced and the pulse sequence was band-pass filtered between 160 and 400 Hz (Brüel & Kjaer spectrum shaper type 5612). This range is somewhat broader than that of the fundamental frequency and was chosen to prevent sounds with a low fundamental frequency from sounding dull. The envelope of the speech signal was used the modulate the amplitude of the filtered pulse sequence. This envelope was detected by full-wave rectification and smoothing with a low-pass filter (cutoff frequency 20 Hz). Henceforth this condition will be notated as S+F0+A.

(2) Speechreading with information only about the fundamental frequency of the speech signal (as in condition 1 except without amplitude modulation). This condition will be notated as S+F0.

(3) Speechreading with information only about the overall sound-pressure level (as condition 1 except that the pulse sequence had a constant frequency of 250 Hz). This condition will be notated as S+A.

(4) Speechreading with binary information about voicing (notated as S+V). A periodic pulse sequence with constant frequency of 250 Hz and with constant amplitude was generated as long as the speech signal was voiced, and was band-pass filtered between 160 and 400 Hz. This signal only provides information about the duration of sequences of voiced sounds.

(5) Speechreading-only (notated as S).

Each condition was presented twice, and one list of sentences was used per presentation. In order to eliminate the influence of learning as much as possible, the conditions were presented according to a counterbalanced design. The second presentation of a condition always took place after all conditions had been presented once. Contrary to our

previous studies no pre-experimental training was given (in fact, this experiment was carried out before the two previous studies, and at that time we considered presenting each condition twice as an appropriate way of adjusting the subjects to the stimuli).

4.2.4 Procedure

The procedure was equal to that of our previous studies. The subject was seated in a soundproof room at a distance of 2 m from the color video monitor. All stimuli were presented binaurally at a comfortable loudness level. Each sentence was preceded by a short audible warning signal and was followed by a 9-s pause during which the subject had to reproduce verbally as much of the sentence as possible. The experimenter, seated in the same room, notated the correct syllables on prepared response sheets. Additionally, the responses were recorded on audiotape so that in case of uncertainty they could be replayed afterwards.

4.3 Results and discussion

Means and standard deviations of the percentage of correctly reproduced syllables are presented in Table 4.1. The mean values are also plotted in Fig. 4.1; the first presentation of a particular condition is indicated by a white bar, the second presentation by a shaded bar. The scores on the first and second presentations of a condition differ considerably; analysis of variance showed this difference to be significant ($F(1,9)= 48.1$, $p<0.005$). In the first presentation only speechreading with fundamental-frequency and overall sound-pressure information (S+F0+A) differed significantly from speechreading-only (S) (t-test on paired observations, $t(9)=3.7$, $p<0.005$). In the second presentation all audiovisual conditions except speechreading with binary-voicing information (S+V) differed significantly from speechreading-only (S) ($t(9)>4.8$, $p<0.001$). Thus, supplementing speechreading with fundamental frequency, overall sound-pressure level, or both led to a significant increase in speech intelligibility.

Speechreading with fundamental-frequency and overall sound-pressure information (S+F0+A) led to the highest score; however, this condition differed only slightly from speechreading with fundamental-frequency

condition	mode of presentation	first presentation		second presentation	
		mean	standard deviation	mean	standard deviation
S+F0+A	audiovisual	31.0	24.2	64.0	25.4
S+F0	audiovisual	20.9	16.4	51.4	27.6
S+A	audiovisual	10.9	13.1	45.7	28.8
S+V	audiovisual	9.6	10.1	29.5	21.8
S	visual only	10.6	15.4	16.7	15.6

Table 4.1. Mean and standard deviation of the percentage of correctly reproduced syllables for the five conditions of Experiment III.

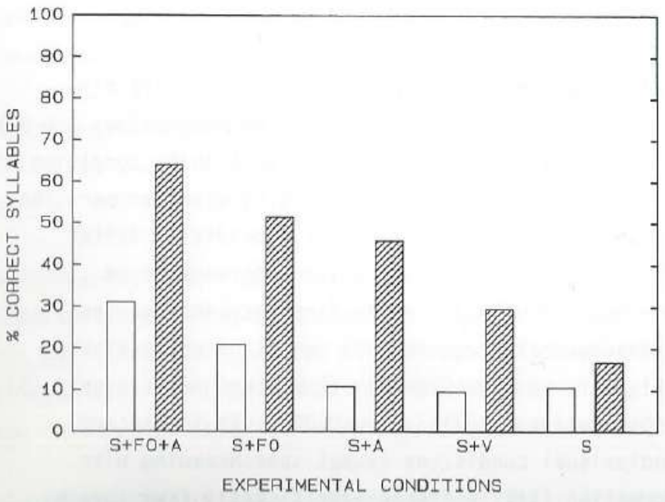

Fig. 4.1. Mean of the percentage of correctly reproduced syllables for the five conditions of Experiment III. The first presentation of a particular condition is indicated by a white bar, the second by a shaded bar.

information (S+F0) (t-test on paired observations, t(9)=1.9, p<0.10). From the values of the standard deviations in Table 4.1 it can be seen that the individual scores varied largely. The scores appeared to be distributed bimodally; two groups of subjects could be distinguished. The first group of five subjects reached a mean score of 83.6% on the second presentation of S+F0+A (st. dev. 7.7%). Four of them also reached a high score on S+F0 (mean 78.6%, st.dev. 11.4%); the fifth scored only 32.4%. The mean speechreading-only score for the five subjects in the first group was 26.6% (st.dev 16.7%). For these subjects the combination of the fundamental-frequency and overall sound-pressure information was a very effective supplement to speechreading. The five subjects in the second group scored only 44.0% (st. dev. 20.9%) on S+F0+A and 33.3% (st. dev. 19.4%) on S+F0. Their mean speechreading-only score was only 6.7% (st. dev. 4.2%), which probably explains their lower scores on speechreading with the auditory supplements.

Risberg and Lubker (1978) used two types of supplements that closely resemble two of the signals of the present study: information about overall sound-pressure level, and information about both overall sound-pressure level and fundamental frequency. As in the present study, a periodic pulse sequence was used to create these signals, but they filtered this pulse sequence in a frequency band around 500 Hz. Furthermore, they first band-pass filtered the speech signal between 100 and 2700 Hz before extracting the sound-pressure level. Rosen et al. (1979, 1981) also used two supplementary signals that resemble two of our signals: the binary-voicing information and the fundamental-frequency information. They also used a periodic pulse sequence, but filtered it between 4 kHz and 20 kHz.

Both in our study and that of Risberg and Lubker the fundamental-frequency plus overall sound-pressure information appears to be a more efficient supplement to speechreading than overall sound-pressure information alone. In our study the difference between speechreading-only and speechreading with overall sound-pressure information (29.0%) is larger than in the study of Risberg and Lubker, who only found an increase of 4.5% in the number of correct responses to unknown questions. This can be due to differences in speech material and method of scoring the responses. When they used (known) sentences as speech material, speechreading-only (40% correct) and speechreading with overall sound-pressure information (72% correct) differed considerably.

Comparison of our data with those of Rosen et al. shows that in both studies fundamental-frequency information is more appropriate for supplementing speechreading than is binary-voicing information. They found an improvement in consonant discrimination and perception of connected discourse on both conditions. In the present study the binary-voicing information proved to be no significant supplement to speechreading unknown sentences. In both studies speechreading with fundamental-frequency information scored significantly higher than speechreading-only.

4.4 Conclusion

From this experiment is can be concluded that for normal-hearing subjects with no experience in speechreading auditorily presented information about the fundamental frequency, the overall sound-pressure level, or the combination of both is a significant supplement to speechreading of short sentences.

CHAPTER 5 - EXPERIMENT IV: A COMPARISON AMONG SUPPLEMENTS

ABSTRACT

In the experiment described in this chapter the best-scoring supplements of the previous experiments (Chapters 2 to 4) were compared. In the previous experiments only normal-hearing subjects without experience in speechreading participated. In the present experiment both inexperienced and experienced speechreaders took part (with normal hearing). Furthermore, in addition to the perception of everyday sentences the discrimination of phonemes (both consonants and vowels) was measured for the inexperienced subjects. Percentage correct responses, confusions among phonemes, and percentage transmitted information about different types of manner and place of articulation and about the feature voicing are presented.

Speechreading was supplemented with: (a) the frequencies of the first and second formants from voiced speech segments, (b) the fundamental frequency, (c) both the fundamental frequency and the overall sound-pressure level, and (d) the sound-pressure levels in two one-octave frequency bands with center frequencies of 500 and 3160 Hz. Sentence-intelligibility scores were measured both for 24 normal-hearing subjects with no experience in speechreading, and for 12 normal-hearing experienced speechreaders. For the inexperienced speechreaders, the sound-pressure levels in the one-octave frequency bands at 500 and 3160 Hz appeared to be the best supplement (87.1% correct syllables). For the experienced speechreaders, the formant-frequency information (88.6% correct) and the fundamental-frequency plus overall sound-pressure information (86.0 % correct) were equally efficient supplements as the sound-pressure information in octave bands at 500 and 3160 Hz (86.1% correct).

Discrimination of phonemes (both consonants and vowels) was measured for the group of 24 inexperienced speechreaders. For the discrimination of vowels only the formant-frequency information appeared to be an effective supplement. All the supplementary signals appeared to be significant supplements for the discrimination of consonants; the gains appeared to be primarily caused by an improvement in the perception of manner of articulation and of the feature voicing.

5.1 Introduction

From Experiments I and II (Chapters 2 and 3) it appeared that information about the sound-pressure levels in two one-octave frequency bands centered at 500 and 3160 Hz, or information about the frequencies of the first and second formants (F1 and F2) are effective supplements to speechreading. Experiment III (see Chapter 4) indicated that also information about the fundamental frequency and amplitude (overall sound-pressure level) can be used as a supplement; however, the highest score (64% correct syllables on S+F0+A) is much lower than the scores obtained previously with sound-pressure information in octave bands at 500 and 3160 Hz, and with formant-frequency information. The set-up of Experiments I and II was similar to that of Experiment III except that pre-experimental training with feedback was given. The lack of training in Experiment III could be partly responsible for the lower sentence intelligibility.

In the experiment reported in this chapter (Experiment IV) the effectiveness of sound-pressure information in octave bands at 500 and 3160 Hz and of formant-frequency information was compared with that of the best-scoring supplements of Experiment III (F0+A and F0) using exactly the same experimental set-up and speech material (sentences) as in Experiments I and II. In order to gain insight into the role of experience in speechreading, the perception of sentences was measured both for 24 subjects with no experience in speechreading and for 12 subjects with experience in speechreading. Additionally, to gain insight into which speech sounds can be distinguished and which speech features can be perceived, the discrimination of phonemes (both vowels and consonants) was measured for the group of 24 inexperienced speechreaders.

5.2 Methods

5.2.1 Speech material

The speech material consisted of both sentences and words. The sentences have already been described in Experiments I,II, and III. Discrimination of consonants was measured using /aCa/ words with one of the eighteen consonants /p/, /b/, /m/, /f/, /v/, /ʋ/, /l/, /r/, /s/, /z/, /t/, /d/, /n/, /j/, /k/, /ŋ/, /x/, and /h/. Discrimination of vowels was measured with /hVt/ words, with one of the twelve vowels /i/, /I/, /e/, /ɛ/, /a/, /ɑ/, /œ/, /ɔ/, /ø/, /o/, /y/, and /u/. Lists of either 36 /aCa/ words (containing each consonant twice) or 24 /hVt/ words (each vowel twice) were used for the experiment. The lists were recorded on videotape under the same conditions and by the same female speaker as in Experiments I to III.

5.2.2 Subjects

Two groups of subjects participated in the present experiment. The first group consisted of 24 normal-hearing individuals with no experience in speechreading, 11 male and 13 female, aged 16 to 27 years, with normal or normally corrected vision (Landolt C-test, visual acuity of 20/20 or better). They were paid for their participation. All except four had pure-tone hearing levels less than 20 dB for their better ear, measured at frequencies of 125, 250, 500, 1000, 2000, 4000, and 8000 Hz. Four had hearing levels between 20 and 30 dB at 8000 Hz (actually, the first group consisted of 25 subjects, but the responses of one female subject were not included in the data analysis because she reported an unpleasant sensation from the sound-pressure information; she could not use this information at all).

The second group consisted of 12 normal-hearing subjects with experience in speechreading, one male and 11 female, aged 25 to 64 years, also with normal or normally corrected vision. All except one had hearing levels of less than 20 dB for their better ear, measured at the above-mentioned frequencies. One had a hearing level of 25 dB at 8000 Hz. Among them were speech therapists who give lessons in speechreading, teachers of the deaf, and coaches of groups that practice speechreading. They were selected only on the basis of their supposed experience in speechreading, not on their speechreading capabilities. As will be seen later, some were good, others poor speechreaders.

5.2.3 Experimental conditions

The sentences were presented under the following five conditions:

(1) Speechreading with (auditorily presented) information about the frequencies of the first and second formants from voiced speech segments. These formant frequencies were presented as a complex speech-like signal, obtained by filtering a periodic pulse sequence of 250 Hz (at voiced moments) by four second-order band-pass filters (with constant bandwidth); the center frequencies of two of these filters followed F1 and F2, whereas the others remained constant at 3500 and 4500 Hz. Furthermore, in order to fill up unvoiced segments, a speech-like noise was added. Further details can be found in Chapter 3 (Experiment II). This condition will be notated as S+F1+F2.

(2) Speechreading with information about the fundamental frequency of the speech signal (condition S+F0 from Experiment III, Chapter 4).

(3) Speechreading with both information about the fundamental frequency and the overall amplitude (sound-pressure level) of the speech signal (condition S+F0+A from Experiment III, Chapter 4).

(4) Speechreading with information about the sound-pressure levels in two one-octave frequency bands at 500 and 3160 Hz. This signal was obtained by using the envelope of the output of two one-octave filter bands at 500 and 3160 Hz to modulate the amplitude of two pure-tone carriers with frequencies of 500 and 3160 Hz, respectively. After modulation the resulting signals were added. This condition resembles the best-scoring condition of Experiments I and II (Chapters 2 and 3), and will be notated as S+SPI.

(5) Speechreading-only (notated as S).

The subjects received one list of sentences in conditions 1 to 4, and two lists in condition 5.

The lists of words were presented under nine conditions (one list per condition): the five above-mentioned conditions and another four conditions in which the acoustic signals of conditions 1 to 4 were presented without visual information. These four conditions were included to get an indication of what information can be perceived by listening only to the auditory signals. They will be notated as F1+F2, F0, F0+A, and SPI. In order to eliminate the influence of learning as much as possible the conditions were presented according to a counterbalanced design. The sentence perception was always measured before the phoneme discrimination.

5.3 Results and discussion

5.3.1 Sentence perception

Figure 5.1-a gives the mean percentages of correctly reproduced syllables for the group of 24 inexperienced subjects; the mean values and standard deviations are also presented in Table 5.1. The intelligibility scores on S+F1+F2 (78.8%) and S+SPI (87.1%) closely resemble the corresponding scores in Experiments I and II (S+F1+F2: 80.8%, S+SPI: 86.7%). The scores on S+F0 (64.0%) and S+F0+A (77.7%) are considerably higher than the scores in Experiment III (S+F0: 51.4%, S+F0+A: 64.0%). The same holds for the speechreading-only scores (Experiment III: 16.7%, Experiment IV: 25.0%).

condition	mode of presentation	24 inexperienced subjects		12 experienced subjects	
		mean	standard deviation	mean	standard deviation
S+F1+F2	audiovisual	78.8	14.5	88.6	10.4
S+F0	audiovisual	64.0	21.1	73.0	15.4
S+F0+A	audiovisual	77.7	14.1	86.0	9.0
S+SPI	audiovisual	87.1	8.7	86.1	12.4
S	visual only	25.0	14.7	33.0	16.6

Table 5.1. Mean and standard deviation of the percentage of correctly reproduced syllables for the five conditions of the sentence-perception task of Experiment IV. In this table both the scores for the group of 24 inexperienced speechreaders and for the group of 12 experienced speechreaders are given.

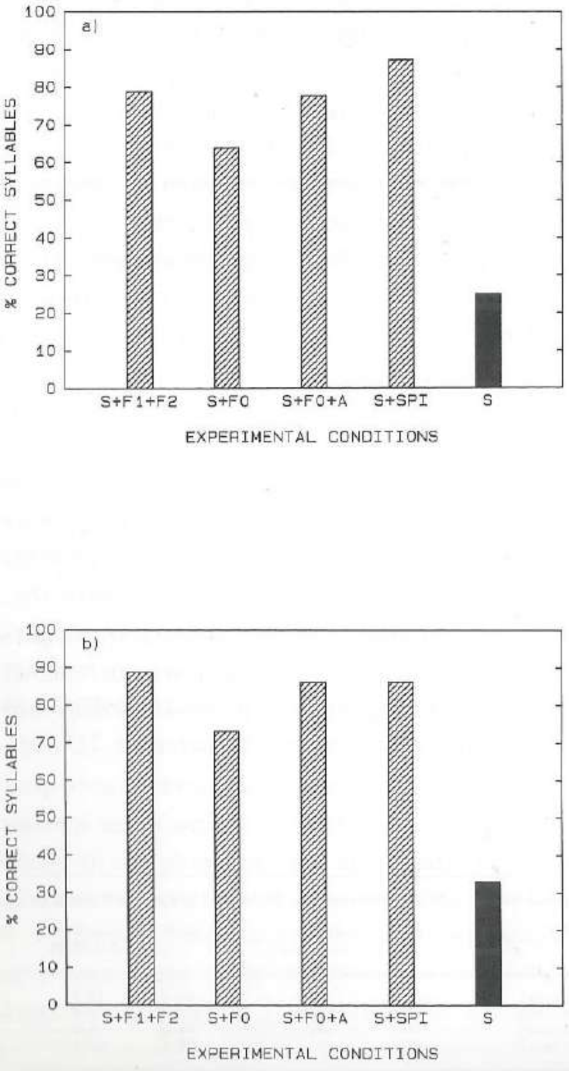

Fig. 5.1. Mean of the percentage of correctly reproduced syllables for the five conditions of Experiment IV: (a) for the group of 24 inexperienced speechreaders, and (b) for the group of 12 experienced speechreaders. The shaded bars indicate the audio-visual conditions, the black bar indicates the visual-only condition.

These differences may be due to the different experimental procedures in these experiments; in the present experiment the subjects received pre-experimental training with feedback, while in Experiment III the subjects had to get accustomed to the stimuli during the experiment (in which no feedback was given). However, in Experiment III two groups of subjects could be distinguished; the subjects in the first group reached much higher scores than the subjects in the second group (83.6% on S+F0+A and 26.6% on speechreading-only). These higher scores resemble the scores of the present experiment (S+F0+A: 77.7%, S: 25.0%).

A t-test on paired observations showed that the sound-pressure information in octave bands at 500 and 3160 Hz is a more effective supplement to speechreading than the other auditory signals (t(23)>3.6, p<0.002). The scores on S+F1+F2 and S+F0+A do not differ significantly (t-test on paired observations, t(23)=0.6, p>0.5), but these scores are significantly higher than the score on S+F0 (t(23)>4.9, p<0.001). Speechreading-only scored 25.0% correct; this score is about the same as in Experiments I and II (22.8%).

Figure 5.1-b gives the mean percentages of correctly reproduced syllables for the group of 12 subjects with experience in speechreading; the mean values and standard deviations are also presented in Table 5.1. The scores on speechreading with F1+F2 (88.6%), with F0+A (86.0%), and with SPI (86.1%) do not differ significantly (t-test on paired observations, t(11)<0.9, p>0.5); however, these scores are significantly higher than the score for speechreading with F0 (t-test on paired observations, t(11)>3.5, p<0.005). Apparently, for the experienced speechreaders F1+F2, F0+A, and SPI are about equally efficient supplements. However, it may well be that the scores reach an asymptotic value that is determined by the type of the task instead of by the auditory supplements that are used. Possibly, other tests, such as the perception of more difficult speech material, can indicate differences between the different supplements.

Except for S+SPI, the scores for the experienced speechreaders are 8-10% higher than the scores for the inexperienced speechreaders. A t-test on the hypothesis H0: m1=m2 against H1: m2>m1 (m1=mean score for the inexperienced subjects, m2=mean score for the experienced subjects) showed that the differences are significant between the 2.5-% and 10-% level (1.31<t(34)<2.08).

On the average, the experienced subjects were better speechreaders than the inexperienced subjects. However, the individual speechreading-only scores varied widely, from 4.1 to 57.8% (inexperienced subjects) and from 8.1 to 53.2% (experienced subjects). All experienced subjects had knowledge of which speech sounds can and which can not be discriminated by speechreading-only, and most of them knew how different speech sounds are produced. However, since they all had normal hearing, they did not have to rely on speechreading for perceiving speech during daily life. Apparently, this knowledge is not a sufficient condition for being a skilled speechreader. Discussions with speech therapists who give lessons in speechreading revealed that even among hearing-impaired and totally deaf persons a wide range of speechreading skills can be found.

5.3.2 Phoneme discrimination

Phoneme discrimination was investigated only for the group of 24 inexperienced speechreaders. Figure 5.2 gives the mean percentage correct responses for consonants (Fig. 5.2-a) and for vowels (Fig. 5.2-b); the mean values and the standard deviations are also presented in Table 5.2.

condition	mode of presentation	consonant discrimination		vowel discrimination	
		mean	standard deviation	mean	standard deviation
S+F1+F2	audiovisual	50.6	7.9	81.9	20.6
S+F0	audiovisual	60.3	16.0	62.5	11.3
S+F0+A	audiovisual	68.6	8.0	59.4	16.7
S+SPI	audiovisual	58.8	15.3	59.0	12.5
F1+F2	auditory only	29.1	6.1	63.4	18.8
F0	auditory only	20.4	8.0	13.5	7.6
F0+A	auditory only	27.0	5.5	14.6	6.5
SPI	auditory only	24.5	8.5	20.8	11.3
S	visual only	38.0	11.1	55.0	16.5

Table 5.2. Mean and standard deviation of the percentage of correctly discriminated consonants and vowels for the nine conditions of the phoneme-discrimination task of Experiment IV.

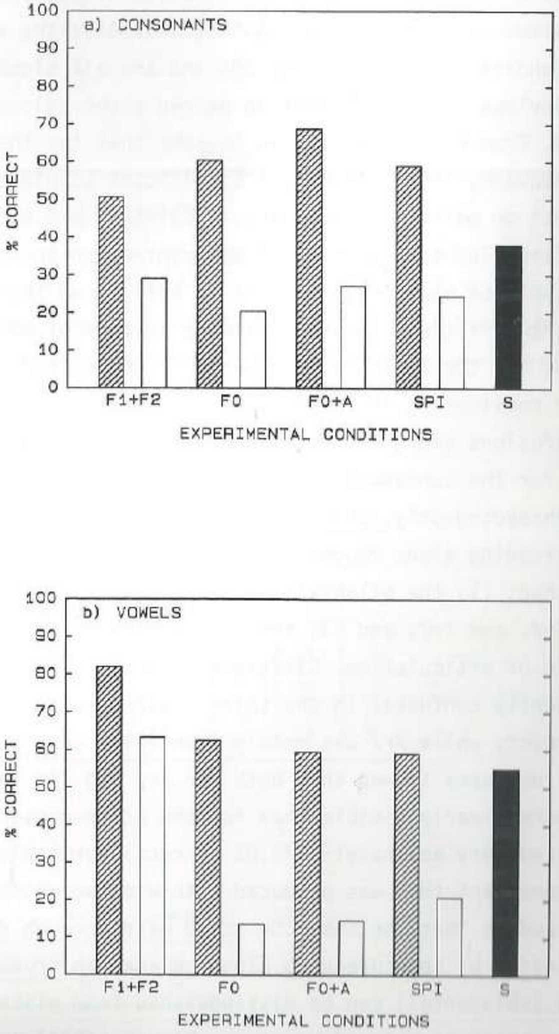

Fig. 5.2. Mean of the percentage of correctly discriminated consonants (Fig. 5.2-a), and vowels (Fig. 5.2-b) for the phoneme-discrimination task of Experiment IV. The shaded bars indicate the audiovisual conditions, the white bars indicate the auditory-only conditions, and the black bar indicates the visual-only condition.

Supplementing speechreading with the four types of auditory information leads to a significant increase in the discrimination of consonants in all cases (t-test on paired observations, $t(23)>5.0$, $p<0.001$); the scores on the auditory-only conditions do not exceed 30% and are all significantly lower than the audiovisual scores (t-test on paired observations, $t(23)>9.7$, $p<0.001$). From Fig. 5.2-b it can be seen that for the vowels only the formant-frequency information is a significant supplement to speechreading (t-test on paired observations, $t(23)>5.4$, $p<0.001$). Listening only to F1+F2 led to a mean vowel-discrimination score of 63.4%. Such a high score could be expected, since it is well known that F1 and F2 are to a great extent sufficient for the characterization of vowels. The auditory-only scores for the other three signals do not exceed 21%.

One method for considering the phoneme discrimination in more detail is the study of confusions among different phonemes. Table 5.3 shows the confusion matrices for the consonants (Table 5.3-a) and vowels (Table 5.3-b) in the speechreading-only condition. From Table 5.3-a it appears that through speechreading alone roughly three clusters of consonants could be distinguished: (1) the bilabials /p/, /b/, and /m/, (2) the labiodentals /f/, /v/, and /ʋ/, and (3) the rest of the consonants with a more posterior place of articulation. Different consonants within these clusters were frequently confused. In the third cluster the /l/ was recognized 100% correct, while /r/ was mostly identified as /l/. Inspection of the videotapes showed that both for /r/ and /l/ the movement of the tongue was more clearly visible than for the other consonants. Also the /h/ was recognized very accurately (75.0% correct), probably because this was the only consonant that was produced both without visible lip and jaw movements. It appears that the subjects could make a rough distinction of place of articulation by speechreading alone: places in front of the mouth (bilabial and labiodental) can be distinguished from places more backwards. Similar results were found by, e.g., Binnie (1974) and Woodward and Barber (1960).

Table 5.3-b shows that through speechreading alone roughly two clusters of vowels could be distinguished: (1) the rounded vowels /œ/, /ø/, /y/, /ɔ/, /o/, and /u/, and (2) the vowels pronounced with more neutral or spread lips. In the second cluster the spread vowels /i/,/I/,/e/,/ɛ/ were more often confused with each other than with the other vowels, except for /ɛ/: /i/, /I/ and /e/ were often responded as

(a)

	p	b	m	f	v	ʋ	r	l	s	z	t	d	n	j	ŋ	k	x	h
p	31	10	7															
b	22	20	6															
m	19	16	13															
f		1		22	19	6												
v				15	26	7												
ʋ				11	23	14												
r							18	25				3		1				1
l								48										
s									17	17	5	4	2			3		
z		1	1			1			22	18	1				1	1	2	
t									8	3	24	10		1	1	1		
d					1				5	5	21	9	2	1		2	2	
n								2	10	4	9	7	11	2		1	2	
j				1					9	11	8	9	2	2		2	3	1
ŋ		1		1		1	1		7	1	8	12	3	5	3	3	2	
k							1	2	7	8	4	4	2	5	2	8	5	
x	1				1	1	5						2	6	5	13	8	6
h														1	3		8	36

(b)

	i	I	e	ɛ	a	ɑ	œ	ɔ	ø	o	y	u
i	16	24	7	1								
I	1	15	7	25								
e	1	3	22	21							1	
ɛ			2	19	14	10	1	1			1	
a			1		41	6						
ɑ					4	44						
œ							39	4	2	1	1	1
ɔ						4	3	25	4	11	1	
ø							2	6	22	13	2	3
o									3	30	3	12
y							5	2	3		12	26
u				1				3	5	4	3	32

(c)

	voiced	unvoiced
voiced	337	191
unvoiced	129	207

Table 5.3. (a) Consonant-confusion matrix, (b) vowel-confusion matrix, and (c) confusion matrix for the feature voicing. All three matrices refer to the speechreading-only condition. The stimuli are indicated vertically on the left side of the matrix, the responses horizontally above the matrix.

/ε/, but /ε/ was more often responded as /a/ or /ɑ/.

The confusion matrices for speechreading with one of the auditory signals look much the same as those for speechreading-only. These matrices are given in Appendix 5.A. Instead of discussing these matrices here, the results are given of a more quantitative analysis of the information that is transmitted by speechreading with each of the auditory signals. In order to gain insight into which information can be perceived, an articulatory feature classification was used to analyze the percentage information transmitted by each feature and by each condition. This analysis was done only for the consonants, since for the vowels only the formant-frequency information could improve the discrimination, which can be understood by the nature of the auditory information. Table 5.4 shows the feature classification that was used: consonants were classified according to their manner and place of articulation and to the feature voicing.

Feature	p	b	m	f	v	ʋ	t	d	n	s	z	l	r	j	ŋ	k	x	h
a) plosion	+	+	-	-	-	-	+	+	-	-	-	-	-	-	-	+	-	-
nasality	-	-	+	-	-	-	-	-	+	-	-	-	-	-	+	-	-	-
friction	-	-	-	+	+	-	-	-	-	+	+	-	-	-	-	-	+	+
lateral	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-
vowel-like	-	-	-	-	-	+	-	-	-	-	-	-	-	+	-	-	-	-
roll	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-
b) bilabial	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
labiodental	-	-	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-
alveolar	-	-	-	-	-	-	+	+	+	+	+	+	+	-	-	-	-	-
palatal	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-
velar	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-
glottal	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+
c) voicing	-	+	+	-	+	+	-	+	+	-	+	+	+	+	+	-	-	-

Table 5.4. Articulatory feature classification for the 18 consonants according to (a) manner of articulation, (b) place of articulation, and (c) voicing. The "+" indicates the presence of a particular feature.

The percentage of information transmitted per feature was calculated using a method described by Miller and Nicely (1955). The percentage transmitted information is defined as

$$T = \frac{H(s,r)}{H(s)} \,.\, 100\ \%$$

where H(s,r) is the transmitted information from s(timulus) to r(esponse) (in bit per stimulus) and H(s) is the information available in the stimuli (in bit per stimulus). H(s,r) and H(s) are defined as

$$H(s,r) \quad = - \sum_i \sum_j p(s_i,r_j).{}^2\log\left(\frac{p(s_i).p(r_j)}{p(s_i,r_j)}\right)$$

$$H(s) \quad = - \sum_i p(s_i).{}^2\log(p(s_i))$$

$p(s_i)$ = probability of occurrence of stimulus feature s_i

$p(r_j)$ = probability of occurrence of response feature r_j

$p(s_i,r_j)$ = probability of joint occurrence of stimulus feature s_i and response feature r_j

The probabilities $p(s_i)$, $p(r_j)$, and $p(s_i,r_j)$ are equal to n_i/n, n_j/n, and n_{ij}/n, where n_i is the frequency of occurrence of stimulus s_i, n_j is the frequency of occurrence of response r_j, n_{ij} is the frequency of the joint occurrence of stimulus s_i and response r_j, and n is the total number of presented stimuli.

For each of the 13 features of Table 5.4 and each of the 9 conditions the stimuli and responses were divided into two groups - consonants that do have the specific feature and consonants that do not have it - and pooled confusion matrices of all 24 subjects' confusions between these groups were calculated, resulting in 117 different two-by-two confusion matrices. The percentage of transmitted information per condition and per feature was calculated from these matrices. For example, from the consonant-confusion matrix for the speechreading-only condition (Table 5.3-a) a two-by-two matrix can be extracted that indicates how the feature voicing was perceived (see Table 5.3-c). The total number of stimuli n is equal to 864 (24 subjects × 2 presentations of each consonant × 18 consonants), consisting of 528 voiced and 336 unvoiced stimuli. Thus $p(s_1)$=p(voiced stimulus)=528/864 and $p(s_2)$=p(unvoiced stimulus)=336/864, which results in H(s)=0.9641 bit/stimulus. From Table 5.3-c it can be seen that voiced stimuli were responded 337 times as voiced and 191 times as unvoiced, whereas unvoiced stimuli were responded 129 times as voiced and 207 times as unvoiced. Thus, $p(r_1)$=p(voiced response)=0.5394, $p(r_2)$=p(unvoiced response)=0.4606, $p(s_1,r_1)$=0.3900, $p(s_1,r_2)$=0.2211, $p(s_2,r_1)$=0.1493, and $p(s_2,r_2)$=0.2396, which gives a H(s,r) of 0.045 bit/stimulus and a T of 4.7%.

Table 5.5 gives the percentage transmitted information for speechreading-only and the gain in transmitted information when speechreading is supplemented with each of the four auditory signals. This gain is simply the difference between the percentage transmitted information for speechreading with an auditory supplement and speechreading-only. Table 5.5-a shows that, except for the feature lateral, very little information could be perceived about manner of articulation by speechreading alone. Almost no information was perceived about nasality and vowel-like, while limited information was obtained about plosion, friction, and roll. The lateral /l/ was perceived very accurately: all subjects responded /l/ when it was presented as stimulus; however, /r/ was often responded as /l/ (37.5%).

Information about place of articulation appeared to be partly perceivable. Bilabial and labiodental place were responded almost perfectly, and a considerable amount of information could be obtained about alveolar and glottal place of articulation. Except for glottal place, the more backwards situated places of articulation (alveolar, palatal, velar) were less well perceived than those in front of the mouth. The glottal /h/ could be perceived very accurately; however, this is probably partly due to the speech material used. Berger (1972) showed that the /h/ as initial consonant preceding a vowel could not be recognized at all.

Table 5.5 further shows that by speechreading alone hardly any information about voicing could be obtained. This, of course, had to be expected, since the resonance of the vocal cords can not be seen by speechreading.

Supplementing speechreading with the four auditory signals led to an increase in the perception of plosion in all cases (see Table 5.5-b). The highest improvement was obtained for those signals that support information about the overall sound-pressure level of the speech signal, i.e. F0+A (38.8% improvement) and SPI (40.8% improvement). The signal F0+A

Condition	Manner of articulation						Place of articulation						Voicing
	plosion	nasality	friction	lateral	vowel-like	roll	bilabial	labio-dental	alveolar	palatal	velar	glottal	
a) Speechreading	17.5	4.2	20.5	72.5	2.5	22.2	94.4	90.8	41.9	0.1	11.7	55.0	4.7
b) S+F1+F2	16.5	39.0	7.5	-7.6	43.9	-9.2	-2.0	-2.4	0.3	71.3	-2.5	-8.4	13.4
S+F0	19.1	40.6	0.0	3.2	5.4	55.2	-8.0	-9.6	-0.4	1.7	13.3	-8.4	46.4
S+F0+A	38.8	53.4	3.5	12.2	7.0	75.2	2.9	3.1	10.1	2.9	22.7	-12.5	55.4
S+SPI	40.8	32.2	11.9	1.9	16.0	44.9	-10.5	-10.8	3.7	7.4	5.7	-12.7	16.5

Table 5.5. (a) Percentage transmitted information for speechreading-only, and (b) the gain in transmitted information when speechreading is supplemented with the four auditory signals.

Condition	Manner of articulation						Place of articulation						Voicing
	plosion	nasality	friction	lateral	vowel-like	roll	bilabial	labio-dental	alveolar	palatal	velar	glottal	
F1+F2	6.9	23.6	3.2	44.3	26.0	12.2	5.0	4.8	9.1	73.2	0.1	1.5	9.6
F0	12.4	15.6	0.4	1.5	0.5	60.7	0.2	1.3	0.5	0.0	0.0	3.5	28.9
F0+A	26.4	21.7	2.1	4.2	3.0	77.8	8.2	3.8	0.7	2.8	1.0	0.6	29.9
SPI	32.7	9.4	16.0	15.2	8.9	50.8	5.6	7.4	1.6	2.6	0.3	0.2	8.0

Table 5.6. Percentage transmitted information when the four auditory signals are presented auditorily-only.

gives information about the overall sound-pressure level, while the signal SPI gives information about the sound-pressure level in the one-octave 500-Hz band. In Chapter 2 (Experiment I) it appeared that these sound-pressure levels correlate highly. The subjects probably used the sudden rise of the sound-pressure level of plosive sounds for identifying them.

Perception of nasality was improved considerably by all auditory signals. This is important, since by speechreading alone hardly any information is transmitted about nasality (only 4.2%). The SPI signal gave the highest improvement on friction. Surprisingly, the improvement is rather small (11.9%); it was expected that more information about friction would be perceivable by the presence of the 3160-Hz tone. Lateral was improved slightly only by F0+A, but this improvement is of minor importance since by speechreading alone lateral was perceived very well. Vowel-like was improved considerably only by F1+F2. This is consistent with the results found for the vowel-discrimination task. All supplementary signals except F1+F2 caused a considerable gain in the perception of roll. The roll /r/ could be clearly heard both by irregularities in the extracted fundamental frequency and by variations in the amplitude of this sound. Inspection of the confusion matrices for speechreading with the formant-frequency information showed that the roll /r/ was mostly identified as /l/.

From Table 5.5 it appears that (except for F1+F2 on palatal) not one of the signals gave a large improvement in the perception of place of articulation, while for some places (e.g. glottal) the amount of transmitted information decreased. The only palatal consonant /j/ is pronounced with a vowel-like manner of articulation. As was discussed above, the signal S+F1+F2 gives much information about vowels and vowel-like sounds, which explains the improvement of palatal for S+F1+F2.

The highest gain on voicing was obtained for those supplementary signals which provide information about fundamental frequency, namely F0 (46.4%) and F0+A (55.4%). This is not surprising, since the fundamental frequency contains the most direct information about voicing. The signals F1+F2 and SPI gave a more limited improvement (13.4% and 16.5%). Both signals do contain accurate, but more indirect, information about voicing. The signal F1+F2 differs from zero only for voiced sounds, while for SPI voicing might be perceived by the intensity ratio of the 500-Hz and

3160-Hz tones. Apparently, the subjects could not use this more indirect information as well as the fundamental-frequency information.

Table 5.6 gives the percentage transmitted information for the four auditory-only conditions. A close agreement can be seen between this table and Table 5.5-b. Plosion is again perceived best by F0+A and SPI. All signals support information about nasality; however, the transmitted information for SPI is somewhat lower than the gain on S+SPI. SPI gives again the highest score on friction. F1+F2 and SPI both provide information about lateral; however, from Table 5.5 it appears that this information was not used to improve the perception of lateral for S+F1+F2 and S+SPI. Vowel-like is again perceived best by F1+F2 and roll by F0, F0+A, and SPI. Also F1+F2 appears to provide some information about roll, but the perception of roll worsened when speechreading was supplemented with F1+F2. Hardly any information (except for lateral on F1+F2) about place of articulation could be obtained by listening only to the auditory signals. Voicing is again perceived best by F0 and F0+A.

Appendix 5.B gives the pooled confusions (of all 24 subjects) among different types of manner of articulations for speechreading with each of the four auditory signals. These matrices were constructed from the consonant-confusion matrices by counting the number of times that consonants with a certain type of manner of articulation were confused with consonants with other types. Thus, the response /b/ to the stimulus /p/ was considered as correct (both plosives), whereas the response /f/ to /p/ would mean confusion of the fricative with the plosive type of manner of articulation.

For S+F1+F2 plosion and friction were often confused, and roll was mostly responded as lateral. For S+F0 and S+F0+A friction was often responded as plosion, and vowel-like as friction. The signals F1+F2, F0 and F0+A do not support information at unvoiced moments; no signal is available for the unvoiced fricatives /f/, /s/, /x/ and /h/. Inspection of the confusion matrices for S+F1+F2, S+F0 and S+F0+A (Appendix 5.A) showed that /s/ was often identified as /t/, and /x/ as /k/. The /f/ and /h/ were not often confused with plosives, probably because these sounds have visible lip movements which are not characteristic for plosives. Apparently, for /s/ and /x/ the subjects confused the artificial pause before the second vowel /a/ with the normally existing pause before plosives. The signal SPI gives information both about voiced and unvoiced sounds. For S+SPI friction and plosion were not often confused; however, vowel-like was often responded as friction. Inspection of the confusion matrices showed that for S+F0, S+F0+A, and S+SPI the vowel-like /ʋ/ was often confused with /v/, and /j/ was often responded as /z/. Thus the four signals share the need for extra information about fricative sounds. Adding this information would improve the perception of plosion, friction and vowel-like. Furthermore, F1+F2 needs extra information about roll.

5.3.3 Comparison between scores on sentences and on phonemes

When the percentage-correct scores for the sentence-perception task (Fig. 5.1, Table 5.1) and the phoneme-discrimination task (Fig. 5.2, Table 5.2) are compared, it appears that the best supplement to speechreading sentences (the sound-pressure information) does not give the highest scores for phoneme discrimination. For phoneme discrimination, consonants appear to be best discriminated by speechreading with fundamental-frequency plus amplitude information, vowels by speechreading with formant-frequency information.

Several factors may account for the differences between these scores. First, the sentences contain information that is not present in isolated words. For example, in the perception of sentences the perception of suprasegmental information (such as stress patterns and intonation patterns) also plays a role. The phoneme-discrimination task can not clarify how well this sentence-bound information is perceived. Second, the information transmission analysis showed that each auditory signal improves the perception of specific speech features; however, no information is available about the relative importance of each feature for the perception of sentences. For example, from Table 5.5 and 5.6 it appeared that both F0 and F0+A give a considerable amount of information about voicing. But in ongoing speech the perception of voicing for cognates such as /t/ and /d/, or /s/ and /z/, probably is of less importance than the perception of features such as plosion and friction.

Relations between the percentage-correct scores on sentences and phonemes were studied by calculating correlations over the 24 subjects between all possible pairs of conditions. The vowel-discrimination score for speechreading-only correlates significantly with the sentence-intelligibility score for speechreading-only (Fisher's test, r=0.678, p<0.01). However, only a weak correlation exists between the consonant-discrimination score and the sentence-intelligibility score for speechreading-only (r=0.447, p<0.05). Erber (1974) also reported that only the visual recognition of vowels is closely related to overall speechreading skill. Speechreading-only of vowels also correlates significantly with the sentence-intelligibility scores for speechreading with F1+F2 (r=0.595, p<0.01), with F0 (r=0.594, p<0.01), and with F0+A (r=0.644,p<0.01). Speechreading-only of vowels and speechreading plus SPI of sentences correlate weakly (r=0.488, p<0.05).

None of the phoneme-discrimination scores and the sentence-intelligibility scores for speechreading with a specific auditory supplement correlate significantly (at the 1-% level). The same holds for the correlations between discrimination scores on the auditory-only presentation of a specific supplement and the sentence-intelligibility score for speechreading with this supplement. It thus appears that, except for speechreading-only of vowels, not one of the phoneme-discrimination scores is closely related to the sentence-intelligibilty scores.

5.4 Conclusions

From this experiment it can be concluded that:

(1) For normal-hearing subjects with no experience in speechreading auditorily presented information about the sound-pressure levels in octave bands at 500 and 3160 Hz appears to be a more effective supplement to speechreading of short sentences than information about the frequencies of the first and second formants, information about the fundamental frequency, or information about the combination of the overall sound-pressure level and the fundamental frequency. For normal-hearing subjects with experience in speechreading information about the sound-pressure levels in the one-octave bands at 500 and 3160 Hz is equally efficient as information about the frequencies of the first and second formants or information about the fundamental frequency and the overall sound-pressure level.

(2) For inexperienced speechreaders only formant-frequency information gives a significant improvement in the discrimination of vowels. All auditory supplements give a significant improvement in the discrimination of consonants. Analysis of the confusions among consonants showed that all auditory supplements improve the perception of manner of articulation and voicing; however, perception of place of articulation is not improved.

APPENDIX 5.A

Confusions among phonemes for speechreading with each of the four auditory supplements. The stimuli are indicated vertically on the left side of the matrix, the responses horizontally above the matrix.

condition S+F1+F2

	p	b	m	f	v	ʋ	r	l	s	z	t	d	n	j	ŋ	k	x	h
p	38	7	3															
b	30	13	4												1			
m			48															
f				33	15													
v				20	22	6												
ʋ				6	11	29							2					
r						1	12	28						2	2		3	
l								48										
s					1			1	19	6	12	5	1			1	2	
z			1			2	1		11	14	4	3	3	3		2	3	1
t		1							8	6	26	3	4					
d	1						1	1	4	6	16	10	3	1		2	3	
n							1	8		1	2	1	28		6	1		
j										1				47				
ŋ			1				2	6					18	12	7	1	1	
k						1			3	5	13	17	1	4		2	1	1
x					1	1	1	1	1		6	1		1	3	19	11	2
h		1			1		1	1					2	1	4	3	4	30

	i	I	e	ɛ	a	ɑ	œ	ɔ	ø	o	y	u
i	42	3	1								2	
I	2	44										2
e	3	1	43	1								
ɛ		2		44		1	1					
a				2	30	16						
ɑ					2	42		4				
œ			1			2	40	2	2		1	
ɔ								37		9		2
ø			2				9		31		6	
o					2					30		16
y							3				45	
u							1		2	1		44

condition S+F0

	p	b	m	f	v	ʋ	r	l	s	z	t	d	n	j	ŋ	k	χ	h
p	45	1								2								
b	2	39	7															
m			46							2								
f				32	11	2					3							
v				4	23	19	2											
ʋ				3	20	23		2										
r		2					46											
l	2							46										
s									15	8	14	4	1			5	1	
z					2	1			4	39								2
t				1					2	2	39	2	1			1		
d							6	1		3	1	31	1	3			2	
n	1						1	8	3	2		3	24	2	2	2		
j					2				9	20		1	6	4	3		2	1
ŋ							2	1	2			5	20	5	11			2
k	1			2					3	1	10	4		1	1	21	4	
χ		1				1					3		1			34	5	3
h								1			2			1		1	11	32

	i	I	e	ɛ	a	ɑ	œ	ɔ	ø	o	y	u
i	18	25	4	1								
I	11	32		5								
e	1	2	43	2								
ɛ				29	3	16						
a			2	1	42	3						
ɑ					1	46	1					
œ				1			39	4	1		1	2
ɔ						2	19	18	2	7		
ø			1				3	2	15	20	1	6
o							1	1	1	26	1	18
y							2	2		2	10	32
u							1	1			4	42

condition S+F0+A

	p	b	m	f	v	ʋ	r	l	s	z	t	d	n	j	ŋ	k	χ	h
p	47	1																
b	1	47																
m		2	46															
f				38	10													
v					31	17												
ʋ				2	24	22												
r							48											
l								48										
s	1								27	8	10	1				1		
z					1				7	38				2				
t								1	4	1	39	2					1	
d												48						
n		1						4	1	1		1	32	3	2		1	2
j						2			9	29	1	1		6				
ŋ						2		6	1	3		1	16	6	12			1
k					1				4	2	8					30	2	1
χ									1		3	1				36	5	2
h							1							5	5		8	29

	i	I	e	ɛ	a	ɑ	œ	ɔ	ø	o	y	u
i	16	26	4	1						1		
I	5	35	1	5		1						1
e			43	3			1					1
ɛ	2	4	3	25	3	10	1					
a	1	1	3		40	2						1
ɑ		1				45				2		
œ		1		1			39	3	1		1	2
ɔ			1			4	17	13	3	5	3	2
ø						1	3	1	13	24		6
o	1						1	1	2	29		14
y		1				3	4	1	2		8	29
u				1			4	2			5	36

condition S+SPI

	p	b	m	f	v	ʋ	r	l	s	z	t	d	n	j	ŋ	k	χ	h
p	37	8	1									2						
b	21	25										2						
m		5	41							2								
f				25	21					2								
v				14	27	3				2	1			1				
ʋ				2	14	30				2								
r							46											2
l								45		1					1			1
s						1			32	10	2			1		2		
z							3		19	24				2				
t							2		3	1	37	5						
d								2	3	6	4	32				1		
n		2		1				8	4	4		4	23	1				1
j				1	2			1	7	20	1	5		11				
ŋ						2			1	7	2		9	12	9		3	3
k		3							4	5	7	4		2	1	19	3	
χ		1	1				21			1			1	1	2	2	14	4
h					1	1			1				1		2		11	31

	i	I	e	ɛ	a	ɑ	œ	ɔ	ø	o	y	u
i	13	29	1	5								
I	6	27	4	11								
e	3		41	4								
ɛ	1	2	1	27	4	12		1				
a	1		3	1	42	1						
ɑ					1	45		2				
œ							39	8			1	
ɔ						2	6	27	2	6		5
ø							1	2	7	35		3
o									2	28	1	17
y				1			3	3	1	1	5	34
u							1	3		3	2	39

APPENDIX 5.B

Confusions among different types of manner of articulation for speechreading with each of the four auditory supplements. The stimuli are indicated vertically on the left side of the matrix, the responses horizontally above the matrix.

S+F1+F2:

	plos	nasa	fric	late	cont	roll
plos	179	16	37	1	6	1
nasa	5	108	2	14	12	3
fric	57	14	197	3	14	3
late	0	0	0	48	0	0
cont	0	2	18	0	76	0
roll	0	2	3	28	3	12

S+F0:

	plos	nasa	fric	late	cont	roll
plos	197	10	22	1	4	6
nasa	11	103	11	9	7	3
fric	67	2	192	1	24	2
late	2	0	0	46	0	0
cont	1	9	57	2	27	0
roll	2	0	0	0	0	46

S+F0+A:

	plos	nasa	fric	late	cont	roll
plos	223	0	16	1	0	0
nasa	5	108	10	10	11	0
fric	53	5	205	0	24	1
late	0	0	0	48	0	0
cont	2	0	64	0	30	0
roll	0	0	0	0	0	48

S+SPI:

	plos	nasa	fric	late	cont	roll
plos	207	2	25	2	2	2
nasa	13	82	26	8	15	0
fric	8	7	239	0	10	24
late	0	1	2	45	0	0
cont	6	0	48	1	41	0
roll	0	0	2	0	0	46

CHAPTER 6 - TOWARDS AN EFFICIENT SPEECHREADING AID FOR THE DEAF

The experiments presented in the previous chapters have provided insight into what information (extracted from the speech signal) is required for supplementing speechreading effectively. These experiments have shown that only two appropriately chosen speech parameters are sufficient for increasing the intelligibility of short sentences from about 25% correct syllables (inexperienced subjects) and 33% (experienced subjects) for speechreading-only to approximately 80 to 90% for speechreading with these parameters. These intelligibility scores are very likely to be sufficient for meaningful conversation in the real-life situation. The choice of parameters does not seem to be very critical; the scores on the sound-pressure levels in octave bands at 500 and 3160 Hz (S+SPI), on the fundamental-frequency plus amplitude information (S+F0+A), and on the formant-frequency information (S+F1+F2) lie within a range of approximately 10% (77.7 to 87.1%) for the inexperienced, and within approximately 3% (86.0 to 88.6%) for the experienced subjects.

In our experiments normal-hearing subjects were used, and the supplementary information was presented through the optimal channel, i.e. the auditory sense. Furthermore, the female speaker that pronounced the speech material had a clear pronunciation, her face was recorded under ideal (i.e. shadow-free) illumination, and the subjects always faced the video monitor from a short distance, in a quiet environment. In the real-life situation the circumstances will be less ideal: for the deaf person the supplementary information has to be transmitted through an alternative sense (which probably means a substantial loss of information), many speakers do not articulate clearly, and often the speaker can not be faced under ideal conditions. However, deaf people may be better speechreaders (on the average) than the subjects in our studies, in the real-life situation they can see more of the speaker than only head and shoulders, and in normal conversation the speech material will contain more contextual information than the sentences in our studies. These factors may partly compensate for the earlier mentioned less ideal situations.

At this moment one important question has to be answered: what further research is needed to develop an efficient speechreading aid for the deaf? As was discussed in Section 1.1, for successfully developing an efficient speechreading aid for the deaf two problems have to be solved: first, it should be clear what information is necessary for effectively supplementing speechreading; then, the problem has to be solved how this information can be optimally transmitted through an alternative sense. The first problem seems to be solved at least partly; our studies have shown that several speech parameters contain sufficient information for achieving a high sentence intelligibility in combination with speechreading. Now research needs to concentrate on the second problem.

First, it should be investigated which alternative senses can be used for transmitting the supplementary information. Several ways of presenting the supplementary information to the deaf person can be considered. Sometimes deaf persons have some residual hearing. In that case it can be attempted to present the supplementary information through this residual hearing. If not, the information may be presented via the tactile sense, via the visual sense, or via a cochlear implant.

Supplementary information can be presented via the visual sense; Cued Speech (Cornett, 1967) may serve as an example. However, the type of supplementary information that was investigated in our studies is completely different to the type that is supported in systems such as Cued Speech. In Cued Speech discrete cues are given about visually indiscriminable phonemes, while in our studies more continuous information about slowly varying acoustic parameters was given. Attempts to supplement speechreading with visually presented acoustic information are scarce, and have not been very successful. Therefore, the tactile sense or a cochlear implant probably are more suited for presenting the more continuous acoustic information of our studies.

Psychophysical research should clarify the information transmission capacities of the different alternative senses. Once these capacities are known, it should be decided what type of information fits best to which sense. Several factors have to be considered. Of course, the capacity of the transmission channel must be large enough for the information to be transmitted. This means that an estimation has to be made of the amount of information that is present in the supplementary speech parameters; the next paragraph will shortly go into the question of how to obtain this amount of information. This value has to be compared with the capacity of the alternative senses. However, a sufficiently large capacity is not enough for achieving speech perception: it also should be investigated

whether an appropriate stimulus-coding can be developed, whether extra information should be added (redundancy) to compensate for a possible loss of information during transmission, and whether the combination of the supplementary information and the chosen channel enables users to build up a stimulus memory and a processing strategy. During these stages of research it may appear that the best-scoring parameters of our studies are not the most appropriate ones for transmission through an alternative sense, e.g., because no appropriate stimulus-coding can be found.

The rate of information that is present in the supplementary speech parameters can only be determined by approximation. First, as an example, the rate of information of the original speech signal will be estimated. A very rough value can be obtained as follows. A conventional telephone channel has a bandwidth of about 3000 Hz. Thus, the speech signal can be represented by 6000 samples per second. If a 64-level (6-bit) quantization is applied, the rate of information is 36,000 bit/s. However, speech synthesis experiments have shown that, for example by using Linear Predictive Coding, intelligible speech with reasonable quality can be synthesized with only 1000-4000 bit/s. A lower bound on the information rate of the speech signal may be calculated by considering the speech signal simply as a stream of independent phonemes with equal probability of occurrence, i.e. aspects such as timbre of the speech sound, intonation and stress patterns are not taken into consideration. For example, for 30 different phonemes and a phoneme rate of 10/s the rate of information is 50 bit/s. However, the above calculations do not account for syntactic and semantic redundancies. Thus, the actual rate of information will be less than 50 bit/s.

The foregoing can also be applied to the supplementary speech parameters. The fluctuations of the sound-pressure levels in one or one-third octave filter bands (Experiment I) or of the overall sound-pressure level (Experiment III) have frequency components that lie below about 20-25 Hz. If again a 6-bit quantization is used the rate of information is roughly 240-300 bit/s. The fundamental frequency and the formant frequencies can also be represented by about 50 samples per second. Thus, with a 6-bit quantization these parameters also contain 300 bit/s. A lower bound on the information rate can be obtained by considering the results of the phoneme discrimination task of Experiment IV (Chapter 5). When the supplementary signals were presented without visual information percentage-correct phoneme-discrimination scores around 20% resulted (except for vowel discrimination on F1+F2). From the phoneme-confusion matrices the percentage of transmitted information (see Section 5.3.2) can be calculated. If each of the 30 phonemes (18 consonants and 12 vowels) is assigned an equal probabilty the percentage transmitted information ranges from 25 to 30%. Thus, the speech parameters contain approximately a quarter of the information of the original speech signal (50 bit/s). However, since also in this calculation syntactic and semantic redundancies were not considered, the actual rate of information will be lower. From studies of blind readers using the Optacon, a device that transforms letters into vibration patterns on the top of the index finger, it can be calculated that experienced readers can receive about 35 bit/s via the tactile sense. This rate is comparable to the rate of information of the supplementary speech parameters. Thus, on the basis of these calculations the capacity of the tactile sense is sufficiently large for transmitting at least one supplementary speech parameter.

The use of a cochlear implant seems to have advantages over the use of a tactile aid. First, psychophysical data suggest somewhat greater information transmission capacities. Comparison of the results on this topic suggests that frequency discrimination is somewhat better, and that for some deaf subjects temporal resolution is better for the cochlear implant (Burian, 1979; Eddington, 1978; Gescheider, 1967; Goff, 1967; Hirsch and Sherrick, 1961; Shannon, 1983). Second, for the postlingually deaf person the cochlear implant addresses the at least partly developed auditory memory, while for the use of a tactile aid a completely new tactual memory has to be developed. Third, of psychological importance is the fact that a cochlear implant enables a person to "hear" again. Fourth, the equipment (electronics, power supply, etc.) needed for a cochlear implant is smaller and less heavy than that for a tactile device, which probably makes the use of a cochlear implant more comfortable. At the moment, more than 400 deaf persons use cochlear implants (Sherrick, 1984), while almost no tactile speech-perception aids are used. However, as long as sufficient quantitative data about the tactile sense are not available, research into the use of this sense must continue.

SUMMARY

For the perception of speech the profoundly deaf person has to rely mainly on speechreading. However, the information that can be obtained by speechreading alone is very limited. Therefore, it has to be supplemented with extra information about the speech signal. It would be most convenient to have a speechreading aid, that automatically extracts the necessary supplementary information from the acoustic speech signal and presents this information to the deaf person through an alternative sense, e.g., through the visual or tactual sense or via a cochlear implant. For a successful development of an automatic speechreading aid two problems have to be solved. First, it has to be clarified which information about the speech signal is the most appropriate as supplement; second, it should be clarified how the supplementary information can be transmitted to an alternative sense most optimally. Very often these two problems were treated as being one and the same. However, confounding these problems makes it difficult to find the causes of the often observed disappointing results. They may lie in information extraction, in information transduction, or in both.

In the research reported in this dissertation the attention is focused on the first problem, namely on investigating which speech parameters contain the most effective information for supplementing speechreading. Parameters that are considered to be candidates are evaluated experimentally, by presenting them auditorily to normal-hearing listeners in combination with speechreading. The difference between speech intelligibility under the conditions of speechreading-only and speechreading in combination with listening to the supplementary speech information is a measure of the effectiveness of the parameters concerned. Three types of parameters were investigated: (1) the sound-pressure levels in one or two frequency bands filtered from the speech signal, (2) the frequencies of the first and second formants from voiced speech, (3) parameters that provide prosodic information, namely the combination of the fundamental frequency and the overall sound-pressure level, the fundamental frequency alone, the overall sound-pressure level alone, or only the duration of sequences of voiced sounds.

In Chapter 1 the problem, the method of-research, and the selected supplementary speech information are described. Furthermore, a review of

the literature on speech-perception aids for the deaf is given, with emphasis on the type of information that was presented.

Chapter 2 reports an experiment in which speechreading was supplemented with information about the sound-pressure levels in one or two frequency bands with center frequencies of 500, 1600, or 3160 Hz, and with one-third or one octave bandwidth, respectively. The sound-pressure information in two octave bands at 500 and 3160 Hz appeared to be the most effective supplement: 18 normal-hearing subjects with no experience in speechreading scored an average of 86.7% correct syllables (from short everyday sentences), speechreading-only scored 22.8% correct.

In Chapter 3 the results are given of an experiment in which speechreading was supplemented with information about the frequencies of the first and second formants from the voiced parts of the speech signal. Eighteen normal-hearing subjects with no experience in speechreading scored about 80% correct syllables for speechreading with this formant-frequency information, speechreading only scored again 22.8% correct.

Chapter 4 gives the results of an experiment in which speechreading was supplemented with prosodic information. Four types of prosodic information were investigated: (1) information about both the fundamental frequency and overall sound-pressure level, (2) information only about the fundamental frequency, (3) information only about the overall sound-pressure level, and (4) information about the duration of sequences of voiced sounds. Information about both the fundamental frequency and the overall sound-pressure level appeared to be the most effective supplement: 10 normal-hearing subjects with no experience in speechreading scored 64.0% correct syllables (speechreading-only scored 16.7% correct).

In the experiment reported in Chapter 5 the best-scoring supplements of the experiments reported in Chapter 2, 3, and 4 were compared. Both 24 inexperienced and 12 experienced normal-hearing speechreaders participated. Furthermore, in addition to the perception of sentences the discrimination of phonemes was investigated.

For the 24 inexperienced subjects information about the sound-pressure levels in octave bands at 500 and 3160 Hz appeared to be the most effective supplement for the perception of sentences (87.1% correct syllables). Speechreading-only scored 25.0% correct. For the 12 experienced subjects this type of information (86.1% correct) was equally

efficient as the formant-frequency information (88.6% correct) or information about the fundamental frequency and the overall sound-pressure level (86.0% correct). Speechreading-only scored 33.0% correct.

Discrimination of phonemes was measured only for the group of 24 inexperienced subjects. Discrimination of vowels was improved significantly only by supplementing speechreading with the formant-frequency information. All auditory supplements appeared to improve the discrimination of consonants. Analysis of the confusions among consonants showed that all signals gave a significant improvement in the perception of manner of articulation and voicing, and that place of articulation was not improved.

Finally, in Chapter 6 the question of what further research is needed for the development of an efficient speechreading aid is discussed. It is concluded that the research reported in this dissertation has provided insight into the question of what information is required for supplementing speechreading and that research now has to be focused on the question of how this information can be transmitted most optimally to an alternative sense.

SAMENVATTING

De volledig dove is voor zijn spraakperceptie grotendeels aangewezen op het spraakafzien (liplezen). De informatie die door middel van spraakafzien verkregen kan worden is echter beperkt. Het is daarom voor het verkrijgen van verstaanbaarheid noodzakelijk het spraakafzien aan te vullen met extra informatie over het gesprokene. De meest ideale situatie zou zijn die, waarbij de benodigde extra informatie automatisch uit het spraaksignaal wordt afgeleid en via een alternatief zintuig, bijvoorbeeld door tactiele of visuele stimulatie of via een electrische binnenoorprothese (cochlear implant), aan de dove wordt aangeboden. Om tot een succesvolle ontwikkeling van een dergelijke lipleeshulp te komen, dienen twee vragen beantwoord te worden: welke aanvullende informatie is nodig en via welke stimulatiemethode kan deze het meest optimaal overgebracht worden. In het verleden zijn deze twee vragen vaak behandeld als zijnde één probleem. Daardoor was het moeilijk de oorzaak van de vaak optredende teleurstellende resultaten aan te wijzen; enerzijds is het mogelijk dat niet de juiste informatie uit het spraaksignaal is afgeleid, anderzijds bestaat de mogelijkheid dat de afgeleide informatie niet op de juiste wijze aan de dove is aangeboden.

De doelstelling van het in deze dissertatie gerapporteerde onderzoek is het beantwoorden van de eerste vraag, nl. het bepalen van die uit het spraaksignaal afleidbare informatie die tezamen met het spraakafzien tot spraakverstaan leidt. Het onderzoek richt zich specifiek op het bepalen van uit het akoestische spraaksignaal afleidbare, langzaam in de tijd variërende parameters zoals de grondfrequentie of het geluiddrukniveau in bepaalde frequentiebanden. Deze parameters dienen dan later via een vervangend zintuig aan de dove aangeboden worden.

De verschillende in aanmerking komende parameters zijn op experimentele wijze geëvalueerd. Spraakmateriaal is aangeboden aan normaalhorende proefpersonen onder de condities van òf alleen spraakafzien òf spraakafzien in combinatie met luisteren naar een gereduceerd spraaksignaal, d.w.z. een auditief signaal dat gesynthetiseerd is door de betreffende parameters op een akoestische draaggolf te moduleren. Vergelijking van de verstaanbaarheidsscores bij deze condities gaf een indicatie van de effectiviteit van de aangeboden aanvullende informatie. Drie typen parameters zijn onderzocht: (1) de geluiddrukniveaus in één of twee uit de spraak gefilterde frequentiebanden, (2) de frequenties van de eerste en tweede formant van stemhebbende spraak, en (3) parameters die prosodische informatie verschaffen, nl. de combinatie van de grondfrequentie en het totale geluiddrukniveau, alleen de grondfrequentie, alleen het totale geluiddrukniveau, of de tijdsduur van stemhebbende segmenten.

In Hoofdstuk 1 worden de onderzoeksproblematiek, de onderzoeksmethode en de gekozen aanvullende informatie nader gespecificeerd. Tevens wordt een beknopt literatuuroverzicht gegeven van het tot nu toe verrichte onderzoek op het gebied van de overdracht van spraak via een alternatief zintuig. De nadruk ligt daarbij op welke uit de spraak afgeleide informatie is gebruikt.

Hoofdstuk 2 behandelt een experiment waarin het spraakafzien werd aangevuld met informatie over het geluiddrukniveau in één of twee uit de spraak gefilterde frequentiebanden met centrale frequenties van 500, 1600 of 3160 Hz en met bandbreedten van 1 of 1/3 octaaf. Uit dit experiment bleek dat de geluiddrukinformatie in octaafbanden rond 500 en 3160 Hz een zeer effectieve aanvulling op het spraakafzien vormt: 18 normaalhorende proefpersonen zonder ervaring in het spraakafzien verstonden gemiddeld 86.7% lettergrepen correct (uit korte zinnen) via spraakafzien aangevuld met deze informatie; spraakafzien alleen scoorde slechts 22.8% lettergrepen correct.

In Hoofdstuk 3 wordt verslag gedaan van een experiment waarin het spraakafzien werd aangevuld met informatie over de frequenties van de eerste en tweede formant van de stemhebbende segmenten van het spraaksignaal. Achttien normaalhorende proefpersonen zonder ervaring in het spraakafzien scoorden rond de 80% correcte lettergrepen voor spraakafzien met deze informatie, terwijl spraakafzien alleen weer 22.8% scoorde.

Hoofdstuk 4 behandelt een experiment waarin het spraakafzien werd aangevuld met informatie over de prosodie van de spraak (lengte van klanken, intonatie en nadruk). Vier verschillende prosodische signalen werden gebruikt: (1) informatie over zowel de grondfrequentie als het totale geluiddrukniveau van het spraaksignaal, (2) informatie over alleen de grondfrequentie, (3) informatie over alleen het totale geluiddrukniveau of (4) informatie over de tijdsduur van stemhebbende spraaksegmenten. De informatie over de combinatie van de grondfrequentie en het totale geluiddrukniveau bleek de beste aanvulling: gemiddeld scoorden tien normaalhorende proefpersonen zonder ervaring in het spraakafzien 64.0% lettergrepen correct (spraakafzien alleen scoorde 16.7%).

Uit de experimenten zoals vermeld in de Hoofdstukken 2, 3 en 4 bleek dat verschillende typen informatie geschikt zijn om het spraakafzien aan te vullen. In Hoofdstuk 5 wordt verslag gedaan van een experiment waarin de beste signalen uit de vorige hoofdstukken met elkaar vergeleken werden. De proefpersonen die deelnamen in de experimenten in de Hoofdstukken 2, 3 en 4 hadden geen ervaring in het spraakafzien. In het in Hoofdstuk 5 beschreven experiment hebben zowel 24 proefpersonen zonder ervaring als 12 personen met ervaring in het spraakafzien deelgenomen (allen hadden normaal gehoor). Verder is behalve de perceptie van zinnen ook de discriminatie van klinkers en medeklinkers onderzocht. Een analyse is gemaakt van welke eigenschappen (zoals stemhebbendheid, friktie, nasaliteit) met de verschillende typen informatie waargenomen kunnen worden.

Voor de 24 proefpersonen zonder ervaring in het spraakafzien bleek de informatie uit de octaafbanden rond 500 en 3160 Hz de meest effectieve aanvulling voor de perceptie van zinnen. Spraakafzien alleen scoorde gemiddeld 25.0% lettergrepen correct, spraakafzien aangevuld met deze informatie 87.1%. Voor de 12 proefpersonen met ervaring bleek deze geluiddrukinformatie (86.1% correct) een even effectieve aanvulling te zijn als de informatie over de formantfrequenties (88.6% correct) en de informatie over de grondfrequentie en het totale geluiddrukniveau (86.0% correct). Spraakafzien alleen scoorde 33.0% lettergrepen correct.

De discriminatie van fonemen is alleen gemeten voor proefpersonen zonder ervaring in het spraakafzien. Alleen aanvullende informatie over de formantfrequenties bleek de discriminatie van klinkers te verbeteren; spraakafzien met deze informatie scoorde 81.9% correct, spraakafzien alleen 55.0% correct. Alle aanvullingen bleken de discriminatie van medeklinkers te verbeteren, informatie over de grondfrequentie en het totale geluiddrukniveau bleek de meest effectieve aanvulling (spraakafzien alleen: 38.0% correct, spraakafzien met deze informatie: 68.6% correct). De andere aanvullingen scoorden 50-60% correct. Een nadere analyse van de verwarringen tussen medeklinkers liet zien dat deze verbeteringen hoofdzakelijk te danken zijn aan een verbeterde waarneming van de manier van articulatie (zoals nasaliteit en plosiviteit) en van het al of niet stemhebbend zijn van de medeklinkers.

Tenslotte worden in Hoofdstuk 6 de resultaten nog eens kort besproken en wordt ingegaan op de vraag welk vervolgonderzoek nodig is om tot de ontwikkeling van een effectieve lipleeshulp voor de totaal dove te komen. Er wordt geconcludeerd dat het in deze dissertatie vermelde onderzoek duidelijkheid heeft verschaft over welke informatie nodig is om het spraakafzien aan te vullen en dat de aandacht nu gericht dient te worden op de vraag hoe deze informatie het best aan de dove aangeboden kan worden. Onderzocht dient te worden welke informatie het best past bij welk zintuig en hoe dit zintuig het meest optimaal gestimuleerd kan worden.

REFERENCES

Atal, B.S., and Hanauer, S.L. (1971). "Speech analysis and synthesis by linear prediction of the speech wave," J. Acoust. Soc. Am. 50, 637-655.

Benedetto, M.D. Di, Destombes, F., Merialdo, B., and Tubach, J. (1982). "Phonetic recognition to assist lip-reading for deaf children," Proc. ICASSP82, IEEE Internal Confer. on Acoustics, Speech and Signal Processing, 2, 739-742.

Berger, K.W. (1972). "Three experiments in speechreading," Journal of Communication Disorders 5, 280-285.

Berliner, K.I. and House, W.F. (1982). "The cochlear implant program: an overview," Ann. Otol. Rhinol. Laryng. suppl. 91, 11-14.

Binnie, C.A., Montgomery, A.A., and Jackson, P.L. (1974). "Auditory and visual contributions to the perception of consonants," J. Speech Hear. Res. 17, 619-630.

Blesser, B. (1972). "Speech perception under conditions of spectral transformation: I. Phonetic characteristics," J. Speech Hear. Res. 15, 5-41.

Blumstein, S.E., Isaacs, E., and Mertus, J. (1982). "The role of gross spectral shape as a perceptual cue to place of articulation in initial stop consonants," J. Acoust. Soc. Am. 72, 43-50.

Breeuwer, M., and Plomp, R. (1984). "Speechreading supplemented with frequency-selective sound-pressure information," J. Acoust. Soc. Am. 76, 686-691.

Breeuwer, M., and Plomp, R. (1985). "Speechreading supplemented with formant-frequency information from voiced speech," J. Acoust. Soc. Am. 77, 314-317.

Burian, K., Hochmair, E., Hochmair-Desoyer, I., and Lessel, M.R. (1979). "Designing of and experience with multichannel cochlear implants," Acta. Otol. 87, 190-195.

Clark, G.M., Patrick, J.F., and Bailey, Q. (1979). "A cochlear implant round window electrode array," J. Laryng. Otol. 93, 107-109.

Clark, G.M., Tong, Y.C., Martin, L.F.A., and Busby, P.A. (1981). "A multiple-channel cochlear implant," Acta Otol. 91, 173-175.

Clements, M.A., Braida, L.D., and Durlach, N.I. (1982). "Tactile communication of speech: II. Comparison of two spectral displays in a vowel discrimination task," J. Acoust. Soc. Am. 72, 1131-1135.

Cole, R.A. and Zue, V.W. (1980). "Speech as eyes see it," in: Attention and Performance VIII, Ed. by R.S. Nickerson, New Jersey.

Cooper, F.S., Delattre, P.C., Liberman, A.M., Borst, J.M., and Gerstman, L.J. (1952). "Some experiments on the perception of synthetic speech sounds," J. Acoust. Soc. Am. 24, 597-606.

Cornett, R.O. (1967). "Cued Speech," American Annals of the Deaf 112, 3-13.

Cornett, R.O., Beadles, R., and Wilson, B. (1977). "Automatic cued speech," Paper from the Research Conference on Speech-Processing Aids for the Deaf, 224-239.

Danhauer, J.L., and Appel, M.A. (1976). "INDSCAL analysis of perceptual judgements for 24 consonants via visual, tactile and visual-tactile inputs," Journal of Speech and Hearing Research 19, 68-77.

Danley, M.J. and Fretz, R.J. (1982). "Design and functioning of the single-electrode cochlear implant," Ann. Otol. Rhinol. Laryng. suppl. 91, 21-26.

DeFilippo, C.L., and Scott, B.L. (1978). "A method for training and evaluating the reception of ongoing speech," J. Acoust. Soc. Am. 63, 1186-1192.

DeFilippo, C.L. (1984). "Laboratory projects in tactile aids to lipreading," Ear and Hearing 5, 211-227.

Eddington, D.K., Dobelle, W.H., Brackmann, D.E., Mladejovsky, M.G., and Parkin, J.L. (1978). "Auditory prosthesis research with multiple channel intracochlear stimulation in man," Ann. Otol. Rhinol. Laryng. Suppl. 53, 5-39.

Eddington, D.K. (1980). "Speech discrimination in deaf subjects with cochlear implants," J. Acoust. Soc. Am. 68, 885-891.

Erber, N.P. (1972). "Speech-envelope cues as an acoustic aid to lipreading for profoundly deaf children," J. Acoust. Soc. Am. 51, 1224-1227.

Erber, N.P. (1974). "Visual perception of speech by deaf children," Scand Audiol. suppl. 4, 97-113.

Flanagan, J.L. (1972). Speech analysis, synthesis and perception. Springer Verlag, New York.

Fourcin, A.J., and Rosen, S.M. (1979). "External electrical stimulation of the cochlea: clinical, psychophysical, speech-perceptual and histological findings," Brit. J. Audiol. 13, 85-107.

Gault, R.H. (1926). "Touch as a substitute for hearing in the interpretation and control of speech," Arch. Otolaryngol. 3, 121-135.

Gengel, R.W. (1976). "Research with Uptons visual speechreading aid," Report Dep. of Communication Disorders, University of Massachusetts.

Gengel, R.W. (1977). "Research with Uptons visual speechreading aid," Paper from the Research Conference on Speech-Processing Aids for the Deaf, 218-223.

Gescheider, G.A. (1967). "Auditory and cutaneous temporal resolution of successive brief stimuli," J. Exp. Psychol. 75, 570-572.

Goff, G.D. (1967). "Differential discrimination of frequency of cutaneous mechanical vibration," J. Exp. Psychol. 74, 294-299.

Green, B., Craig, J.C., Wilson, A.M., Pisoni, D.B., and Rhodes, R.P. (1983). "Vibrotactile identification of vowels," J. Acoust. Soc. Am. 73, 1766-1778.

Hirsch, I.J., and Sherrick, C.E. (1961). "Perceived order in different sense modalities," J. Exp. Psychol. 64, 423-432.

Hochmair-Desoyer, I.J., Hochmair, E.S., and Stiglbrunner, H.K. (1983). "Psychoacoustic temporal processing and speech understanding in cochlear implant patients," Paper presented at the 10-th Anniversary Conference on Cochlear Implants, San Francisco.

House, W.F. (1982). "Surgical considerations in cochlear implantation," Ann. Otol. Rhinol. Laryng. suppl. 91, 15-20.

Kirman, J.H. (1973). "Tactile communication of speech: a review and an analysis," Psychological Bulletin 80, 54-74.

Kirman, J.H. (1974). "Tactile perception of computer-derived formant patterns from voiced speech," J. Acoust. Soc. Am. 55, 163-169.

Klatt, D.H. (1976). "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence," J. Acoust. Soc. Am. 59, 1208-1221.

Klatt, D.H. (1980). "Software for a cascade/parallel formant synthesizer," J. Acoust. Soc. Am. 67, 971-995.

Kringlebotn, M. (1968). "Experiments with some visual and vibrotactile aids for the deaf," Amer. Ann. Deaf 113, 311-317.

Lamoré, P.J.J., Verweij, C., and Brocaar, M.P. (1980). "The sensitivity of the tactile sensory system to amplitude variations," Abstract Proc. 15-th Int. Congres of Audiology, 107.

Lehiste, I. (1970). Suprasegmentals. M.I.T. press, Cambridge, Massachusetts and Londen.

Lowell, E.L. (1974). "Perceptibility of vocalic nuclei," Scand. Audiol. Suppl. 4, 136-152.

Markel, J.D., and Gray, A.H. (1976). Linear prediction of speech. Springer Verlag, New York.

Martony, J. (1974). "On lipreading with visual and tactual lipreading aids," Scand. Audiol. Suppl. 4, 114-127.

Millar, J.B., Tong, Y.C., and Clark, G.M. (1984). "Speech processing for cochlear implant prostheses," J. Speech Hear. Res. 27, 280-296.

Miller, G.A., and Nicely, P.E. (1955). "An analysis of perceptual confusions among some English consonants," J. Acoust. Soc. Am. 27, 338-352.

Moore, B.C.J., Douek, E., Fourcin, A.J., Rosen, S.M., Walliker, J.R., Howard, D.M., Abberton, E., and Frampton, S. (1984). "Extracochlear electrical stimulation with speech patterns: experience of the EPI group (UK)," Adv. Audiol. 2, 148-162.

Oller, D.K., Payne, S.L., and Gavin, W.J. (1980). "Tactual speech perception by minimally trained deaf subjects," J. Speech Hear. Res. 23, 769-778.

Pialoux, P., Chouard, C.M., Meyer, B., and Fugain, C. (1979). "Indications and results of the multichannel cochlear implant," Acta Otol. 87, 185-189.

Pickett, J.M. (1963). "Tactual Communication of speech sounds to the deaf: Comparison with lipreading," J. Speech Hear. Disorders 28, 315-330.

Plant, G.L. (1982). "Tactile perception by the profoundly deaf," British Journal of Audiology 16, 233-244.

Plant, G.L., and Risberg, A. (1983). "The transmission of fundamental frequency variations via a single channel vibrotactile aid," STL-QPRS 2-3, 61-84.

Plomp, R. (1978). "Auditory handicap of hearing impairment and the limited benefit of hearing aids," J. Acoust. Soc. Am. 63, 533-549.

Plomp, R., and Mimpen, A.M. (1979). "Improving the reliability of testing the speech reception threshold for sentences," Audiology 18, 43-52.

Pols, L.C.W., Tromp, H.R.C., and Plomp, R. (1973). "Frequency analysis of Dutch vowels from 50 male speakers," J. Acoust. Soc. Am. 53, 1093-1101.

Potter, R., Kopp, G., and Green, H. (1947). Visible speech. New York: Van Nostrand.

Remez, R.E., Rubin, P.E., Pisoni, D.B., and Carrell, T.D. (1981). "Speech perception without traditional speech cues," Science 212, 947-950.

Risberg, A. (1974). "The importance of prosodic speech elements for the lipreader," Scand. Audiol. Suppl. 4, 153-164.

Risberg, A., and Lubker, J.L. (1978). "Prosody and speechreading," Report nr. STL-QPSR 4, Dept. of Linguistics, University of Stockholm, 1-16.

Rosen, S.M., Moore, B.C.J., and Fourcin, A.J. (1979). "Lipreading with fundamental frequency information," Proc. of the Institute of Acoustics, paper 1A2, 5-8.

Rosen, S.M., Fourcin, A.J., and Moore, B.C.J. (1981). "Voice pitch as an aid to lipreading," Nature 291, 250-252.

Rothenberg, M., Verillo, R.T., Zahorian, S.A., Brackman, M.L., and Bolanowski, S.J. (1977). "Vibrotactile frequency for encoding a speech parameter," J. Acoust. Soc. Am. 62, 1003-1012.

Rothenberg, M., and Molitot, R.D. (1979). "Encoding voice fundamental frequency into vibrotactile frequency," J. Acoust. Soc. Am. 66, 1029-1038.

Saunders, F.A., Hill, W.A., and Simpson, C.A. (1976). "Speech perception via the tactile mode: Progress report," IEEE Internat. Conf. on Acoust. Speech and Signal Process. 1976, 594-597.

Shannon, R.V. (1983). "Multichannel electrical stimulation of the auditory nerve in man. I. Basic psychophysics," Hear. Research 11, 157-189.

Sherrick, C.E. (1984). "Basic and applied research on tactile aids for deaf people: progress and prospects," J. Acoust. Soc. Am. 75, 1325-1342.

Sparks, D.W., Kuhl, P.K., Edmonds, A.E., and Gray, G.P. (1978). "Investigating the MESA (Multipoint Electrotactile Speech Aid): The transmission of segmental features of speech," J. Acoust. Soc. Am. 63, 246-257.

Sparks, D.W., Ardell, L.A., Bourgeois, M., Wiedmer, B., and Kuhl, P.K. (1979). "Investigating the MESA (Multipoint Electrotactile Speech Aid): The transmission of connected discourse," J. Acoust. Soc. Am. 65, 810-815.

Thielemeir, M.A., Brimacombe, J.A., and Eisenberg, L.S. (1982). "Audiological results with the cochlear implant," Ann. Otol. Rhinol. Laryng. suppl. 91, 27-34.

Tong, Y.C., Clark, G.M., Dowell, R.C., Martin, L.F.A., Seligman, P.M., and Patrick, J.F. (1981). "A multiple-channel cochlear implant and wearable speech-processor," Acta Otol. 92, 193-198.

Traunmüller, H. (1975). "Lippenablesehilfe fur Gehorlose: visuelle oder taktile Darbietung von Erganzungsinformation," Report nr. STL-QPSR 4, Dept of Linguistics, University of Stockholm, 27-34.

Upton, H.W. (1968). "Wearable eyeglass speechreading aid," Amer. Ann. Deaf 113, 222-229.

Vogten, L.M.M. (1983). Analyse, zuinige codering en resynthese van spraakgeluid. Dissertatie Technische Hogeschool Eindhoven.

Voiers, W.D. (1973). "Experimental investigation of the consonant information structure of the visible speech signal," in: Symp. Speech Intelligibility Liège, 325-334.

White, R.L. (1982). "Review of current state of cochlear prostheses," IEEE Trans. on Biom. Eng. BME-29, 233-238.

Woodward, G.A., and Barber, C.G. (1960). "Phoneme perception in lipreading," J. Speech Hear. Res. 3, 212-222.