SPEECH INTELLIGIBILITY AND SOUND QUALITY IN HEARING AIDS

Settings and Characteristics
of Hearing Aids
Evaluated Psychoacoustically

Ronald A. van Buuren

Offset: Drukkerij Elinkwijk bv, Utrecht

VRIJE UNIVERSITEIT

SPEECH INTELLIGIBILITY AND SOUND QUALITY IN HEARING AIDS

Settings and Characteristics
of Hearing Aids
Evaluated Psychoacoustically

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor aan
de Vrije Universiteit te Amsterdam,
op gezag van de rector magnificus
prof.dr E. Boeker,
in het openbaar te verdedigen
ten overstaan van de promotiecommissie
van de faculteit der geneeskunde
op vrijdag 16 mei 1997 om 13.45 uur
in het hoofdgebouw van de universiteit,
De Boelelaan 1105

door

Ronald Alexander van Buuren

geboren te Jutphaas

Promotor: prof.dr.ir. T. Houtgast
Copromotor: dr.ir. J.M. Festen

Aan de beoordeling van dit proefschrift is door prof.dr.ir. W.A. Dreschler, verbonden aan de Universiteit van Amsterdam, een belangrijke bijdrage geleverd.

Het in dit proefschrift beschreven onderzoek en de uitgave van dit proefschrift werden financieel ondersteund door Philips Hearing Instruments.

aan mijn ouders

zonder jullie steun was dit nooit gelukt

Contents

Chapter 1. Introduction

To all those new to the research of hearing

Hearing has fascinated mankind since what can be considered to be the onset of science, the ancient Greek civilisation. We know that already Pythagoras noticed that strings that he thought to "sound well" when played in combination were characterised by simple length-ratios, a phenomenon now known as *consonance*. Alternatively, adjacent keys (e.g., on a piano keyboard), when struck at the same time, will produce *dissonance*. The same argument that Pythagoras used is valid here, since the lengths of "adjacent" piano strings are not related by simple ratios. In Pythagoras' days, however, and for a long time afterwards, such rational explanations for what was subjectively experienced were not always welcome. Plato, for example, is known to have characterised Pythagoras' observations as 'torture of strings'.

Importance

An important ability of any new-born human being is the use of the hearing organ. Without hearing, speech will hardly develop and education will be very difficult, which is one of the reasons that *hearing impairment* should be diagnosed at as early an age as possible. In addition to educational problems, there may be biological consequences of hearing impairment in new-borns. It has recently become clear from animal experiments that when the hearing functions are disturbed at a very young age, certain neural developments in the brain are different or do not take place at all (e.g. Harrison *et al.*, 1993). It is not clear whether the neural developments may still occur at a later age, adding weight to the importance of diagnosing hearing impairment at an early age. However, the biological aspects of hearing research, which are the subject of *physiology*, are *not* covered in this thesis.

A much larger part of the population will face hearing impairment at a later age. The occurrence of hearing impairment is 7%, averaged for the whole Dutch population. By comparison, only 1 out of every 1000 new-borns is estimated to have a hearing impairment (Kapteyn, 1994). A substantial part of the hearing-impaired people cannot do without

prosthetic devices: *hearing aids*. It is with this device in mind that the research which is described in this thesis was formulated.

Sound

At this point, it is useful to introduce some terms that are central to hearing research. This may be easiest when we think of a specific *sound*, for which we will adopt the 'simple tone'. A simple tone can be produced by striking a tuning fork; the sound of a recorder also closely resembles a simple tone. Physically, a simple tone can be characterised by its *amplitude*, which relates to our sense of loudness, and by its *frequency*, which relates to what we experience as pitch.
Mathematically, a simple tone can be described by means of a *sinusoidal* function (see Figure 1).

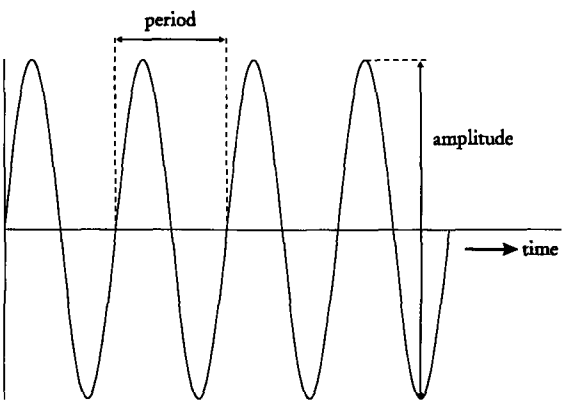

Figure 1. Sinusoid

The sinusoidal function can be regarded as an infinite repetition of identical building blocks with a finite length in time. This length is called the *period* of the sinusoid. The shorter the period, the more periods will fit into, say, one second. The number of periods that fit into one second is called the *frequency* of the sinusoid; it is expressed in cycles (periods) per second (s^{-1} or Hz).
Sound will only travel in media such as air, water etc. Unlike light, it cannot travel in a vacuum (e.g., outside the earth's atmosphere). Physically, sound is a variation of the *density* or *pressure* in a medium. Since sound is based on *variations* in pressure, there is no resultant transport in the medium. For the simple tone, the regularity of these variations corresponds to the frequency of the tone, whilst the amount by which the pressure varies (with respect to an average) corresponds

to its amplitude. A sound source may introduce these variations in pressure by moving parts that are in direct contact with the medium, such as the strings and the wooden body of a violin, which are surrounded by air. Reversely, the pressure variations themselves can induce motions in other structures, such as the membrane in a microphone.

Hearing

From the anatomical point of view, the hearing organ can be described quite accurately. Figure 2 shows what would become visible in an appropriate cross section of the human head. Coming from outside, sound enters the ear canal and induces vibrations in the tympanic membrane. The vibrations are transferred to the inner ear by the three middle-ear ossicles ('ossicle' is Latin for 'small bone'). Inside the inner ear, the motion of the ossicles is converted to electrical stimuli that are sent into the auditory nerve. The signals travel through the auditory nerve to the brain, where the higher centres of the auditory system are located. It is there that the actual experience of sound is believed to take place.

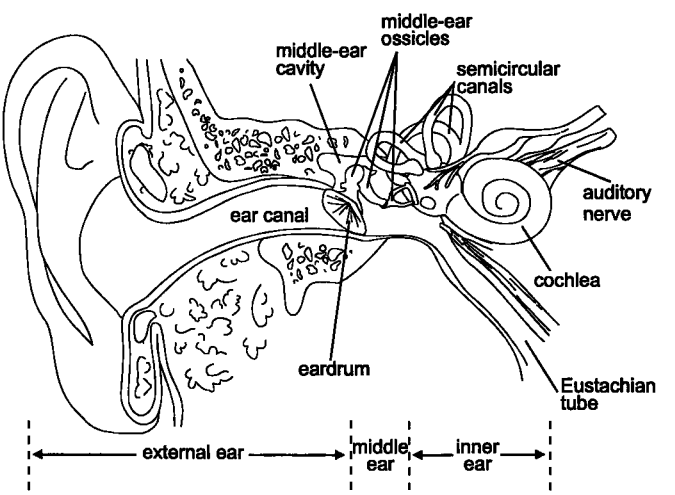

Figure 2. Anatomical structure of the human ear.

When we look at this route, from air-pressure variations outside through the middle and inner ear to the experience of sound in the higher auditory centres in the brain, then it seems that the further we get along the route, the less we understand of what really happens. The

function of the pinna, the ear canal, and the middle-ear ossicles can be well understood using elementary acoustics and mechanics. The pinna performs a 'spatial selection', which is a filtering operation (preserving some frequencies and attenuating others) that depends on the direction from which sound reaches the pinna. The ear canal subsequently performs another filtering operation. At the end of the ear canal, the tympanic membrane and the one ossicle attached to it respond to the sound by vibrating. The motions then travel through the chain of middle-ear ossicles to the oval window, which is the location where the third ossicle is flexibly connected to the inner ear. The middle-ear ossicles increase the amplitude of the tympanic membrane's motion by a frequency-dependent factor of maximally 26 (Pickles, 1982). As a result, the minimum vibrational displacement of the eardrum that will still produce audible results is about 0.03 nm (at a frequency of 3 kHz; see Moore, 1982). This displacement roughly corresponds to the diameter of the hydrogen atom, which illustrates the very high sensitivity of the (healthy) ear.

What exactly occurs inside the inner ear is still subject to debate, and a lot of research is currently being done to clarify this. Roughly speaking, we may regard the inner ear as a *frequency separator*. The inner ear sends electrical activity into certain nerve fibres depending on the frequency content of the sound. For a simple tone with a high frequency, most electrical activity is sent into other fibres than for a simple tone with a low frequency. Further, it appears that, to a first-order approximation, the level of the sound determines the amount of electricity sent into a nerve fibre; the higher the sound level, the more electrical activity it will receive. This information is further processed in the higher auditory centres in the brain, where it is transformed into what we know as the *experience* of sound.

Impaired hearing; Types & solutions

Until not very long ago, hearing research was concentrated on the normally functioning ear. To some extent, this may have been due to the lack of understanding of the healthy hearing system, without which the analysis of impaired hearing may be difficult. However, from the 1970s a great number of scientific publications has reported about which aspects of hearing are affected by common diseases such as middle-ear inflammation (otitis media) and Menière's Disease. Another well-known cause of deterioration of the hearing capabilities is the exposition to excessively high sound levels. Research in this area has led

to specific sound absorbers for various noisy situations, which will prevent permanent damage to hearing. Since ageing is among the best-known 'causes' of impaired hearing, it has received great attention in hearing research. Even without disease or excessive noise exposure, hearing will gradually deteriorate with increasing age, although severity and age of onset vary among the ageing population. With maximum age increasing, especially in the Western countries, hearing impairment and its compensation may very well become of even greater importance in the future.

From the diagnostic point of view, hearing impairment may be subdivided into two types; *conductive* and *sensorineural* hearing loss. A conductive hearing loss is characterised by a less efficient conduction of sound, and it often occurs with a middle-ear inflammation. Thus, it is generally a matter of time for the hearing loss to disappear. However, repeated middle-ear inflammations can cause the bones in the middle ear to become less flexible or even disconnected, as a consequence of which the hearing loss becomes permanent. Such problems may be solved by reconstructing the (flexibility of the) chain of bones.

The type considered in this thesis, *sensorineural hearing impairment*, is one of the most prevalent in the ageing population. As may be deduced from the term, sensorineural hearing impairment finds its causes in the sensor (i.e., the inner ear) and/or in the neurone (the auditory nerve). This type of hearing impairment is sometimes accompanied by "sounds inside the head" (tinnitus), often described as whistling or hissing, which may be extremely annoying to those suffering from it. Sensorineural hearing loss is characterised by a reduced sensitivity, the reduction generally being greatest at the highest audible frequencies. Although sensitivity is lower, loud sound will be unpleasant at about the same level as for normal hearing. In other words, the loudness range from "just audible" to "too loud" is reduced with regard to normal hearing; this is called *recruitment*. Unlike in some cases of conductive hearing loss, there are presently no operative techniques for alleviating sensorineural hearing loss.

Hearing impairment may be compensated for in several ways. In the case of a mild impairment, a hand behind the ear, or moving to a quieter surrounding (i.e., with less disturbing sound), may suffice. Many people with impaired hearing rely on watching the lips of the person they are listening to (speechreading), which is a very effective supplement to hearing (e.g. Breeuwer, 1986). Of course, this ability can be trained (and it often is) for the further enhancement of a listener's

speechreading capabilities. It will evidently not work in a situation where the talker's lips are not visible, for example in a telephone conversation or when listening to a radio.

In the case of more severe hearing impairment for which an operation does not seem appropriate (e.g., the pathology is of the sensorineural type), a technical compensation will be of great help, although it will hardly ever compensate for all aspects of hearing impairment. Such a solution usually takes the form of a *hearing aid*, which is essentially a miniaturised combination of a microphone, an amplifier and a loudspeaker. A more severe impairment requires a greater amount of amplification, which translates into a relatively large amplifier about the size of a credit card (though a bit thicker); a special small loudspeaker in the outer ear (usually in the *concha*; see Figure 2) is cable-connected to the *body-worn* amplifier-microphone combination. The moderate hearing losses, requiring less bulky amplifiers, can be compensated with a hearing aid worn *behind* the ear (the well-known "banana"-shaped versions), in the ear (at the entrance of the ear canal), or even *inside* the ear canal. The latter is called an "in the canal" (ITC) hearing aid, which has recently been further miniaturised into a variant that sits close to the tympanic membrane, the *peritympanic* hearing aid. The peritympanic hearing aid is completely invisible from outside. It is therefore very attractive to all those people with mild hearing impairments who would like to use a hearing aid, but are afraid of being stigmatised as "aged" in case they wear a more visible hearing aid.

There is a category of very severe impairments for which even the most powerful of body-worn hearing aids is not sufficient. In some cases, it is feasible to operate upon such people and implant a special transducer inside the inner ear (alongside the auditory nerve), which is linked to a transmitter outside the head containing a microphone and special signal-processing electronics. The implanted apparatus will stimulate the auditory nerve directly; it is intended as an artificial inner ear (cochlea). To a much greater extent than with conventional hearing aids, listeners with such *cochlear implants* will have to re-educate themselves in hearing, since the sounds they receive through the implant are radically different from what the healthy ear transmits.

What this thesis is about

Since purely conductive hearing loss is a mere attenuation of sound, amplification by means of a hearing aid will in fact be a perfect compensation of this type of hearing loss. The other diagnostic type of

hearing impairment, the sensorineural variant, will generally be compensated only to a certain extent when a hearing aid is used. Several questions are considered in this thesis, related to sensorineural hearing impairment, to the present knowledge of it, and to hearing aids, to hopefully fill gaps in our understanding of this hearing handicap and to further enhance its compensation.

Now that you're familiar with the research of hearing

In everyday situations, a person with sensorineural hearing impairment will not only experience difficulties in *hearing* (as a consequence of the reduced sensitivity of the hearing system), but also in *discriminating* sounds. The latter phenomenon has been demonstrated in many experiments where speech was presented in combination with noise or other disturbing sounds (e.g., another speaker). When such experiments are carried out for listeners with sensorineural hearing impairment, it turns out that they have more difficulties to understand speech than listeners with normal hearing, even when the reduced sensitivity of the impaired listener's hearing is compensated for.

A descriptive model for speech intelligibility in noise

Plomp (1978) reviewed a great number of such speech-in-noise experiments, on which he based a model which distinguishes two components in sensorineural hearing loss; a *sensitivity* component (attenuation) and a *discrimination* component (distortion). He argues that a hearing aid, being an amplifier, will compensate for attenuation but not for distortion.

Plomp developed a special speech test which he described one year later (Plomp & Mimpen, 1979). In the test, which is one of the tests applied in the experiments described in this thesis, speech and speech-shaped noise are mixed and presented to the listener. The speech material consists of short (eight to nine syllables) everyday Dutch sentences, grouped in lists of thirteen sentences. The speech-shaped noise is obtained by filtering Gaussian noise according to the long-term average frequency spectrum of the speech. Listeners subjected to the test are asked to repeat the sentences as accurately as they can. The S/N ratio is varied according to an adaptive procedure (generally known as a 'simple up-down procedure') and converges to the S/N ratio at which 50% of the sentences in the list is reproduced correctly: the speech-reception threshold (SRT). Not surprisingly, the test became known as the "SRT test".

When the SRT test is carried out for listeners with normal hearing, the resulting SRT is about -6 dB (Plomp & Mimpen, 1979), when the signals are monaurally presented through headphones. The SRT turns out to be independent of the presentation level of the speech-and-noise combination, for the range of sound levels encountered in everyday life. For listeners with sensorineural hearing impairment, the SRT is higher than -6 dB, even when their reduced sensitivity is compensated for by a higher presentation level. This confirms that they have more trouble in separating speech from noise. The higher the SRT for a listener with hearing impairment, the more distortion is apparently present in the inner ear.

Speech intelligibility and sound quality

Sensorineural hearing impairment very often causes the greatest loss of sensitivity at high frequencies, roughly above 1 kHz. Since hearing impairment develops over the years, people suffering from it will have become accustomed to not hearing those frequencies. With the purpose of restoring the hearing capabilities, an audiologist helping such a person will select a hearing aid that provides ample amplification in the area of reduced sensitivity. One can imagine what happens when this hearing aid is first switched on in the patient's ear: there will be loud complaints about the sharp "sound" of the hearing aid, even though the now aided listener will be better able to interpret speech.

This apparent contradiction between the optimum hearing-aid setting for speech intelligibility (e.g., ample high-frequency amplification) and for sound quality (e.g., little high-frequency amplification) has been one of the motivations for writing this thesis. As a consequence, the effects of settings or characteristics of hearing aids on speech intelligibility and on sound quality had to be measured.

Speech intelligibility was evaluated with a slightly adapted version of the SRT test described previously; instead of thirteen sentences, a lower number of sentences per list was used to accomodate the number of experimental conditions. Since Plomp equalised intelligibility per sentence, the principal effect of using less than the original number of sentences per list was expected to be a somewhat lower reliability of each individual SRT (Plomp & Mimpen, 1979). Additionally, shorter lists will have a larger variation in the frequency of occurrence of each phoneme. Both effects are compensated by (a) the relatively large number of listeners in each experiment and (b) the assignment, for different listeners, of different lists to the same experimental condition.

Sound quality was measured through judgements of speech and music, using both magnitude estimations (on a five-point rating scale) and paired comparisons.

The first question, which is considered in Chapter 2 of this thesis, is about where -within the residual dynamic range of a listener with hearing impairment- to situate the frequency spectrum of amplified speech. This has involved the generation of 25 individually shaped frequency spectra, all within the residual dynamic range of the listener "under test". All these 25 frequency spectra, which may be considered to be 25 different settings of a single hypothetical hearing aid, were evaluated for both speech intelligibility and sound quality.

The second question put forward has to do with limitations of the microphone-amplifier-loudspeaker combination which makes up a hearing aid. Since size is constrained, qualities common to high-fidelity stereo sets such as a smooth frequency response may be given up in favour of small size or high acoustic efficiency. In Chapter 3, the effects of peaks in a hearing aid's frequency response will be considered for speech intelligibility and sound quality.

The final question considered in this thesis has to do with recruitment in the sensorineurally impaired ear. Since recruitment may be regarded as a faster-than normal growth of perceived loudness, reducing loudness variations externally (i.e., by applying compression in a hearing aid) has been often suggested as a compensation for recruitment. However, following a different line of reasoning, speech may be considered to be a meaningful signal in which the information is carried by the *modulations* (level variations) of the temporal (time) envelope. Especially for listeners with hearing impairment, enlarging these modulations (i.e., expansion) might prove successful in overcoming the apparent distortion in their hearing. In Chapter 4, various compression and expansion variants are tested for their effect on speech intelligibility and on sound quality.

Chapter 2. Evaluation of a wide range of amplitude-frequency responses for the hearing impaired

The long-term average frequency spectrum of speech was modified to 25 target frequency spectra in order to determine the effect of each of these spectra on speech intelligibility in noise and on sound quality. Speech intelligibility was evaluated using the test as developed by Plomp and Mimpen (1979), whereas sound quality was examined through judgements of loudness, sharpness, clearness, and pleasantness of speech fragments. Subjects had different degrees of sensorineural hearing loss and sloping audiograms but not all of them were hearing-aid users. The 25 frequency spectra were defined such that the entire dynamic range of each listener, from 5 dB above threshold to 5 dB below UCL, was covered. Frequency shaping of the speech was carried out on-line by means of Finite Impulse Response (FIR) filters. The tests on speech reception in noise indicated that the Speech-Reception Thresholds (SRTs) did not differ significantly for the majority of spectra. Spectra with high levels, especially at low frequencies (probably causing significant upward spread of masking), and also those with steep negative slopes, resulted in significantly higher SRTs. Sound quality judgements led to virtually identical conclusions as the SRT data: frequency spectra with an unacceptably low sound quality were in most of the cases significantly worse on the SRT test as well. Because the SRT did not vary significantly among the majority of frequency spectra, it was concluded that a wide range in the dynamic range of listeners with hearing losses is suitable for the presentation of speech energy. This is very useful in everyday listening, where the frequency spectrum of speech may vary considerably.

Based on *Journal of Speech and Hearing Research* **38** 211-221 (1995)

Introduction

In hearing-aid fitting, two factors can be considered to be responsible for the satisfaction of the person with a hearing impairment. One is the ability to achieve good *understanding* when listening to speech that is presented through the aid, and the other is the extent to which the hearing aid's *sound* is considered acceptable. In order to achieve optimum performance, many amplification rules have been designed, e.g. the "half-gain" rule, Prescription Of Gain and Output (POGO, e.g. Lyregaard, 1986), National Acoustic Laboratories (NAL, e.g. Byrne & Dillon, 1986), etc. The amplitude-frequency responses that result from applying these rules tend to differ most notably for extreme hearing-loss configurations (such as steeply sloping audiograms) but, for many real-life hearing losses, differences are less pronounced (Hamill & Barron, 1992).

Because amplification rules simply relate the hearing aid's amplitude-frequency response to the pure tone audiogram, and because many do not pay any attention to sound quality, the hearing aid can possibly give a "sharp" or otherwise "unpleasant" sound impression. In such a situation, an audiologist could be tempted to modify the amplitude-frequency response to one that sounds more pleasant but might have degrading effects on speech understanding (because some frequency regions are no longer sufficiently amplified). However, if no attention is paid to the patient's complaints regarding sound quality, the hearing aid might be never used. Some indications as to the contrast between speech intelligibility and sound quality were given by Lutman and Clark (1986), who compared "flat" and "rising" frequency responses with respect to speech intelligibility and subjective preference. They found that flat frequency responses are preferred to rising responses, whereas speech intelligibility was slightly better for the latter. Leijon, Lindkvist, Ringdahl, and Israelsson (1991), however, did not find any significant intelligibility differences between four frequency responses, of which some had high-frequency emphasis. They did confirm that flat frequency responses are preferred subjectively, which is found in many other investigations on this subject (e.g. Thompson & Lassman, 1970). Further experiments regarding the perceived sound quality with different amplitude-frequency responses by subjects with hearing loss were conducted by Gabrielsson, Schenkman, and Hagerman (1988). They performed tests on speech intelligibility and sound quality using five different amplitude-frequency responses and concluded that

differences between amplitude-frequency responses are best reflected by quality judgements, although there are small differences in speech intelligibility among the responses. From this, it can be expected that quality judgements will impose the toughest restrictions on the freedom of choice for a hearing aid's amplitude-frequency response. Byrne (1986) evaluated six different amplitude-frequency response calculation methods for hearing aids through tests of speech discrimination, subjective intelligibility and pleasantness. He found that none of the amplification rules tested was superior to all others: the best amplitude-frequency response was not the same from listener to listener. However, there was a group of three amplitude-frequency responses that, after averaging, performed better than the others on the subjective-intelligibility and pleasantness tests.

The present study was set up to examine systematically the effect of a wide range of amplitude-frequency responses on speech understanding and sound quality. In general, signal processing in hearing aids should keep the speech signal below the level of uncomfortable loudness (UCL) and above the hearing threshold level at all frequencies (e.g. Skinner, Pascoe, Miller, & Popelka, 1982). With this in mind, all amplitude-frequency responses investigated in the present research were chosen such that the long-term average frequency spectra of the processed speech were within the listener's dynamic range. This necessitated a careful determination of the dynamic range of each listener. Based on the results from that test, a set of frequency responses was calculated for each listener to optimally fill the dynamic range. Each frequency response was evaluated by means of an SRT-test (Experiment 1) according to Plomp and Mimpen (1979), and by means of a set of tests for the evaluation of *clearness* and *pleasantness* of speech (Experiment 2), and of *loudness* and *sharpness* of speech (Experiment 3).

General Method

Equipment and Listeners

All experiments were carried out in a sound-proof booth using a PC-hosted digital signal processor (Texas Instruments TMS 320C25 on OROS "AU21" DSP board) with a single-channel 16-bits D/A converter and additional analogue equipment, connected to a pair of circumaural headphones (Sony MDR-CD999). In all experiments, care was taken to compensate for the amplitude-frequency response of the headphones, that had a slope of about -2 dB/oct above 100 Hz. Signals

were presented monaurally to the better ear, or to the preferred ear when the hearing losses for the two ears were similar, without masking of the contralateral ear. During the experiments, listeners did not use their hearing aids, because all speech stimuli were presented above the unaided threshold of hearing.

Subjects with hearing loss were selected from the files of the University Hospital VU. They were all free of persistent tinnitus and had sensorineural hearing losses in both ears, and sloping audiograms in the test ear. The listeners had average pure-tone hearing losses at 500, 1000 and 2000 Hz in the range of 18 to 52 dB HL, whereas at 8000 Hz, hearing loss ranged from 40 to 105 dB HL (HL *re* ISO, 1975). Performance functions for monosyllables in quiet reached at least 90% intelligibility. Age ranged from 27 to 82 years, with an average of 59 years; there were 16 men and 10 women in the group. Not all of the subjects with hearing loss were hearing aid users.

As a reference for the dynamic-range measurements, 10 university students with normal hearing were invited. Their age ranged from 19 to 27 years, with an average of 23. In this group, there were 7 men and 3 women.

Determination Of Dynamic Range

Both threshold and uncomfortable loudness levels were determined in an interactive computer-controlled procedure, using 15 one-third octave noise bands that covered the frequency range from 200 to 6400 Hz. For the threshold measurements, we used an adaptive procedure, where the listener was asked to push a button for as long as he/she could hear a pulsating noise burst (length: 310 ms; rise/fall times: 10 ms; repetition frequency: 2.4 Hz). Pushing the button (as long as the noise was audible) caused the computer to decrease the level of the noise burst, whereas releasing it (when the noise had become inaudible) would increase the level. The level of the burst was 70 dB SPL at the start of the experiment. From the first time the button was pushed until it was first released, the level was decreased in steps of 2.0 dB, to rapidly arrive at the approximate threshold level. Hereafter, the step size was reduced to 1.0 dB to increase the method's sensitivity. After either 11 reversals or when the standard deviation of the turning-points was less than 3.0 dB, whichever occurred first, the measurement of a threshold was finished. The threshold was then computed by averaging all but the first turnpoint levels.

The determination of the UCL consisted of two steps (see Walker, Dillon, Byrne, & Christen, 1984), of which the first consisted of a series of narrow-band UCL measurements, and the second was a broadband UCL measurement. Before each step, the listeners were carefully instructed to indicate that level at which the stimulus was considered uncomfortably loud ('when the stimulus is too loud for you'), rather than the level at which their ear was hurt. In the first step, the level of a narrow-band noise burst (duration: 310 ms; rise/fall times: 10 ms) was increased by 3.0 dB at a rate of 1.4 Hz. We chose a somewhat lower repetition rate than in the threshold measurement, because (a) the step size was greater, and (b) the listener should have ample time to react. This was repeated until the listener pushed a button, indicating that the level was considered uncomfortably loud. The noise level was then decreased by a random amount between 21 and 31 dB (to prevent the listener from simulating consistency by counting presentations of the noise burst), after which it was increased again in 3.0 dB steps. The measurement was stopped after either six reactions or when the standard deviation of the indicated levels was below 4.0 dB, whichever occurred first. At this point, UCL was computed by averaging the levels at which the subject had pushed the button. In the second step of the UCL measurement, we used a wideband noise burst that was spectrally shaped according to the narrow-band UCL levels. This noise burst (length: 3.7 s, which is indicative of the length of the signals in the speech-reception experiment) was then generated at gradually increasing levels, and after each presentation the listener was asked whether the signal had been uncomfortably loud. If so, the corresponding level in each of the third-octave bands was taken as the real UCL. This extended procedure was used because a wideband masking noise was to be used in following experiments, and since the combined narrow-band UCLs are a questionable measure of wideband UCL (Bentler & Pavlovic, 1989, Walker *et al.*, 1984).

In order to be able to compare the subjects with hearing loss to listeners with normal hearing, ten university students with normal hearing were invited to participate as paid volunteers in the dynamic-range measurements.

Definition Of The 25 Amplitude-Frequency Responses

Figure 1a shows average threshold and uncomfortable loudness levels as a function of the centre frequency of the one-third octave bands of

noise, for subjects with normal (dotted curves) and impaired (thick drawn curves) hearing.

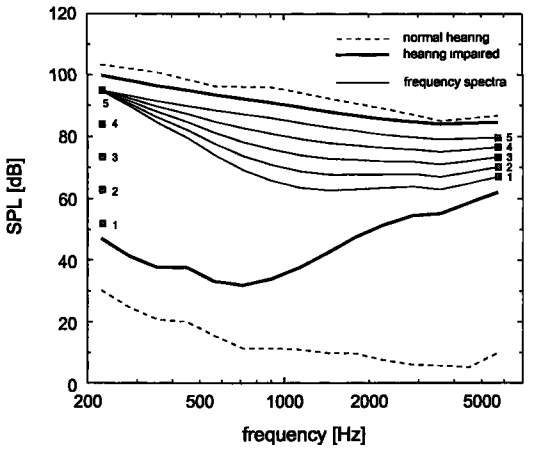

Figure 1a. Construction of the desired frequency spectra. Squares at the extreme frequencies represent "anchor points". Heavy and dotted curves represent average data from groups with impaired and normal hearing, respectively; the upper curves represent the average uncomfortable loudness levels and the lower curves represent the average threshold levels.

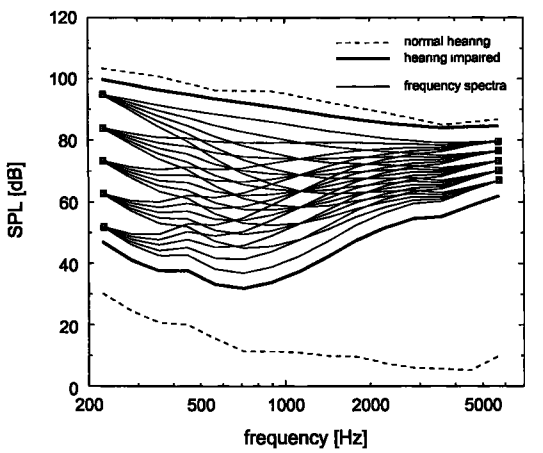

Figure 1b. All 25 frequency spectra used in the experiments.

All levels are long-term RMS levels, expressed in dB SPL as measured on a Brüel & Kjær type 4152 artificial ear (i.e. the exact levels at the eardrum may have been different). For the group with impaired hearing, average hearing loss ranged from about 17 dB at low frequencies to over 50 dB at high frequencies, as can be estimated from

the difference between the data from the two subject groups. UCLs were not significantly different between the two groups at any frequency, as was concluded from a series of t-tests ($p > 0.15$ for each test). The UCLs that are depicted in Figure 1a were measured with a wide-band noise burst that was spectrally shaped according to the narrow-band UCL levels. Narrow-band UCLs were 18 dB higher than broadband UCLs, on average; the maximum difference was 27 dB. This difference can be explained from loudness summation in the broadband case.

After registration of the listener's threshold and uncomfortable loudness levels, 25 different amplitude-frequency responses were defined, each of which brought the long-term average frequency spectrum of our speech material within the dynamic range. The frequency spectra were defined in such a way that the entire dynamic range (apart from a 5-dB margin at the threshold and UCL levels) was filled. First, we defined five levels for each of the two extreme third-octave centre frequencies (i.e. 224 and 5706 Hz) dividing the range from 5 dB above threshold to 5 dB below UCL into 4 equal parts (see Figure 1a). These levels were used as "anchor points", indexed 1 to 5, for the 25 frequency spectra. Then, in the range from 5 dB above threshold to 5 dB below UCL, the 15 third-octave levels for each of the 25 desired frequency spectra were calculated according to:

$$level_{L,H,B} = uclM_B - (uclM_B - thrM_B) \cdot \left(\frac{(5-L)}{4} + \frac{(B-1)}{14} \cdot \frac{(L-H)}{4} \right) \qquad \begin{pmatrix} L = 1..5 \\ H = 1..5 \\ B = 1..15 \end{pmatrix}$$

where

$level_{L,H,B}$ = level in dB SPL at third-octave band B for the frequency spectrum with low-frequency index L and high-frequency index H (see Figure 1a);

$thrM_B$ = threshold level at third-octave band B, plus 5 dB;

$uclM_B$ = uncomfortable loudness level at third-octave band B, minus 5 dB.

The curves that result (see Figure 1b) represent the desired long-term average frequency spectra of the speech that was used in the experiments. Throughout this paper, the frequency spectra will be referred to according to the indices of their anchor points, i.e. "$spec_{11}$" is the frequency spectrum that has $L=1$ and $H=1$ (see the equation above and Figure 1a). Because the spectra were re-calculated individually, they are equivalent in their relative position inside the

dynamic range for each listener (i.e. $spec_{33}$ is exactly halfway between threshold and UCL levels in all 1/3-octave bands for each listener).
The amplitude-frequency responses for the filters were calculated by subtracting the long-term average frequency spectrum of speech from the desired frequency spectra. Then, an inverse Fast Fourier Transform was applied to these amplitude-frequency responses. Finite Impulse Response (FIR) filters were constructed by time-windowing the impulse responses with a Kaiser window (see Rabiner & Gold, 1975). The number of filter coefficients that was retained was 256. Deviations of the FIR-filter's amplitude-frequency response from the desired response, as analysed by applying an FFT to the windowed impulse response, were negligible.

Experiment 1: Speech Intelligibility

Method

For each of the 25 frequency spectra, as outlined in the previous section, speech intelligibility in noise was measured using short everyday Dutch sentences and speech-spectrum shaped noise, according to the procedure as developed by Plomp and Mimpen (1979). The Speech Reception Threshold (SRT), which is defined as the S/N ratio at which 50% of the test sentences can be reproduced correctly, is a useful measure for comparing different signal processing strategies regarding their effects on speech intelligibility (e.g. Ter Keurs, Festen, & Plomp, 1992, and Van Dijkhuizen, Festen, & Plomp, 1989). When a certain type of signal processing results in a higher-than-normal SRT value, then this signal treatment will, in everyday practice, lead to intelligibility problems more often than when no processing had been carried out (the normal situation).
In the present implementation of the test, both speech and noise were filtered on-line, i.e. while the test was run. During the testing of a frequency spectrum, the level of the masking noise was constant, and its frequency spectrum was equal to the spectrum that was tested (see Figure 1b); the S/N ratio was varied by changing the speech level. The sentences were grouped into sets of nine, of which the first four were used to obtain an initial estimate of the SRT. For each new condition in the test, a sample sentence was presented (more than once, if desired) at a S/N-ratio of 0 dB, to let the listener get accustomed to the "sound" of this specific frequency spectrum. Then, the first sentence from the set would be presented at a S/N ratio of -10 dB. The level of this

sentence was increased by 4 dB at each successive presentation, until it was reproduced correctly by the listener. From then on, each next sentence was presented at a level that was 2 dB lower if the preceding sentence had been correctly reproduced, and 2 dB higher if not. Listeners were encouraged to guess the words they did not understand completely. The S/N ratios specified after the presentation of sentences four through nine were averaged to produce the SRT for the frequency spectrum under concern.

Throughout the experiment, speech of a male and a female talker was used. In order to prevent learning effects or fatigue from systematically influencing the experimental results of one or more conditions, the order in which the frequency spectra were presented to each listener was balanced over all listeners. Sentences were presented in a fixed order for all listeners.

Results and discussion

For each listener, the SRT values that were measured have been corrected for the individual average SRT, i.e. the per-listener average of the SRTs of the 25 conditions. Throughout the rest of this paper, these "relative SRTs" will be referred to as "SRT".

Figure 2 shows the SRTs, averaged over listeners, for each of the frequency spectra.

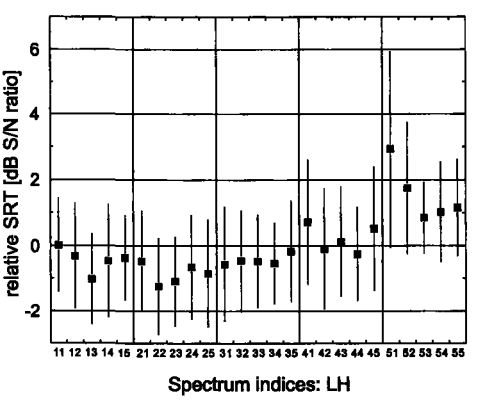

Figure 2. Average SRTs (of 26 listeners with hearing impairment) for sentences in noise as a function of frequency spectrum. Numbers on abscissa are the low-frequency and high-frequency indices of the spectra. Vertical bars represent standard deviations. SRTs as depicted are deviations from the average of each subject's 25 SRTs, i.e. "relative".

Note that a higher value of the SRT indicates that speech reception is more difficult under that condition. For speech shaped according to $spec_{22}$, the SRT is lowest (-1.2 dB). In order to be statistically significant, a difference in SRT between spectra must exceed 1.7 dB, as was found by analysing the data with a repeated-measures ANOVA ($F(24,600) = 8.54$, $p < 0.0005$), combined with a Tukey HSD post-hoc analysis of significance (Hays, 1988). This means that only those spectra with an average SRT higher than 0.5 dB are significantly different from $spec_{22}$, which is the case for $spec_{51}$ through $spec_{55}$, $spec_{41}$, and $spec_{45}$.

The spectra that have a higher SRT can be classified as having a relatively high low-frequency level and/or a high overall sound-pressure level. A possible explanation for the observed high SRTs could be Upward Spread Of Masking (USOM), whereby low-frequency signal components mask high-frequency signal components. USOM is expected to be most prominent when the low-frequency level is high (e.g. Kryter, 1962), as is the case for the spectra that have a significantly higher SRT.

Listeners with hearing loss have been reported to show a higher masked threshold than listeners with normal hearing, when a low-frequency masker is present. Van Dijkhuizen, Festen, and Plomp (1989) varied spectra of speech in noise with respect to $spec_{33}$. They found higher SRTs in listeners with hearing loss especially when they shaped speech according to a steeply sloping (approx. -11 dB/oct.) frequency spectrum. Jerger, Tillman, and Peterson (1960) report that an 87 dB SPL low-frequency masker gives rise to masking toward higher frequencies with a slope of about -5 dB/oct in subjects with sensorineural hearing loss. In the present study, the average slope of $spec_{51}$ (which has the steepest negative slope of all spectra tested) has been steeper than -5 dB/oct for most of the subjects, whereas the SPL of speech and noise in the lowest frequency band in $spec_{51}$ was frequently above 87 dB. This, in combination with the results from Jerger *et al.*'s experiments (as mentioned above) on masked thresholds in subjects with hearing loss, indicates that USOM probably caused the poor SRT of $spec_{51}$. Further, the SRT of $spec_{51}$ seems to be related to its spectral slope, because the correlation coefficient between the spectral slope and the SRT was -0.47 ($p = 0.008$, one-tailed, all listeners included). The average for all spectra of the corresponding correlation coefficients (i.e. one correlation coefficient for each spectrum) was -0.05, which indicates that the spectral slope is of greater-than-average

importance to the SRT of $spec_{51}$. USOM may have influenced the SRTs of the other spectra as well, depending on their slope and SPL.

For a more quantitative explanation of the speech intelligibility results, the Speech Intelligibility Index (SII, formerly known as Articulation Index) was computed for each of the 25 frequency spectra as depicted in Figure 1b, according to the procedure that is described in the proposed revision for ANSI S3.5-1969 (ANSI, 1992). For these calculations, we assumed a S/N ratio of -5 dB, because this is close to the expected SRT for listeners with normal hearing. The SII accounts for audibility as well, but because all signals were presented well above the absolute threshold of hearing, this did not affect its value; replacing the threshold data from the listeners with hearing loss by standard data for listeners with normal hearing, and re-computing the SIIs confirmed this. Further, we did not include the 'level distortion factor' that is prescribed in the procedure, because the speech we used had been spoken at a normal level and was electronically amplified to higher levels. The computed SIIs show a good correlation (-0.91) with the measured SRTs, indicating that the supposed underlying mechanism that caused the variations in SRT, USOM, is indeed a good candidate. Comparing the measured SRTs and the calculated values of the SII, it is apparent that the listeners with hearing loss do worse than the SII predicts. This is not surprising, because the scope of the SII is limited to otologically normal listeners (ANSI, 1992, p. 1) and does not include the kind of distortion that is introduced by the sensorineurally impaired ear.

To investigate whether the observed behaviour of the SRT as a function of frequency spectrum varies with the degree of hearing loss of the listeners, two subgroups of 13 listeners were formed: in one subgroup, all listeners had average 1/3-octave thresholds between 33 and 43 dB SPL, and in the other subgroup, all listeners had average 1/3-octave thresholds between 43 and 63 dB SPL. The magnitude of the differences in the average SRTs between the subgroups was smaller than 1 dB for each condition. This leads to the conclusion that, for the range of hearing impairment that was covered in this experiment, hearing loss does not have a significant influence on the SRT results.

Experiment 2: Clearness and Pleasantness Judgements

Method

A small set of speech fragments was used to let the listeners judge several sound quality aspects of the filtered speech. The qualities that were to be judged in Experiments 2 and 3 have been adopted from existing research in the area of subjective evaluations of sounds (e.g. Plomp, 1976), except for clearness, which was used because of its applicability to speech in noise. Only speech of the female talker was used in these experiments, in order to avoid the introduction of extra variance (due to inter-speaker differences) into the results.

Clearness and pleasantness were judged with a paired comparisons set-up. Because mutual comparison of all 25 speech spectra would simply have taken too much time, and because variations were expected to be gradual, only 13 selected frequency spectra were used in these experiments. These spectra were indexed 11, 13, 15, 22, 24, 31, 33, 35, 42, 44, 51, 53, and 55 (see Figure 1a). In a pilot experiment, in which full-length sentences were presented, listeners were very rapid in making their decisions. Therefore, it was concluded that it was not necessary to present full-length sentences, because listeners could perform the experimental task equally well when only a speech fragment was used as the test signal. Ten fragments were taken from the original speech material, that had been used in the speech intelligibility experiment, by taking about the first half of ten sentences. Fragment duration was between 250 and 310 ms. For each comparison, a single speech fragment was filtered twice to produce the two frequency spectra that were to be compared.

In the first test, speech fragments were presented in noise at a S/N ratio equal to the listener's SRT for the frequency spectrum under concern, thus equalising speech intelligibility within a comparison. Using this S/N ratio will not result in an intelligibility of exactly 50% for all fragments, because the redundancy, and thus the intelligibility, of the original sentences is affected by fragmenting them. To ensure that each fragment would be completely intelligible, its contents were displayed on the monitor of the experiment computer during headphone presentation. Also, at the start of each paired comparisons test, some examples were presented to familiarise the listener with the experimental task. These examples comprised the complete set of fragments, so that the listeners knew their contents. Listeners were asked to indicate which of the fragments they thought was *more clear*, i.e. from

which of the two they could extract the text more easily. In the second test, speech fragments were presented in quiet and *reversed* in time, to prevent the contents of the fragment from having an influence on the judgement. Listeners were asked to indicate which fragment sounded *more pleasant*.

Comparisons were set up such that each of the 13 selected frequency spectra was compared to every other spectrum once. In each comparison, the "score" of the preferred spectrum was raised by 1. In the case of no preference for either spectrum, both scores were raised by ½. In this way, the maximum score for any particular spectrum was 12. Spectra were ordered within comparisons in such a way that each spectrum was presented equally often as the first as it was the second (Phillips, 1964). Thereby, any response bias associated with the order of presentation was cancelled out. Also, a number of unfiltered sentences was presented in both normal and time-reversed conditions before the start of the paired comparisons experiments, to let the listener get accustomed to the "funny" sound of time-reversed speech and to demonstrate that it was completely unintelligible, even though the signal remained clearly speech-like. Each of the experiments took about 15 min.

Results and discussion

Figure 3 shows the results of the paired comparisons experiments on clearness and pleasantness, for each of the 13 frequency spectra that were evaluated.

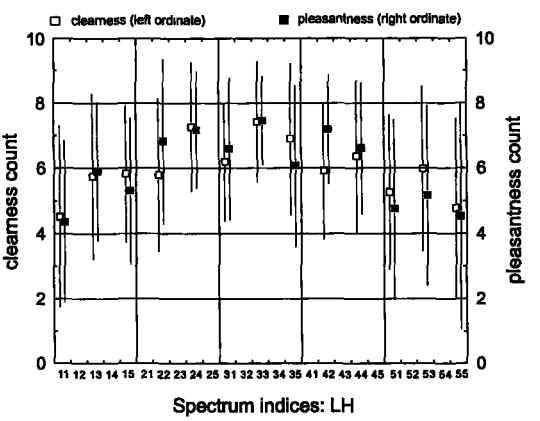

Figure 3. Mean clearness (left ordinate) and pleasantness (right ordinate) preference counts, for 26 listeners with hearing impairment, as a function of frequency response. Bars represent standard deviations. Parameters on abscissa are as in Figure 2.

The spectrum that was preferred most often for clearness is $spec_{33}$ (7.42 times out of 12, averaged over the 26 listeners). Statistical analysis was carried out in two steps:

1. a repeated-measures ANOVA, that revealed a significant effect of frequency spectrum ($F(12,300) = 3.47, p < 0.0005$), and
2. Tukey's HSD procedure, that showed that average clearness scores have to be below 5.29 in order to differ significantly from $spec_{33}$.

The latter is the case for $spec_{11}$ and $spec_{55}$, with average scores of 4.52 and 4.79, respectively.

The spectrum that was preferred most often for clearness was also preferred most often for pleasantness (7.46 times out of 12, on the average). Pleasantness scores were analysed in the same way as the clearness scores, i.e.

1. a repeated-measures ANOVA, that again revealed that the effect of frequency spectrum was significant ($F(12,300) = 4.70, p < 0.0005$), and
2. Tukey's HSD procedure, that showed that pleasantness scores should be below 5.26 to differ significantly from $spec_{33}$.

The latter is the case for four spectra: $spec_{11}$, $spec_{51}$, $spec_{53}$, and $spec_{55}$, with average scores of 4.37, 4.77, 5.17 and 4.54, respectively. The spectrum that was preferred most often in both experiments is halfway between the threshold and uncomfortable loudness level, i.e. it bisects the dynamic range. Further, it seems that the more a spectrum differs (in terms of third-octave levels) from the best spectrum, the less it is preferred for clearness and pleasantness.

When the results of these experiments are compared to those of the SRT experiments, it can be seen that the majority of the spectra that were preferred significantly less had a significantly higher SRT as well. This is not a trivial result, because in the test on clearness, all speech signals had been equated for intelligibility, justifying an a priori expectation that clearness, which is a measure of subjective intelligibility, would be judged equal for all spectra. Nevertheless, differences in clearness between spectra were found in the listeners' judgements. Apparently, clearness differs from intelligibility as determined by the SRT. Other factors, like overall SPL and sound impression, may have had an additional effect on the clearness judgements. This indicates that judgements of speech intelligibility, as sometimes applied in fitting procedures for hearing aids, are not interchangeable with objective SRT results.

Figure 3 shows a striking similarity between clearness and pleasantness judgements: differences are never greater than 1.3, on the average. Judging from this similarity among the results of the two tests, it is possible that pleasantness has been a major determinant in the clearness test as well, particularly because the differences in objective intelligibility had been removed in that test. However, because of the difference between the stimuli in the two experiments (speech in noise and time-reversed speech in quiet, respectively), and the different aspects of the stimuli that were to be judged, it seems rather unlikely to assume that the listeners somehow judged the same quality in both experiments.

Experiment 3: Loudness and Sharpness Judgements

Method

Loudness and sharpness were judged in two rating-scale tests, that were similar in set-up. In both tests, the speech fragments that had been used in Experiment 2 were presented in quiet and *time-reversed*, to prevent the contents of the fragments from having an influence on the judgements. Each fragment was filtered to obtain the desired frequency spectrum and then presented to the listener. Judgements were made on a five-point rating scale, described to the listener with the adjectives "too soft", "soft", "good", "loud", and "too loud" in the case of loudness rating, and the adjectives "too dull", "dull", "good", "sharp", and "too sharp" in the case of sharpness rating. In the explanation of the test it was stressed that the extreme scale points "too soft", "too loud", "too dull", and "too sharp" should be used for unacceptable conditions. The remaining points on the scale were to be used for acceptable, though possibly not optimal, conditions. Listeners were asked to give their judgements by pressing one of five marked keys on a personal-computer keyboard. In both tests, all filter characteristics were judged twice to increase reliability. At the start of each session, some conditions were presented to the listener to indicate the range of values that could be expected and to familiarise the listener with the task to be performed. The order of the conditions was balanced over listeners, as in the SRT experiment. Each of the tests took about 6 min.

Results and discussion

Figure 4 shows mean values for the loudness and sharpness judgements as a function of frequency spectrum.

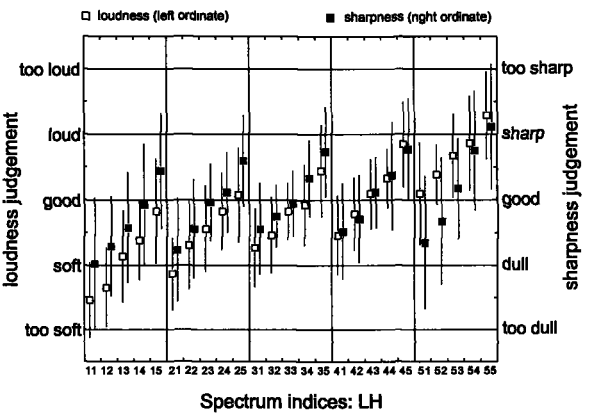

Figure 4. Mean loudness (left ordinate) and sharpness (right ordinate) judgements, for 26 listeners with hearing impairment, as a function of frequency response. Bars represent standard deviations. Parameters on abscissa are as in Figure 2.

A clear trend is visible for both measures, corresponding to the frequency content of the processed stimuli. Marked deviations from the judgement "good" (i.e. differences from the middle of the available scale range) were, in the case of both loudness and sharpness, found for various spectra. For the analysis of variance, we assigned numerical values to the rating-scale items. The variance was analysed with the same two-step procedure that had been used for the clearness and pleasantness results. Repeated-measures ANOVAs showed a significant effect of frequency spectrum on both loudness ($F(24,600) = 46.83$, $p < 0.0005$) and also on sharpness ($F(24,600) = 15.32$, $p < 0.0005$). Then, Tukey's HSD procedure was used to determine whether an average judgement was significantly above "loud" or "sharp", or significantly below "soft" or "dull". This applies to $spec_{11}$, $spec_{12}$ and $spec_{21}$ (based solely on the loudness evaluation) and also to $spec_{55}$ (which is rejected for both extreme loudness and extreme sharpness, and was rejected as well because of its poor SRT).

The loudness judgement of a frequency spectrum is likely to be influenced primarily by its SPL, as was confirmed by the correlation between subjective loudness and SPL: the average of all listeners' individual correlation coefficients was 0.73. Sharpness judgements were

expected to depend on the slope of the frequency spectrum. However, correlations between the slopes of the frequency spectra and subjective sharpness were of varying sign; the average correlation was -0.02. From Figure 4, one might conclude that subjective sharpness depended primarily on the high-frequency level; however, the average correlation between these variables was only 0.56. Sharpness judgements for the five spectra with $H=5$, i.e. with identical high-frequency levels, became higher as the low-frequency level increased, even though the slope of the spectra decreased. However, this trend is statistically non-significant, since it falls within the minimum range that Tukey's HSD requires for significance. Concluding, it seems that the sharpness judgements have been influenced by both the high-frequency level and the SPL of the frequency spectra, although the results are less correlated to the SPL than those of the loudness judgements.

General Discussion And Conclusions

Of the 25 frequency spectra that were tested, 18 could not be distinguished statistically in terms of speech intelligibility (see Figure 5).

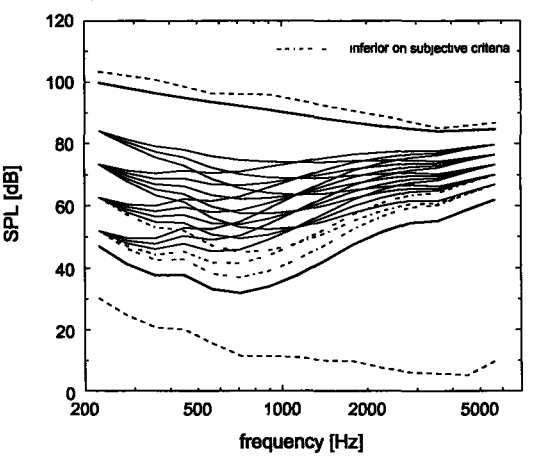

Figure 5. Frequency spectra that are equivalent in terms of relative SRT (see text). Dashed curves represent spectra that were inferior to the others on subjective criteria only.

However, the statistical criterion that was used depends on the structure of the results of the present experiment. For instance, in case the number of listeners had been greater, the significance criterion would probably have indicated a narrower range of equivalent spectra.

Thus, the question is whether the present significance criterion corresponds to relevant differences in SRT. In our data, differences in SRT had to exceed 1.7 dB in order to be significant, which corresponds to a change in sentence intelligibility of about 26% (Plomp & Mimpen, 1979). In the case of a single speaker in ambient noise, a reduction of the S/N ratio by 1.7 dB corresponds to an increase of the speaker-listener distance by 22% in free-field conditions.

An interesting question is, of course, whether the SRT would have behaved accordingly for listeners with normal hearing. From literature, we know that spectral slopes of -12 dB/oct and less will increase the SRT for sentences (Van Dijkhuizen, Anema, & Plomp, 1987). From Figure 1b, we can estimate the spectral slopes; for $spec_{51}$, $spec_{52}$, and $spec_{41}$, the estimated slopes for listeners with normal hearing are below -12 dB/oct. This suggests that, like for listeners with hearing loss, a higher SRT would have been found for these conditions. This suggestion is confirmed by the results of Speech Intelligibility Index (SII) calculations that were performed with the dynamic range data from the 10 listeners with normal hearing.

From the reduced set that results from the SRT experiments, three additional spectra should be removed because they are, on average, judged "too soft" (dashed curves in Figure 5). This is in accordance with the results of Gabrielsson, Schenkman, and Hagerman (1988), who also found that subjective criteria are more restrictive than those based on speech intelligibility. Still, a wide range of spectra, i.e. the major part of the dynamic range, is available for the presentation of speech. Frequency responses according to amplification rules such as POGO, NAL, etc., will also put speech inside this range, provided that the overall amplification (which is set by the hearing-aid user) is sufficient.

In the present results, there is no preference for a "flat" frequency response; the response that produced the most preferred frequency spectrum $spec_{33}$ had a slope of about +6 dB/oct (producing an approximately flat frequency spectrum). This result is contrary to what is generally reported (e.g. Lutman & Clark, 1986); an explanation may be the fact that all frequency responses that were used in the current experiments presented the entire speech spectrum above the threshold of hearing, causing every spectrum to contain suprathreshold high-frequency energy. Other investigators might have included responses that did not satisfy this condition, and because listeners with hearing loss have become accustomed to hardly hearing anything at high

frequencies, they might prefer responses subjectively that do not present high-frequency energy at all.

A remarkable result of the subjective tests was the outcome of the clearness judgement. Even though speech intelligibility had been equalised according to each listener's own SRTs, the 13 conditions were not rated "equally intelligible" at all. Subjective intelligibility ratings are sometimes used in hearing-aid fitting procedures (e.g. Kuk & Pape, 1992), with the aim of selecting a frequency response that delivers optimum speech intelligibility. In view of the present results, such tests should be interpreted with some caution, and preferably be compared to objective intelligibility measurements.

For practical application, the present results indicate that the hearing aid's amplitude-frequency response should put the speech spectrum within the range of 15 frequency spectra that produced equivalent speech intelligibility and were not rejected on subjective criteria either. Because the spectrum of speech will, in everyday life, contain considerably more variation than is present in the test materials that have been used in the present study, a good startoff for an amplitude-frequency response might be one that puts the average everyday speech spectrum in the middle of the range in Figure 5. This will keep the processed speech spectrum within the desired range most of the time. In order to cope with extreme input frequency spectra (e.g. steep negative slopes or very high sound levels), it is important that the hearing aid contains a kind of Automatic Gain Control (AGC) that should operate only in extreme situations, in order to keep the spectrum of the output signal within the desired range. The exact parameters of the AGC (attack time, release time, onset level, and compression ratio, in each of the n frequency bands) should be chosen carefully, because AGC can affect speech intelligibility in a negative way (e.g. Plomp, 1994). Further research in this direction is needed to determine the optimum for these parameters.

Acknowledgements

This research was financially supported by Philips Hearing Instruments. We would like to thank Theo S. Kapteyn and Martin H.P. Stollman, of the University Hospital VU, Audiology Centre, for their kind assistance in selecting the listeners with hearing loss, and Tammo Houtgast for his useful suggestions during the preparation of the manuscript.

Ruth A. Bentler, Denis Byrne, Dianne J. Van Tasell, and Timothy D. Trine reviewed an earlier draft of this paper. Their suggestions are gratefully acknowledged.

Chapter 3. Peaks in the hearing aid's frequency response: Evaluation of their effect on speech intelligibility and sound quality

In a series of experiments, we introduced peaks of 10, 20, and 30 dB, in various combinations, onto a smooth reference frequency response. For each of the conditions, we evaluated speech intelligibility in noise, using a test as developed by Plomp and Mimpen (1979), and sound quality (for both speech and music), using a rating-scale procedure. We performed the experiments with 26 listeners with sensorineurally impaired hearing, and 10 listeners with normal hearing. Signal processing was accomplished digitally; for each listener, the stimuli were filtered and subsequently amplified so that the average speech spectrum was well above the threshold of hearing at all frequencies. The results show that, as a result of the introduction of peaks onto the frequency response, speech intelligibility is affected more for the listeners with impaired hearing than for the listeners with normal hearing. Sound-quality judgements tend to be less different between the listener groups. Especially conditions with 30-dB peaks show serious effects on both speech intelligibility and sound quality.

Based on *Journal of Speech and Hearing Research* **39** 239-250 (1996)

Introduction

In hearing aids, especially in the high-power models, we often find a rather irregular frequency response, that might be modelled as a smooth response with several peaks and troughs on it. Apart from peaks that are introduced by the hearing aid itself, earmold ventings (e.g. Studebaker & Zachman, 1970) and tubing (e.g. Carlson, 1974; Killion, 1980; Lybarger, 1985) may have significant adverse effects on the smoothness of the hearing aid's frequency response. By intuition, irregularities in the frequency response of sound-reproduction equipment are often regarded as undesirable, because they would lead to deterioration of the fidelity of the sound. The magnitude of the effects on speech intelligibility is difficult to predict; we might expect intuitively that, because of Upward Spread Of Masking (USOM), speech energy at frequencies above the peak will easily be masked. Because listeners with sensorineural hearing impairment have been reported to show greater-than-normal amounts of USOM in laboratory tests (e.g. Gagné, 1988; Jerger, Tillman, & Peterson, 1960; Trees & Turner, 1986), peaks in a hearing aid's frequency response might be especially disadvantageous to them, with regard to speech intelligibility. Another effect that peaks in the frequency response might have on speech intelligibility is in the area of formant detection. Because formant extraction is linked to the presence of maxima in the envelope of the frequency spectrum, the peaks in a hearing aid's frequency response might interfere with the spectral envelope's maxima (i.e. new maxima are introduced, or existing ones reduced). Especially for vowels, of which the spectra are relatively constant, there could be pronounced effects on recognition; this depends on the location (in the frequency domain) of the peaks in the frequency response, and also on the speaker (because the formants are shaped by the vocal tract).

Apart from the effects on speech intelligibility and on sound quality, problems such as acoustical feedback and loudness discomfort may occur more often with peaky frequency responses. In both cases, especially at high levels of amplification, the hearing aid might start to oscillate at a frequency that corresponds to the centre frequency of one of the peaks, or cause uncomfortably high sound levels at that frequency. A hearing aid with a smoother frequency response might not require the level of amplification at which the peaky aid shows these undesirable effects, because all frequencies are already sufficiently amplified at a lower overall gain setting.

In the literature, one can find a few reports about the effects of irregularities in the frequency response on speech intelligibility. Bücklein (1981) superimposed peaks and troughs onto a flat frequency response and examined their effect on the intelligibility of nonsense syllables in listeners with normal hearing. When he introduced ten peaks (width: about one-fifth octave) of about 20 dB in the frequency range from 0.2 to 3.2 kHz, the percentage-correct score for syllables decreased by 11.3%, as compared to the score for a flat response. All other configurations of peaks or troughs that he tested were of less influence on intelligibility; he found troughs to be generally less harmful to speech intelligibility than peaks of approximately equal shape. Jerger and Thelin (1968) analysed the frequency responses of 21 commercially available hearing aids; one of their tests was the measurement of speech intelligibility both in listeners with normal and with impaired hearing. They found some relation between the so-called Index of Response Irregularity (IRI, an indicator for the irregularity of a hearing aid's frequency response) and the identification score for synthetic sentences (meaningful words in a second order approximation of English syntax) in the presence of a competing voice. However, this relation was weaker for listeners with hearing impairment than for listeners with normal hearing, which is contrary to what one would expect. Smaldino (1979) correlated a number of electroacoustic hearing-aid characteristics (including distortion, bandwidth, saturation sound level, and the IRI) to the identification of key words in sentences. He found that the IRI was only one out of five parameters in a regression equation, that accounted for about 59% of the variance in his data. Cox and Gilmore (1986) used earhook dampers and real hearing aids to evaluate the effects on speech clarity. Their results indicated that the difference between damped and undamped frequency responses was hardly noticeable for most of their listeners (significance level on χ^2-test: 0.09), which may be explained by noting that

1. the presence of dampers in the earhook does not always completely remove a peak in the response, and
2. the peaks in the undamped responses were not higher than about 10 dB.

There have been quite a few investigations into the *audibility* of irregularities in the frequency response, especially in the area of the development and evaluation of loudspeakers. Bücklein (1981) presented

1. a series of 13 different fragments (11 containing music, and 2 containing speech by a male and by a female speaker, respectively), and

2. fragments of white noise,

through a system that introduced peaks and dips onto the frequency response; he determined the percentage of listeners (with normal hearing) that could reliably detect the irregularities in a paired-comparisons paradigm (i.e. one stimulus was processed, the other was not). Peaks were perceived more easily than dips with the same shape; with all stimuli, subjects start to make mistakes at a peak height of roughly 5 dB. Toole and Olive (1988) measured the detection threshold for resonances in listeners with normal hearing; they varied the quality factor (Q, the relative width) of the resonances, the type of stimulus (speech, music, pink noise, pulse trains), and the acoustical environment (headphones, a single loudspeaker in an anechoic chamber, a living room, and a small hall). They found a decreasing sensitivity for resonance detection when they varied their stimuli from pink or white noise, via speech, to music. Relatively broad peaks (low Q) were detected more easily than narrow ones (high Q), and resonance detection was easier at low frequencies than at high frequencies. The various acoustical environments did not introduce systematic shifts in the detectability of resonances.

In order not to complicate the experimental conditions, we chose to introduce artificial irregularities, *like* those found on 2cc-coupler responses of real hearing aids, onto the frequency response; we tested these in a well-controlled environment and we included a well-established reference condition that was derived from previous research (van Buuren, Festen, & Plomp, 1995). We are aware that, in everyday situations, other sources (e.g. acoustic feedback through leaks or vents in the earmold, different types and lengths of tubing) might introduce even more peaks onto the frequency response, or enlarge existing ones. We have not specifically considered those, because their characteristics depend very much on the exact configuration of the earmold, tubing, and vents. Further, we did not include the effects of limited headroom (i.e. harmonic distortion), which may occur at high levels of amplification. Our aim was to test conditions that produce varying effects on both speech intelligibility and sound quality, in order to establish an upper limit to peak-like frequency response irregularities in hearing aids. This implies that not all of the conditions (e.g. peak heights) that we tested will be found in real hearing aids. For the current experiments, we invited listeners with sensorineural hearing loss, as well as listeners with normal hearing.

General Method

Equipment and Listeners

Twenty-six listeners with sensorineural hearing loss were selected from the files of the University Hospital's Audiology Centre. The pure-tone hearing losses at the test ears, averaged for 0.5, 1 and 2 kHz, ranged from 23.3 to 60.0 dB HL (*re* ISO, 1975). The losses in all these ears can be classified as *sloping* to various degrees. As an estimate for the overall slope of the audiogram (from 0.25 to 8.0 kHz), we computed straight-line approximations to the thresholds by means of linear regression analyses; in 3 ears, slopes were shallow (between 0 and 5 dB/octave), in 12 ears, slopes were moderate (between 5 and 10 dB/octave), in 8 ears, slopes were steep (between 10 and 15 dB/octave), and in 3 ears, the slopes were extremely steep (above 15 dB/octave). In quiet, these listeners could reach at least 70% intelligibility for monosyllables and they were free of persistent tinnitus; age ranged from 61 to 88 years, with an average of 66 years. Ten listeners with normal hearing participated in the experiments; in this group, age ranged from 16 to 27 years, with an average of 22.6 years.

The experiments were carried out in a double-walled, sound-treated booth. All test signals were presented monaurally, without masking of the contralateral ear. If the two ears had equal thresholds (e.g. for all listeners with normal hearing), we presented the tests to the preferred ear; otherwise, the better ear was selected. We used a PC-hosted Digital Signal Processor card (OROS "AU21", featuring Texas Instruments' TMS 320C25) with a 16-bit single-channel D/A converter, to process and generate the experimental stimuli. The stimuli were presented to the listeners through Sony MDR-CD999 circumaural headphones. The experiments took about three hours per listener, including breaks.

Determination Of Dynamic Range

For each listener, we used an individually adapted reference frequency response, well within the dynamic range. To determine this reference frequency response, we started the experimental session by measuring the threshold and uncomfortable loudness (UCL) levels for one-third octave bands of noise. We used 15 noise bands that covered the frequency range from 0.2 to 6.4 kHz.

For the threshold measurements, each noise band was presented repeatedly (stimulus duration: 310 ms; rise/fall times: 10 ms; repetition frequency: 2.4 Hz) and the listener was instructed to keep the space bar

of the PC keyboard pressed as long as the noise burst could be heard. The level of the noise band was decreased at each next presentation, until the space bar was released, indicating that the noise was now inaudible. Then, the level of the noise was increased again and the listener had to push the space bar as soon as the noise was audible again. During the first run for each noise band, the level of the noise band was decreased in 2-dB steps to quickly arrive at the threshold level; all subsequent level changes were 1 dB. The threshold level was estimated continuously by averaging all but the first turnpoint levels. The measurement was terminated after the standard deviation of the turnpoints had become less than 3.0 dB, or after 11 turnpoints had been registered, whichever occurred first. To familiarise the listener with the experimental task, some extra measurements were performed prior to the actual runs.

For the measurements of the UCL levels, we used only 8 of the 15 third-octave noise bands (every second band was left out) to restrict fatigue and irritation; for the absent frequency bands we computed the UCL levels by interpolation of the levels of the two adjacent bands. Since UCL levels have been shown to vary much more gradually over frequency than threshold levels (van Buuren, Festen, & Plomp, 1995), this was not considered to be a great loss of accuracy. The listeners were instructed to listen to a series of noise bursts (burst duration: 310 ms; rise/fall times: 10 ms; repetition frequency: 1.4 Hz), in which the level of each next burst was increased by 3.0 dB as long as the listeners thought they could tolerate it. When the bursts had become too loud (we stressed the fact that we were interested in a level of uncomfortable loudness, rather than the onset of pain), the subject was to push the space bar, causing the level of the next burst to be decreased by a random amount between 21 and 31 dB. After this decrease, the level was increased again by 3.0 dB at each presentation. The UCL level was estimated by averaging the levels at which the space bar was pressed; the measurement was terminated after either six reactions or when the standard deviation of the levels had become less than 4.0 dB, whichever occurred first.

After these narrow-band measurements, we filtered noise to obtain a spectrum equal to the combined narrow-band UCL levels. This wide-band noise signal (duration: 3.7 s, typically the duration of a sentence-in-noise in the speech-intelligibility test) was then presented to the listener repeatedly at increasing levels; after each presentation the listener was asked to indicate whether the noise had been 'too loud'. If

not, the noise would be presented again at a higher level. The result of this experiment was used to determine the wide-band UCL level, which was used in the definition of the experimental conditions.
Figure 1 shows the average results of the dynamic-range experiments, for both listener groups.

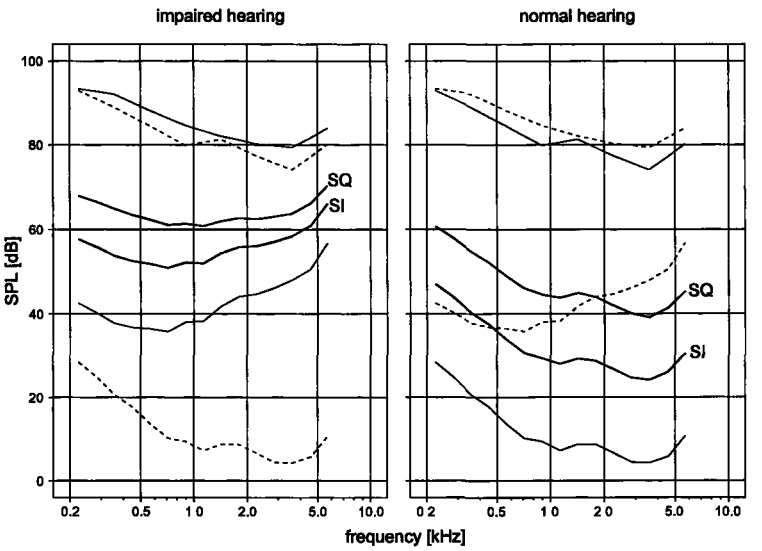

Figure 1. Average dynamic ranges for the two listener groups. The lower light curves represent the threshold levels; the upper ones represent the UCL levels. For each group, the dynamic-range data from the other group have been reproduced with dashed lines.
The heavy curves are examples of the reference spectra for speech intelligibility (marked "SI") and sound quality (marked "SQ"). For each listener, these spectra were computed from the individual *threshold and UCL data.*

All levels are long-term RMS levels, expressed in dB SPL as measured on a Brüel & Kjær type 4152 artificial ear (i.e. the exact levels at the eardrum may have been different). Between groups, thresholds differ from 14 dB at the lowest frequency to 46 dB at the highest frequency; as was already suggested by the classification of the hearing losses mentioned above, most of the subjects had sloping hearing losses, with the greatest losses at the highest frequencies. The differences between the average UCL levels are small and statistically non-significant ($F(1,34) = 1.68, p > 0.2$).

Experimental Conditions

In previous experiments (van Buuren, Festen, & Plomp, 1995), we scanned the available dynamic range in listeners with hearing impairment with respect to intelligibility and sound quality of speech. The best frequency spectra (out of 25) found in those experiments for speech intelligibility and sound quality, respectively, have been used as the reference conditions for the current tests. The reference spectrum for the speech intelligibility experiments was computed from the individual threshold and UCL levels by adding 5 dB to the threshold, and then adding a quarter of the distance from 5 dB above the threshold to 5 dB below UCL, according to:

$$spec_{SI,B} = (thr_B + 5) + \frac{(ucl_B - 5) - (thr_B + 5)}{4}$$

$$= \frac{3 \cdot thr_B + ucl_B + 10}{4} \qquad (B = 1..15)$$

In Figure 1, this spectrum is labelled 'SI', for both listener groups. The reference spectrum for the sound-quality experiments is labelled 'SQ' in Figure 1; it is positioned halfway between the individual threshold and UCL levels, according to:

$$spec_{SQ,B} = \frac{thr_B + ucl_B}{2} \qquad (B = 1..15)$$

where, in both formulas,

$spec_{xx,B}$ = level of the reference spectrum in dB SPL at third-octave band B;
thr_B = threshold level at third-octave band B;
ucl_B = uncomfortable loudness level at third-octave band B.

The reference spectra represent the long-term RMS levels of speech. For several sound-quality experiments, we used music fragments. Because these stimuli obviously differed in spectral content from the speech, their average spectra were not exactly halfway between the threshold and UCL levels, as was the case for speech in the sound-quality experiments. Because we were interested in the effects of each fragment's frequency spectrum on sound-quality judgements, the music fragments were amplified to a level that corresponded to the level of

the speech, but their spectra were not changed. In comparison to the speech-intelligibility measurements, we used a shallower low-frequency cut-off (about 6 dB/oct up to 200 Hz) in the sound-quality experiments, because this is a better representation of a hearing aid's low-frequency behaviour than the very steep low-frequency cut-off (about 80 dB/octave up to 200 Hz) that we used in the speech-intelligibility experiments. The steep low-frequency cut-off was used to eliminate spectral differences at low frequencies between the female and the male speakers.

Onto the reference frequency spectrum, we superimposed peaks with predefined heights, widths, and centre frequencies. The values for these parameters were chosen such that they correspond to those that may be found in commercially available hearing aids, although we did impose some restrictions, based on the desired experimental design. We chose three frequencies, 1.3, 2.8, and 5.5 kHz (based on the location of peaks encountered in real hearing aids), at which to centre the peaks. The shape of the peaks was a scaled version of one period of a raised cosine (i.e. from $-\pi$ to π), defined on a logarithmic frequency scale; the outer edges of the peaks were either 0.3 octaves apart ("narrow") or 0.6 octaves apart ("wide"). Although, in real hearing aids, peaks tend to become narrower (on a logarithmic frequency scale) with increasing centre frequency, we did not include that in our experiments to prevent the confounding of factors. The height of the peaks was either 10, 20, or 30 dB; height was defined as the distance from the top of the peak to the reference spectrum. For *single* peaks, we tested all possible combinations of two widths and three heights, producing six experimental conditions for each of the three centre frequencies. In the case of *multiple* peaks, we restricted ourselves to conditions with three peaks of equal height and width, producing another six experimental conditions (three heights times two widths). Together with the reference condition, this sums up to a total of 25 experimental conditions.

A real hearing aid will, apart from providing extra amplification at the peak frequency, also apply a phase shift to the input signal. Therefore, we carried out a pilot experiment in which we evaluated speech intelligibility for peaked responses with and without the accompanying phase shift. The results of this experiment did not reveal a clear influence of phase shifts on speech intelligibility, suggesting that the main effect of a peak in the frequency response corresponds to its amplitude. Informal listening to filtered stimuli with and without phase

shifts applied to them only revealed the ringing-like effects caused by the peaks in the frequency response; once again, no noticeable differences were introduced by the phase shifts. We therefore chose to use Finite Impulse Response (FIR) filters with linear phase.

The introduction of the peaks required 512-tap FIR filters. We applied these filters to the stimuli before each experimental session. The realisation of the listener-specific reference spectra was accomplished on-line by means of 256-tap FIR filters. For each listener, these filters were computed by

1. subtraction of the long-term RMS spectra of the speech from the desired reference spectra,
2. an inverse FFT to create an impulse response, and
3. multiplication of the impulse response with a 256-coefficient Kaiser window.

As had been confirmed in previous experiments (van Buuren, Festen, & Plomp, 1995), a filter length of 256 taps was sufficient for our purposes. Computation of the resonance filters was carried out only once; except for the length of the Kaiser window, the computation scheme was identical to that of the listener-specific filters.

Tests of Statistical Significance

We used two types of statistical analysis to analyse the experimental results of our experiments.

1. For the results of the speech-intelligibility tests, we used a repeated-measures analysis of variance (ANOVA), after having transformed the data to obtain a normal distribution of the individual values for each experimental condition. From the results of this analysis, and the number of conditions involved, we computed Tukey's Honestly Significant Difference (Tukey's HSD; see Hays, 1988) for a confidence level of 95%. Tukey's HSD-criterion can be applied to the difference between any two conditions and, for the comparisons and significance level involved in the present tests, to differences between theoretical values and conditions as well (Miller, 1981).
2. The results of the speech-intelligibility and sound-quality *ratings* were analysed by means of non-parametric statistical tests, because the data were categorical and could not be transformed into a normally distributed data set. We used a Friedman ANOVA to test whether there was an overall effect of the experimental conditions on the ratings, and a Wilcoxon Matched-Pairs Signed-Ranks test to detect differences between the reference condition and each of the

other conditions. Because there were 24 other conditions, a 5% error rate for the whole set of comparisons requires a significance level of 0.002 (5% divided by 24) or less for each of the individual comparisons (this procedure is known as the Bonferroni approach for multiple comparisons; see Altman, 1992).

A small group of listeners with normal hearing was included in the experiments for the sake of making overall comparisons to the listeners with hearing impairment. Because of the relatively small (N = 10) group size, the data from the listeners with normal hearing should be regarded with some caution. This is also expressed by the fact that the statistical tests either required greater differences between conditions for significance, or did not show significant differences at all.

Experiment 1: Speech Intelligibility

Method

For the evaluation of speech intelligibility, we determined the Speech-Reception Threshold (SRT; see Plomp & Mimpen, 1979), which is the S/N ratio at which 50% of short, everyday Dutch sentences can be reproduced without error. The SRT was measured in an adaptive procedure, in which the S/N ratio of the sentences was varied by changing the speech level (i.e. the noise level was fixed). The long-term average frequency spectra of speech and noise were always equal for each experimental condition. The original speech material and the corresponding noise signals (with long-term RMS spectra equal to those of the two speakers) had been recorded on analogue tape, after which they were digitised with 16-bit resolution at a sampling frequency of 15625 Hz. We used lists of nine sentences for each experimental condition, preceded by a sample sentence at a S/N ratio of 0 dB. After this sample sentence, which was intended to familiarise the listener with the sound of the upcoming set of sentences, the first sentence of the list was presented at a S/N ratio of -10 dB. The S/N ratio of this sentence was increased by 4 dB at each next presentation, until it was reproduced without error. Each next sentence was presented only once, at a S/N ratio that was computed from that of the previous sentence. The S/N ratio was lowered by 2 dB if the previous sentence had been reproduced correctly, and raised by 2 dB if not.

Listeners were allowed to guess the words they did not understand completely, as is often the case in everyday situations. The SRT of an experimental condition was computed from the S/N ratios specified

after the presentation of sentences 3 through 9. Half of the sentences we used had been spoken by a female talker, the other half by a male talker. Each listener would first hear the sentence lists by the female talker; the second half of the lists were always those by the male talker. The sentence lists were presented in identical order for each listener. The order in which the experimental conditions were applied to the sentence lists was balanced over all listeners.

Results and discussion

The average results of the speech intelligibility tests, for both listener groups, are depicted in Figure 2.

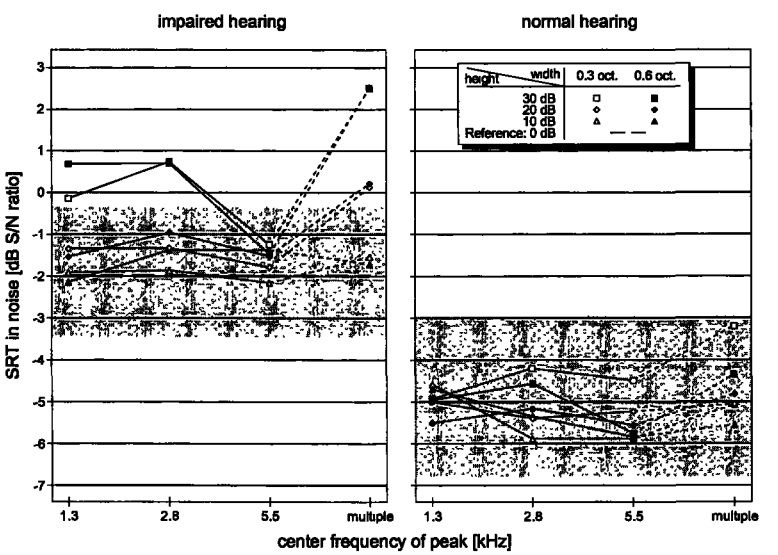

Figure 2. SRTs in noise as a function of the centre frequency of the peaks on the response ('multiple' indicates multiple peaks*). Peak height and width are the parameters, as explained in the legend. Symbols outside the shaded area represent results that are significantly different from the SRT for the reference condition.*

All statistical computations (i.e. averaging & ANOVA) were based on *transformed* data, to compensate for the deviation from the normal distribution in the raw data; the transformation we used was 500/(SRT+30), after which normality was confirmed with a Kolmogorov-Smirnov test. The averages in Figure 2 were obtained by applying the inverse transformation to the averages of the transformed data. Upon combining the data from both listener groups, a repeated-measures ANOVA revealed both a significant overall effect of group

($F(1,34) = 32.69$, $p < 5{\cdot}10^{-4}$), and a significant interaction between group and experimental condition ($F(24,816) = 2.28$, $p < 5{\cdot}10^{-4}$). Therefore, we decided to analyse the data from both groups separately. As was expected, the listeners with hearing impairment had more problems with the reference condition than the listeners with normal hearing. The average S/N-ratio that the group with hearing impairment needed to correctly reproduce 50% of the sentences is -2.0 dB, whereas the group with normal hearing could do the same at a S/N-ratio of -5.0 dB. The effect of the width of the peaks on speech intelligibility is small in both listener groups; averaged over all conditions, it is 0.2 dB for the listeners with impaired hearing, and 0.3 dB for the listeners with normal hearing.

For the listeners with hearing loss (N = 26), the ANOVA indicated significance for the effect of peak configuration on the transformed SRT ($F(24,600) = 15.75$, $p < 5{\cdot}10^{-4}$); Tukey's HSD is 5.2 dB^{-1}. After inverse transformation, the grey area in Figure 2 results; it comprises the interval in which conditions do not differ significantly from the reference condition. Because there are no conditions that have SRTs below -3.5 dB, this implies that only those conditions that result in SRTs above -0.4 dB are significantly different from the reference condition. This is the case for conditions with 30-dB peaks, except those with single peaks at 5.5 kHz, and for the conditions with three 20-dB peaks.

For the group with normal hearing (N = 10), the effect of peak configuration on the transformed SRT was also significant in the ANOVA ($F(24,216) = 2.77$, $p < 5{\cdot}10^{-4}$). The accompanying value for Tukey's HSD is 1.5 dB^{-1}; there are pairs of averages in this transformed data set that differ by more than the HSD, which is in accordance with the significance found in the ANOVA. However, because the reference condition is roughly halfway within the range of the transformed data, not a single average differs more than the HSD from the reference condition. In Figure 2, this is symbolised by the fact that the grey area contains all averages for the listeners with normal hearing.

Experiment 2: Speech Intelligibility Rating

Method

Five sentences from the set that had been used for the speech intelligibility measurements were used for ratings of speech intelligibility. These sentences were all pronounced by the female speaker, because we

were not interested in the effects of different speakers on subjective speech intelligibility. Each of the five sentences, plus the masking noise, had been filtered according to the 25 experimental conditions. During the experiment, sentences and masking noise were filtered (to obtain the listener-specific frequency spectrum $spec_{SI}$), mixed at a S/N ratio according to the listener's SRT for the condition under concern, and presented to the listener. At the same time, the orthographic representation of the sentence was presented on a PC monitor. Thereby, the listener always knew exactly what text was spoken. The task was to judge how *clear* this sentence sounded; that is, how easily the text could be extracted from the noise. Judgements were made on a five-point rating scale, on which the categories were labelled with the Dutch equivalents of "very UNclear", "UNclear", "average", "clear", and "very clear"; the listeners pushed one out of five marked keys on the PC keyboard to report their judgements. For the statistical analysis of the data, we transformed the judgements into numbers in the range of -2.0 to 2.0 (i.e. each discrete point on the scale was assigned an integer value).

Although the S/N ratio was always set according to the SRT, intelligibility may have varied slightly from sentence to sentence, because each SRT had been measured with a list of nine sentences. But we expect these variations to be averaged out over listeners, because the order of presentation of the conditions was balanced over all listeners (as in the speech intelligibility experiments). The five sentences were always presented in the same order, returning to the first sentence after the fifth sentence had been presented. To improve reliability, each condition was presented twice; at the second presentation, we used a different sentence than at the first presentation, to avoid systematic sentence effects on the judgements of the listeners. The results from these two judgments-per-condition were averaged before the data were analysed.

Results and discussion

For both listener groups, the median results of the speech intelligibility ratings are depicted in Figure 3. Already from these graphs, one would conclude that the effect of the peak configurations that we tested is very limited; the medians of all judgements, especially in the group with hearing impairment, are close to "average". For the listeners with normal hearing, Friedman's ANOVA did not show a significant effect of peak configuration ($\chi^2(24) = 28.38$, $p = 0.244$) at the 5%-level. For

the listeners with hearing impairment, the effect of peak configuration was just below 5% significance in Friedman's ANOVA ($\chi^2(24) = 36.74$, $p = 0.046$); however, the Wilcoxon tests, carried out subsequently, indicated that not one condition differed significantly from the reference condition in this group.

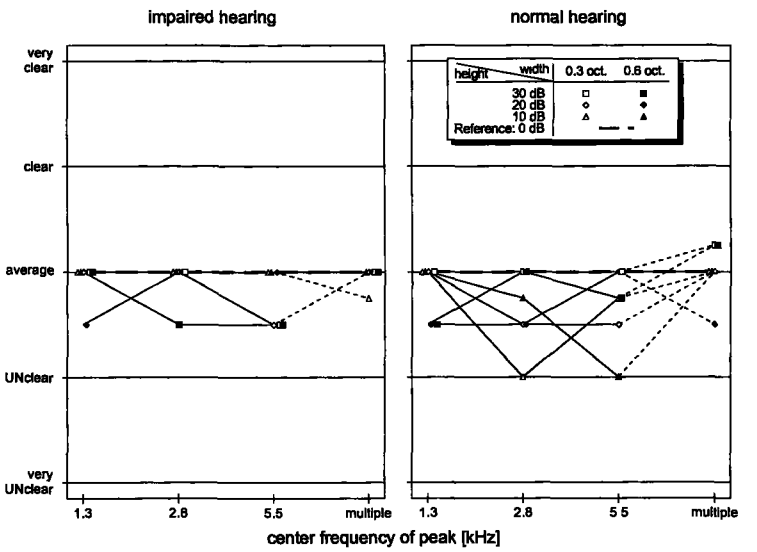

Figure 3. Median intelligibility judgements for speech in noise; S/N ratio according to SRT. Overlapping symbols have been offset horizontally.

Concluding, it seems that when intelligibility is equalised according to the results from the SRT test, the introduction of peaks in the frequency spectrum of speech does not present an extra problem to the listeners with hearing impairment.

Experiment 3: Sound-Quality Ratings

Method

The sound quality ratings were carried out with much the same set-up as the speech intelligibility ratings. The stimuli were (a) a set of five different sentences by the female talker (the set that had been used in the speech intelligibility ratings), and (b) four different fragments of music, with each tested separately; this amounts to five sound-quality rating sessions. In each of these sessions, we used the same five-point rating scale; the categories on the scale were labelled with the Dutch equivalents of "very UNpleasant", "UNpleasant", "average", "pleasant",

and "very pleasant". For the statistical analysis, we assigned numbers in the range of -2.0 to 2.0 to the rating-scale items (as with the intelligibility judgements).

Speech

In the sound-quality rating of the sentence set, we presented speech in quiet. During the experiment, the sentences were filtered to obtain the desired listener-specific spectrum (in this case: $spec_{SQ}$, which is halfway between the threshold and UCL levels). As was the case in the intelligibility judgements, the presentation order of the conditions was balanced over all listeners; each condition was presented twice with different sentences. The listeners were asked to judge the *pleasantness* of the voice they heard, that is, intelligibility or intonation should not be taken into account.

Music

The music fragments were taken from the following compositions:
1. "Opzij" by Herman van Veen (German flute, piano, and voice),
2. "Te Deum" by M.A. Charpentier (trumpet and orchestra),
3. "Drive" by The Cars (drums, synthesizer, and voice), and
4. "Mazurka in C", op. 56 no. 2 by F. Chopin (piano).

The average length of the fragments was 3.8 s and they were cut from Compact-Disc tracks after resampling at 15625 Hz. As with the speech fragments, the music fragments had been pre-processed in order to introduce the peaks into their frequency spectra. During the experiment, the filter that had been used in the speech quality experiment was applied to the music fragments as well. Thereby, the differences between the average spectra of the music fragments, and between the music fragments and speech, were maintained. This enabled us to study differences in quality between speech and music, and also mutual differences between the music fragments. The music fragments had been scaled digitally in order to put their average spectra at the approximate level of the average speech spectrum.

Results and discussion

Speech

For both listener groups, the results of the sound-quality judgements of speech are depicted in Figure 4.

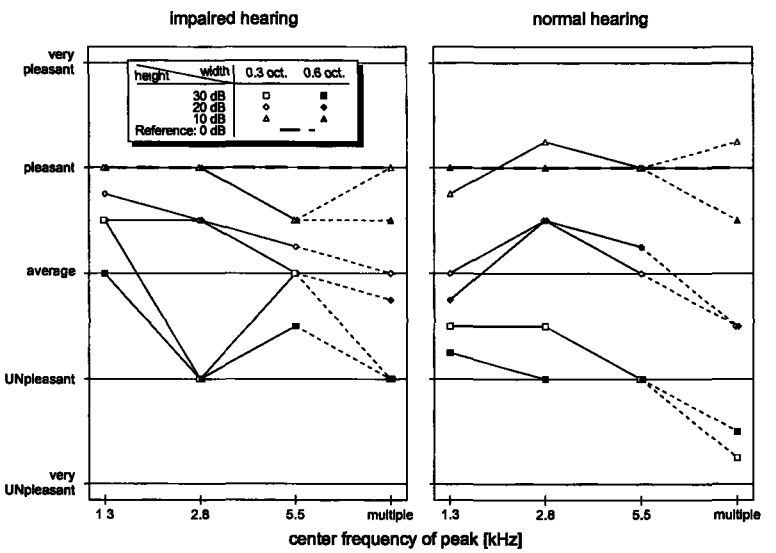

Figure 4. Median pleasantness judgements for speech in quiet. Overlapping symbols have been offset horizontally.

As can be seen clearly from the graphs, the general trend in the data is very much the same for the two groups. The effect of peak configuration is significant in both listener groups (normal hearing: $\chi^2(24) = 166.6$, $p < 5{\cdot}10^{-4}$; impaired hearing: $\chi^2(24) = 296.4$, $p < 5{\cdot}10^{-4}$). For the group with impaired hearing, the Wilcoxon test indicated that all conditions with 30-dB peaks, the conditions with three 20-dB peaks, and the condition with one "wide" 20-dB peak at 2.8 kHz, were judged significantly less pleasant than the reference condition. Because of the small number of listeners with normal hearing, the Wilcoxon test did not detect a significant difference between the reference condition and any other condition in this group. The reference condition turns out to be among the best frequency spectra for both listener groups, as far as pleasantness of speech is concerned.

Music

The results of the sound-quality judgements of music fragments, for the two listener groups, are depicted in Figure 5.

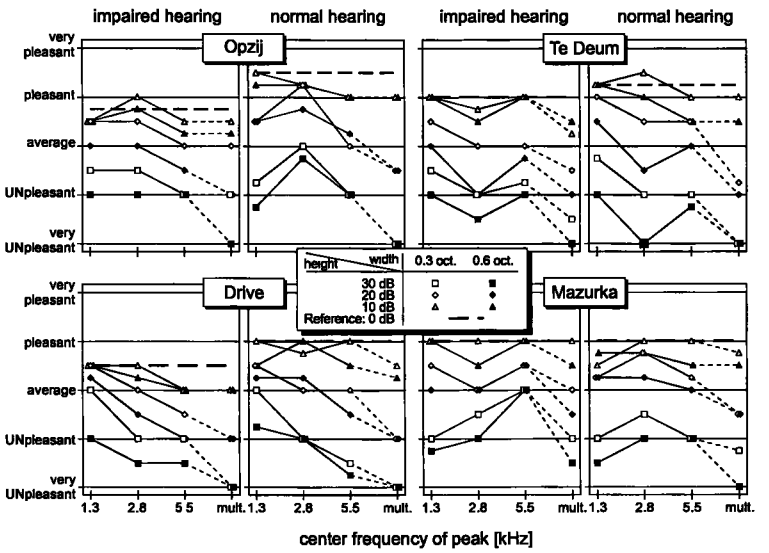

Figure 5. Median pleasantness judgements for music for each of the four fragments. Overlapping symbols have been offset horizontally.

The results of the judgements of the three fragments "Opzij," "Te Deum," and "Drive," are very alike; with increasing centre frequency of the single peaks, the stimuli are generally judged more unpleasant. The results for the fourth fragment, "Mazurka," show less influence of peaks at higher frequency, especially in the group with hearing impairment. This difference can be explained from the average frequency spectra of the fragments: because the fourth fragment consists of piano music, it contains less energy at the higher frequencies than the other three fragments (at 5.5 kHz, the average level in the "Mazurka" fragment is 20 dB lower than in the other fragments), which makes peaks in the frequency response less noticeable there. Statistically, the effect of the peak configuration is significant in the results of each of the judgement sessions, for both listener groups (group with normal hearing: "Opzij", $\chi^2(24) = 175.7$; "Te Deum", $\chi^2(24) = 181.2$; "Drive", $\chi^2(24) = 164.6$; "Mazurka", $\chi^2(24) = 160.1$; group with impaired hearing: "Opzij", $\chi^2(24) = 335.9$; "Te Deum", $\chi^2(24) = 421.8$; "Drive", $\chi^2(24) = 422.3$; "Mazurka", $\chi^2(24) = 353.4$; for

all fragments and both groups, $p < 5 \cdot 10^{-4}$). As in the results of the speech judgements, the results of the two listener groups are similar.
Again, for each individual fragment and in both listener groups, the reference condition is among the most pleasant conditions. For the listeners with hearing impairment, the conditions that always differ significantly from the reference condition are the same as in the speech judgements, with one exception: the 20-dB "wide" peak at 2.8 kHz does not introduce a significant difference for the "Opzij" fragment. Depending on the specific fragment, conditions with single 20-dB peaks and with three 10-dB peaks are also significantly less pleasant than the reference condition.
In comparison with the speech pleasantness judgements, the outcome of the music pleasantness judgements seems to point at a lower acceptability of peaks in the frequency response, most notably by the listeners with hearing impairment. In the case of speech, peaks are judged less pleasant only at a height of 20 dB or more, while in the case of music, this already occurs at a height of 10 dB (although only for three simultaneous peaks).

General Discussion

When we consider the results of the speech intelligibility tests, it is clear that listeners with impaired hearing need a larger S/N ratio in every test condition to understand 50% of our sentences, as compared to the listeners with normal hearing. In addition to the significant effect of listener group that was found in a repeated-measures ANOVA of the combined SRT data, this was confirmed by performing a t-test for each condition, that showed all SRTs from the group with impaired hearing to be significantly higher ($p < 5 \cdot 10^{-4}$ for each condition) than the corresponding values from the group with normal hearing. Moreover, for some conditions, the *increase* in SRT (*re* the reference condition) as a consequence of adding peaks to the amplitude-frequency response is significantly greater in the group with impaired hearing ($p \leq 0.001$). All this seems to be in accordance with the greater effects of Upward Spread Of Masking (USOM) that have been measured in listeners with impaired hearing (Gagné, 1988; Jerger, Tillman, & Peterson, 1960; Trees & Turner, 1986). Because the hearing losses in our group of listeners can be classified as mild to moderate, it is possible that listeners with more severe hearing losses will experience difficulties in speech understanding already at smaller peak heights. In the data from our group of listeners with hearing impairment, however, we could not

identify a significant relation between the average thresholds (both the overall average and the average of the thresholds above 1 kHz) and the adverse effects of peaks on speech intelligibility. Neither was there an indication of a relationship between the *slope* of the hearing loss and the change in SRT *re* the reference condition.

Because vowel identification is based on the discrimination of maxima (formants) in the envelope of the frequency spectrum, the deliberate introduction of peaks into the frequency spectrum of speech can be regarded as a source of confusion in formant extraction, possibly producing vowel identification errors. To gain some insight into the mechanisms that cause the degradation of speech intelligibility as found in the current experiments, we performed an informal experiment in which we asked listeners with normal hearing to identify filtered nonsense-CVCs that had been processed by our filters. The procedure for this experiment has also been employed by Steeneken (1992). Only a limited set of experimental conditions (those with three simultaneous "wide" peaks of heights 0, 10, and 30 dB) was considered. The nonsense-CVCs were embedded in a carrier phrase, and had been pronounced by a female and a male speaker. The results show that in the 30-dB condition, especially for the male speaker, the vowel score is affected more seriously than the consonant score. This differs from conditions with unfavourable S/N ratios, band limiting in the frequency domain, etc. (as described by Steeneken), in which both vowels and consonants are about equally affected.

As a more quantitative check for the explanation of the speech-intelligibility results, we computed the Speech Intelligibility Index (SII; as proposed to ANSI, 1992) for each of the 25 experimental conditions, and for both listener groups. Because the SII incorporates the effect of USOM, it can be used as a validation for the suggestion that the introduction of peaks in the frequency response will increase the masking of higher-frequency speech components by components at the peak. We performed two sets of calculations, one for the listeners with normal hearing, and one for the listeners with impaired hearing. For all calculations, we assumed a S/N ratio of -5 dB, which corresponds to about 50% sentence intelligibility under normal conditions and in normal ears (Plomp & Mimpen, 1979). The two sets of SIIs (for both listener groups, there is a SII for each condition) show a fairly good correlation with the sets of average SRTs that we measured in the respective groups; for the listeners with normal hearing, this correlation is -0.69, and for the listeners with impaired hearing, it

amounts to -0.92 (in both cases, $p < 5 \cdot 10^{-4}$). Because USOM is the only factor that causes differences in the SII, it seems a plausible explanation for the results of the speech-intelligibility experiments.

When we consider the results of the pleasantness judgements, it turns out that the judgements of processed music were generally lower than those for speech after the same processing, although differences are not dramatic. A different result was obtained by Toole and Olive (1988), who determined the detection threshold (i.e. audibility, not acceptability) of resonances. They concluded that "In general, pink or white noise are the most sensitive indicators of these resonances, with speech and music progressively less sensitive" (p. 138), suggesting that speech is more sensitive to coloration by resonances than music. Taking into account that their experimental task was resonance *detection* rather than *acceptability judgement*, this seems to indicate that, although resonances are detected more easily with speech stimuli, they are less annoying in that situation.

From an informal inspection of hearing-aid specifications, we have the impression that peak heights of up to 20 dB do, although not often, occur in the frequency responses (as measured on a coupler) of today's hearing aids, especially in those aids that have been designed for the compensation of severe hearing impairments. The peaks in the response of these hearing aids enable them to deliver the desired high output levels. From the present results, we can conclude that this is achieved at the cost of speech intelligibility and sound quality. Carlson (1974) described a quite elaborate solution to the irregularity problem, using an extra tube in parallel to the "actual" sound-conducting tube. But quite often, reducing the peaks' heights can be as simple as placing a small piece of absorbing material inside the tubing between the hearing aid and the earmold (Killion, 1981; Lybarger, 1985), which should certainly be considered in view of the present results. In the case of a high-power hearing aid, reducing the peaks may result in insufficient amplification for certain listeners with severe hearing impairments, which should stimulate hearing-aid manufacturers to provide smooth frequency responses in their high-power models as well. Another encouragement to this effect is that in Norway and Sweden, hearing aids with peak-to-valley ratios above 8 dB (for peaks and valleys less than 0.67 octave apart, and below 3.5 kHz) will not be accepted for sale (Nordic Committee on Disability, 1994); preferably, these peak-to-valley ratios should not exceed 5 dB. In view of the results of our research, these requirements are "on the safe side".

Conclusions

In this paper, we reported the effect of peaks in the frequency response of a hearing aid on speech intelligibility and sound quality, for listeners with normal hearing and listeners with impaired hearing. The results can be summarised as follows:

1. In some extreme conditions, speech intelligibility is affected more seriously for listeners with impaired hearing, as compared to listeners with normal hearing. For the reference condition, the difference between the average SRTs of the two listener groups is 3.0 dB, whilst for some of the conditions with 30-dB peaks, it increases to more than 6 dB.
2. Sound quality is influenced in much the same way for listeners with hearing impairment and listeners with normal hearing.
3. A smooth frequency response is consistently among the best of the tested responses, for both speech intelligibility and sound quality.
4. Peaks are less annoying to speech than they are to music.

Acknowledgements

This research was financially supported by Philips Hearing Instruments. We would like to thank Theo S. Kapteyn, of the University Hospital VU, Audiology Centre, for his kind assistance in selecting the listeners with hearing loss. We thank Mr Ten Hoeve (at Philips Hearing Instruments, Eindhoven, The Netherlands) and Björn Hagerman (at Teknisk Audiologi, Stockholm, Sweden) for clarifying discussions on the issue of peaks in the hearing aid's frequency response.

Carol A. Sammeth and two anonymous reviewers are acknowledged for their useful suggestions after reading an earlier draft of this paper.

Chapter 4. Compression and expansion of the temporal envelope: Evaluation of speech intelligibility and sound quality

Sensorineural hearing loss is accompanied by loudness recruitment, a steeper-than-normal rise of perceived loudness with presentation level. To compensate for this abnormality, amplitude compression is often applied (e.g., in a hearing aid). Alternatively, since speech intelligibility has been modelled as the perception of fast energy fluctuations, enlarging these (by means of *expansion*) may improve speech intelligibility. Still, even if these signal-processing techniques prove useful in terms of speech intelligibility, practical application might be hindered by unacceptably low sound quality. Therefore, both speech intelligibility and sound quality were evaluated for syllabic compression and expansion of the temporal envelope. Speech intelligibility was evaluated with an adaptive procedure, based on short everyday sentences either in noise or with a competing speaker. Sound quality was measured by means of a rating-scale procedure, for both speech and music. In a systematic set-up, both the ratio of compression or expansion and the number of independent processing bands were varied. Audibility of the stimuli was guaranteed by a listener-specific filter and amplification. Both listeners with normal hearing and listeners with sensorineural hearing impairment have participated as paid volunteers. The results show that, on average, both compression and expansion fail to show better speech intelligibility or sound quality than linear amplification.

Introduction

In many of today's hearing aids, some type of non-linear amplification is employed. In practically all cases, amplification is linear up to a certain input level (the *knee point*), above which the input is amplified by a level-dependent amount. An extreme case of non-linear amplification is clipping, where the knee point equals the maximum output level. But in many other applications of non-linear amplification, the aim is to compensate for the narrower dynamic range of listeners with sensorineural hearing impairment.

According to Caraway and Carhart (1967), non-linear amplification has been applied in hearing aids since about 1936, but these applications were essentially output limiters. Until they wrote their paper, there were no commercially available hearing aids "with compression functioning over all or at least a substantial part of the operating range" (p. 1426). Already at that time, there had been several investigations into the effects of compression limiting on speech intelligibility, but wide-range level-dependent amplification had not been tested. Therefore, they tested speech intelligibility after processing their material with a three-band compressor circuit, in which compression ratios of 2 and 3 were applied to the input signals. The tests were all performed at low sensation levels (up to 24 dB SL) without any masking signals. They found small advantages for non-linear processing in their group with normal hearing (on the order of 15% spondee identification) but smaller (and hardly significant) advantages for their three groups with hearing impairments. This led them to the conclusion that "the aberrations in the loudness function brought about by recruitment do not make the auditor more capable of abstracting information from compressed speech than from uncompressed speech" (p. 1432).

Villchur (1973), however, reported significant improvements in speech intelligibility for six listeners with sensorineural hearing losses, using a system which applied compression in two independent frequency bands with (individually adapted) compression ratios between 2 and 3. Besides improved intelligibility for CVCs, he found that his six listeners preferred the compressed speech to the unprocessed condition. The problem, however, in interpreting Villchur's results is that he only considered the *combination* of compression and frequency shaping, and compared it to linear amplification *without* frequency shaping. Villchur supports this choice by noting that the listeners would simply not have

tolerated frequency shaping without compression, since it will amplify high-frequency noises into the discomfort region.

The results of the authors mentioned above can be seen as exemplary for many others, in that the evaluation of speech intelligibility for compression processing has produced varying results over time. Lippman *et al.* (1981) found compression to generally result in slightly reduced speech intelligibility when compared to linear amplification; compression was superior only when the speech material contained significant level variations or when the input level was low. Nábělek (1983) tested a wide variety of compression ratios and attack/release times, in combination with reverberation, noise masking, and peak clipping. He found, for example, positive effects of compression on nonsense syllable intelligibility in quiet for listeners with hearing impairment, but could not reproduce this on a word intelligibility test, where scores were equivalent for the linear and non-linear conditions. Overall, he concludes that compression is advantageous to speech intelligibility only for certain compression settings (values of ratio, attack/release times) and at larger S/N ratios. Walker *et al.* (1984) compared linear amplification to a combination of compression and expansion (compression above the knee point, and expansion below it), in a system with six independent frequency bands. Time constants ranged from 2 to 5 ms for attack, whilst they were from 30 to 100 ms for release (shorter time constants for the higher-frequency bands). Their results are "mainly negative" (as they put it), since positive effects of their processing on nonsense syllable intelligibility are found only for some listeners in specific listening conditions. Bustamante and Braida (1987) tested various compression algorithms and linear amplification; at best, the compression conditions resulted in speech-intelligibility scores comparable to the condition with linear amplification, although compression processing maintained its score over a greater range of input levels. Levitt and Neuman (1991) also found that none of their principal-component compression algorithms performed better than linear amplification, except (obviously) for the lowest input level (55 dB SPL) in their experiments, where the fixed amount of linear amplification is insufficient. Maré *et al.* (1992) compared three different compression curves to linear amplification. Their results show that, for listeners with normal and with sensorineurally impaired hearing, effects of compression on speech intelligibility in noise are comparable. Advantages of the compression conditions over linear amplification were maximally 9% (considering

initial and final consonants only) for the listeners with hearing impairment. Additionally, a sentence-intelligibility test (according to Plomp and Mimpen, 1979), which estimates the S/N ratio for 50% performance, was used. It showed that the most advantageous compression condition results in an average S/N ratio of only 0.4 dB below the value for linear amplification, indicating that speech intelligibility is only marginally easier in the compression condition. Verschuure *et al.* (1994) evaluated the effects of frequency shaping and compression on speech intelligibility by listeners with impaired hearing. Their one-band compression system provided a maximum gain of about 7% in CVC intelligibility, compared to linear processing, but only for some compression factors. Surprisingly, frequency shaping did *not* have an independent effect for the majority of conditions, which they attribute to anti-upward spread of masking logic in the compressor algorithm. Yund and Buckles (1995a, 1995b) evaluated a compression system in which the compression ratio in each frequency band was fitted individually, based on each listener's threshold. In this way, their compression ratios ranged from 1 to 7; attack and release times were short (on the order of 4 ms). They carried out nonsense-syllable tests in quiet and at S/N ratios from -5 to 15 dB. In one study (Yund and Buckles, 1995a), where compression configurations were compared, they found maximum performance in noise for 8- and 16-band systems, although the increase in intelligibility with respect to a 4-band system (which gave the lowest correct scores at each S/N ratio) was always smaller than 10%. In the second study (Yund and Buckles, 1995b), where linear amplification (with and without frequency shaping) was compared to the 8-band compression system from their first study, they found increasing advantages from the compression system with decreasing S/N ratio; as in the first study, intelligibility gains with respect to the linear-amplification system were small (below 10%).

For listeners with normal hearing, syllabic amplitude compression is not expected to have any beneficial effect on speech intelligibility at all. This is in accordance with views that have led to the Modulation Transfer Function (MTF) concept, upon which the Speech Transmission Index (STI) is based. According to Houtgast and Steeneken (1985), temporal intensity modulations in relevant frequency bands are the actual carriers of speech information. Several well-known disturbances of speech communication (e.g. interfering noise, reverberation) can elegantly be described in terms of a reduction of the modulation

depth in some or all of the frequency bands. For speech-intelligibility experiments in various signal-processing conditions (e.g. reverberation, peak clipping, automatic gain control), Steeneken and Houtgast (1980) found a good correlation between word scores and the STI, thus confirming the importance of intensity modulations for speech intelligibility. The STI will predict *reduced* intelligibility for listeners with normal hearing when syllabic amplitude compression is applied, since all level variations (including intensity modulations) are reduced. It is interesting to see whether the 'opposite' operation, syllabic amplitude *expansion* (i.e. enlarging intensity modulations), will *enhance* speech intelligibility, because the STI so predicts.

Apart from speech intelligibility, sound quality will also be affected by compression and expansion. Agreeable sound quality is very important for the successful implementation of any signal processing scheme in a hearing aid; poor sound quality is likely to prevent any hearing aid from being accepted by listeners with hearing impairment. As opposed to analyses of speech intelligibility, sound quality has rarely been evaluated for amplitude compression and expansion. Already before the results for nonsense-syllable intelligibility were published (Walker *et al.*, 1984), Byrne and Walker (1982) have evaluated their system, in which compression and expansion were combined, for sound quality as well. Their three listeners gave paired-comparison judgements of intelligibility, pleasantness and naturalness of linearly versus nonlinearly processed speech at three SPLs, in quiet and in noise. The vast majority of preferred conditions were those that had been processed linearly, although one of the three listeners, when judging pleasantness of speech in noise, preferred the non-linear processing at all SPLs. Neuman *et al.* (1994) have asked 20 listeners with sensori-neural hearing impairment to give paired-comparison judgements of sound quality for speech in three types of noise, processed by six compression algorithms (attack time: 5 ms, release time: 200 ms) varying only in compression ratio. The results show a monotonically decreasing preference score for increasing compression ratios, with linear amplification being preferred about 70% of the trials, and the highest compression ratio (10) only about 25%. Differences between linear amplification and compression ratios of 1.5 and 2 were not statistically significant. In other words, from the limited experimental results available for sound quality, one can deduce that non-linear processing may not be preferred strongly to linear amplification.

For the experiments reported in this paper, we invited 26 listeners with sensorineural hearing loss, as well as 26 listeners with normal hearing. We have measured speech intelligibility and sound quality for several conditions with syllabic compression and expansion, for which the number of independent frequency bands and the compression (or expansion) ratio were varied systematically. Speech intelligibility was measured in steady-state noise and with a single competing speaker, using short everyday sentences. Sound quality was evaluated for speech in quiet and for four fragments of music, using a rating-scale procedure. Because the signal processing was carried out off-line, we were able to create a system *without* delays. Therefore, the amplification was optimally matched to the actual input envelope level at each time sample, instead of slightly lagging behind (as in most practical implementations of compression).

General Method

Equipment and Listeners

Twenty-six listeners with sensorineural hearing loss were selected from the files of the University Hospital's Audiology Centre. The pure-tone hearing losses at the test ears, averaged for 0.5, 1 and 2 kHz, ranged from 10.0 to 60.0 dB HL (*re* ISO, 1975). The losses in all these ears can be classified as *sloping* to various degrees. As an estimate for the overall slope of the audiogram (from 0.25 to 8.0 kHz), we computed straight-line approximations to the thresholds by means of linear regression analyses. Four ears showed essentially flat losses (slope between -5 and 5 dB/oct), in nine ears, slopes were moderate (between 5 and 10 dB/oct), in ten ears, slopes were steep (between 10 and 15 dB/oct), and in three ears, the slopes were extremely steep (above 15 dB/oct). In quiet, these listeners could reach at least 70% intelligibility for monosyllables and they were free of persistent tinnitus; age ranged from 25 to 75 years, with an average of 58 years. Twenty-six listeners with normal hearing participated in the experiments; in this group, age ranged from 17 to 28 years, with an average of 22 years. For these listeners, the threshold of hearing was maximally 15 dB HL at any test frequency between 0.125 and 8.0 kHz.

The experiments were carried out in a double-walled, sound-proof booth. All test signals were presented monaurally, without masking of the contralateral ear. If the two ears had equal thresholds (e.g. for all listeners with normal hearing), we presented the tests to the preferred

ear; otherwise, the better ear was selected. Off-line signal processing was carried out on a Tucker-Davis AP2 (with an AT&T DSP32C). We used a PC-hosted Digital Signal Processor board (OROS "AU21", featuring Texas Instruments' TMS 320C25) with a 16-bits single-channel D/A converter, to generate the experimental stimuli. The stimuli were presented to the listeners through Sony MDR-CD999 circumaural headphones. The experiments took about 2½ hours per listener, including breaks.

Determination Of Dynamic Range

At the beginning of the experimental session, each listener completed a test in which the dynamic range was measured. The frequency range of interest was from 0.1 to 6.4 kHz, in which we defined 16 adjacent frequency bands with logarithmically equal widths. For each of these bands, the threshold and uncomfortable loudness (UCL) levels were determined. The stimuli were 16 noise bands, each corresponding to a frequency band in the range from 0.1 to 6.4 kHz, and an individually shaped wideband noise (see below) in the second stage of the UCL determination.

The measurement of the threshold level, for each of the frequency bands, consisted of repeated presentation (repetition frequency: 2.4 Hz; stimulus duration: 310 ms; rise/fall times: 10 ms) of the noise band of interest. The listener was instructed to push the space bar of a PC keyboard as long as the noise burst was audible (upon which the level was decreased), and to release it as soon as the noise had become inaudible (whereafter the level was increased again). Starting at a clearly audible level, the noise was attenuated by 2 dB at each next presentation, until the noise first became inaudible. After the listener had released the space bar, the level was increased in 1-dB steps until the spacebar was pressed again. Then, the level was decreased again (in 1-dB steps) until the space bar was released, and so on, until 11 reversals had been registered. At that point, the measurement was terminated and the threshold level was computed by averaging the 10 last turnpoint levels.

UCL levels were measured only for the odd-numbered frequency bands, plus band 16, to restrict fatigue; the UCL levels for the nonmeasured bands were interpolated from the results of the two adjacent bands. We urged our listeners to react as soon as they considered the stimulus uncomfortably loud, rather than using pain sensation as a criterion. For each of the measured frequency bands, a

noise burst was presented repeatedly (burst duration: 310 ms; rise/fall times: 10 ms; repetition frequency: 1.4 Hz). At each next presentation, its level was increased by 3 dB, until the listener pushed the space bar (indicating that the noise had become too loud). At that point, the level was decreased by a random amount between 21 and 31 dB, and the procedure was run again. After six "too loud" reactions, the measurement was terminated and the UCL level was computed by averaging the six levels at which the space bar had been pressed. In succession to the nine narrowband UCL measurements, we shaped a wideband noise such that its frequency spectrum equalled the combined results of the narrowband stimuli. This noise (duration: 3.7 s, representative for the duration of a sentence-in-noise in the speech-intelligibility test) was then used as a stimulus in a second UCL determination, since a wideband stimulus will generally be considered uncomfortably loud already at lower narrowband levels (because of loudness summation). The wideband stimulus was presented repeatedly at increasing levels; after each presentation, the listener had to indicate whether or not the stimulus had been too loud. If not, the level was increased and the stimulus was presented again; otherwise, the last presentation level was taken as the UCL that was used in the computation of the reference frequency spectrum (see below).

The results of the dynamic-range experiments are displayed in Figure 1. In the group with impaired hearing, the average threshold levels are higher in comparison to the group with normal hearing at all frequencies. The difference ranges from 13.6 dB at 114 Hz to 45.7 dB at 5620 Hz (both frequencies are at the centre of a noise band). The average UCL levels are less different; here, differences range from -2.0 dB at 114 Hz to maximally 7.3 dB at 4334 Hz. An ANOVA of the UCL data reveals that this difference is not significant ($F(1,50) = 3.49$; $p = 0.068$) at the 5% level. As a result of the UCL determination with the shaped wideband noise, the UCL levels drawn in Figure 1 are lower than the narrowband results. In the group with hearing impairment, this shift is 26.8 dB, on average; the corresponding value for the group with normal hearing is 30.7 dB.

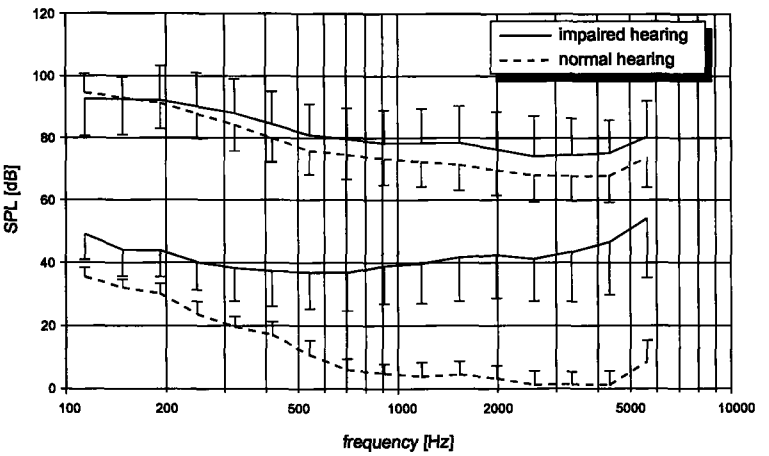

Figure 1. Average dynamic ranges of both listener groups. All levels are for noise bands centred at the frequencies indicated. Upper curves denote the UCL levels (as measured with the wideband stimulus); lower curves are for the threshold levels. Vertical bars are the standard deviations (shown in only one direction) corresponding to the curve they are connected to.

Signal Processing

In the frequency range from 0.1 to 6.4 kHz, we performed band-filtering, envelope detection, envelope compression/expansion, and resynthesis of speech in noise (see Figure 2).

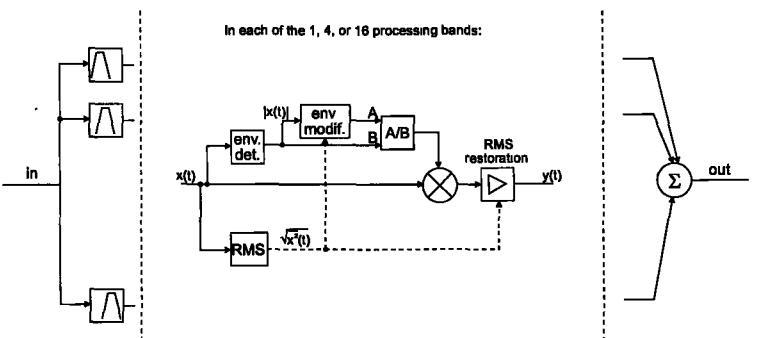

Figure 2. Flow diagram of signal processing. Left: filter bank; centre: non-linear processing (dashed lines indicate the use of the signal's average level, not the time-varying level); right: summation of processed bands.

Since our processing is clearly non-linear, and since we were interested in critical S/N ratios at the input of an imaginary hearing aid, we summed speech and masker signals *before* processing. After the addition

of noise, the signal was always split up into 16 frequency bands by means of a bank of elliptical bandpass filters. The filters corresponded to the 16 frequency bands in the dynamic-range experiments. These filters were applied twice; once to the wideband signal, and once again to the filtered, *time-reversed* signal, to remove any phase shifts introduced in the first pass. In the range of about -3 downto -30 dB of a filter band's frequency response, the resulting slopes were about 96 dB/oct. Depending on the experimental condition, the number of independent processing bands was 16, 4, or 1; in the case of 4 or 1 processing bands, the 16 bands were combined (by summation) into the desired number of processing bands. Next, the signal was fed to an envelope detector, comprising a Hilbert transformer (the envelope was defined as the magnitude of the *analytic* signal; see Rabiner and Gold, 1975) and a 32-Hz lowpass filter (to prevent higher-frequency envelope components, such as the pitch, from controlling the processing). The lowpass filter, like the bandpass filters, was applied twice, effecting a system without phase shifts. The resulting envelope signal was processed (i.e. compressed or expanded) and then divided, sample by sample, by the original envelope, resulting in a multiplication factor for each sample in the band signal. Finally, these multiplication factors were applied to the band signal and the input RMS level was restored. For conditions with more than one independent processing band, there was of course a final summation.

The expansion and compression of the lowpass-filtered temporal envelope were carried out on a logarithmic amplitude scale (see Figure 3). The expansion/compression factor determines to how many dB a one-dB change in input level is expanded/reduced. In the case of compression, only envelope levels down to 30 dB below RMS (the compressor's "knee point") were compressed; lower envelope levels were linearly amplified.

Because expansion of the temporal envelope will increase the crest factor (the ratio of maximum amplitude and RMS level), all signals were digitally attenuated by 12.0 dB prior to processing; this attenuation was compensated for by analogue amplification. The processing could not be accomplished in real time and was carried out off-line, before an experimental session. Computation of the stimuli took about 21 hours for each listener.

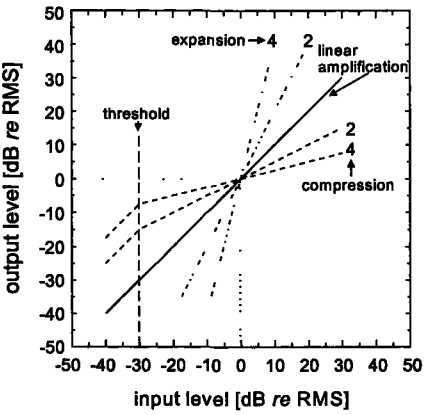

Figure 3. Output level as a function of input level for each of the compression and expansion conditions.

During the experiments, all signals were filtered such that the spectrum of the steady-state noise, expressed in band levels, was halfway between the threshold and UCL levels;

$$spec_B = \frac{thr_B + ucl_B}{2} \qquad (B = 1,...,16)$$

where

$spec_B$ = level of the target spectrum in dB SPL at band B;
thr_B = threshold level at band B;
ucl_B = uncomfortable loudness level at band B.

This on-line filtering was achieved by means of a 256-tap Finite Impulse Response (FIR) filter, that was computed individually from the speech spectrum and the listener's threshold and UCL data. Because the RMS level is restored to its original value after processing (see Figure 2), the average frequency spectrum of the processed speech was the same in all processing conditions.

The speech-intelligibility experiments were carried out for 26 signal-processing conditions. The number of processing bands was 1, 4, or 16; there were four compression ratios (0.25, 0.50, 2.0, and 4.0, with ratios below 1.0 for expansion), and there was of course linear amplification. This adds up to a total of thirteen conditions, each of which was evaluated with the two maskers (steady-state noise and a competing speaker). Linear amplification was realised by processing the speech in one band (i.e. before envelope manipulation, all 16 filter bands were

summed) with compression factor 1.0; in other words, this condition has undergone the same filtering as all other conditions. Because the sound-quality ratings were done for speech and music in *quiet*, there were only thirteen conditions in those experiments.

Tests of Statistical Significance

All data have been analysed by means of non-parametric statistical tests, because the prime assumption for using parametric tests (i.e. a normally distributed data set) was not met. For the evaluation of overall effects, we used a Friedman ANOVA. In case the result of this test was significant, each of the conditions was compared to the linear-amplification condition by means of a Wilcoxon Matched-Pairs Signed-Ranks test. Because there were 12 other conditions, a 5% error rate for the whole set of comparisons requires a significance level of 0.0042 (5% divided by 12) or less for each of the individual comparisons (this procedure is known as the Bonferroni approach for multiple comparisons; see for example Altman, 1992). The results of these comparisons turned out such (in almost all figures) that we were able to separate conditions that were significantly different from linear amplification from those that were not, by means of a horizontal dashed line. In only one case (Figure 6, see below), the dashed line crosses two data points of which one is and the other is not significantly different from linear amplification.

Speech Intelligibility

Method

Speech intelligibility was evaluated by means of a sentence test (Plomp and Mimpen, 1979) in which short, everyday Dutch sentences are presented in a noise background at S/N ratios that are chosen according to a simple up-down procedure. This procedure converges to a S/N ratio at which 50% of the sentences is correctly reproduced, which is defined as the Speech Reception Threshold (SRT) in noise. For each of our experimental conditions, the SRT was determined with a list of ten sentences; the first three sentences in a list were used to obtain an initial estimate of the SRT, while the S/N ratios specified after the remaining seven sentences were averaged to produce the SRT for the condition. Starting at a low S/N ratio (-10 dB in steady-state noise, and -20 dB in a competing speaker), the first sentence was presented as often as necessary, increasing the S/N ratio by 4 dB at each

next presentation. This was continued until the sentence was reproduced correctly. For each next sentence, the S/N ratio was 2 dB lower after a correct response, and 2 dB higher after an incorrect response. Listeners were urged to reproduce the sentences as accurately as they could; at the same time, we encouraged them to guess those words they could not extract.

We used speech from a female and a male speaker in our speech-intelligibility experiments. The test would always start with speech from one speaker, masked by a spectrally matched steady-state noise for all thirteen conditions. After this, speech from the other speaker was presented, but now masked by the speech from the first speaker. The masker signals for the second speaker were chosen randomly from the speech set of the first speaker. Half of the listeners first heard the female speaker; the other half first listened to the male speaker. The order of the sentences within one speaker was identical for all listeners; to prevent order and list effects from systematically influencing the results, we balanced the signal processing conditions over the sentence lists.

Results and discussion

Figure 4 shows the median SRTs for both listener groups and the two masker types.

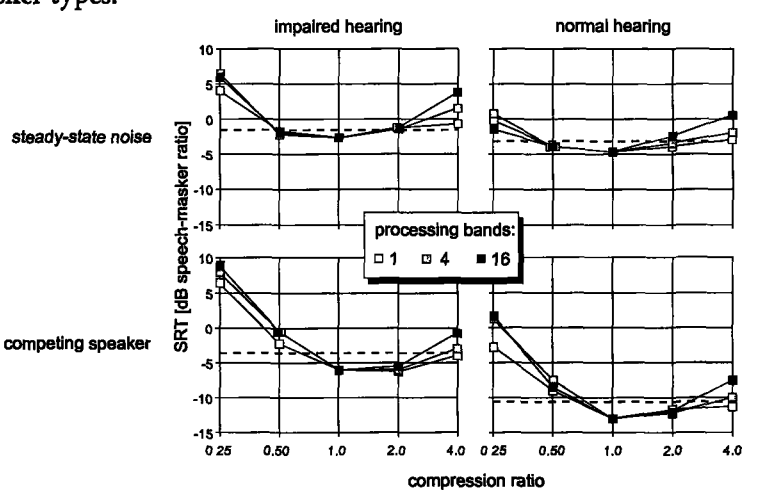

Figure 4. Median SRTs, in steady-state noise (upper panels) and in the presence of a competing speaker (lower panels), as a function of compression ratio (ratios below 1.0 indicate expansion*). The left panels are for the listeners with hearing impairment; the right panels are for the listeners with normal hearing. Dashed lines indicate the level* above *which conditions differ significantly from linear amplification.*

From these graphs, it is immediately clear that linear amplification is generally the best choice for speech intelligibility. Only one condition produces a slightly lower median SRT (i.e. better speech intelligibility) in comparison to linear amplification, i.e. compression by a factor of 2.0 in 1 band with a competing speaker. This small advantage of -0.25 dB occurs only for the listeners with hearing impairment.

Statistical analyses (Friedman's ANOVA) have shown that the main effect of compression/expansion is highly significant in all four situations (listeners with normal hearing: steady-state noise, $\chi^2(12)=165.4$; competing speaker, $\chi^2(12)=218.8$; listeners with impaired hearing: steady-state noise, $\chi^2(12)=204.0$; competing speaker, $\chi^2(12)=234.4$; in all cases, $p < 5 \cdot 10^{-5}$). In order to discriminate between linear amplification and the conditions with compression or expansion, pairwise comparisons were performed. The results of these analyses have been incorporated in Figure 4: those medians that lie above the dashed line in a panel differ significantly from the corresponding condition with linear amplification (see the section "Tests of Statistical Significance").

From Figure 4, it can be seen that the listeners with impaired hearing, compared to those with normal hearing, always need a higher S/N ratio for 50% intelligibility. This observation was confirmed by a series of Mann-Whitney tests, that reported 24 out of 26 conditions to result in significantly higher SRTs for the listeners with hearing impairment (the exceptions are, in the case of steady-state noise, for a compression ratio of 2.0 in either 4 or 16 independent bands). In the case of linear amplification, the difference between the medians is 2.0 dB with steady-state noise; for a competing speaker, it amounts to 7.0 dB. For all other conditions, the difference is of roughly the same order as in the corresponding condition with linear amplification, except for the 0.25 compression factor (i.e. expansion with a factor of 4.0) with steady-state noise, where it rises to as much as 7.5 dB in the 16-band condition.

Sound-Quality Ratings

Method

For two types of stimulus, speech and music (both in quiet), sound quality was judged in rating-scale experiments. These stimuli had been processed off-line according to the same scheme as for the speech-intelligibility experiments. During the experiments, the same frequency

shaping as in the speech-intelligibility experiments was applied to both speech and music. Since we were not interested in the effect of masker signals, there were only 13 conditions in this experiment. The presentation order of the conditions was balanced over the listeners to prevent order effects from having a systematic influence; this sequence was repeated three times to improve reliability. After having listened to a stimulus, the listener had to judge the sound quality by pressing one out of five keys, labelled with the Dutch equivalents (including capitalisation) of "very UNpleasant", "UNpleasant", "average", "pleasant", and "very pleasant". In order to familiarise the listener with the experimental task, and to give an idea of what range the conditions were in for the stimulus under test, ten judgements were carried out prior to the start of the actual test.

Speech

Five sentences from the speech set by the female speaker were used as stimuli for the speech judgement. To prevent processing artefacts (i.e. zero input envelope), spectrally matched noise was added, at a S/N ratio of -30 dB, prior to processing. Since all relevant information in the speech signal is well above this level, this is not considered a problem. Additionally, it effectively masked the background noise of the analogue source tape. In the experiment, the five sentences were always presented in fixed order, returning to the first sentence after having judged the fifth. Because there were 13 conditions, each condition was judged three times with different sentences.

Music

Four different fragments of music were judged in four judgement sessions (i.e. one session per fragment). The music fragments were taken from the following compositions:

1. "Opzij" by Herman van Veen (German flute, piano, and voice),
2. "Te Deum" by M.A. Charpentier (trumpet and orchestra),
3. "Drive" by The Cars (drums, synthesizer, and voice), and
4. "Mazurka in C", op. 56 no. 2 by F. Chopin (piano).

The average length of the fragments was 3.8 s and they were cut from Compact-Disc tracks after resampling at 15625 Hz. To prevent processing artefacts because of a zero input envelope, the input envelope was set to a constant value (of at least 80 dB below the RMS of the band, which was possible because of the very low background noise level) whenever it was below that value.

Results and discussion

For the purpose of statistical analyses, each rating-scale item was assigned an integer value from -2 ("very UNpleasant") to 2 ("very pleasant"); after averaging the trials, these values were analysed.

Speech

Figure 5 shows median values for the sound-quality judgements of speech. The general trend is the same for both listener groups; the more the envelope is compressed or expanded, the worse the sound quality. The overall effect of processing condition is highly significant in both listener groups (impaired hearing: $\chi^2(12) = 195.7$, normal hearing: $\chi^2(12) = 207.6$; in both cases, $p < 5 \cdot 10^{-5}$). Further analysis showed that the only condition for which the judgement is *not* significantly lower than in the case of linear amplification, is compression with a factor of 2.0 in 1 frequency band. Therefore, it is the only condition which lies above the dashed lines in Figure 5 (which separate the significantly different conditions from those not statistically distinguishable from linear amplification). This result is identical for the two listener groups; from a comparison of the two panels in Figure 5, one would conclude that the data from the two listener groups are in good agreement.

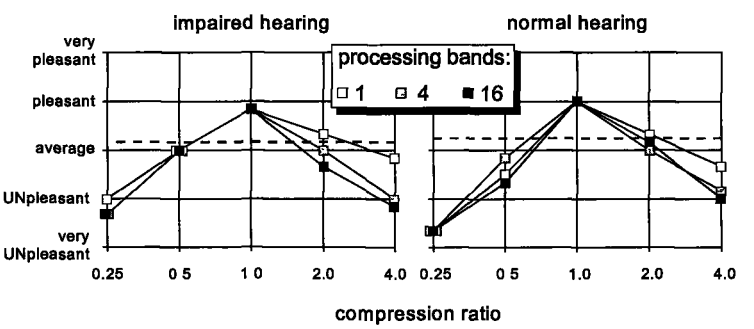

Figure 5. Median pleasantness judgements for speech. Overlapping symbols have been offset horizontally. The left panel is for the listeners with hearing impairment; the right panel is for the listeners with normal hearing. Dashed lines indicate the level below *which conditions differ significantly from linear amplification.*

Music

For each of the four fragments, median judgements are depicted in Figure 6, with the listener groups side-by-side for easy comparison. Just like in the results from the speech judgements, the general pattern is the same in both listener groups, in the sense that the *differences* between processing conditions are of the same sign and of (roughly) the same order of magnitude. But it is also evident that the listeners with hearing impairment are generally "less positive" in their judgements; for all fragments, the data points from the listeners with hearing impairment are at lower-or-equal levels than those from the listeners with normal hearing, except for two conditions of the "Opzij" fragment.

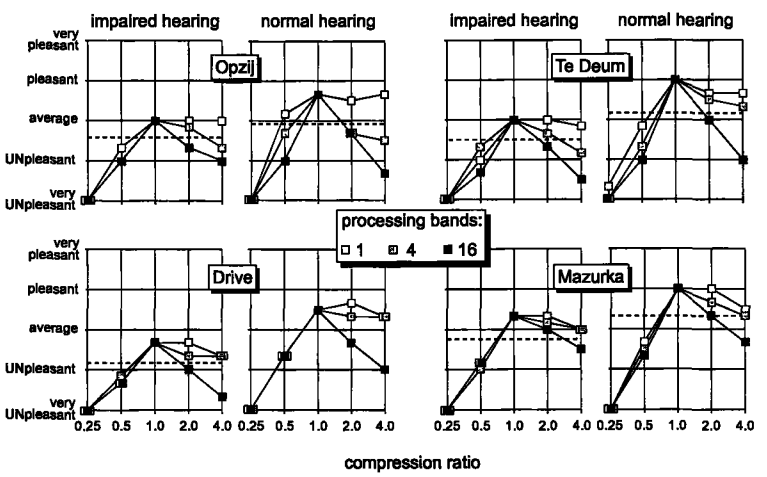

Figure 6. Median pleasantness judgements for music for each of the four fragments. Overlapping symbols have been offset horizontally. The left panels are for the listeners with hearing impairment; the right panels are for the listeners with normal hearing. Dashed lines indicate the level below *which conditions differ significantly from linear amplification.*

The effect of processing condition is highly significant for both listener groups and all fragments (group with impaired hearing: "Opzij", $\chi^2(12) = 170.0$; "Te Deum", $\chi^2(12) = 213.0$; "Drive", $\chi^2(12) = 183.0$; "Mazurka", $\chi^2(12) = 195.4$; group with normal hearing: "Opzij", $\chi^2(12) = 193.9$; "Te Deum", $\chi^2(12) = 197.9$; "Drive", $\chi^2(12) = 195.8$; "Mazurka", $\chi^2(12) = 210.1$; in all cases, $p < 5{\cdot}10^{-5}$). The dashed lines in Figure 6 indicate the level below which conditions are significantly

different from linear amplification. In the lower right panel ("Mazurka", normal hearing) there are two conditions of which one (compression with a ratio of 4.0 in 4 bands) is significantly different from linear amplification, and another (compression with a ratio of 2.0 in 16 bands) is *not*, although the medians are exactly the same (0.33). This may be understood by noting that the graphs show median values, whilst the statistical test compares the entire data sets.

Wideband compression (i.e. a single processing band) with a factor of 2.0 never causes a significant degradation of sound quality. This may be caused by the relative "gentleness" of wideband compression, since only overall level variations are reduced, whilst the relative levels of sounds at different frequencies are preserved. In both listener groups, the conditions in which expansion was applied tend to be judged less pleasant than those with compression. Also, the appreciation of expansion seems to be less dependent on the number of independent processing bands than in the case of compression (where an increase in the number of independent bands results in lower appreciation). However, this is partly caused by the floor effect inherent to the experimental set-up. In the case of expansion with a factor of 4.0, the median judgements are (in seven of the eight panels) at the negative extreme of the rating scale for all variants in processing bands. This does not imply that those conditions are equally unpleasant to the listeners; had the rating scale been extended further into the negative, then the listeners might have used those points to discriminate between the conditions.

General Discussion

The narrow dynamic range in sensorineurally impaired ears is accompanied by a steeper-than-normal growth of perceived loudness (e.g., Hellman and Meiselman, 1990). Intrinsically, it has been supposed that the external compensation of this abnormality by means of a compressing amplifier would also, at least to a substantial degree, restore speech intelligibility to normal levels. This does not necessarily follow, however. First, it might be revealing to measure just-noticeable differences (jnd's) for SPL in listeners with sensorineural hearing impairment, to check whether the steeper growth of loudness is indeed accompanied by increased sensitivity for level *differences* (such as those occurring in modulated signals). If *smaller* jnd's are found, then there is reason to expect that compression of the input to the impaired ear will restore normal loudness- and speech perception. Zwislocki and Jordan

(1986) measured intensity jnd's in both normal and sensorineurally impaired ears. They found that the jnd's were essentially *equal* for the two groups. Comparable results have been reported by Buus *et al.* (1995), who found difference limens for level to be normal for some but *enlarged* in other listeners from a group with predominantly sensorineural hearing impairment. Second, since speech intelligibility is based on the perception of *rapid* energy fluctuations (i.e., modulations of the temporal energy envelope), it is of interest to test the *temporal* acuity of the sensorineurally impaired ear. Results from such experiments have not been equivocal. For example, Moore *et al.* (1992) measured the modulation detection threshold for a sinusoidally modulated noise band in normal and cochlearly impaired ears; they did not find significant differences between the results from normal and impaired ears, both at equal sound-pressure levels (SPLs) and at equal sensation levels (SLs). On a similar modulation detection task, but using a *wideband* noise stimulus, Bacon and Gleitman (1992) found that at the lowest SL at which they tested both of their listener groups, performance of the listeners with hearing impairment tended to be slightly *better* than that of the listeners with normal hearing. This difference disappeared or was even reverse at higher SLs. The authors stress the importance of using wideband stimuli for demonstrating such effects; further, they suggest that the magnitude of the hearing loss or the exact etiology can be decisive for temporal acuity of the impaired ear (see also Florentine and Buus, 1984). Experiments by Moore and Glasberg (1988) showed that, for the detection of temporal gaps in *sinusoids*, impaired ears often perform better than normal ears at equal SL, while performance is similar at equal SPL. The listeners with hearing impairment showed worse performance than those with normal hearing when temporal gaps in *noise bands* were to be detected. This difference was explained from the intrinsic level variations (modulations) that are present in noise bands. In combination with loudness recruitment (which may enlarge level variations), these modulations may be mistaken for the deterministic gap that is to be detected.

The experiments referenced above do not show a very distinct difference between listeners with normal hearing and listeners with hearing impairment. The same goes for the results of our experiments; the effects of most of the processing conditions are of roughly the same order of magnitude in both listener groups. Furthermore, neither

compression nor expansion were beneficial to either speech intelligibility or sound quality.

The fact that the present data do not show improved speech intelligibility with compression for listeners with sensorineural hearing impairment should be taken for an indication that their abnormal loudness growth is *not* accompanied by an enhanced ability to resolve (compressed) modulations. In other words, the concept of compensating loudness recruitment, which assumes that a steeper loudness growth implies an improved level discrimination, does not seem to be supported by the present experiments either. Plomp (1994) stated that recruitment is not a characteristic of the sensorineurally impaired ear only. He refers to measurements by Hellman and Zwislocki (1964) of the loudness of pure tones in noise-masked normal ears, where recruitment-like curves were found. Recruitment, according to Plomp, "reduces the interaction between sounds," and therefore "should be considered positively, not to be destroyed by compression" (p. 7). In a discussion paper on the effects of amplitude compression on speech intelligibility, Plomp (1988) argued that besides deleterious effects on temporal modulations, spectral contrasts would also be affected by multichannel syllabic amplitude compression. These effects will be greater as the number of independent frequency bands or the compression ratio increases, leading to a completely stationary frequency spectrum in the extreme case of a very large number of narrow frequency bands, in which high compression ratios are applied. Plomp notes that the loss of spectral contrasts is a problem for listeners with hearing impairment, because their frequency resolution is generally *lower* than for listeners with normal hearing.

The fact that the expansion of level variations does not improve speech intelligibility for either listener group also indicates that the concept of regarding modulations as the basic carriers of information has its limits. The Speech Transmission Index (STI, e.g. Houtgast and Steeneken, 1985), which is based upon this assumption, would certainly have predicted improved speech intelligibility after the application of expansion, whilst our results clearly show the opposite. The STI, however, was never verified for conditions like those applied in the present experiments. As stated before, it is a good predictor of speech intelligibility in practical situations with noise or reverberation. The fact that the effects of non-linear processing on speech intelligibility cannot be predicted accurately with the STI has also been found by Drullman *et al.* (1994), although their results show only small errors in

the magnitude of the STI, and by Hohmann and Kollmeier (1995). In our case, the prediction from the STI is in the wrong direction. An explanation for this prediction error may lie in the nature of the modulations as they appear in the summed speech and noise that we used in the intelligibility experiments. Both speech and noise contain modulations, of which those in the noise cause part of the reduction of intelligibility (the other part being associated with the fine structure of the noise; see Drullman, 1995). By enlarging the modulations in the summed signal, both the information-carrying modulations (speech) and the disturbing modulations (noise) are enhanced. Therefore, one cannot guarantee that speech intelligibility will be improved by expansion. Another effect of expansion is that weak components in the speech signal will be made even weaker. Because of temporal masking, these sounds may now be masked, whilst they were not in conditions with linear amplification. The result of this masking may be that the number of useful cues for speech intelligibility decreases, resulting in lower scores.

We have not used a frequency-dependent compression ratio in our experiments. Since recruitment is most notable at those frequencies where the hearing loss is most severe, it might have been feasible to adapt the compression ratio to the available dynamic range. Although we did not implement this in our experiments, our data may allow some insight into the success of adapting the compression ratio to the residual dynamic range. The idea here is that since the majority of our listeners with hearing impairment has a sloping loss, recruitment is more pronounced at the higher frequencies. At these frequencies, a higher compression ratio would be needed for optimal compensation. Supposed this is true, then one would expect speech that was processed with a certain compression ratio to yield better intelligibility for some losses (i.e., for some listeners), but not for others, depending on how well the residual dynamic range matches the level variations that result after compressing the speech. To examine whether this is actually the case in our data, we investigated the relationship between high-frequency hearing loss (i.e., the pure-tone average for 2.0, 4.0, and 8.0 kHz, which ranges from 15 to 78 dB HL) and speech intelligibility for the compression conditions. In our data, we could *not* identify distinct maxima for speech intelligibility, for the range of hearing losses present in our listener group. This is not encouraging for the actual implementation of frequency-dependent compression ratios; it serves as an indication that the problems of listeners with hearing impairment

are more complicated than to be fully compensated by syllabic amplitude compression. Nevertheless, performing experiments which *do* incorporate frequency-dependent amplitude compression will probably be the only acceptable argument in this discussion.

The results of the sound-quality ratings are in accordance with those from the speech-intelligibility experiments, in the sense that linear amplification is always judged to be among the most pleasant conditions. Byrne and Walker (1982), who compared a linear amplifier to a non-linear-amplification system in which compression and expansion were *combined*, report comparable results at this point. We found that, for compression, a greater number of processing bands is generally associated with poorer sound quality. This is similar to what we found in the speech-intelligibility experiments, where a greater number of independent processing bands causes a degradation of intelligibility. Also, increasing the compression ratio causes sound quality to be degraded. This resembles to what Neuman *et al.* (1994) found in their paired-comparison experiments. They report that small compression ratios (i.e., maximally 2) do not cause a significant degradation of sound quality, whilst linear amplification is preferred in most of the comparisons.

So far, we have concentrated on group averages. But even though our data show that compression or expansion is not beneficial to speech intelligibility at the group level, individual listeners may have experienced benefit from certain types of non-linear processing. To analyse whether this has in fact occurred, we will now zoom in on the speech-intelligibility results from the group with hearing impairment, since they are the only real candidates for the practical application of non-linear amplification. The analysis consisted of comparisons of individual SRT values, with their uncertainty not compensated for by taking many listeners together; therefore, the results presented here should be interpreted with some caution. For the group with hearing impairment, there are five conditions with steady-state noise in which not a single listener achieves better speech intelligibility than with linear amplification (viz. expansion with a factor of 4.0 in all three bandwidth variants, and compression with a factor of 4.0 in 4 and 16 independent bands). For the remaining conditions with steady-state noise, at most nine listeners do better with non-linear processing (viz. expansion with a factor of 2.0); their SRTs are lower than in the case of linear amplification by an average 1.2 dB. The competing-speaker conditions show a slightly different result. Here, there are four conditions where not a single listener achieves better speech

intelligibility than with linear amplification (viz. expansion with a factor of 4.0 in all three bandwidth variants, and expansion with a factor of 2.0 in 16 independent bands). At most 14 listeners (some of which are in the above-mentioned group of nine as well) do achieve better results in non-linear-processing conditions; for compression with a factor of 2.0 in either 1 or 4 independent bands, their SRTs are lower than for linear amplification by 2.2 or 2.6 dB, respectively. The dilemma is now clear: in steady-state noise, one non-linear amplification strategy provides benefit to some listeners, whilst in a situation with a single competing speaker, another type of non-linear amplification is of help to the same listeners. If possible at all in a practical hearing aid, and if the modest improvements in SRT are of practical interest (considering the limited reliability of this analysis), this calls for a very advanced signal analysis prior to processing. For the time being, since such analysis has not been implemented in a practical hearing aid, a multiple-programme hearing aid could be of help. But a uniform solution for all listeners with sensorineural hearing impairment cannot, at present, be derived from our results.

A positive aspect of the experimental results reported here is that there are processing conditions (e.g., wideband compression with a factor of 2.0) in which both speech intelligibility and sound quality hardly suffer, as compared to linear amplification. Should the decrements in performance be counterbalanced by enhanced comfort, for example because the listener will not any longer have to manually change the amplification in situations with large loudness variations, then non-linear processing of the type tested in our experiments may be considered for practical application.

Conclusions

1. Neither compression nor expansion, for the variants tested, provide a consistent improvement for speech intelligibility or sound quality.
2. Given the measured effect of expansion on speech intelligibility, the STI (in its present form) should not be used for predictions of speech intelligibility in conditions with such signal processing.

Acknowledgements

This research was financially supported by Philips Hearing Instruments. We would like to thank Rick Aretz and Theo S. Kapteyn, of the University Hospital VU, Audiology Centre, for their kind assistance in selecting the listeners with hearing loss.

Chapter 5. Concluding remarks; a personal note

After working for some five years in the research of hearing, there are of course many issues which were not covered by the specific experiments I was involved in. I would like to indicate which directions I expect to be most revealing for future research; further, I will try to indicate the limitations of the experiments that are reported in this thesis and touch upon some recent developments as they appear in the literature.

It seems correct to state that psychophysics has contributed greatly to the understanding of many aspects of human hearing. It may at present not have delivered a concise model of the hearing system as a whole, but it has made us aware of a great number of interesting capabilities, such as there is the ability to 'hear out' a harmonic from a multitone complex or a single instrument from an orchestra, the ability to extract one voice from a mix of many, etc. Many of these abilities turn out to be degraded in sensorineurally impaired hearing. Until our understanding is complete and the model is there, solutions to overcome hearing impairment will necessarily cover only a few of its symptoms, or apply to only a limited number of (acoustical) situations.

A model of human hearing

The understanding of sensorineural hearing impairment would be helped very much by a well-founded model of the healthy hearing system. There are many scientific papers about aspects that should be incorporated into this model, but as yet a conclusive model has not been presented. The approaches taken towards the development of such a model can be divided into two categories, one being the *physiological* approach, and the other the *psychophysical* approach. To me, physiology seems the more fundamental approach, since it is concerned with the hearing system itself, at the level of signals inside the system. By contrast, psychophysical research regards the hearing system as a black box and is concerned only with the outputs at a behavioural level, in reaction to acoustical inputs. This approach has the advantage of only measuring effects that are subjectively noticed by the listener, as opposed to physiologically measured signals (e.g.,

spontaneous activity in nerve fibers) that may not be 'heard' at all. As is often the case in other areas of science, the combination of the two approaches may prove most efficient in building the model.

As an illustration of how complex the model may be (and as an excuse for not having produced the model myself) I would like to mention results by Robinson and Gatehouse (1995). They tested intensity discrimination in experienced, unilaterally aided (but bilaterally impaired) listeners. Their results show that the aided ear performed better at higher levels, whilst the unaided ear did so at lower levels. The explanation they provide for this phenomenon is that the aided ear has optimised its intensity coding to the higher levels it normally receives (from the hearing aid), whilst the unaided ear has accordingly optimised to a lower level. Thus, the hearing system (the inner ear and/or the central auditory processing) seems to adapt itself over time, to the effect that intensity coding becomes optimally matched to the levels frequently encountered. Harrison *et al.* (1993) found that when the inner ear does not function correctly at a very young age, the central functions develop abnormally. They also state that such degeneration may occur after long-term hearing loss, as a consequence of the lack of electrical stimuli being sent into the auditory nerve. Thus, in case this degeneration is reversible, it may be that the compensation for an impaired inner ear (e.g., by a hearing aid) would also, after a certain period, restore the central auditory functions.

In view of these results, the experiments reported in this thesis may produce different results when such acclimatisation over time is allowed to take place. Because the changes seem to involve an optimisation to sound levels most frequently encountered, it might be that speech intelligibility would improve in many of the tested conditions. One wonders to which condition the hearing system would optimise itself most advantageously in the end. Although it may be a difficult and time consuming task to track such developments of the auditory system, it could have great value in the clinical practice of hearing-aid fitting.

Solutions of limited value

Nowadays' hearing aids are good sensitivity enhancers. They can even provide non-linear processing to compensate for abnormally fast loudness growth (recruitment) associated with sensorineural hearing impairment, but still they will not restore the impaired listener's speech-reception capabilities back to normal. As is clear from this

thesis, the SRTs that listeners with sensorineural hearing impairment generally achieve are some dBs higher than those for listeners with normal hearing in equivalent acoustical conditions. To overcome this handicap, it may be worthwhile to evaluate techniques for increasing signal-to-noise ratios of speech that is immersed in noise. One example of such techniques is the use of highly directional microphones, which will be effective when speech and noise originate from spatially separated sources. Recently, Soede *et al.* (1993a) developed such a system, in which high directivity was realised by processing the signals from an *array* of microphones. They report (Soede *et al.*, 1993b) that a group of listeners with sensorineural hearing impairment, when using the experimental microphone array, achieved an SRT which was about 7 dB *lower*, on average, in comparison to the situation where they used their own (conventional) hearing aid. This means that the array seems to compensate for some of the effects of sensorineural hearing impairment. Interestingly, an improvement of about 5 dB was shown for listeners with normal hearing, under the same experimental conditions (a single speaker in a diffuse noise field).

Another option for separating speech and noise can be applied when the noise and the speech differ in frequency content. By using signal processing which will only attenuate noisy frequency regions (e.g., Van Dijkhuizen, 1991, and Van Tasell *et al.*, 1988), the overall S/N-ratio will be improved and speech intelligibility will be enhanced.

However, the applicability of the solutions mentioned above is limited: with the array instrument, no improvement is to be expected in situations with speech and noise from one source (e.g., a single loudspeaker), whilst frequency-specific attenuation will evidently not work when speech and noise have roughly the same frequency spectrum (which can be the case in situations with many talkers of which one is to be understood).

Speech intelligibility & sound quality: Controversy?

The experiments on sound quality presented in this thesis have not revealed the expected controversy between sound quality and speech intelligibility. Rather, the main conclusion for the conditions that have been tested must be that sound quality generally pointed in the same direction as speech intelligibility, in the sense that when sound quality was significantly affected in a certain experimental condition, speech intelligibility was significantly affected as well in most cases. In other

words, sound quality has been the most restricting property to the conditions described.

In view of these results, one might be tempted to replace all speech-intelligibility tests by sound-quality ratings, under the assumption that judging sound quality will be far easier than measuring speech intelligibility. However, the accordance of speech intelligibility and sound quality that was found in the specific conditions reported in this thesis may be invalid in other situations in which a listener with hearing impairment may find him- or herself.

Laboratory studies: what do they tell?

The goal of the experiments presented in this thesis has been to examine groups of listeners and to make general statements about their hearing abilities. In order to guarantee reproducible results, the experiments were performed in a laboratory where surrounding noise levels etc. could be carefully controlled. Nevertheless, because of the large number of conditions involved and the restricted amount of test material, the reliability of each individual's results is limited. To gain more insight into this aspect, a different approach would be necessary in which a limited number of conditions is tested with a larger amount of test material. Although assumptions about individuals' results may be (and have been) made from the data in this thesis, it has never been the experimenters' prime goal to do so.

In the audiologist's everyday work, rehabilitating each individual with hearing loss is of prime importance, rather than doing sort of an "average" job. The individual seeking help from an audiologist will demand the best solution possible and will not easily accept something that has proved to do good "on average". The results from the experiments in this thesis should therefore be seen as guidelines; they show how the individual will probably behave, but they do not guarantee this. Furthermore, since an audiologist will not send his patients home with the PC, DSP and headphones we used for the experiments, but rather with a necessarily compromised miniature apparatus, some caution would be advisable in making the application of the results to the audiologist's practice.

Chapter 6. Summary

Throughout this thesis, effects related to hearing impairment and to hearing aids were considered. The criteria for the evaluation of these effects were *speech intelligibility* and *sound quality*. In this section, special attention will be paid to the relations between these criteria.

Frequency responses (Chapter 2)

In practical hearing-aid fitting, the audiologist's adjustment time is limited and hearing-aid settings cannot be adapted infinitely. Therefore, this was done in an experimental situation, with the frequency response of a (computer simulated) linear hearing aid as the sole parameter. The results of these experiments are described in Chapter 2. For both speech intelligibility and sound quality, the sensitivity to the frequency-response variations was nonsignificant within a wide range of frequency responses. Moreover, there was no clear difference between the results for speech intelligibility and sound quality, in the sense that the ranges of 'equivalent' frequency responses show a large overlap. In other words, the apparent contradiction between the optimum settings for speech intelligibility versus sound quality, as they are sometimes found in practical hearing-aid fitting, could not be reproduced. This may have been caused by the choice of frequency responses, which may not replicate the frequency responses of actual hearing aids with sufficient accuracy. Alternatively, the fact that all tests were carried out with speech stimuli may have obscured an eventual difference; if this is the case, then the difference for other stimuli is expected to be small.

Peaks (Chapter 3)

In the world of high-fidelity sound reproduction, one of the most fundamental requirements to all equipment is a so-called 'flat' frequency response. The main reason for this requirement is contained already in the words 'high fidelity'; the *sound quality* of a reproduction should be as close to the original as possible and manufacturers take great trouble to approximate this ideal. For some types of hearing aid, however, other factors (such as high acoustic output levels from tiny loudspeakers) have prevailed over high fidelity, resulting in frequency responses far from 'flat'. In Chapter 3, the presence of *peaks* in a

hearing aid's frequency response was evaluated for both speech intelligibility and sound quality. These experiments were carried out for listeners with impaired hearing as well as listeners with normal hearing; besides speech, several fragments of music served as stimuli in the sound-quality judgements. A clear difference between the two groups of listeners emerged in the speech-intelligibility experiments; the results for the listeners with impaired hearing showed significantly poorer speech intelligibility in several conditions with high peaks, whilst for the listeners with normal hearing, these effects were smaller and not significant. Nevertheless, because the peaks causing speech intelligibility to suffer significantly are higher than those which appear in real hearing-aid frequency responses, speech intelligibility is not expected to be affected dramatically by actual hearing aids.

The situation is different for sound quality. Here, results from the two listener groups are less different than is the case for speech intelligibility. Moreover, significantly lower sound quality is found, in the group with impaired hearing, for peak heights that *do* occur in actual hearing aids. In other words, the difference between the results for speech intelligibility versus the appreciation of a certain frequency response, as it had been expected to emerge from the experiments in Chapter 2, now turns up for the peaky responses. Although peaks as they appear in some real hearing aids are not harmful to speech intelligibility, they do affect sound quality; thereby, they may make the wearer of such a hearing aid less happy.

Compression & expansion (Chapter 4)

One of the characteristics of sensorineural hearing impairment is *loudness recruitment*, the abnormal growth of perceived loudness which occurs especially at levels close to the impaired threshold of hearing. It has very often been hypothesised that the compensation of this abnormality, by decreasing the loudness variations before sound reaches the sensorineurally impaired ear (i.e., by applying *amplitude compression* in a hearing aid), would essentially restore the impaired ear's capabilities. Since experimental evidence could be found for both confirmation and negation of this hypothesis, it was carefully re-examined in Chapter 4. As an alternative to amplitude compression, and stimulated by ideas from the Speech Transmission Index (STI, e.g. Houtgast & Steeneken, 1985), amplitude *expansion* was tested as well in Chapter 4. The idea behind expanding (i.e., enlarging) amplitude modulations was that, since reducing these modulations resulted in

decreased speech intelligibility (e.g., Drullman *et al.*, 1994), enlarging them might result in the opposite. Although it was envisioned that amplitude expansion might not prove beneficial to listeners with normal hearing, because the normal ear (supposedly) already makes optimal use of the available speech information, it could still be of help to the listener with sensorineural hearing impairment who apparently has more problems in 'picking out the modulations'. Like in the other chapters, both speech intelligibility and sound quality were tested for all signal-processing variants.

The results from the experiments in Chapter 4 were disappointing in the sense that none of the applied signal-processing strategies proved beneficial to either speech intelligibility or sound quality. The signal processing itself may be considered rather ideal, since there was no phase-shift distortion from the bank of filters used, and there was no time delay between the computed signal envelopes and the signal itself. These distortions will generally occur in a practical non-linear hearing aid. Therefore, the reasons for not finding any improvement may need to be sought in our present conception of sensorineural hearing impairment, which is apparently more than just loudness recruitment, and also more than a simple reduction in the ability to resolve temporal modulations.

Literature

Altman, D.G. (1992). *Practical statistics for medical research* (2nd ed.). London: Chapman & Hall.

American National Standards Institute (1992). "American National Standard methods for the calculation of the Speech Intelligibility Index," ANSI S3.79-199X, Draft V3.0 12/17/92.

Bacon, S.P., and Gleitman, R.M. (1992). "Modulation detection in subjects with relatively flat hearing losses," *Journal of Speech and Hearing Research* **35**, 642-653.

Bentler, R.A., & Pavlovic, C.V. (1986). "Comparison of discomfort levels obtained with pure tones and multitone complexes," *Journal of the Acoustical Society of America* **86**, 126-132.

Breeuwer, M., and Plomp, R. (1986). "Speechreading supplemented with auditorily presented speech parameters," *Journal of the Acoustical Society of America* **79**, 481-499.

Bücklein, R. (1981). "The audibility of frequency response irregularities," *Journal of the Audio Engineering Society* **29**, 126-131; a translation of a 1962 original paper in German.

Bustamante, D.K., and Braida, L.D. (1987). "Principal-component amplitude compression for the hearing impaired," *Journal of the Acoustical Society of America* **82**, 1227-1242.

Buus, S., Florentine, M., and Zwicker, T. (1995). "Psychometric functions for level discrimination in cochlearly impaired and normal listeners with equivalent-threshold masking," *Journal of the Acoustical Society of America* **98**, 853-861.

Byrne, D., and Walker, G. (1982). "The effects of multichannel compression and expansion on perceived quality of speech," *Australian Journal of Audiology* **4**, 1-8.

Byrne, D. (1986). "Effects of frequency response characteristics on speech discrimination and perceived intelligibility and pleasantness of speech for hearing-impaired listeners," *Journal of the Acoustical Society of America* **80**, 494-504.

Byrne, D., & Dillon, H. (1986). "The National Acoustics Laboratories' (NAL) new procedure for selecting the gain and frequency response of a hearing aid," *Ear and Hearing* **7**, 257-265.

Caraway, B.J., and Carhart, R. (1967). "Influence of compressor action on speech intelligibility," *Journal of the Acoustical Society of America* **41**, 1424-1433.

Carlson, E.V. (1974). "Smoothing the hearing aid frequency response," *Journal of the Audio Engineering Society* **22**, 426-429.

Cox, R.M., and Gilmore, C. (1986). "Damping the hearing aid frequency response: Effects on speech clarity and preferred listening level," *Journal of Speech and Hearing Research* **29**, 357-365.

Drullman, R., Festen, J. M., and Plomp, R. (1994). "Effect of temporal envelope smearing on speech reception," *Journal of the Acoustical Society of America* **95**, 1053-1064.

Drullman, R. (1995). "Temporal envelope and fine structure cues for speech intelligibility," *Journal of the Acoustical Society of America* **97**, 585-592.

Festen, J.M., & Plomp, R. (1990). "Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing," *Journal of the Acoustical Society of America* **88**, 1725-1736.

Florentine, M., and Buus, S. (1984). "Temporal gap detection in sensorineural and simulated hearing impairment," *Journal of Speech and Hearing Research* **27**, 449-455.

Gabrielsson, A., Schenkman, B.N., & Hagerman, B. (1988). "The effects of different frequency responses on sound quality judgments and speech intelligibility," *Journal of Speech and Hearing Research* **31**, 166-177.

Gagné, J.-P. (1988). "Excess masking among listeners with a sensorineural hearing loss," *Journal of the Acoustical Society of America* **83**, 2311-2321.

Hamill, T.A., & Barron, T.P. (1992). "Frequency response differences of four gain-equalized hearing aid prescription formulae," *Audiology* **31**, 87-94.

Harrison, R.V., Stanton, S.G., Ibrahim, D., Nagasawa, A., Mount, R.J., (1993) "Neonatal cochlear hearing loss results in developmental abnormalities of the central auditory pathways," *Acta Otolaryngologica* **113**, 296-302.

Hays, W.L. (1988). *Statistics*. New York, NY: Holt, Rinehart, and Winston.

Hellman, R.P., and Zwislocki, J. (1964). "Loudness function of a 1000-cps tone in the presence of a masking noise," *Journal of the Acoustical Society of America* **36**, 1618-1627.

Hellman, R.P., and Meiselman, C.H. (1990). "Loudness relations for individuals and groups in normal and impaired hearing," *Journal of the Acoustical Society of America* **88**, 2596-2606.

Hohmann, V., and Kollmeier, B. (1995). "The effect of multichannel dynamic compression on speech intelligibility," *Journal of the Acoustical Society of America* **97**, 1191-1195.

Houtgast, T., and Steeneken, H.J.M. (1985). "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *Journal of the Acoustical Society of America* **77**, 1069-1077.

International Organization for Standardization (1975). "Acoustics - Standard reference zero for the calibration of pure-tone audiometers," ISO 389-1975(E).

Jerger, J.F., Tillman, T.W., & Peterson, J.L. (1960). "Masking by octave bands of noise in normal and impaired ears," *Journal of the Acoustical Society of America* 32, 385-390.

Jerger, J., & Thelin, J. (1968). "Effects of electroacoustic characteristics of hearing aids on speech understanding," *Bulletin of Prosthetics Research* 10, 159-197.

Kapteyn, T.S. (1994). *Slechthorendheid* (in Dutch), Inmerc bv, Wormer, The Netherlands.

Killion, M.C. (1980). "Problems in the application of broadband hearing aid earphones," in G.A. Studebaker & I. Hochberg (Eds.), *Acoustical factors affecting hearing aid performance* (pp. 219-264). Baltimore, MD: University Park Press.

Killion, M.C. (1981). "Earmold options for wideband hearing aids," *Journal of Speech and Hearing Disorders* 46, 10-20.

Kryter, K.D. (1962). "Methods for the calculation of the Articulation Index," *Journal of the Acoustical Society of America* 34, 1689-1697.

Kuk, F.K., & Pape, N.M.C. (1992). "The reliability of a modified Simplex procedure in hearing aid frequency-response selection," *Journal of Speech and Hearing Research* 35, 418-429.

Leijon, A., Lindkvist, A., Ringdahl, A., & Israelsson, B. (1991). "Sound quality and speech reception for prescribed hearing aid frequency responses," *Ear and Hearing* 12, 251-260.

Levitt, H., and Neuman, A.C. (1991). "Evaluation of orthogonal polynomial compression," *Journal of the Acoustical Society of America* 90, 241-252.

Lippman, R.P., Braida, L.D., and Durlach, N.I. (1981). "Study of multichannel amplitude compression and linear amplification for persons with sensorineural hearing loss," *Journal of the Acoustical Society of America* 69, 524-534.

Lutman, M.E., & Clark, J. (1986). "Speech identification under simulated hearing-aid frequency response characteristics in relation to sensitivity, frequency resolution, and temporal resolution," *Journal of the Acoustical Society of America* 80, 1030-1040.

Lybarger, S.F. (1985). "Earmolds," in J. Katz (Editor), *Handbook of clinical audiology*, 3rd edition, pp. 885-910. Baltimore: Williams & Wilkins.

Lyregaard, P.E. (1986). "On the practical validity of POGO," *Hearing Instruments* 37, 13-16, 147.

Maré, M.J., Dreschler, W.A., and Verschuure, H. (1992). "The effects of input-output configuration in syllabic compression on speech perception," *Journal of Speech and Hearing Research* 35, 675-685.

Miller, R.G., Jr. (1981). *Simultaneous statistical inference* (2nd ed.). New York: Springer Verlag.

Moore, B.C.J. (1982). *An introduction to the psychology of hearing* (2nd ed.). Academic Press Inc., London, United Kingdom.

Moore, B.C.J., and Glasberg, B.R. (1988). "Gap detection with sinusoids and noise in normal, impaired, and electrically stimulated ears," *Journal of the Acoustical Society of America* **83**, 1093-1101.

Moore, B.C.J., Shailer, M.J., and Schooneveldt, G.P. (1992). "Temporal modulation transfer functions for band-limited noise in subjects with cochlear hearing loss," *British Journal of Audiology* **26**, 229-237.

Nábělek, I.V. (1983). "Performance of hearing-impaired listeners under various types of amplitude compression," *Journal of the Acoustical Society of America* **74**, 776-791.

Neuman, A.C., Bakke, M.H., Hellman, S., and Levitt, H. (1994). "Effect of compression ratio in a slow-acting compression hearing aid: Paired-comparison judgments of quality," *Journal of the Acoustical Society of America* **96**, 1471-1478.

Nordic Committee on Disability (1994). "Requirement specification for hearing aids," 4th edition, 1994-10-26.

Phillips, J.P.N. (1964). "On the presentation of stimulus-objects in the method of paired comparisons," *American Journal of Psychology* **77**, 660-664.

Pickles, J.O. (1982). *An introduction to the physiology of hearing*. Academic Press Inc., London, United Kingdom.

Plomp, R. (1976). *Aspects of Tone Sensation*, London: Academic Press.

Plomp, R. (1978). "Auditory handicap of hearing impairment and the limited benefit of hearing aids," *Journal of the Acoustical Society of America* **63**, 533-549.

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

Plomp, R. (1988). "The negative effect of amplitude compression in multichannel hearing aids in the light of the modulation-transfer function," *Journal of the Acoustical Society of America* **83**, 2322-2327.

Plomp, R. (1994). "Noise, amplification, and compression: Considerations of three main issues in hearing aid design," *Ear and Hearing* **15**, 2-12.

Rabiner, L.R., & Gold, B. (1975). *Theory and Application of Digital Signal Processing*. Englewood Cliffs, NJ: Prentice-Hall.

Robinson, K., and Gatehouse, S. (1995). "Changes in intensity discrimination following monaural long-term use of a hearing aid," *Journal of the Acoustical Society of America* **97**, 1183-1190.

Skinner, M.W., Pascoe, D.P., Miller, J.D., & Popelka, G.R. (1982). "Measurements to determine the optimal placement of speech energy within the listener's auditory area: A basis for selecting amplification characteristics," in G.A. Studebaker & F.H. Bess (Eds.), *The Vanderbilt Hearing-Aid Report* (pp. 161-169). Upper Darby, PA: Monographs in Contemporary Audiology.

Smaldino, J. (1979). "A multivariate strategy for prediction of psychoacoustic performance from electroacoustic characteristics of hearing aids," *Journal of the American Auditory Society* **5**, 130-137.

Soede, W. (1990). *Improvement of speech intelligibility in noise: Development and evaluation of a new directional hearing instrument based on array technology*. Ph.D. thesis, Gebotekst, Zoetermeer, The Netherlands.

Soede, W., Berkhout, A.J., and Bilsen, F.A. (1993a). "Development of a directional hearing aid based on array technology," *Journal of the Acoustical Society of America* **94**, 785-798.

Soede, W., Bilsen, F.A., and Berkhout, A.J. (1993b). "Assessment of a directional microphone array for hearing-impaired listeners" *Journal of the Acoustical Society of America* **94**, 799-808.

Steeneken, H.J.M., and Houtgast, T. (1980). "A physical method for measuring speech-transmission quality," *Journal of the Acoustical Society of America* **67**, 318-326.

Steeneken, H.J.M. (1992). *On measuring and predicting speech intelligibility*, Ph.D. thesis. Amsterdam, The Netherlands: University of Amsterdam.

Studebaker, G.A. & Zachman, T.A. (1970). "Investigation of the acoustics of earmold vents," *Journal of the Acoustical Society of America* **47**, 1107-1115.

Ter Keurs, M., Festen, J.M., & Plomp, R. (1992). "Effect of spectral envelope smearing on speech reception. I," *Journal of the Acoustical Society of America* **91**, 2872-2880.

Thompson, G., & Lassman, F. (1970). "Listener preference for selective vs flat amplification for a high-frequency hearing-loss population," *Journal of Speech and Hearing Research* **13**, 670-672.

Toole, F.E., & Olive, S.E. (1988). "The modification of timbre by resonances: perception and measurement," *Journal of the Audio Engineering Society* **36**, 122-141.

Trees, D.E., & Turner, C.W. (1986). "Spread of masking in normal subjects and in subjects with high-frequency hearing loss," *Audiology* **25**, 70-83.

Van Buuren, R.A., Festen, J.M., and Plomp, R. (1995). "Evaluation of a wide range of amplitude-frequency responses for the hearing impaired," *Journal of Speech and Hearing Research* **38**, 211-221.

Van Dijkhuizen, J.N., Anema, P.C., & Plomp, R. (1987). "The effect of varying the slope of the amplitude-frequency response on the masked speech-reception threshold of sentences," *Journal of the Acoustical Society of America* **81**, 465-469.

Van Dijkhuizen, J.N., Festen, J.M., & Plomp, R. (1989). "The effect of varying the amplitude-frequency response on the speech-reception threshold of sentences for hearing-impaired listeners," *Journal of the Acoustical Society of America* **86**, 621-628.

Van Dijkhuizen, J.N. (1991). *Studies on the effectiveness of multichannel automatic gain-control in hearing aids.* Ph.D. thesis, Vrije Universiteit, Amsterdam, The Netherlands.

Van Tasell, D.J., Larsen, S.Y., and Fabry, D.A. (1988). "Effects of an adaptive filter hearing aid on speech recognition in noise by hearing-impaired subjects," *Ear and Hearing* 9, 15-21.

Verschuure, H., Prinsen, T.T., and Dreschler, W.A. (1994). "The effects of syllabic compression and frequency shaping on speech intelligibility in hearing impaired people," *Ear and Hearing* 15, 13-21.

Villchur, E. (1973). "Signal processing to improve speech intelligibility in perceptive deafness," *Journal of the Acoustical Society of America* 53, 1646-1657.

Walker, G., Byrne, D., and Dillon, H. (1984). "The effects of multichannel compression/expansion on the intelligibility of nonsense syllables in noise," *Journal of the Acoustical Society of America* 76, 746-757.

Walker, G., Dillon, H, Byrne, D., & Christen, R. (1984). "The use of loudness discomfort levels for selecting the maximum output of hearing aids," *Australian Journal of Audiology* 6, 23-32.

Yund, E.W., and Buckles, K.M. (1995a). "Multichannel compression hearing aids: Effect of number of channels on speech discrimination in noise," *Journal of the Acoustical Society of America* 97, 1206-1223.

Yund, E.W., and Buckles, K.M. (1995b). "Enhanced speech perception at low signal-to-noise ratios with multichannel compression hearing aids," *Journal of the Acoustical Society of America* 97, 1224-1240.

Zwislocki, J.J., and Jordan, H.N. (1986). "On the relations of intensity jnd's to loudness and neural noise," *Journal of the Acoustical Society of America* 79, 772-780.

Samenvatting

In dit proefschrift worden de resultaten beschreven van experimenten aan spraakverstaan en geluidskwaliteit. Deze experimenten zijn uitgevoerd met zowel slechthorenden (met perceptieve gehoorverliezen) als normaalhorenden. Drie hoofdvragen zijn daarbij achtereenvolgens de leidraad geweest:

1. In welke mate hangen spraakverstaan en geluidskwaliteit af van de globale vorm (niveau, spectrale helling) van de frequentiekarakteristiek van een hoortoestel? (hoofdstuk 2)
2. In welke mate hebben onregelmatigheden (pieken) in de frequentiekarakteristiek van een hoortoestel invloed op spraakverstaan en geluidskwaliteit? (hoofdstuk 3)
3. In welke mate worden spraakverstaan en geluidskwaliteit hersteld door syllabische amplitudecompressie of -expansie? (hoofdstuk 4)

In hoofdstuk 2 worden de experimenten aan de frequentiekarakteristiek van een hoortoestel beschreven. Spraakverstaan (in ruis) en geluidskwaliteit zijn gemeten voor 25 verschillende frequentiekarakteristieken, die alle resulteerden in bovendrempelige spraakspectra; de verschillen betroffen de niveaus en spectrale hellingen van de spraakspectra. De belangrijkste conclusie die uit deze experimenten kan worden getrokken is dat, in een breed gebied tussen de (verhoogde) gehoordrempel en het niveau van onaangename luidheid, noch spraakverstaan noch geluidskwaliteit significant worden beïnvloed door de keuze van een bepaalde frequentiekarakteristiek. Bij heel lage niveaus is de geluidskwaliteit significant slechter, terwijl bij steil negatieve hellingen in het spraakspectrum de spraakverstaanbaarheid significant lager is. Bij heel hoge niveaus worden zowel spraakverstaan als geluidskwaliteit significant slechter.

In hoofdstuk 3 wordt de aanwezigheid van pieken in de frequentiekarakteristiek geëvalueerd voor spraakverstaan en voor geluidskwaliteit. De pieken zijn gesuperponeerd op een gladde frequentiekarakteristiek, die als referentie diende. De resultaten van deze experimenten laten zien dat voor slechthorenden de spraakverstaanbaarheid alleen wordt aangetast door pieken van 30 dB, en door drie gelijktijdig

aanwezige pieken van 20 dB. Normaalhorenden ondervinden geen significante hinder in elk van de condities. De resultaten voor geluidskwaliteit stemmen ruwweg overeen voor de twee groepen luisteraars; bij de slechthorenden wordt de geluidskwaliteit het meest aangetast in het geval van muziek, waar (afhankelijk van de soort muziek) soms zelfs pieken van 10 dB een significante achteruitgang veroorzaken.

In hoofdstuk 4 worden syllabische amplitudecompressie en -expansie toegepast op de stimuli, om de effecten op spraakverstaan en geluidskwaliteit te kunnen testen. Voor beide groepen luisteraars lijkt een kleine hoeveelheid compressie (factor 2) geen significante invloed te hebben op spraakverstaan, maar alle overige condities met compressie of expansie resulteren in significant slechtere prestaties. De resultaten voor geluidskwaliteit stemmen ruwweg overeen met die voor spraakverstaan, hoewel soms (afhankelijk van welke stimulus beschouwd wordt) een grotere hoeveelheid compressie wordt getolereerd; dit geldt in het bijzonder wanneer de compressie breedbandig wordt toegepast. Lineaire versterking is consistent bij de best presterende, zo niet de beste, van de beschouwde condities.

Nawoord

Het zit er op! Ongeveer zeven jaar nadat ik bij de Vrije Universiteit aantrad en ruim een jaar nadat ik er vertrok is dit het tastbare resultaat van vele uren in geluiddichte kamers, in bibliotheken en achter beeldschermen met steeds hogere resoluties en grotere diameters. Na in de vaste-stoffysica te zijn afgestudeerd was de psychofysica een aangename verrassing voor iemand die de fysica een beetje was gaan zien als een vakgebied met veel serieuze collegae maar weinig direkt praktische relevantie. Dat beeld is inmiddels wel bijgesteld.

In de zeven jaren heeft zich een wordingsproces voltrokken waarin velen een aandeel hebben gehad. Het heeft mij mede beïnvloed, gevormd, en daarmee ook deze dissertatie. Een aantal namen mogen niet onvermeld blijven. Allereerst (chronologisch) de beide hoogleraren, Reinier Plomp en Tammo Houtgast, die de grote lijn hebben aangegeven. Vooral Reinier heeft hierbij, als instigator van het onderzoek, een belangrijke rol gespeeld. Tammo heeft dit vloeiend overgenomen, natuurlijk met eigen accenten. Als dagelijks klankbord was er Joost Festen, altijd bereid om mij te assisteren bij het uitzetten van de iets minder grote lijnen. De waarde hiervan kan nauwelijks worden overschat. Voor de techniek was daar Hans van Beek, die mij juist dan hielp als het echt niet meer wilde: heel vreemde foutmeldingen op het beeldscherm, bromstoringen op de hoofdtelefoon. Je liet je niet gek maken door die onderzoeker(s). Dan zijn er de collegae: soortgenoten en studenten, logopedisten en biologen, secretaresses en baliemedewerkers (m/v waar van toepassing), die op wetenschappelijk en intermenselijk vlak het leven inkleurden. Ik denk met plezier terug aan de tijd dat ik samen met jullie op De Boelelaan 1118 werkte, in een al met al zeer divers gezelschap. Ik had het beslist niet willen missen. Allen, goed- of slechthorend, die als „proefpersoon" hebben meegedaan aan de soms 4 uur (!) durende luisterexperimenten, ben ik zeer erkentelijk. Heel veel medische of verwante wetenschap zou niet vooruit komen zonder deze vrijwilligers.

Het heeft allemaal iets gehad van het „volgende hoofdstuk" in mijn wiskundeboek op de middelbare school. Ik keek toen vaak vooruit, benieuwd naar wat er komen ging, en kon me nooit goed voorstellen dat ik dát allemaal straks zou begrijpen. Toch ging het telkens zo; een

maand later was dat volgende hoofdstuk gesneden koek geworden. Nu is „gesneden koek" een wat al te simpele karakterisering voor het vakgebied waarin ik aktief ben geweest; er blijft altijd genoeg over dat niet is beschouwd, laat staan begrepen. Maar voortgang is er wel degelijk geweest en ik hoop dat dit proefschrift daarvan blijk geeft.

Bodegraven, maart 1997

Curriculum Vitae

Ronald van Buuren werd geboren te Jutphaas (tegenwoordig Nieuwegein) op 24 juni 1967. Hij groeide op in Bodegraven en doorliep het VWO aan „De Driestar" in Gouda, dat werd afgesloten met een diploma op 30 mei 1985. In dat jaar werd begonnen met de studie Technische Natuurkunde aan de Technische Universiteit Delft. In de laatste fase van de studie werd gespecialiseerd in de fysica van de vaste stof; het afstudeeronderzoek werd verricht in de groep Metaalfysica van prof. dr ir A. van den Beukel. Op 27 februari 1990 werd de studie bekroond met de ingenieurstitel. Direkt aansluitend werd aan de Vrije Universiteit te Amsterdam, faculteit der Geneeskunde, vakgroep Keel-, Neus- en Oorheelkunde, sectie Experimentele Audiologie, onder leiding van prof. dr ir R. Plomp (later opgevolgd door prof. dr ir T. Houtgast) begonnen aan het onderzoek dat de basis vormt voor deze dissertatie.
Sinds ongeveer 1½ jaar is Ronald van Buuren als wetenschappelijk medewerker verbonden aan TNO Fysisch en Elektronisch Laboratorium te 's Gravenhage.

Stelt er niemand vragen meer

is het niet veelbetekenend
hoe vaak je denkt: wat moet ik nu
diep van binnen regent het
en je hebt geen paraplu
stelt er niemand vragen meer
stelt er niemand vragen meer
die een kind graag stelt
hoe ver is de zon
hoe groot is de maan
en waarom maken de mensen oorlog
mama waarom
papa waarom
wat doet die man daar aan dat kruis

tekst: Rikkert Zuiderveld