INTELLIGIBILITY OF TEMPORALLY DEGRADED SPEECH

A study on the significance of narrowband temporal envelopes

Rob Drullman

VRIJE UNIVERSITEIT

INTELLIGIBILITY OF TEMPORALLY DEGRADED SPEECH

A study on the significance of narrowband temporal envelopes

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 woensdag 14 december 1994 te 13.45 uur
in het hoofdgebouw van de universiteit, De Boelelaan 1105

door

ROB DRULLMAN

geboren te Haarlem

Promotor : prof.dr ir T. Houtgast
Copromotor : dr ir J.M. Festen
Referent : dr ing. H.J.M. Steeneken

Dit onderzoek werd gesteund door de Stichting voor Taal, Spraak en Logica (v.h. Stichting Taalwetenschap), die wordt gesubsidieerd door de Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO).

Voorwoord

Dit proefschrift doet verslag van een onderzoek dat werd uitgevoerd in de periode 1-10-1990 tot 1-7-1994 bij de afdeling Experimentele Audiologie van de vakgroep Keel-, Neus- en Oorheelkunde van het Academisch Ziekenhuis van de Vrije Universiteit te Amsterdam.

Prof.dr ir Reinier Plomp, initiatiefnemer van het onderzoeksproject, wil ik bedanken voor de wijze waarop hij het onderzoek gedurende de eerste twee jaar heeft begeleid. Zijn opvolger en mijn promotor, prof.dr ir Tammo Houtgast, heeft die taak zeer voorspoedig overgenomen. Bij hen en bij dr ir Joost Festen heb ik vele malen steun ondervonden bij het voorbereiden van de experimenten en het schrijven van de artikelen.

Verder dank ik alle (vroegere) collega's voor de plezierige werksfeer en de meer of minder wetenschappelijke discussies.

Contents

CHAPTER 1

General Introduction

Speech perception involves the process of decoding a message from the stream of sounds coming from the speaker. This process can be described at a number of levels, among which audibility ("something is being said") and intelligibility ("that is being said") are two important stages. Audibility implies that the speech sound has to be louder than a certain (masking) threshold; intelligibility implies it contains sufficient relevant information to understand the message. Particularly for intelligibility, we need to know which aspects of the acoustic speech signal are essential for the identification of phonemes, syllables, words, and sentences. In general, loudness and timbre constitute two of the main perceptual features of complex sounds, highly correlated with the physical properties of intensity and frequency spectrum, respectively. With time as a third parameter, we can make up a three-dimensional pattern (intensity, frequency, time) of the speech sound. This pattern can be visualized by means of the spectrogram, which has been a major tool in speech research for many years.

The spectrogram does not display the instantaneous amplitude of the signal, but gives the intensity envelope both in time (horizontal axis) and in frequency (vertical axis). These envelopes are the keys to the acoustic and perceptual study of speech in terms of modulations (cf. Plomp, 1984). Details about the spectral modulations can be obtained by taking vertical cross sections. This provides information about the spectral contrasts that are present in the speech signal. The importance of spectral contrasts to speech intelligibility and the effect of reducing these contrasts was studied by ter Keurs (1992). By taking a horizontal cross section from the spectrogram, we get the temporal envelope for a limited frequency region.

In the latter approach, speech is considered as a summation of a number of frequency bands with amplitude-modulated signals. Each frequency band consists of a fine structure (carrier) and a time-varying envelope. This envelope describes the temporal fluctuations reflecting the sequence of the different speech sounds and may be regarded to represent the information-bearing characteristics of speech. The envelope itself is a complex signal which can be analyzed in terms of temporal modulation frequencies. Such a

emporal modulation spectrum shows the magnitude of the different luctuation rhythms that occur in the envelope. For normal everyday speech connected discourse) the dominant modulation frequency is found to be bout 4 Hz, reflecting the sequence rate of words/syllables.

In transferring the speech signal from talker to listener, temporal nodulations should be preserved faithfully to ensure good intelligibility. The xtent to which the temporal envelope of a series of octave bands remains ntact is given by the modulation transfer function (MTF, Houtgast and Steeneken, 1973; Houtgast and Steeneken, 1985). The MTF typcially letermines the degree of reduction of a range of temporal modulation requencies, when comparing input and output speech. It has successfully een applied as a physical measure for the quality of speech transmission. For the evaluation or prediction of speech intelligbility with various types of distorted) transmission channels, the MTF concept has led to the levelopment of a speech transmision index (STI). Several studies (e.g, Steeneken and Houtgast, 1980; Duquesnoy and Plomp, 1980; Steeneken, 1992) have shown that MTF and STI are reliable measures for the performance of sound systems under disturbances that occur in practice, e.g., noise (affecting all modulations) and reverberation (affecting primarily higher modulation frequencies).

The aim of the present study is to determine which features of the emporal envelope are important for speech intelligibility. More specifically, we want to asses the limits within which temporal modulations can be reduced before having a detrimental effect on intelligibility. Investigating the importance of temporal modulations in the speech signal as presented in this study is rather fundamental, but it can be related to more application-like aspects such as speech coding, vocoder design, and amplitude compression in hearing aids.

For the signal processing, a basic analysis-resynthesis algorithm was developed, in which the temporal envelopes of a series of frequency bands can be manipulated. In this way the envelope can be treated separately from the fine structure. Degradation of the temporal envelope includes low- and highpass filtering to measure the effect of reducing specific modulation frequencies, and nonlinear processing to assess separately the contributions of the envelope peaks and troughs to intelligibility. The effect of these types of temporal degradation on speech reception is studied with (young) normal-hearing listeners.

Throughout this thesis intelligibility will often be expressed in terms of the speech-reception threshold (SRT) for sentences in noise (Plomp and

Mimpen, 1979). The SRT is defined as the speech-to-noise ratio in dB at which 50% of short everyday sentences (representing conversational speech) can be reproduced correctly. The method to estimate the SRT consists of presenting a list of (usually) 13 sentences against a background of steady-state noise of a fixed level. The level of the sentences is changed according to a simple up-down adaptive procedure. To ensure a constant signal-to-noise ratio over the entire frequency range, the masking noise has a frequency spectrum equal to the long-term spectrum of the sentences. With a standard deviation of about 1 dB for normal-hearing listeners, the SRT is an accurate measure for speech intelligibility in critical listening situations. Due to its fixed performance criterion, comparisons between SRTs from different listening experiments can be made easily.

The first part of this study, presented in chapters 2 and 3, addresses the question of which frequencies in the temporal envelope of the speech signal are important for intelligibility. We performed a series of experiments to asses the effect of reducing certain temporal modulation frequencies on sentence intelligibility and phoneme recognition. In chapter 2, the speech processing involves lowpass filtering (smearing) of the temporal envelope in a number of consecutive frequency bands. This is done for various cutoff frequencies (0 to 64 Hz) and bandwidths (¼, ½, or 1 oct). To some extent this can be seen as a temporal counterpart of the work by ter Keurs (1992) on spectrally smeared speech. In continuation of the lowpass envelope filtering, chapter 3 describes a similar series of experiments in which the narrowband envelopes are highpass filtered. Comparison and combination of the low- and highpass filtering results will also be discussed.

In chapters 4 and 5 the importance of temporal modulations is addressed in a different perspective. Instigated by discussions in the literature about the effect of multichannel amplitude compression in hearing aids and the role of the MTF (Plomp, 1988; Villchur, 1989), the relative contribution of temporal envelope and fine structure cues to intelligibility is investigated. In chapter 4, intelligiblity is measured for several manners of nonlinear envelope processing in ¼-oct bands in order to assess the importance of the envelope peaks and troughs. Also, a critical light is shed on the relation between the MTF and intelligibility scores. For the case of speech presented against a noise background, chapter 5 reports some additional experiments to evaluate in more detail the importance of the presumably masked weak speech elements.

Finally, in chapter 6 we focus on the effects of temporal modulation reduction on spectral modulations. A method is presented to quantify the

amount of spectral modulation reduction in general, and the intelligibility of spectrally reduced speech is measured. With these data an attempt is made to interpret the results of the temporal envelope filtering experiments of chapters 2 and 3 in terms of the associated reduction of spectral contrasts.

The respective chapters in this thesis are based on papers that have been published (chapters 2 and 3), accepted (chapter 4), or submitted (chapters 5 and 6) for publication in the Journal of the Acoustical Society of America.

CHAPTER 2

Effect of temporal envelope smearing on speech reception*

Abstract

The effect of smearing the temporal envelope on the speech-reception threshold (SRT) for sentences in noise and on phoneme identification was investigated for normal-hearing listeners. For this purpose, the speech signal was split up into a series of frequency bands (width of ¼, ½, or 1 oct) and the amplitude envelope for each band was lowpass filtered at cutoff frequencies of 0, ½, 1, 2, 4, 8, 16, 32, or 64 Hz. Results for 36 subjects show (1) a severe reduction in sentence intelligibility for narrow processing bands at low cutoff frequencies (0-2 Hz); and (2) a marginal contribution of modulation frequencies above 16 Hz to the intelligibility of sentences (provided that lower modulation frequencies are completely present). For cutoff frequencies above 4 Hz, the SRT appears to be independent of the frequency bandwidth upon which envelope filtering takes place. Vowel and consonant identification with nonsense syllables were studied for cutoff frequencies of 0, 2, 4, 8, or 16 Hz in ¼-oct bands. Results for 24 subjects indicate that consonants are more affected than vowels. Errors in vowel identification mainly consist of reduced recognition of diphthongs and of confusions between long and short vowels. In case of consonant recognition, stops appear to suffer most, with confusion patterns depending on the position in the syllable (initial, medial, or final).

*Paper published in: J. Acoust. Soc. Am. **95**, 1053-1064 (1994a)

2.1 INTRODUCTION

The speech signal is characterized by a spectrum that varies in time. This is clearly illustrated by the spectrogram: The distribution of light and dark spots in vertical direction (frequency) changes continuously in horizontal direction (time). These variations contain the information that is essential for the identification of phonemes, syllables, words, and sentences. For this identification we need a detector which is able to perceive the spectrotemporal differences. Our ear is such a detector.

The ear's resolution in both frequency and time is sufficiently high to perceive the essential acoustical features of the various speech sounds. Depending on the speech material, we even have a reserve capacity. This reserve capacity is rather small for isolated phonemes, but large for sentences. For normal-hearing listeners, the speech-reception threshold (SRT) in noise, defined as the speech-to-noise ratio at which 50% of short everyday sentences are reproduced correctly, is about −5 dB (Plomp, 1986).

An interesting question is: How critical is the resolution in frequency and time for the intelligibility of speech? Recently, ter Keurs *et al.* (1992, 1993a) investigated the effect of smearing in the frequency domain, as a way to reduce spectral contrast. They smeared the envelope of the spectrum over bandwidths varying from ⅛ to 4 oct. The effect of this operation can be considered as a blurring of the formant structure. The results indicate that the SRT for sentences in noise increases as spectral energy is smeared over ½ oct and more, thus exceeding the ear's critical bandwidth.

In the present study we focus on the *temporal* envelope. Temporal modulations of the speech signal have been described in terms of the modulation index (Houtgast and Steeneken, 1985). In all octave bands the most important modulation frequencies (i.e., where the modulation index reaches its peak value) are 3-4 Hz, reflecting the syllable rate in speech. Taking the frequency for which the modulation index is reduced to half its peak value (comparable to the −6-dB point of a filter), one can find relevant modulation frequencies up to about 15-20 Hz in undisturbed speech. The ear's sensitivity for temporal modulations shows a lowpass characteristic, with a 6-dB down point corresponding to a frequency roughly between 25 Hz (Festen and Plomp, 1981; Plomp, 1984) and 100 Hz (Rodenburg, 1977; Viemeister, 1979). From these data we may conclude that for normal hearing the ear's capacity to detect temporal modulations is not a limiting factor in speech perception.

What is the effect of reducing the degree to which temporal fluctuations are present in the speech signal? In case of reverberation, resulting in

attenuation of fast temporal modulations (due to 'filling' of the minima in the waveform by reflected speech), experiments have demonstrated a reduction in intelligibility for sentences (Duquesnoy and Plomp, 1980). With regard to multichannel amplitude compression, Plomp (1988) argues that with small time constants intensity fluctuations (particularly at low modulation frequencies) are attenuated in every channel, resulting in reduced intelligibility. Plomp states that this reduction increases as the compression ratio and the number of channels increase. In a comment on Plomp's paper, Villchur (1989) claims that infinite peak clipping (i.e., 100% compression) in a two-channel compression system would hardly affect intelligibility for normal-hearing listeners. One of the goals of the present study is to quantify this effect as a function of the number of frequency bands.

The significance of the various modulation frequencies for speech communication can be compared with the significance of the various audio-frequencies. For example, in designing channel vocoders, we need to know not only the frequency range (e.g., up to 4 kHz) to be covered by the channels, but also the upper limit of the envelope frequencies required to preserve intelligible speech. Similarly, in applying alternative presentation of speech information to the deaf, we need to know up to which envelope frequency the (tactile, visual) channel must transfer the signal faithfully. The range of modulation frequencies most relevant for speech, as mentioned above, has been determined by means of physical/acoustical measurements, and not by any formal perceptual evaluation. In much the same way, a 25-Hz limit for temporal modulations in up to 100 filter bands was applied in early channel vocoders (cf. Flanagan, 1972). There have been measurements of consonant and vowel intelligibility scores, mostly for diagnostic purposes. But, as far as we know, the limit for temporal modulations has never been determined explicitly by means of intelligibility tests.

As to phoneme perception, several investigators have studied the information contained in the temporal envelope for consonant recognition in cases of limited spectral cues. With noise stimuli modulated by the amplitude envelope of /aCa/ syllables, Van Tasell *et al.* (1987) found poor identification scores, with highly variable performance across (untrained) subjects. Several features (voicing, amplitude, and burst) could be derived from the envelope, but for modulation frequencies up to 20 Hz, they accounted for only 19% of the transmitted information. When /aCa/ stimuli are masked by white noise with the same temporal envelope as the speech waveform, Freyman *et al.* (1991) have shown that nonlinear amplification of the envelope (a 10-dB increase of the consonant portion) has no effect on overall consonant recognition, but it can alter confusion patterns for specific

consonant groups. In cases without limited spectral information, Behrens and Blumstein (1988) found that interchanging the amplitude of various voiceless fricatives in CV syllables resulted in few or inconsistent place of articulation errors. They concluded that, at least for voiceless fricative noise, compatibility of the spectral properties and of formant transitions dominated the effect of amplitude manipulations. It should be noted that the results of all these studies are based on *wideband* amplitude envelopes.

With the present perception experiments we investigated the extent to which speech intelligibility depends on the fast modulations in the temporal envelope of the signal. In a first experiment the intelligibility for sentences in quiet and the SRT for sentences in noise were measured as a function of temporal smearing. In a second experiment the effects on vowel and consonant identification in nonsense syllables were studied.

2.2 METHOD

2.2.1 Temporal envelope smearing

Before describing the actual signal processing applied in this study, we have to consider the possible methods for envelope smearing. In general, there are two ways to reduce envelope variations: Convolution applied to the fine structure (carrier signal) on the one hand, or convolution (lowpass filtering in case of smearing) applied to the envelope on the other.

The essential feature of the former method is energy splatter. An example of this is reverberation, characterized by convolution of the carrier signal (within a frequency band) with a causal one-sided exponentially decaying impulse response. In signal processing, however, one could apply any impulse response, also symmetrical ones. A consequence of convolving the fine structure is that low amplitude and silent intervals are filled with carrier energy imported from adjacent phonemes. The same is true for the troughs of the fast (and weaker) amplitude modulations. As a result of this, the higher modulation frequencies in the temporal envelope are attenuated (typically by 6 dB/oct; Plomp, 1984), yielding a smeared envelope.

In the second method the temporal envelope is smeared directly, viz. by lowpass filtering the original envelope and modulating the carrier signal according to this modified envelope (i.e., multiplying the fine structure by the ratio between the filtered and the original envelope at each point in time). With this method there is no energy splatter. So, contrary to the previous method, there is no filling in case of silent intervals (zero carrier amplitude). However, as the digitized speech material we used was originally

recorded on analog tape (with a signal-to-noise ratio of about 50 dB), tape noise provided just enough carrier signal in 'silent' intervals.

We adopted the second method for smearing the temporal envelope. It has the advantage that the fine structure remains intact and that we can control the process of filtering the envelope by selecting the cutoff frequency and the slope of the filter. In this way it is known how the temporal modulation spectrum changes, and the intelligibility can be evaluated as a function of the temporal envelope cutoff frequency.

Smearing of the temporal envelope should not be done on the wideband speech signal. It is known that those modulations can be reduced considerably, without affecting the intelligibility in a major way. Since the temporal fluctuations of speech are only partly correlated over frequency (the more two frequency bands are separated, the lower their correlation, cf. Houtgast and Verhage, 1991), the wideband signal does not include all temporal amplitude variations in the different frequency bands. Therefore, the speech signal has to be split up into several frequency bands, so that the temporal envelope of each individual band can be modified.

2.2.2 Signal processing

For the signal processing, an analysis-resynthesis scheme for smearing the amplitude envelope of digitized speech was developed. A block diagram of the processing is shown in Fig. 2.1. First, the original wideband speech signal (digitized into 16 bits at a sampling rate of 15,625 Hz) is led through a filter bank with linear-phase FIR bandpass filters, covering the range 100-6400 Hz. The slopes of these filters are at least 80 dB/oct. The amplitude envelope from each band is computed by means of a Hilbert transform (Rabiner and Gold, 1975). The next step is a lowpass filtering of this original envelope to get the modified envelope. In the experiment the cutoff frequencies (−6-dB points) of these lowpass FIR filters ranged from 0.5 to 64 Hz. These filters have to be sufficiently steep at very low cutoff frequencies and, at the same time, manageable for implementation, i.e., their impulse response should have only a few hundred points instead of several thousand. In order to achieve that, the envelope is downsampled by a factor 64 (to a sampling rate of 244 Hz) before filtering.[1] The slopes of the lowpass filters were empirically set to approximately −40 dB/oct, so that the modified envelope will not become negative. After filtering, the envelope is upsampled again. The modified band signal is obtained by multiplying the original band (fine structure) by the ratio of the filtered envelope and the original envelope at each corresponding point in time.

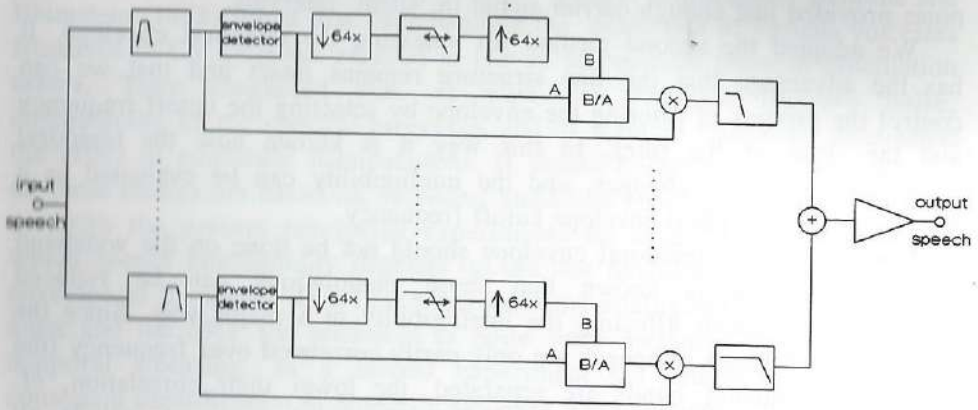

FIG. 2.1. Block diagram of the speech processing. The wideband input speech signal is split up into several frequency bands. For each band the amplitude envelope is determined, downsampled by a factor 64, lowpass filtered, and upsampled. The new band signal is obtained by modulating the original band signal according to the modified envelope. Each new band signal is lowpass filtered to eliminate undesired high frequency noise. After adding all modified bands, the wideband signal is rescaled to match the rms level of the original speech.

As a result of the envelope filtering (especially at low cutoff frequencies), parts of the original band signal having low amplitude are amplified in the modified band signal, particularly just before and after periods with a high amplitude. These modified parts sometimes contain amplified quantization noise of high frequencies (not belonging in the frequency band), causing sharp, clicking sounds. To eliminate these, the modified band signal is lowpass filtered, using a FIR filter with a cutoff frequency 5% above the upper cutoff frequency of the corresponding bandpass filter.[2] Finally, all modified band signals are added and the level of the new wideband signal is adjusted to have the same rms as the original input signal.

All signal manipulations are performed (non-real-time) on an Olivetti PCS 286 computer, using an OROS-AU21 card with TMS320C25 signal processor. Figure 2.2 shows an example of the various stages during the processing of one ¼-oct band of a short sentence.

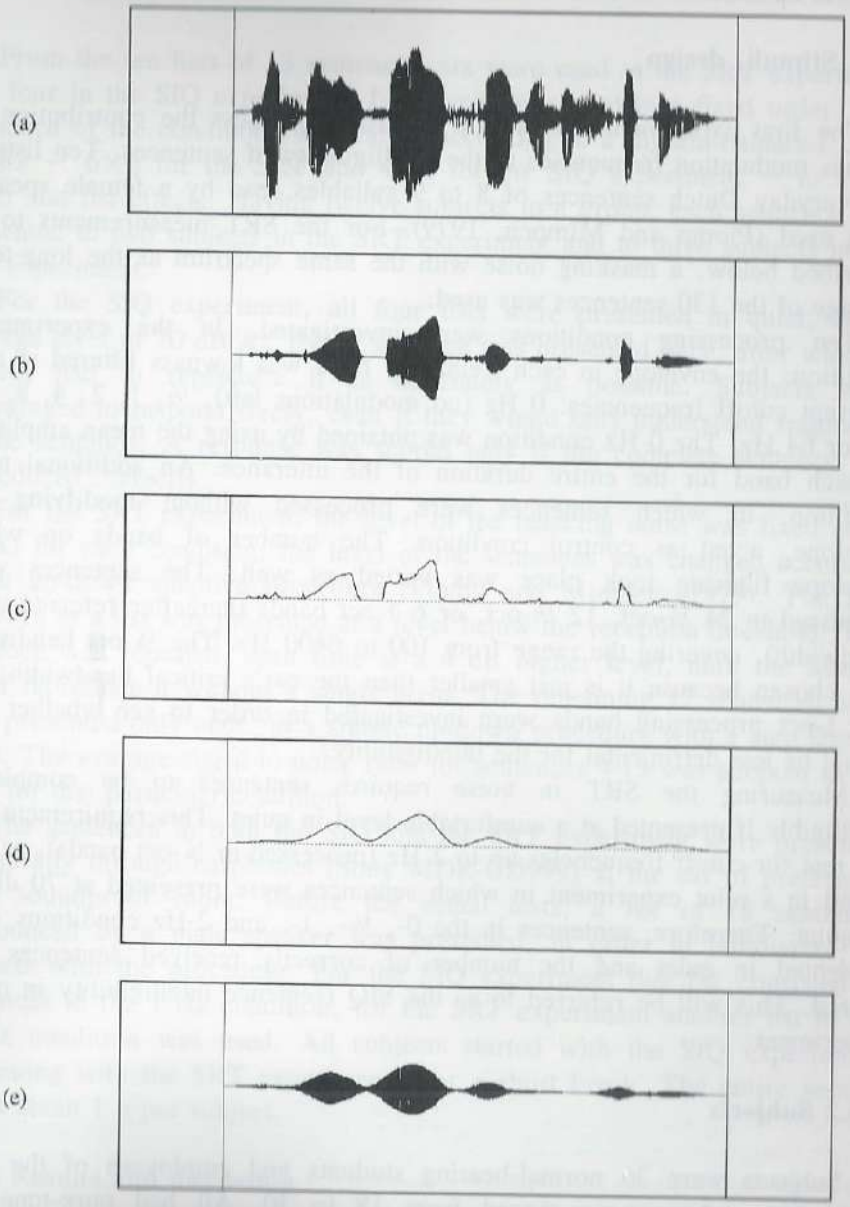

FIG. 2.2. Example of temporal smearing for one sentence (2.2 s duration) in a single frequency band. (a) original wideband signal; (b) ¼-oct band signal (283-336 Hz); (c) amplitude envelope of (b); (d) envelope (c) lowpass filtered with 4-Hz cutoff frequency; (e) resulting modified band signal, amplitude modulated according to (d).

2.3 EXPERIMENT 1: SENTENCE INTELLIGIBILITY

2.3.1 Stimuli, design

The first experiment was set up in order to assess the contribution of various modulation frequencies to the intelligibility of sentences. Ten lists of 13 everyday Dutch sentences of 8 to 9 syllables read by a female speaker were used (Plomp and Mimpen, 1979). For the SRT measurements to be described below, a masking noise with the same spectrum as the long-term average of the 130 sentences was used.

Ten processing conditions were investigated. In the experimental conditions the envelope in each frequency band was lowpass filtered at nine different cutoff frequencies: 0 Hz (no modulations left), ½, 1, 2, 4, 8, 16, 32, or 64 Hz. The 0-Hz condition was obtained by using the mean amplitude of each band for the entire duration of the utterance. An additional tenth condition, in which sentences were processed without modifying the envelope, acted as control condition. The number of bands on which envelope filtering took place was varied as well. The sentences were processed in 24 ¼-oct, 12 ½-oct, or 6 1-oct bands (hereafter referred to as bandwidth), covering the range from 100 to 6400 Hz. The ¼-oct bandwidth was chosen because it is just smaller than the ear's critical bandwidth; ½- and 1-oct processing bands were investigated in order to see whether this would be less detrimental for the intelligibility.

Measuring the SRT in noise requires sentences to be completely intelligible if presented at a comfortable level in quiet. This requirement was not met for cutoff frequencies up to 2 Hz (processed in ¼-oct bands), as was found in a pilot experiment in which sentences were presented at 70 dB(A) in quiet. Therefore, sentences in the 0-, ½-, 1-, and 2-Hz conditions were presented in quiet and the number of correctly received sentences was scored. This will be referred to as the SIQ (sentence intelligibility in quiet) experiment.

2.3.2 Subjects

Subjects were 36 normal-hearing students and employees of the Free University, whose ages ranged from 18 to 30. All had pure-tone air-conduction thresholds less than 15 dB HL in their preferred ear at octave frequencies from 125 to 4000 Hz and at 6000 Hz. They were divided into three groups of twelve, each group receiving the ten conditions for one of the three processing bandwidths.

2.3.3 Procedure

From the ten lists of 13 sentences, six were used in the SRT experiment and four in the SIQ experiment. Lists were presented in a fixed order. The sequence of the conditions was varied according to a digram-balanced Latin square – 6×6 for the SRT and 4×4 for the SIQ experiment – to avoid order and list effects. Having twelve subjects in a group, each sequence was presented to two subjects in the SRT experiment and to three subjects in the SIQ experiment.

For the SIQ experiment, all four lists were presented in quiet, at an average level of 70 dB(A). Every sentence was presented once, after which a subject had to reproduce it as accurately as possible. Subjects were encouraged to respond freely, even if they would only understand fragments of the sentence. A response was scored only if the complete sentence was reproduced correctly.

For the SRT experiment, the level of the masking noise was fixed at 70 dB(A) for each condition; the level of the sentences was changed according to an up-down adaptive procedure (Plomp and Mimpen, 1979). The first sentence in a list was presented at a level below the reception threshold. This sentence was repeated, each time at a 4 dB higher level, until the listener could reproduce it without a single error. The remaining 12 sentences were then presented only once, in a simple up-down procedure with a step size of 2 dB. The average signal-to-noise ratio for sentences 4-13 was adopted as the SRT for that particular condition.

The sentences in both the SIQ and the SRT experiment were presented monaurally through earphones (Sony MDR-CD999) at the ear of preference in a soundproof room. Before the actual tests, a list of 13 sentences pronounced by a male speaker was presented, in order to familiarize the subjects with the procedure. For the SIQ experiment this list consisted of sentences in the 1-Hz condition; for the SRT experiment another list in the 16-Hz condition was used. All subjects started with the SIQ experiment, continuing with the SRT experiment after a short break. The entire session lasted about 1 h per subject.

2.3.4 Results and discussion

The mean results of the SIQ experiment for the four filtering conditions in the three bandwidths are plotted in Fig. 2.3. As expected, the overall performance improves with increasing cutoff frequency. With the ¼- and ½-oct bands, however, the scores vary widely among subjects, resulting in

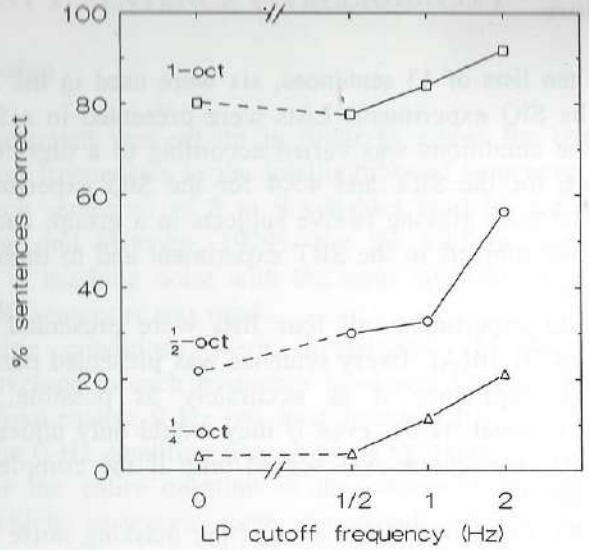

FIG. 2.3. Mean score of sentences in quiet as a function of cutoff frequency, with processing bandwidth as parameter.

large standard deviations. A two-way analysis of variance with repeated measures on the factor conditions, using arcsine transformed scores (Studebaker, 1985), revealed that both the effect of cutoff frequency and the effect of bandwidth were highly significant ($p < 0.001$), whereas the interaction was not. Pairwise comparisons (Tukey HSD test) of the mean scores for the three bandwidths showed the latter were all significantly different ($p < 0.01$). So, for very low cutoff frequencies, intelligibility increases when the frequency bands become broader. As to the four cutoff frequencies, scores in the 2-Hz condition are significantly better than in the other three conditions ($p < 0.01$); the 0-, ½-, and 1-Hz conditions do not differ significantly.

The mean SRT for sentences in noise as a function of filtering condition and bandwidth is listed in Table 2.1. For all bandwidths, the thresholds in the 4- and 8-Hz conditions are clearly higher than in the other conditions. A constant threshold is reached for the 16-, 32-, and 64-Hz conditions, although it is still about 1 dB higher than for the control conditions. The raw data suggest a slightly different threshold for different processing bandwidths in all filtering conditions. Since this difference is also present in the control conditions, we inspected the stimuli more closely. It appeared that processing

TABLE 2.1. Mean SRT in dB with standard deviations in parentheses for the six filtering conditions in three bandwidths.

	Condition					
Bandwidth	4 Hz	8 Hz	16 Hz	32 Hz	64 Hz	Control
¼-oct	−0.1 (1.8)	−3.8 (0.9)	−5.0 (1.2)	−5.6 (1.1)	−5.5 (0.8)	−6.2 (0.8)
½-oct	0.5 (2.5)	−3.3 (1.6)	−4.7 (1.4)	−4.7 (1.4)	−4.7 (1.4)	−5.8 (1.1)
1-oct	−0.7 (1.4)	−2.6 (1.5)	−3.8 (1.2)	−4.4 (1.0)	−4.0 (1.6)	−5.1 (1.2)
Mean	−0.1 (2.0)	−3.3 (1.4)	−4.5 (1.3)	−4.9 (1.3)	−4.7 (1.4)	−5.7 (1.1)

in ¼- and ½-oct bands had resulted in slightly tilted long-term spectra. Because the same unprocessed masking noise was used in all conditions, actual signal-to-noise ratios were slightly more favorable for narrow processing bands. Therefore, we expressed the SRT in all experimental conditions relative to the mean SRT in the control condition for the corresponding bandwidth. This relative SRT is shown in Fig. 2.4. Although it would have been better if we had processed the noise in the same way as the sentences, the results are not essentially affected.

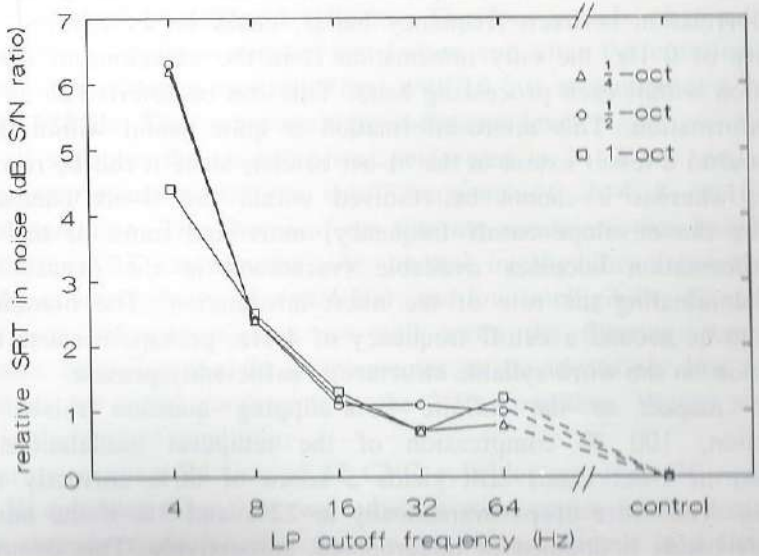

FIG. 2.4. Mean relative SRT in noise as a function of cutoff frequency, with processing bandwidth as parameter.

A two-way analysis of variance on the relative SRT with repeated measures on the factor conditions revealed a highly significant effect of cutoff frequency ($p < 0.001$); there were no bandwidth or interaction effects. *Post hoc* tests (Tukey HSD) showed that increasing the cutoff frequency from 4 to 8 Hz and from 8 to 16 Hz significantly improved the subjects' performance ($p < 0.01$), while the further increase to 16, 32, and 64 Hz did not. The 1-dB threshold shift between the 64 Hz and control condition is just significant ($p < 0.05$). This difference can be explained from our envelope definition, viz. the Hilbert envelope. As a consequence, very high modulation frequencies can be found in the output of broad frequency bands (e.g., modulations by the pitch periods). Lowpass filtering causes these details to disappear. Although modulations above 64 Hz are very small, omitting them apparently results in a 1-dB increase of the SRT.

The results of the experiments indicate that the intelligibility increases progressively with lowpass cutoff frequency up to about 16 Hz. In other words, modulation frequencies in the amplitude envelope above 16 Hz (with lower modulation frequencies present) do not really contribute to understanding ordinary sentences. For cutoff frequencies above 4 Hz, the intelligibility appears to be independent of the processing bandwidth; the beneficial effect of a larger bandwidth is only demonstrated for cutoff frequencies below 4 Hz. An explanation for this, considering there is no 100% correlation between frequency bands, could be as follows. For the limit case of 0 Hz, the only information is in the variations of the energy distribution *within* each processing band. This can be referred to as spectral micro-information. This micro-information is quite useful within the 1-oct bands (and to a lesser extent in the ½-oct bands), since it can be resolved by the ear, whereas it cannot be resolved within the ¼-oct bands. When increasing the envelope cutoff frequency, more and more of the spectral macro-information becomes available (variations in the overall spectral shape), dominating the role of the micro-information. The breaking-point appears to be around a cutoff frequency of 4 Hz; perhaps because then the information on the word/syllable structure is sufficiently present.

With respect to the infinite peak-clipping question raised in the introduction, 100 % compression of the temporal modulations (0-Hz condition) in 1-oct bands still yields a score of 80% correctly received sentences. The score drops dramatically to 22% and 3% if the number of frequency bands is doubled or quadruppled, respectively. This demonstrates clearly that zero-crossing information alone (no modulations in ¼-oct bands) is insufficient for speech intelligibility.

The above experiments do not answer the question of how individual phonemes are affected by temporal smearing. Therefore, vowel and consonant identification were studied in a second experiment.

2.4 EXPERIMENT 2: VOWEL AND CONSONANT IDENTIFICATION

2.4.1 Stimuli, design

The speech material consisted of two types of meaningless syllables. CVC syllables were used for the identification of initial consonants (C_i), vowels (V), and final consonants (C_f); VCV syllables were used for the identification of medial consonants (C_m).

The CVC words were obtained from 24 existing lists of 12 different syllables each (Bosman, 1989), read by the same speaker who had produced the sentences. Vowels and consonants were chosen from three sets of 12 phonemes, all of which appeared once in a list. C_i was choosen from /b, d, χ, h, j, k, l, n, t, v, w, z/, V from /ɑ, a, ɛ, e, I, i, ɔ, o, ø, u, ɑu, ɛi/, and C_f from /f, χ, j, k, l, m, n, ŋ, p, s, t, w/.

For the identification of C_m, we used VCV syllables spoken by the same speaker. Each syllable consisted of one of 16 consonants /b, d, f, χ, h, j, k, l, m, n, p, s, t, v, w, z/ surrounded by one of four vowels /a, i, u, œ/, yielding a total of 64 different syllables. The four vowels were selected to induce different envelope courses just before and after the consonant. Both CVC and VCV syllables were digitized with 16 bits resolution at a sampling rate of 15,625 Hz. They were normalized for rms level.

For all syllables the smearing was performed in 24 ¼-oct bands. There were six experimental conditions, cutoff frequencies 0, 2, 4, 8, or 16 Hz and a control condition. The choice of the filtering conditions was based on the results of the SRT experiments, viz. normal intelligibility (control and 16 Hz), reduced intelligibility (8 and 4 Hz), and low intelligibility (2 and 0 Hz). For the sake of convenience, we will write the filtering condition in parentheses following the set of phonemes to be identified. For example, $C_i(2)$ stands for initial consonants in the 2-Hz condition, V(con) stands for vowels in the control condition.

From the original 24 lists of 12 CVC syllables we made 36 randomized lists of 50 syllables. The first two syllables were copies of the last two and acted as dummy trials, so that there were 48 test stimuli in a list for the identification of C_i, V, and C_f. These lists were constructed in such a way that each initial consonant, vowel, and final consonant appeared four times in different contexts.

From the original four lists of 16 VCV syllables we constructed 36 randomized list of 66 syllables for the identification of C_m. Each list contained all 64 VCV syllables and again the first two syllables were copies of the last two.

2.4.2 Subjects

Subjects were 24 normal-hearing students of the Free University, whose ages ranged from 19 to 28. All had pure-tone air-conduction thresholds less than 15 dB HL in their preferred ear at octave frequencies from 125 to 4000 Hz and at 6000 Hz. They were divided into two groups of twelve, one group for the identification of C_i and C_f, the other for the identification of V and C_m.

2.4.3 Procedure

For both identification of C_i and C_f and of V and C_m, the 36 lists were assigned to the filtering and control conditions according to a 6×6 digram-balanced Latin square to avoid effects of measurement order. With twelve subjects per group, each sequence of conditions was presented to two subjects.

All stimuli were presented in quiet, monaurally through headphones (Sony MDR-CD999) at the ear of preference in a soundproof room. The level of presentation was 70 dB(A). The subject was seated in front of a video monitor and a response box. On this box a number of buttons (12 in case of vowels and 17 in case of consonants) were labeled with phonemes, in orthographic notation. After presentation of a stimulus (only once), the subject could take as much time needed to give a response by pressing one of the labeled buttons. The response was displayed on the monitor. After a response was given, there was a 1-s interval before the next stimulus was presented. If a mistake was made (which occurred only sporadically), the 1-s interval could be used to correct the reponse.

The order of the tests (i.e., whether the subject started with the C_i or C_f part, or with the V or C_m part) was counterbalanced. Before each test, two lists of 20 stimuli in the 4- and 2-Hz conditions were presented to familiarize the subjects with the experimental task. Subjects participating in the C_f test were told that the CVC syllables followed the Dutch phonological rules, e.g., they would not end in /b/ or /v/.

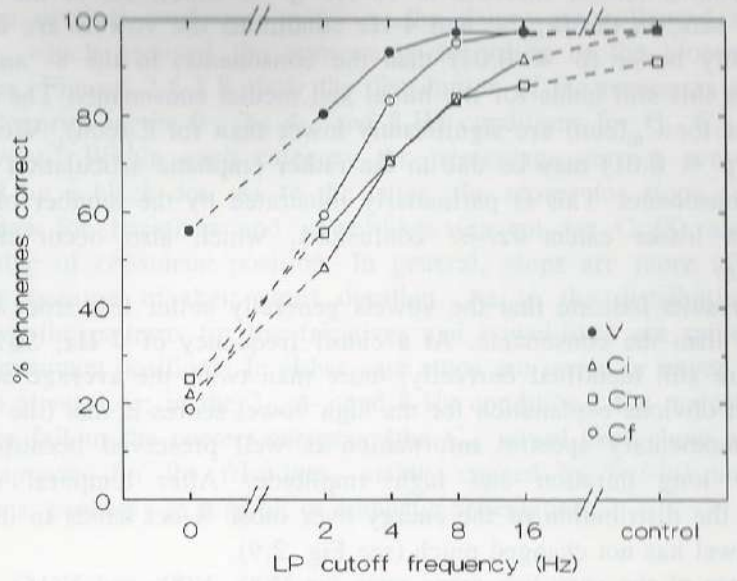

FIG. 2.5. Overall vowel- and consonant-identification score as a function of cutoff frequency, with phoneme type as parameter.

2.4.4 Results and discussion

In total, 48 identifications (12 subjects × 4 utterances) were obtained for each vowel and consonant in each condition. Due to an error in the automatic registration of the responses, however, we could not recover two subjects' V(con), one subject's V(16), and one subject's C_m(con). So, for these conditions we only had 40, 44, and 44 identifications per phoneme, respectively. The mean score for consonants and vowels in each filtering condition is plotted in Fig. 2.5. Confusion matrices for the four phoneme sets in all conditions are given in Appendix A. A two-way analysis of variance on the arcsine transformed scores (Studebaker, 1985) with repeated measures on the factor condition showed significant effects of phoneme set, filtering condition, and interaction ($p < 0.001$). Because of the significant interaction, separate analyses were carried out for each of the six conditions and for each of the four phoneme sets. The results of these analyses and subsequent *post hoc* tests (Tukey HSD) can be summarized as follows. The mean recognition improves for all phoneme sets as the cutoff frequency increases up to 8 Hz ($p < 0.01$ for all pairwise comparisons, except for the V(4) and V(8) comparison, where $p < 0.05$). Only the initial consonants

benefit from a further increase to 16 Hz ($p < 0.01$). As to the different phoneme sets, in the 0-, 2-, and 4-Hz conditions the vowels are identified significantly better ($p < 0.01$) than the consonants; in the 8- and 16-Hz conditions this still holds for the initial and medial consonants. The fact that the scores for C_m(con) are significantly lower than for C_i(con), V(con), and C_f(con) ($p < 0.01$) may be due to the rather emphatic articulation of some medial consonants. This is particularly illustrated by the number of /w/-/v/ and to a lesser extent /z/-/s/ confusions, which also occur the other conditions.

The results indicate that the vowels generally suffer less from temporal smearing than the consonants. At a cutoff frequency of 0 Hz, 56% of the vowels are still identified correctly, more than twice the average consonant score. An obvious explanation for the high vowel scores is that (the majority of) the momentary spectral information is well preserved because of the relatively long duration and high amplitude. After temporal envelope filtering, the distribution of the energy over most ¼-oct bands in the center of the vowel has not changed much (see Fig. 2.9).

Because of the very low error rates for V(4), V(8), and V(16), we will restrict ourselves to V(0) and V(2) for a discussion of the vowel confusions. The greater part of the errors can be attributed to two factors: diphthong confusions (/ɛi/-/ɛ/ and /ɑu/-/ɑ/ or /ɑu/-/a/) and long-short/short-long confusions (/a/-/ɑ/, /e/-/I/, and /o/-/ɔ/ and vice versa). Of all errors in the 0-Hz condition (253), these factors account for 33% and 28%, respectively. In the 2-Hz condition (113 errors) these figures are 42% and 37%, respectively. The percentages in the 0-Hz condition are slightly less, since there is also a tendency to respond to /ø/ (15% of all errors), which is the most neutral vowel in the set. It is clear that the rapid spectral changes in a diphthong (the /u/- and /i/-states take only about a quarter of the entire duration of /ɑu/ and /ɛi/, respectively) are not modelled properly for these low cutoff frequencies. As to the confusions among long-short vowel pairs, the blurring of the temporal structure makes it difficult to perceive the begin and/or end of a vowel and hence the vowel duration, so that listeners have to rely more on the spectral contents, which often leads to the observed confusions. It is also possible that smearing causes a reduction in the perceptually important vowel-inherent spectral change, i.e., slowly varying changes in formant frequencies. In the absence of these dynamic spectral changes long-short confusions may occur (cf. Nearey and Assman (1986), who found this for isolated monophthongs in English).

For an evaluation of the C_i, C_m, and C_f scores, the set of consonants was divided into three subsets (cf. Steeneken, 1992):[3] stops (/t, k, p, b, d/),

fricatives (/f, s, χ, v, z/), and vowel-like consonants (/m, n, ŋ, l, w, j, h/). From the original confusion matrices, the data were collapsed into 3×3 matrices, which grouped the consonants according to the aforementioned categories. Figures 2.6-2.8 show the distribution of the responses across the three categories in the 0-, 2-, 4-, and 8-Hz conditions for C_i, C_m, and C_f, respectively.[4] Within each category the percentage correct consonants is indicated by a black dot. As to the latter, the scores for stops are clearly lower than for fricatives and vowel-likes (except for C_m(8) and C_f(8)), irrespective of consonant position. In general, stops are more affected by smearing because of their short duration. As to the distribution of the responses, the patterns for the fricatives and vowel-likes are rather similar across consonant positions: In either case stops are rarely or never used as a response alternative; in the 2-, 4-, and 8-Hz conditions, the majority of the responses fall in the correct category (the C_m vowel-likes show a constant 17% responses for the fricatives, mainly caused by /w/-/v/ and /h/-/v/ confusions, probably as a result of emphatic articulation).

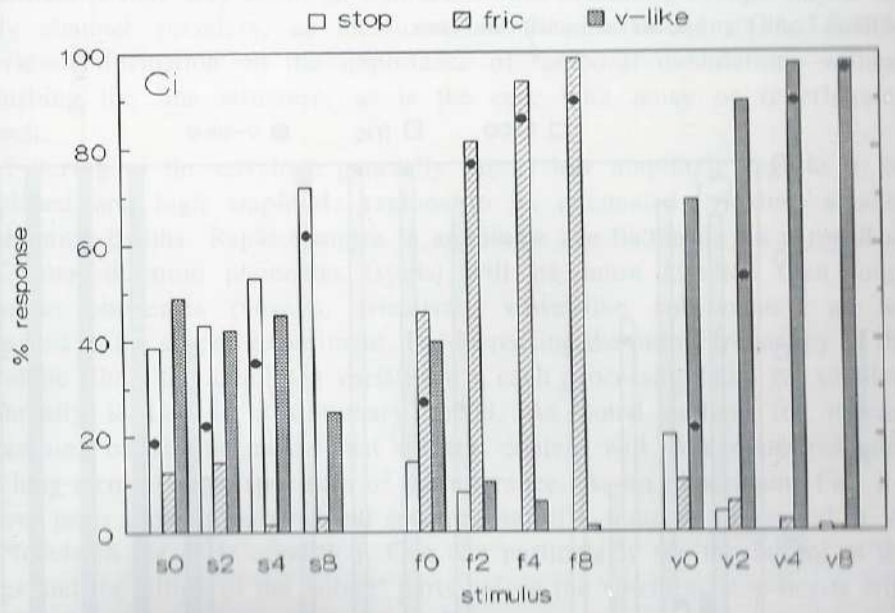

FIG. 2.6. Distribution of the responses for the initial consonants across the three categories stop, fricative, and vowel-like as a function of stimulus category and cutoff frequency (s0 = stop/0-Hz condition, s2 = stop/2-Hz condition, f4 = fricative/4-Hz condition, v8 = vowel-like/8-Hz condition, etc.). The black dots indicate the percentage of correctly identified consonants per category/condition.

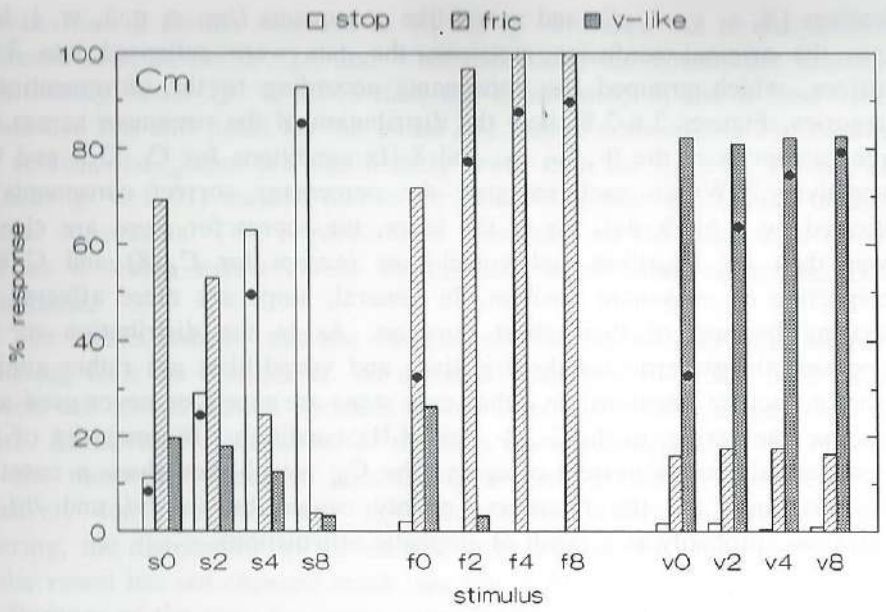

FIG. 2.7. As Fig. 2.6, for the medial consonants.

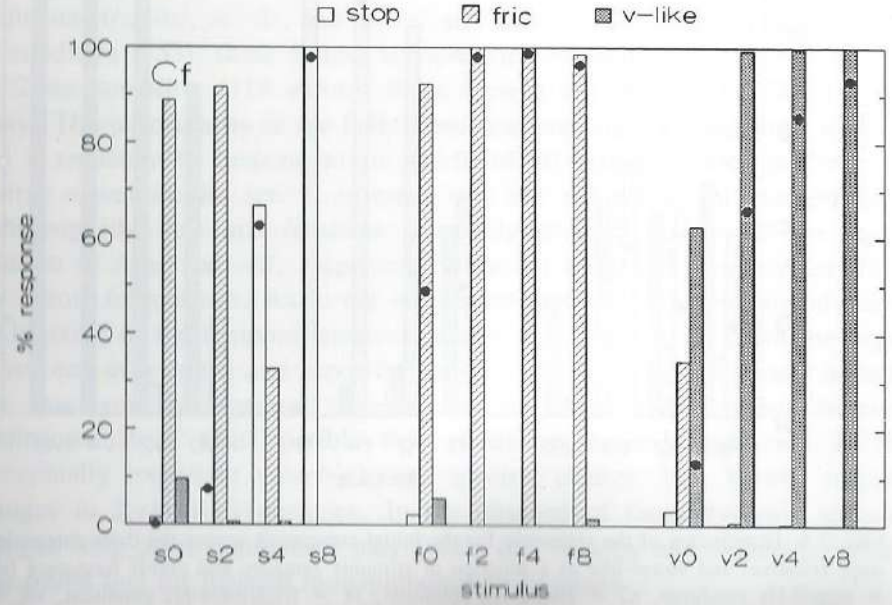

FIG. 2.8. As Fig. 2.6, for the final consonants.

The confusions of the stop consonants show more variation, and depend on their position in the syllable. In initial position, stops are confused with other stops (place errors) and with vowel-likes. Besides the bias towards /h/ responses in the 0- and 2-Hz conditions, we find /b/-/w/ and /d/-/w/ confusions. Particularly the perception of /w/ for /b/ could be expected to result from temporal smearing. In medial and final position, stops are mainly confused with fricatives. In the 0-Hz condition the responses are rather scattered, but in the 2- and 4-Hz conditions we can distinguish some confusions that could be expected: /b/-/v/, /b/-/w/, /t/-/s/, /k/-/χ/, and /p/-/f/ (the last three especially in final position).

2.5 GENERAL DISCUSSION

In the experiments described above, we tried to assess the contribution of temporal modulations to intelligibility and identification. The manipulations of the speech signal are quite artificial and are not intended to model reduced temporal resolution by hearing-impaired listeners. Nor do they reflect any disturbances that may occur in communication channels, except maybe for early channel vocoders, as mentioned in the introduction. The method provides information on the importance of temporal modulations without disturbing the fine structure, as is the case with noisy or reverberated speech.

Filtering of the envelope generally causes low amplitude regions to be amplified and high amplitude regions to be attenuated, yielding smaller modulation depths. Rapid changes in amplitude are flattened. As a result of this, short-duration phonemes (stops) will be more affected than long-duration phonemes (vowels, fricatives, vowel-like consonants), as we observed in the second experiment. By decreasing the cutoff frequency of the envelope filter, the amplitude variations in each processing band get smaller, eventually leading to a stationary sound. As noted earlier, for narrow processing bands this means that spectral content will ever more resemble the long-term average spectrum of the utterance. As an illustration, Fig. 2.9 (lower part) shows a narrowband spectrogram of a sentence processed in ¼-oct bands in the 4-Hz condition. One can particularly see the fading of the stops and the filling of the 'silent' parts before the voiceless stop-bursts with amplified noise.

It is worthwhile to view the results of the first experiment in the light of the MTF (Houtgast and Steeneken, 1985). The MTF gives the extent to which the frequency components of the *intensity* envelope are transferred.

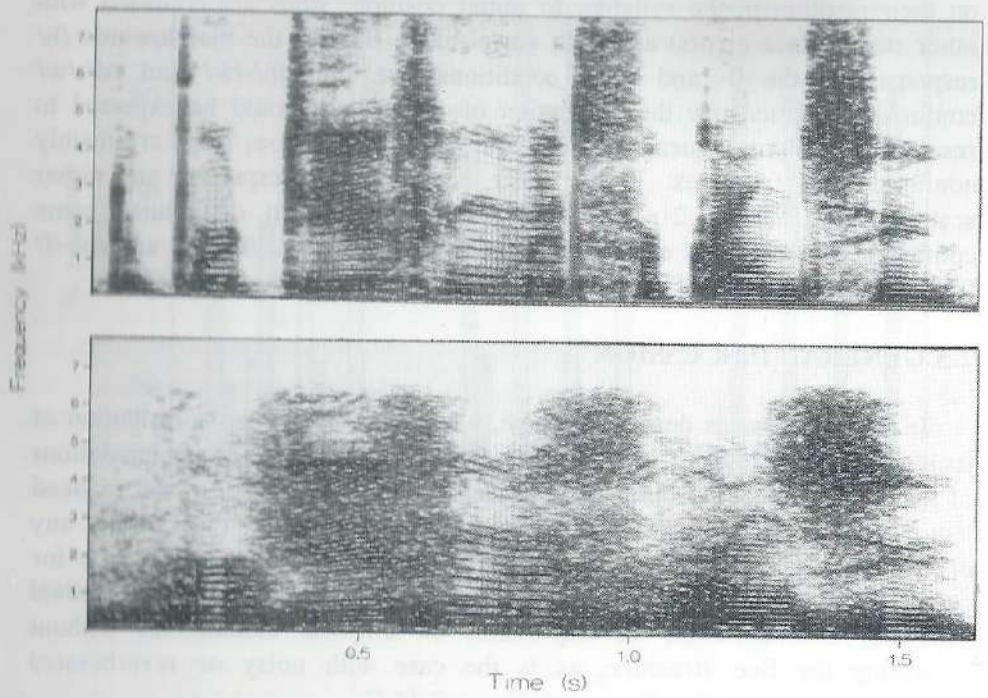

Fig. 2.9. Narrowband spectrograms of the sentence "de portier ging met vakantie" (the porter went on holiday), original (above) and smeared in ¼-oct bands and 4-Hz cutoff frequency (below).

Since we filtered the *amplitude* envelope, the reduction of amplitude modulations should be translated into a reduction of intensity modulations.[5] It turns out that the filter slope in the intensity domain remains unchanged (approximately −40 dB/oct), but the point at which the modulation is reduced to half (the cutoff frequency) is about ⅓ oct higher.

An interesting point is the relation between the intelligibility scores in the first experiment and the speech-transmission index (STI), based on the MTF. The STI can be calculated from the relative reduction of intensity modulations in 14 ⅓-oct intervals of modulation (ranging from 0.63 to 12.5 Hz) within octave bands. The STI is a numerical index between 0 and 1, which bears a linear relation to signal-to-noise ratios from −15 to +15 dB (for a review of the STI-concept, see Houtgast and Steeneken, 1985). We computed the STI for each of the nine experimental conditions.[6] The results are given in Table 2.2. The STI values apply to the 'clean' processed signal,

i.e., without noise. In order to get an estimate of the SRT for each filtering condition, the signal-to-noise ratio which would give a STI of 0.33 was computed. A STI of 0.33 corresponds to a signal-to-noise ratio of −5 dB, which is a reasonable threshold for normal (unprocessed) speech. The estimated SRTs are shown in the second row of Table 2.2. It is clear that no SRT could be computed for the 0-, ½-, and 1-Hz conditions, since in these cases the STI is lower than 0.33. However, one can argue that these computed values are rather low, since an 80% correct score for sentences (0-Hz condition, 1-oct bandwidth) would already yield a STI of about 0.4. According to the STI in Table 2.2, measuring the SRT in the 2-Hz condition would be possible; this seems plausible, given the 92% intelligibility score for sentences in quiet (see Fig. 2.3, 1-oct bandwidth). For the other conditions, the estimated SRT corresponds well to the actually measured SRT in the first experiment (see Table 2.1, 1-oct bandwidth); although the estimated SRTs are systematically lower, the deviation is maximally 1.5 dB (in the 8-Hz condition). Since we found no effect of processing bandwidth for cutoff frequencies above 4 Hz, the esimated SRTs are also comparable with the measured SRTs in ¼- and ½-oct bands (the deviation for the 4-Hz condition is about 2 dB). This is consistent with the octave-band specific STI-concept, which does not account for effects within ½- or ¼-oct bands (there is a bandwidth dependence for cutoff frequencies below 4 Hz, though). For cutoff frequencies of 16 Hz and higher, the good correspondence between the estimated and measured SRT is not surprising, since the modulations involved in the calculation of the STI do not exceed 12.5 Hz. In summary, realizing that the STI was developed for time-domain distortions like reverberation, the SRT estimates down to 4 Hz are reasonable.

The results of the second experiment indicate that consonants are affected most by temporal smearing. Gelfand and Silman (1979) investigated the

TABLE 2.2. Computed STI and estimated SRT for the nine filtering conditions without noise.

	Condition								
	0 Hz	½ Hz	1 Hz	2 Hz	4 Hz	8 Hz	16 Hz	32 Hz	64 Hz
STI	0.00	0.09	0.25	0.47	0.68	0.88	0.98	1.00	1.00
est. SRT	...	...	...	+4.7	−1.1	−4.1	−4.9	−5.0	−5.0

effect of reverberation (T = 0.8 s) upon consonant recognition in CVC syllables. As far as the reduction of fast modulations is concerned, this corresponds roughly to a situation between our 4- and 8-Hz condition. Gelfand and Silman found that initial consonants are on average less affected than final consonants, whereas our study does not show this difference. In reverberant speech, final consonants are masked by delayed energy of preceding segments, which does not hold for initial consonants. In our processing however, smearing of the envelope causes segments to integrate with preceding and following segments. Initial consonants will thus be corrupted by the following vowel, and will therefore also deteriorate.

As to the confusions, place of articulation errors did not occur regularly in fricatives, which seems to be in agreement with the findings of Behrens and Blumstein (1988). Place errors were primarily found in initial stops, which may be related to a reduction of changes in the distribution of spectral energy from burst onset to vowel onset. In both natural and synthetic stop-V stimuli the relative amplitude of the burst and its (rapidly) changing spectrum have been shown to affect the perception of place of articulation (Ohde and Stevens, 1983; Lahiri *et al.*, 1984). We found manner feature recognition for fricatives and vowel-likes to be (far) less reduced and less position-dependent than for stops (Figs. 2.6-2.8). The position-dependence of the stop confusions (vowel-like responses for C_i and fricative responses for C_m and C_f) may be explained as follows. In initial position, smearing increases the integration of the stop with the immediately following vowel. This evokes the perception of a vowel-like consonant (strictly speaking, this accounts for the voiced stops /b/ and /d/). Final stops are more isolated from the preceding vowel, and tend to be longer in duration (word-final lengthening). Smearing will therefore not lead to integration with the preceding vowel. Because there is no speech sound following, the stop can be 'spread out' entirely, resulting in a fricative perception. According to the above reasoning, smearing of the medial stops would evoke more vowel-like than fricative confusions. The opposite is true, however. The medial stops were pronounced rather emphatically, making them longer than usual, so that there was less influence of the following vowel after smearing, yielding fricative perception.

Finally, it is important to consider the effect of narrowband smearing upon perception, compared to a wideband approach. For natural speech, Nittrouer and Studdert-Kennedy (1986) investigated the effect of interchanging the wideband amplitude envelope of /b/-V and /w/-V stimuli. They found that this had little effect on the consonant recognition (97% correct). Their results suggest that increasing the amplitude of occlusion and

burst of /b/ does not automatically evoke the perception of /w/. The fact that we do find /b/-/w/ confusions with the initital consonants must be ascribed to the narrowband approach, smearing the envelopes of 24 ¼-oct bands.

2.6 CONCLUSIONS

The most important conclusions of this chapter are

(1) Amplitude fluctuations in successive ¼-, ½-, or 1-oct frequency bands can be limited to about 16 Hz without substantial reduction of speech intelligibility for normal-hearing listeners.
(2) Listeners can only partially understand speech in quiet when the amplitude fluctuations are limited to 2 Hz; performance improves as broader frequency bands are used. In case of 100% compression within ¼-oct bands (beyond the critical bandwidth), intelligibility drops dramatically. For envelope cutoff frequencies above 4 Hz, intelligibility is independent of the processing bandwidth.
(3) SRT values obtained for envelope cutoff frequencies above 4 Hz appear to correspond well (maximal deviation of 1.5 dB) to those predicted by the speech-transmission index. However, for low cutoff frequencies (0-2 Hz) the computed STI is rather low and does not account for the observed dependence of the processing bandwidth.
(4) Phoneme identification with nonsense syllables shows that consonants are more affected by temporal smearing than vowels. Stops appear to suffer most, due to their short duration, with confusion patterns depending on the position in the syllable.

Notes

[1] A factor 64 was chosen for practical reasons. The OROS-AU21 signal processing card enables the use of fast decimator and interpolator routines by a factor 4, with automatic filtering. By calling these routines three times in a row, a factor of 64 is obtained. The sampling rate is then brought down to 244 Hz, which enables the implementation of sufficiently steep lowpass FIR filters with cutoff frequencies as low as 0.5 Hz.

[2] Formally, subsequent lowpass filtering of the fine structure may restore the original envelope to some extent, depending on the envelope cutoff frequency and the processing bandwidth. However, it was checked that this had no noticeable influence. In the worst case of a completely flat envelope within a 1-oct band (in which case there is a maximum

difference from the original envelope), dominant modulations were still sufficiently attenuated by about 33 dB.

[3] The subdivision of the stimuli is based on Steeneken's study (chapter 3) on the perceptual similarity of phonemes at various transmission conditions (among which distortion in the time domain).

[4] The percentages are based on the total number of stimuli within each category. The total number of stops, fricatives, and vowel-likes are for C_i: 192, 144, and 240; for C_m: 240, 240, and 288; for C_f: 144, 144, and 288, respectively.

[5] In order to get the modulation transfer in the intensity domain, the following procedure was undertaken: The modulation spectrum of the squared original amplitude envelope was compared with the modulation spectrum of the squared filtered amplitude envelope (for all cutoff frequencies used in the experiment). The original amplitude envelope was taken of a 30-s 1-oct filtered speech fragment (viz. 13 concatenated sentences used in the first experiment). The MTF was obtained by computing the ratio between the original and the modified spectrum.

[6] The calculations were verified by actual measurement of the STI with the so-called STITEL procedure (cf. Steeneken, 1992, pp. 133-139). Differences between the computed and measured STI values for the same condition were within 0.04.

CHAPTER 3

Effect of reducing slow temporal modulations on speech reception*

Abstract

The effect of reducing low-frequency modulations in the temporal envelope on the speech-reception threshold (SRT) for sentences in noise and on phoneme identification was investigated. For this purpose, speech was split up into a series of frequency bands (¼, ½, or 1 oct wide) and the amplitude envelope for each band was highpass filtered at cutoff frequencies of 1, 2, 4, 8, 16, 32, 64, or 128 Hz, or ∞ (completely flattened). Results for 42 normal-hearing listeners show: (1) a clear reduction in sentence intelligibility with narrow-band processing for cutoff frequencies above 64 Hz, and (2) no reduction of sentence intelligibility when only amplitude variations below 4 Hz are reduced. Based on the modulation transfer function of some conditions, it is concluded that fast multichannel dynamic compression leads to an insignificant change in masked SRT. Combining these results with previous data on lowpass envelope filtering [chapter 2] shows that at 8-10 Hz the temporal modulation spectrum is divided into two equally important parts. Vowel and consonant identification with nonsense syllables were studied for cutoff frequencies of 2, 8, 32, 128 Hz, and ∞, processed in ¼-oct bands. Results for 12 subjects indicate that, just as for lowpass envelope filtering, consonants are more affected than vowels. Errors in vowel identification mainly consist of reduced recognition of diphthongs and of durational confusions. For the consonants there are no clear confusion patterns, but stops appear to suffer least. In most cases, the responses tend to fall into the correct category (stop, fricative, or vowel-like).

Slightly modified version of a paper published in: J. Acoust. Soc. Am. **95**, 2670-2680 (1994b)

3.1 INTRODUCTION

One way to describe a speech signal is as a summation of a number of amplitude modulated narrow frequency bands. In this view, every frequency band can be considered to consist of a carrier signal (fine structure) and a time-varying envelope. The envelope in turn contains a variety of modulation frequencies, the amplitude of which can be illustrated by the temporal modulation spectrum. These modulations play an important role in the transmission of information in speech (cf. Houtgast and Steeneken, 1985). When processing and/or transmitting speech in some way, a faithful transfer of these modulation frequencies seems necessary. In terms of perception, we would like to know how attenuation of the details (fast amplitude variations) or of the gross movements (slow amplitude variations) affect the understanding of everyday speech. In other words, the question is: Within which limits can specific amplitude modulations be reduced before having a detrimental effect on intelligibility?

In an earlier study [chapter 2], the effect of temporal envelope smearing on sentence intelligibility and phoneme recognition was investigated. In this approach, the wideband signal was subdivided into a series of frequency bands (¼, ½, or 1 oct wide) and the amplitude envelope of each band was lowpass filtered at a variable cutoff frequency. The results showed that preserving only modulations up to about 16 Hz yields almost the same speech-reception threshold (SRT) for sentences in noise as obtained for unprocessed speech. For lower cutoff frequencies the SRT increases, and for cutoff frequencies as low as 0-2 Hz sentence intelligibility in quiet is heavily affected if envelope filtering takes place in narrow frequency bands. Consonants, especially the stops, suffer more from severe temporal smearing than vowels.

In continuation of this, the present study was set up to investigate the effect of reducing low-frequency temporal modulations. By highpass filtering the temporal envelope in a series of frequency bands, the extent to which intelligibility depends on the slow amplitude variations can be established. The applied signal processing and the experimental design are closely related to the lowpass filtering in the previous study. In chapter 2 we discussed the significance of temporal modulations for intelligibility, referring to related issues in the literature, such as the modulation transfer function (MTF), vocoders, and phoneme perception based on envelope information. Apart from the mere effect of highpass envelope filtering, results from the present study can shed light on the issue of speech intelligibility in multichannel compression systems (cf. Plomp, 1988; Hohmann and Kollmeier, 1990).

Together with the earlier lowpass filtering results, more insight in the contribution of temporal modulations can be given.

In this chapter we will describe two perception experiments. In experiment 1, the intelligibility for sentences in quiet and the SRT for sentences in noise were measured as a function of envelope highpass cutoff frequency and processing bandwidth. In experiment 2, the effects on vowel and consonant identification in nonsense syllables were studied.

3.2 METHOD

For the signal processing, the analysis-resynthesis scheme used in the previous experiments with lowpass filtered envelopes was slightly modified. A block diagram of the processing is shown in Fig. 3.1. Since the details of the method have been described in chapter 2, this section primarily contains information on the modifications (viz. the changes in envelope filtering).

The wideband speech (sampling rate of 15,625 Hz) is led through a linear-phase FIR digital filter bank, and from the output of each channel the Hilbert envelope is determined. In order to meet the requirements of the envelope filter's low cutoff frequencies, the envelope is downsampled before filtering. The actual downsampling factors used were 64 for envelope cutoff frequencies below 80 Hz, and 16 for envelope cutoff frequencies above 80 Hz. Due to downsampling (with preceding lowpass filtering), the upper modulation frequency still present in the envelope is lowered. Therefore, direct highpass filtering of the downsampled envelope would in fact be bandpass filtering with an upper cutoff frequency below half the sampling frequency after downsampling. To overcome this problem, the downsampled envelope is lowpass filtered, upsampled, and subtracted from the original envelope (in the time domain). Taking the −6-dB point as cutoff frequency and a lowpass filter slope of about −80 dB/oct, this results in a modified envelope that is effectively highpass filtered at the same cutoff frequency, with a slope of approximately +40 dB/oct. In order to maintain a sufficient level, the mean level (dc) of the original envelope is added to the filtered version. Parts of the modified envelope that are still negative are set to zero (eventually resulting in a short silent interval in that particular frequency band).

For each frequency band, the modified signal is obtained by multiplying the fine-structure, sample by sample, by the ratio of the modified and the original envelope. To eliminate any spectral distortion, the modified signal is lowpass filtered with cutoff frequency 5% higher than the upper limit of the corresponding bandpass filter. Finally, all modified signals are added and the

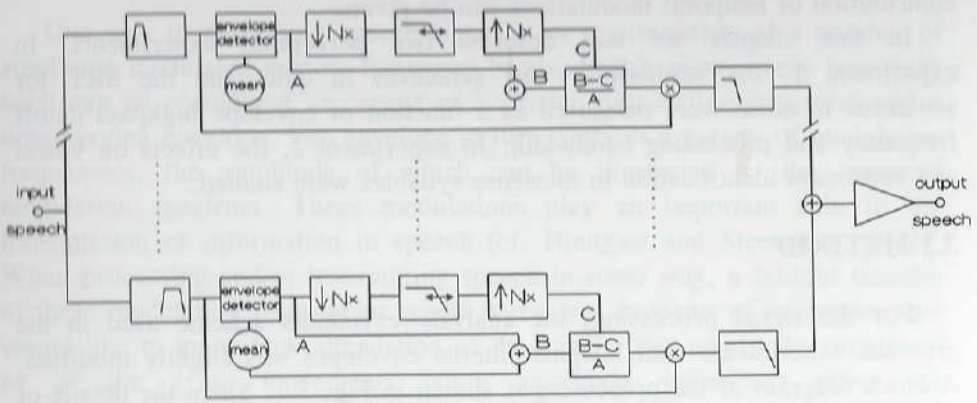

FIG. 3.1. Block diagram of the speech processing. The wideband input speech signal is split up into several frequency bands. For each band the amplitude envelope is determined, downsampled by a factor N (16 or 64), lowpass filtered, and upsampled. The highpass filtered envelope is obtained by substracting the lowpass filtered envelope from the original envelope and adding the dc component. Modulating the fine structure according to the modified envelope then yields the new band signal. Each new band signal is lowpass filtered to eliminate undesired high-frequency noise. After adding all modified bands, the wideband signal is rescaled to match the rms of the input speech.

level of the new wideband signal is adjusted to have the same (wideband) rms as the input signal.

All signal manipulations were performed (non-real-time) on an Olivetti PCS 286 computer, using an OROS-AU21 card with TMS320C25 signal processor. Figure 3.2 shows an example of the various stages in the processing of one ¼-oct band for a short sentence.

3.3 EXPERIMENT 1: SENTENCE INTELLIGIBILITY

3.3.1 Stimuli, design

Ten lists of 13 Dutch sentences of eight to nine syllables read by a female speaker were used as basic material (Plomp and Mimpen, 1979). All sentences were digitized at a sampling rate of 15,625 Hz and 16 bits resolution. Ten processing conditions were investigated, in which the envelope in each frequency band was highpass filtered at the following cutoff frequencies: 0 Hz (all modulations intact, control condition), 1, 2, 4, 8, 16, 32, 64, 128 Hz, or ∞ (no modulations left). The last condition was obtained

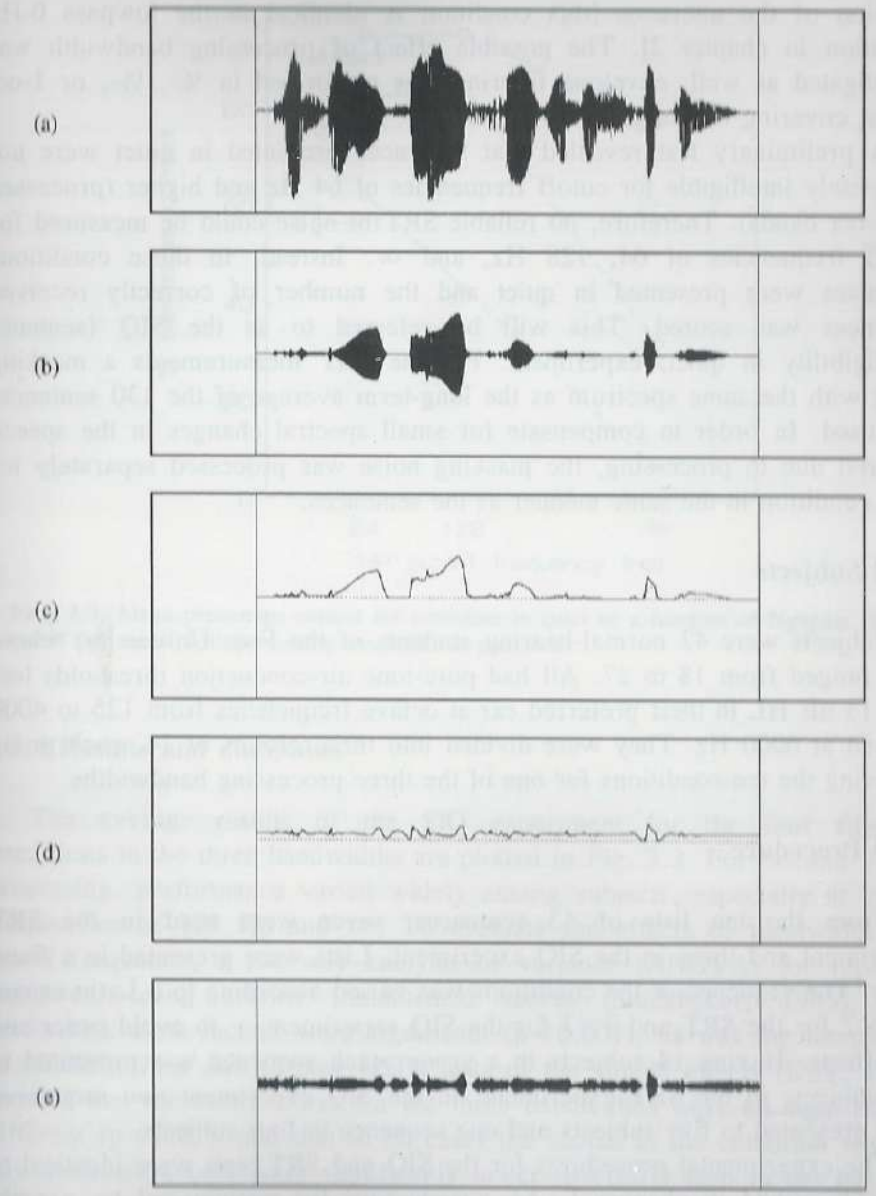

FIG. 3.2. Example of the processing for one sentence (2.2-s duration) in a single frequency band. (a) original wideband signal; (b) ¼-oct band signal (283-336 Hz); (c) amplitude envelope of (b); (d) envelope (c) highpass filtered with 8-Hz cutoff frequency; (e) resulting modified band signal, amplitude modulated according to (d).

by using the mean amplitude level of each frequency band for the entire duration of the utterance [this condition is identical to the lowpass 0-Hz condition in chapter 2]. The possible effect of processing bandwidth was investigated as well; envelope filtering was performed in ¼-, ½-, or 1-oct bands, covering the range 100-6400 Hz.

A preliminary test revealed that sentences presented in quiet were not completely intelligible for cutoff frequencies of 64 Hz and higher (processed in ¼-oct bands). Therefore, no reliable SRT in noise could be measured for cutoff frequencies of 64, 128 Hz, and ∞. Instead, in those conditions sentences were presented in quiet and the number of correctly received sentences was scored. This will be referred to as the SIQ (sentence intelligibility in quiet) experiment. For the SRT measurements a masking noise with the same spectrum as the long-term average of the 130 sentences was used. In order to compensate for small spectral changes in the speech material due to processing, the masking noise was processed separately for each condition in the same manner as the sentences.

3.3.2 Subjects

Subjects were 42 normal-hearing students of the Free University, whose ages ranged from 18 to 27. All had pure-tone air-conduction thresholds less than 15 dB HL in their preferred ear at octave frequencies from 125 to 4000 Hz and at 6000 Hz. They were divided into three groups of 14, each group receiving the ten conditions for one of the three processing bandwidths.

3.3.3 Procedure

From the ten lists of 13 sentences, seven were used in the SRT experiment and three in the SIQ experiment. Lists were presented in a fixed order. The sequence of the conditions was varied according to a Latin square – 7×7 for the SRT and 3×3 for the SIQ experiment – to avoid order and list effects. Having 14 subjects in a group, each sequence was presented to two subjects in the SRT experiment; in the SIQ experiment two sequences were presented to five subjects and one sequence to four subjects.

The experimental procedures for the SIQ and SRT tests were identical to those described in chapter 2. An introductory list pronounced by a male speaker was presented to familiarize the subjects with the procedure. For the SIQ experiment this list consisted of sentences in the 128-Hz condition; for the SRT experiment another list in the 4-Hz condition was used.

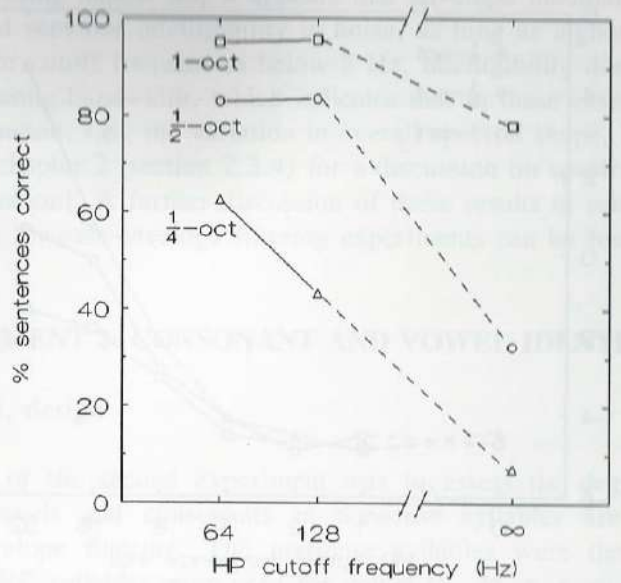

FIG. 3.3. Mean percentage correct for sentences in quiet as a function of highpass (HP) cutoff frequency, with processing bandwidth as parameter.

3.3.4 Results and discussion

The average results of the SIQ experiment for the four filtering conditions in the three bandwidths are plotted in Fig. 3.3. For ¼- and ½-oct processing, performance varied widely among subjects, especially at cutoff frequencies of 128 Hz and ∞. To evaluate the effects of bandwidth and cutoff frequency, a two-way analysis of variance (ANOVA) for repeated measures, using arcsine transformed scores (Studebaker, 1985), was performed. Both factors were significant ($p<0.001$), as was the interaction ($p<0.005$). *Post hoc* (Tukey HSD) tests of the simple effects (Kirk, 1968) showed that the mean scores for the three bandwidths were all significantly different ($p<0.01$), and that in all cases the scores in the condition without any modulation (∞) were significantly lower ($p<0.01$) than in the 64- and 128-Hz conditions. Only for the ¼-oct bandwidth was there a significant difference ($p<0.05$) between cutoff frequencies of 64 and 128 Hz. The results of the ∞ conditions of 7, 32, and 78% for processing bandwidths of ¼-, ½-, and 1-oct, respectively, are in good agreement with earlier findings (3, 22, and 80%, respectively; see lowpass 0-Hz condition in chapter 2).

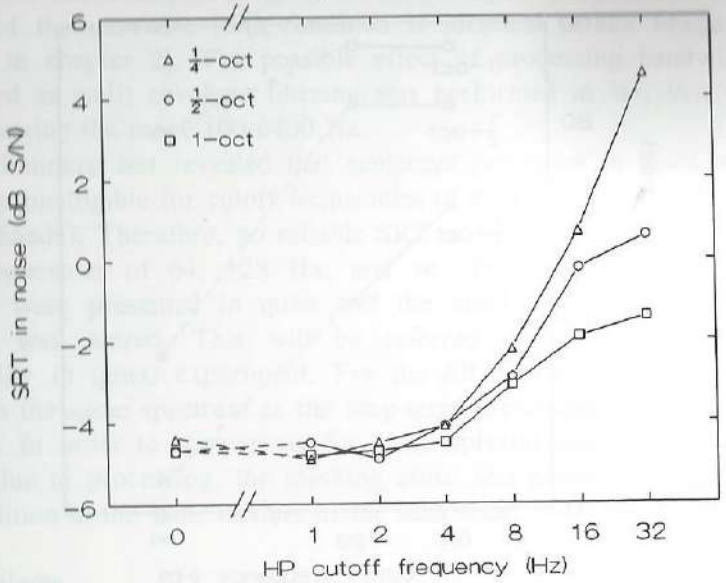

FIG. 3.4. Mean speech-reception threshold for sentences in noise as a function of highpass (HP) cutoff frequency, with processing bandwidth as parameter.

Variations in the scores must be attributed to the earlier mentioned difference in subjects' performance.

The mean SRT for sentences in noise as a function of highpass cutoff frequency and processing bandwidth is plotted in Fig. 3.4. A similar ANOVA as for the SIQ data revealed significant effects for bandwidth, cutoff frequency, and the interaction between them ($p<0.001$). Tests for simple effects (Kirk, 1968) and *post hoc* tests showed that there was no significant bandwidth effect up to 8 Hz; in the 16-Hz condition the SRT for the 1-oct bandwith differs from the other two ($p<0.01$), while in the 32-Hz condition the SRT for all three bandwidths are different ($p<0.01$). As for the filtering conditions, there are no significant differences in SRT below 4 Hz, where there appears to be a constant threshold of on average −4.6 dB (standard deviations 1.0 to 1.4 dB). The SRT for cutoff frequencies of 8 Hz and upward differs significantly from the control (0 Hz) condition ($p<0.01$).

The results of these experiments indicate that the intelligibility of everyday sentences with highpass filtered envelopes remains at the same level as unprocessed speech up to a cutoff frequency of 4 Hz. For higher cutoff frequencies intelligibility decreases progressively, particularly for

narrow processing bands. So, it appears that envelope modulations below 4 Hz do not aid sentence intelligibility in noise, as long as higher modulations are intact. For cutoff frequencies below 8 Hz, intelligibility does not depend on the processing bandwidth, which indicates that in these cases the spectral macro-information, i.e., the variation in overall spectral shape, is sufficiently present [see chapter 2 (section 2.3.4) for a discussion on spectral micro- and macro-information]. A further discussion of these results in relation to those of the earlier lowpass-envelope filtering experiments can be found in section 3.5.

3.4 EXPERIMENT 2: CONSONANT AND VOWEL IDENTIFICATION

3.4.1 Stimuli, design

The aim of the second experiment was to assess the degree to which individual vowels and consonants in nonsense syllables are affected by highpass envelope filtering. The nonsense syllables were the same as in chapter 2. CVC syllables were used for vowel (V, /ɑ, a, ɛ, e, I, i, ɔ, o, ø, u, ɑu, ɛi/) identification, and VCV syllables for consonant (C, /b, d, f, χ, h, j, k, l, m, n, p, s, t, v, w, z/) identification.[1]

For all syllables the envelope filtering was performed in 24 ¼-oct bands. There were six experimental conditions: A control condition (0 Hz) and cutoff frequencies of 2, 8, 32, 128 Hz, or ∞. The choice of the filtering conditions was based on the results of the SRT experiments, viz. normal intelligibility (control and 2 Hz), reduced intelligibility (8 and 32 Hz), and low intelligibility (128 Hz and ∞).

As in chapter 2, a total of 36 randomized lists of 50 CVC syllables and 36 randomized lists of 66 VCV syllables were made. The first two syllables of each list were copies of the last two and acted as dummy trials, so that there were actually 48 or 64 test stimuli, respectively.

3.4.2 Subjects

Subjects were 12 normal-hearing students of the Free University, whose ages ranged from 18 to 29. All had pure-tone air-conduction thresholds less than 15 dB HL in their preferred ear at octave frequencies from 125 to 4000 Hz and at 6000 Hz.

3.4.3 Procedure

For both identification of C and V, the 36 lists were assigned to the filtering conditions according to a 6×6 digram-balanced Latin square to avoid effects of measurement order. Each sequence of conditions was presented to two subjects. All stimuli were presented in quiet at a level of 70 dB(A), monaurally through a headphone (Sony MDR-CD999) to the subject's ear of preference. Subjects made their response by means of labeled buttons on a box connected to a PC. The entire procedure is described in detail in chapter 2.

Half of the subjects started with the vowel identification, the other half with the consonant identification. Before each test, two lists of 20 stimuli in the 8- and 128-Hz conditions were presented to familiarize the subjects with the experimental task.

3.4.4 Results and discussion

In total, 48 identifications (12 subjects × 4 utterances) were obtained for each vowel and consonant per condition. The mean scores are plotted in Fig. 3.5. As in chapter 2, the filtering condition will be written in parentheses following the set of phonemes to be identified; e.g, C(8) stands for consonants in the 8-Hz condition. Confusion matrices for vowels and consonants for all conditions are given in Appendix B. A repeated-measures analysis of variance on the arcsine transformed scores (Studebaker, 1985) was performed, with phoneme (vowels versus consonants) and filtering condition as within-subjects factors. The analysis showed significant effects of phoneme ($p<0.001$), condition ($p<0.001$), and the interaction between them ($p<0.002$). *Post hoc* (Tukey HSD) tests indicated that for a highpass cutoff frequency above 2 Hz, the decreasing scores for vowels and consontants are not significantly different; for cutoff frequencies of 32 Hz and higher, however, consonant scores are significantly lower than vowel scores ($p<0.01$).

Like in the previous identifications of temporally smeared syllables [chapter 2], the present results show that consonants suffer more from envelope filtering than vowels. The scores of 58% and 33% for the extreme conditions V(∞) and C(∞), respectively, closely match earlier results (56% and 25%, respectively).

Analysis of the confusion matrices for V(8), V(32), V(128), and V(∞) showed that the majority of the vowel errors are of the same type as in the previous lowpass-filtering experiments: diphthong-monophthong confusions

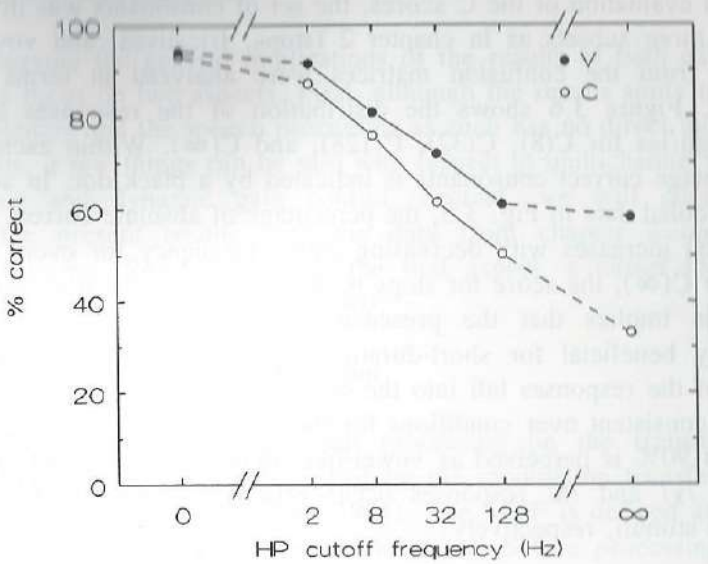

FIG. 3.5. Overall vowel- and consonant-identification score as a function of highpass (HP) cutoff frequency, with phoneme type as parameter.

(/ɛi/-/ɛ/ and /ɑu/-/a/ or /ɑu/-/ɑ/), long-short/short-long confusions (/a/-/ɑ/, /e/-/I/, and /o/-/ɔ/, and vice versa), and incorrect /ø/ responses (neutral vowel). An explanation for the former two error types is that a reduction of the modulations (but a present 0-Hz component) causes the speech signal to be more steady-state; therefore, diphthongs tend to sound like monophthongs, and, due to the blurring of the temporal structure, the perception of vowel duration is hampered [see also chapter 2]. The percentage of the total number of errors made in the different conditions that these three factors account for are listed in Table 3.1. Together they yield 58 to 68% of the errors, depending on the condition. Among the other errors one can find some /ɔ/-/u/ and /i/-/I/ confusions.

	Condition			
	8 Hz	32 Hz	128 Hz	∞
Diphthong	30	35	28	27
Long-short	26	31	18	28
/ø/ responses	4	1	12	13

TABLE 3.1. Percentage of vowel errors due to diphthong-monophthong confusions, long-short/short-long confusions, and incorrect /ø/ responses. The total number of errors (out of 576 stimuli) for V(8), V(32), V(128), and V(∞) are 110, 162, 224, and 240, respectively.

For an evaluation of the C scores, the set of consonants was divided into the same three subsets as in chapter 2 (stops, fricatives, and vowel-likes). The data from the confusion matrices were analyzed in terms of these categories. Figure 3.6 shows the distribution of the responses across the three categories for C(8), C(32), C(128), and C(∞). Within each category the percentage correct consonants is indicated by a black dot. In accordance with the pooled data in Fig. 3.5, the percentage of absolute correct responses (black dots) increases with decreasing cutoff frequency for every category. Except for C(∞), the score for stops is higher than for fricatives and vowel-likes. This implies that the presence of high-frequency modulations is particularly beneficial for short-duration consonants. In all conditions the majority of the responses fall into the correct category. This is most distinct and most consistent over conditions for the vowel-like consonants, in which case about 90% is perceived as vowel-like. With C(32), C(128), and C(∞) erroneous /χ/ and /h/ responses occur relatively often for fricative and vowel-like stimuli, respectively.

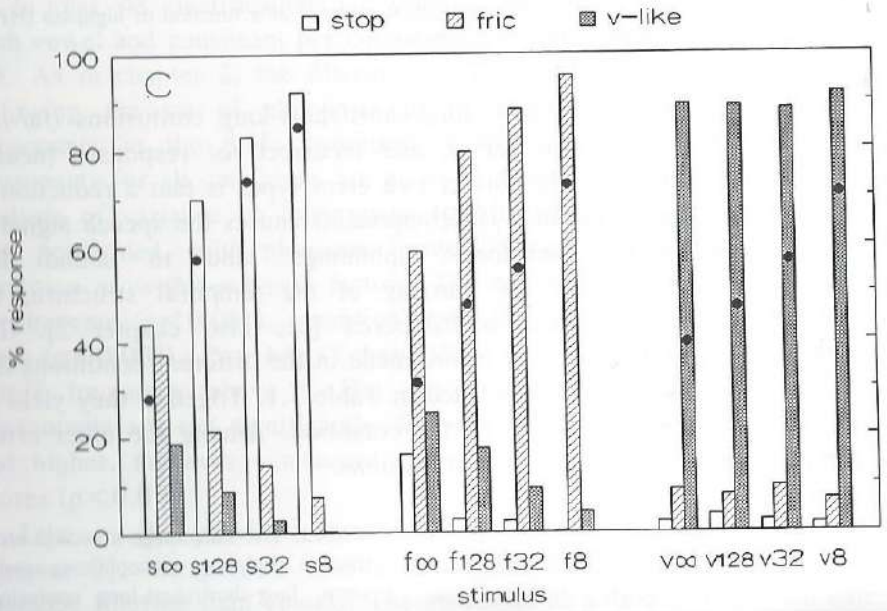

FIG. 3.6. Distribution of the responses for the consonants across the three categories stop, fricative, and vowel-like as a function of stimulus category and highpass cutoff frequency (s∞ = stop/∞ condition, s32 = stop/32-Hz condition, f128 = fricative/128-Hz condition, v8 = vowel-like/8-Hz condition, etc.). The black dots indicate the percentage of correctly identified consonants per category/condition.

3.5 GENERAL DISCUSSION

In discussing the general implications of the results of both experiments we want to focus on two aspects. First, although the results apply to normal-hearing listeners and the speech processing as such has no direct meaning for hearing aids, a few things can be said with respect to multichannel amplitude compression and dynamic gain control. Second, we will compare and combine the present results with the data from chapter 2 on lowpass envelope filtering. Before adressing the first aspect, a closer look at the effects of the signal processing is needed.

3.5.1 The modulation transfer function

The effect of the present signal processing on the transmission of temporal modulations can be described by the modulation transfer function (MTF; cf. Houtgast and Steeneken, 1985). The MTF is defined as the ratio between the *intensity*-envelope spectra after and before processing, usually applied in an octave band. A direct measurement of the MTF could thus be performed by comparing the modulation spectra of the squared input and output amplitude envelopes. However, the signal processing we used contains a nonlinear element, viz. the elimination of the negative parts (clipping) of the filtered envelope after addition of the mean amplitude (dc). For the calculation of the speech-transmission index (STI), Ludvigsen *et al.* (1990) proposed a method to deal with nonlinear distortions, based on the correlation between the entire input and output envelopes. Because we need an estimate of the transmission as a function of modulation frequency, we propose a different method, closely related to the MTF. This method takes into account the phases of the modulations in the original and modified envelope. We will name it the phase-locked MTF, referred to as MTF_{pl}. Details of the derivation and computation of the phase-locked MTF can be found in Appendix C.

As an example, Fig. 3.7 shows the normal MTF (solid lines) and the phase-locked MTF_{pl} (dashed lines) for the 2- and 16-Hz conditions, averaged over the intensity envelopes of four 1-oct bands with center frequencies of 0.25, 1, 2, and 4 kHz. As can be seen in Fig. 3.7, the MTF_{pl} for both conditions is lower than the MTF, indicating that the nonlinearities from the processing have been removed. It is also clear that the index for the lower modulations in the 2-Hz condition is still relatively high, despite the steep slope (40 dB/oct) of the highpass filter used. This can be explained as follows. Clipping occurs more often with low than with high envelope cutoff

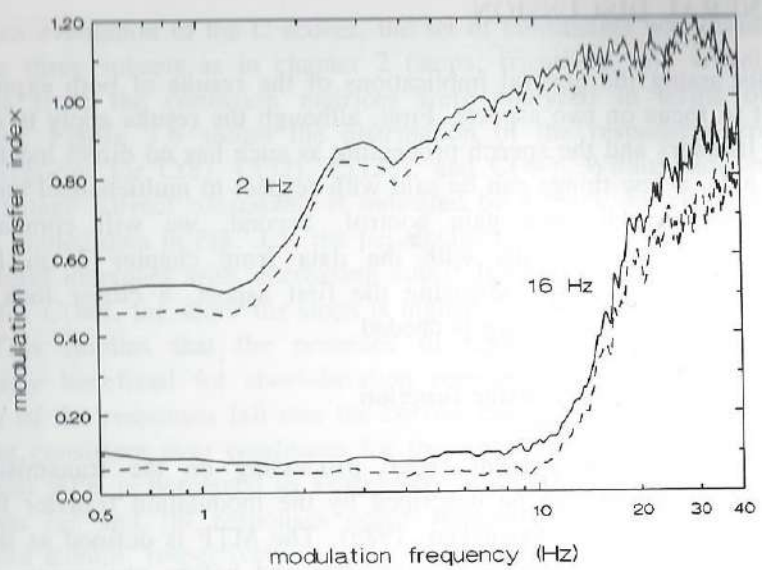

FIG. 3.7. Normal MTF (solid lines) and phase-locked MTF_{pl} (dashed lines) for the 2- and 16-Hz conditions, averaged over four oct bands with center frequencies of 0.25, 1, 2, and 4 kHz.

frequencies. It can in fact be seen as a partial restoration of the reduced low-frequency modulations. The low-frequency modulations that are reintroduced are not completely uncorrelated with the original modulations and will thus show up in the MTF_{pl}. So, in practice it is impossible to severely reduce low-frequency modulations when using low highpass envelope cutoff frequencies. However, if it were possible, the effect on (reduced) intelligibility would be minor, because other, slightly higher modulation frequencies are present and contain sufficient information.

3.5.2 Relation with amplitude compression

The method used to process the speech signal acts directly on the temporal envelope. In view of the MTF concept discussed above, it bears some relation to multichannel amplitude compression. As has been measured by Plomp (1988), short compression times reduce the temporal contrasts in the speech signal. The MTF in case of multichannel amplitude compression shows a (weak) highpass characteristic, the slope of which depends on the compression ratio and on the attack and release times. However, the reduction of the modulations is not as severe as in most of our conditions.

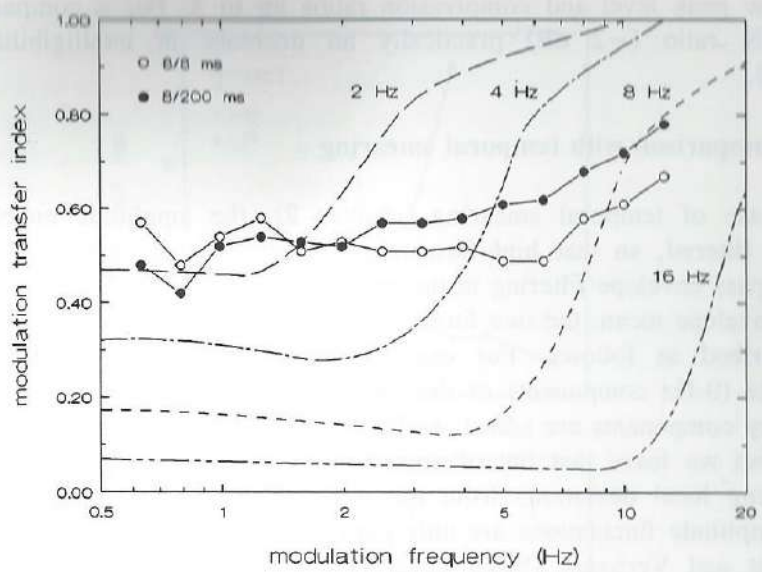

FIG. 3.8. MTF_{pl} for the 2-, 4-, 8-, and 16-Hz conditions, compared with the MTFs for amplitude compression, as measured by Plomp (1988). All values are averaged over oct bands with center frequencies of 0.25, 1, 2, and 4 kHz.

To illustrate this, Fig. 3.8 shows the phase-locked MTF_{pl} for the 2-, 4-, 8-, and 16-Hz conditions, averaged over 1-oct bands with center frequencies of 0.25, 1, 2, and 4 kHz (dashed lines). In the same figure the MTFs of the two compression systems discussed by Plomp (1988) are given. The latter are based on a compression threshold of 30 dB below peak level, a compression ratio of 4, and attack/release times of 8/8 and 8/200 ms, respectively. Roughly speaking, both compression curves show MTFs that are in between our 2- and 16-Hz condition; most of the time the modulation index is even higher than in our 8 Hz condition. The results of experiment 1 (1-oct bandwidth) show the same SRT for the 2- and 4-Hz conditions compared to unprocessed speech, whereas in the 8- and 16-Hz conditions the SRT increases by 1.5 and 2.8 dB, respectively. These data imply that, although amplitude compression does not improve intelligibility (at least not in normal hearing), the increase in SRT is limited to about 1 dB, which is less than the 5 dB suggested by Plomp (1988) for the 8/8-ms compression. The idea of only limited loss in intelligibility is supported by data from Hohmann and Kollmeier (1990), who investigated for normal-hearing listeners the effect of amplitude compression over 23 channels on CVC words. They used a fast 5/5 ms system with a compression threshold of 60

dB below peak level and compression ratios up to 3. For a comparatively high S/N ratio (−2 dB) practically no decrease in intelligibility was observed.

3.5.3 Comparison with temporal smearing

In case of temporal smearing [chapter 2], the amplitude envelope is lowpass filtered, so that high-frequency modulations are reduced. Because the highpass envelope filtering in the present study is accompanied by adding of the envelope mean, the two forms of degrading the temporal envelope can be described as follows. For every channel, the baseline is the mean amplitude (0-Hz component) of the entire utterance. Then other successive frequency components are added, with full or reduced amplitude (and certain phase, but we leave that out of consideration). Adding components means introducing local deviations from the mean envelope along the time-axis. Since amplitude fluctuations are only partly correlated over frequency bands (Houtgast and Verhage, 1991), deviations from the average spectrum are also introduced along the frequency axis. Adding all frequency components with their original amplitude yields, of course, the original spectrogram. Adding only low-frequency components gives global envelope variations. Conversely, adding only high-frequency components gives information on the rapid changes in the envelope. Apparently, lack of fast fluctuations can be compensated for by sufficient global information (components up to 16 Hz); on the other hand, lack of global information can be compensated for by giving enough fast fluctuations. In some way, the two types of information (although they are on a continuum) seem to be interchangeable. It must be emphasized that the 16-Hz limit for lowpass filtering and the 4-Hz limit for highpass filtering do not automatically imply that intelligibility is unaffected if only amplitude modulations between 4 and 16 Hz are transferred.

An interesting point on the scale is the crossover frequency, which divides the modulation-spectrum range into two parts that are equally important for intelligibility. Figure 3.9 shows the relative SRT (i.e., the measured SRT relative to the control condition) as a function of low- and highpass cutoff frequency, separately for the three processing bandwidths. Virtually independent of the bandwidth, the crossover frequency is about 8-10 Hz. At this frequency the masked SRT for sentences has increased by 2 dB. This means that, in a critical signal-to-noise ratio condition, reducing amplitude modulations below or above 8-10 Hz results in a loss in sentence intelligibility of 30-40%. This crossover frequency for temporal modulations

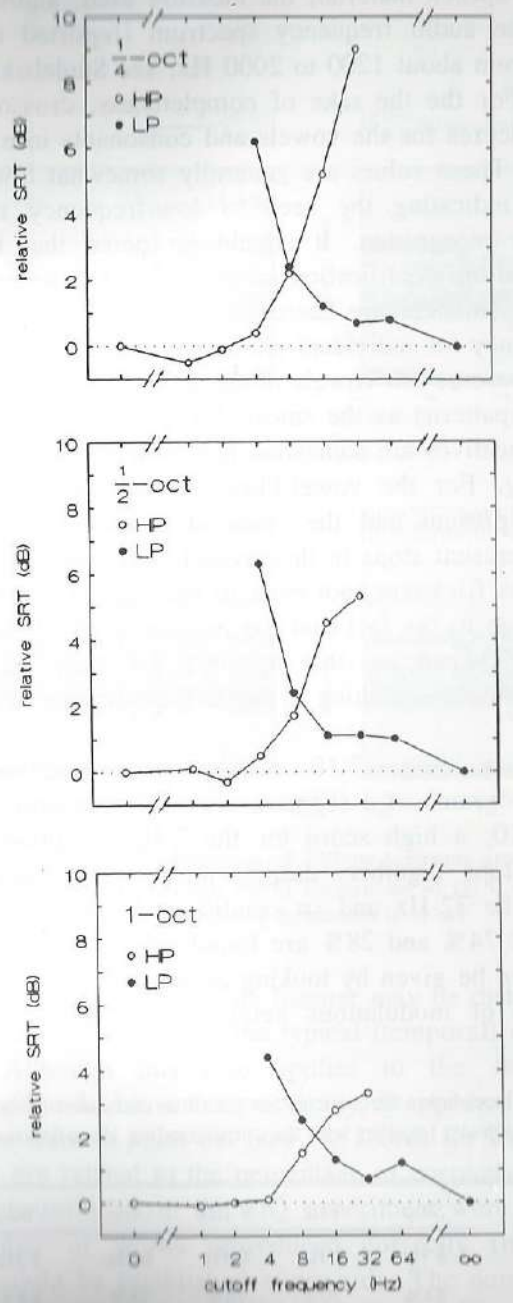

FIG. 3.9. Comparison of the SRT relative to the control condition as function of envelope lowpass (LP) and highpass (HP) cutoff frequency, processed in ¼-, ½-, and 1-oct bands.

may depend on the speech material; the measure used, and/or the talker, as is the case with the audio frequency spectrum (reported audio crossover frequencies range from about 1200 to 2000 Hz; see Studebaker *et al.* (1987) for an overview). For the the sake of completeness, crossover frequencies and corresponding scores for the vowels and consonants in experiment 2 are listed in Table 3.2. These values are generally somewhat lower, particularly for the fricatives, indicating the need of low-frequency modulations for individual phoneme recognition. It should be noted that these crossover frequencies are based on identification scores in quiet, whereas the crossover frequencies for the sentences are based on the SRTs in noise. On average, the crossover frequency for individual phonemes in quiet tends toward 4 Hz.

As mentioned before, the vowels in the present conditions show almost the same confusion patterns as the smeared vowels. As for the consonants, the scores for the fricatives are somewhat lower, which also accounts for the category consistency. For the vowel-likes a high category consistency is found in both the present and the smeared conditions. Contrary to the smeared stops, the present stops in the severely degraded conditions are not as much perceived as fricatives, not even in the case of 100% compression. The latter may be due to the fact that the envelopes of all stimuli contained at least a 0-Hz component, so that listeners got more familiar with the speech sounds this evokes, resulting in better performance for the 'difficult' stops.

As an illustration, Fig. 3.10 displays wideband waveforms and corresponding spectrograms of a segment around /k/ in different conditions. In view of Fig. 3.10, a high score for the 2-Hz condition [panel (b)] is plausible. For all stops together, despite the seeming correspondence in waveform between the 32-Hz and ∞ conditions [panel (c) and (d)], mean recognition scores of 74% and 28% are found, respectively. An explanation for this difference can be given by looking at the spectrograms. One can see that the suppression of modulations below 32 Hz does not remove any

TABLE 3.2. Crossover modulation frequencies for the three consonant categories, overall consonants, and overall vowels, together with the corresponding identification scores.

	Stop	Fric	V-like	Cons	Vowel
Frequency	8 Hz	2 Hz	6 Hz	6 Hz	3 Hz
Score	84%	79%	76%	78%	88%

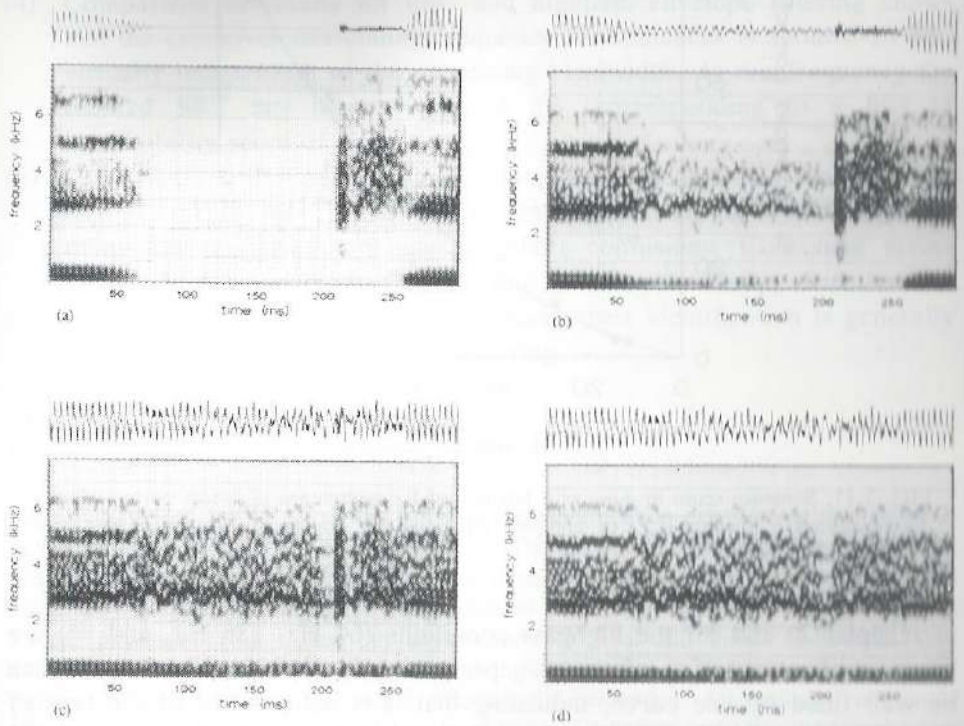

FIG. 3.10. Waveforms and spectrograms of 300-ms segments around /k/ in the nonsense syllable /iki/, with highpass envelope cutoff frequencies of (a) 0 Hz (unprocessed), (b) 2 Hz, (c) 32 Hz, and (d) ∞. The processing bandwidth is ¼ oct.

spectral cues for the /k/ burst. The listener may be disturbed by the "filling" of the occlusion part, by which the typical (temporal) character of a plosive is affected. Although this also applies to the ∞ condition, spectral information for the /k/ burst has practically disappeared in that case.

Finally, we want to point out how the scores for sentences in quiet (SIQ, experiment 1) are related to the percentage of correctly received words. The responses of the subjects in the SIQ experiment were recorded on tape, so that the number of words reproduced correctly (including articles and prepositions) could be established afterwards. The number of words per list of 13 sentences varies between 76 and 84. In Fig. 3.11 this relation is visualized. The mean scores for the lowpass conditions (0, ½, 1, and 2 Hz;

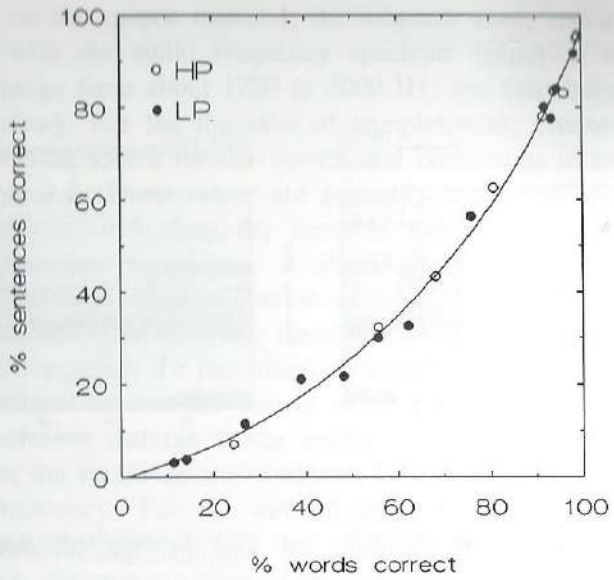

FIG. 3.11. Sentence score in quiet as a function of the percentage of words correct. Every dot represents a lowpass (LP) or highpass (HP) filtering condition for each of the three processing bandwidths.

cf. chapter 2) and for the highpass conditions (64 Hz, 128 Hz, and ∞) are given, without specific reference to processing bandwidth. The relation can be well fitted by one curve, indicating that it is independent of the type of filtering. Figure 3.11 shows that a low sentence score does not necessarily correspond to a low word score. For example, missing half of the sentences means on average that still about 75% of all words were heard correctly.

3.6 CONCLUSIONS

In summary, the most important conclusions of the present study are

(1) Temporal amplitude variations in succesive ¼-, ½-, or 1-oct frequency bands below 4 Hz can be reduced without reducing speech intelligibility for normal-hearing listeners.
(2) Sentences in quiet cannot be completely understood when only fast amplitude fluctuations above 32 Hz are present, but performance improves for broader frequency bands. For highpass envelope cutoff frequencies below 8 Hz intelligibility is not dependent on the processing bandwidth.

(3) The results suggest that multichannel amplitude compression with small time constants reduces intelligibility, but by no more than about 1 dB in terms of the masked speech-reception threshold.
(4) Comparison of results for low- and highpass envelope filtering shows that the crossover modulation frequency for sentences is about 8-10 Hz, virtually independent of the processing bandwidth. At this frequency the masked SRT has increased by 2 dB, corresponding to a loss in intelligibility score of 30-40%.
(5) Consonants suffer more from reducing slow modulations than vowels. Vowel errors are characterized by misclassifying diphthongs as monophthongs and short-long/long-short confusions. Consonant errors mainly consist of confusions within a category (stop, fricative, or vowel-like). Stops appear to suffer least; their identification is generally better than in case of temporal smearing.

Note

[1] We investigated medial phonemes only. In the identification experiments in chapter 2, initial and final consonants were also studied. The reason for not considering initial and final consonants this time is that the present signal processing causes a sudden onset and offset of the syllable, due to the reintroduction of the dc component in the highpass filtered envelope. This makes the identification of final and initial consonants very difficult. One could solve this problem by giving the processed syllable a smooth rise and fall, but it is unclear how that might influence the perception. Alternatively, one could add a few hundred milliseconds of low level noise just before and after the unprocessed syllable. But since this noise will be amplified by the processing, closely attached to both ends of the syllable, one cannot speak of initial and final consonants in their proper sense anymore.

CHAPTER 4

Temporal envelope and fine structure cues for speech intelligibility*

Abstract

This chapter describes a number of listening experiments to investigate the relative contribution of temporal envelope modulations and fine structure to speech intelligibility. The amplitude envelopes of 24 ¼-oct bands (covering 100-6400 Hz) were processed in several ways (e.g., fast compression) in order to assess the importance of the modulation peaks and troughs. Results for 60 normal-hearing subjects show that reduction of modulations by the addition of noise is more detrimental to sentence intelligibility than the same degree of reduction achieved by direct manipulation of the envelope; in some cases the benefit in speech-reception threshold (SRT) is almost 7 dB. Two crossover levels can be defined in dividing the temporal envelope into two equally important parts. The first crossover level divides the envelope into two perceptually equal parts: removing modulations either x dB below or above that level yields the same intelligibility score. The second crossover level divides the envelope into two acoustically equal peak and trough parts. The perceptual level is 9-12 dB higher than the acoustic level, indicating that envelope peaks are perceptually more important than troughs. Further results showed that 24 intact temporal speech envelopes with noise fine structure retain perfect intelligibility. In general, for the present type of signal manipulations, no one-to-one relation between the modulation-transfer function (MTF) and the intelligibility scores could be established.

*Paper accepted for publication in: J. Acoust. Soc. Am. **96** (1994c)

4.1 INTRODUCTION

The relevance of the temporal envelope in evaluating the quality of speech transmission has been elaborated in the concept of the modulation-transfer function (MTF; Houtgast and Steeneken, 1985). According to this concept, the detrimental effects of noise and reverberation on speech are adequately measured in terms of the reduction in modulation depth they produce in each of a series of frequency bands. The disappointing results in several studies on the benefit of amplitude compression in hearing aids were discussed by Plomp (1988) in terms of the importance of temporal modulations (intensity contrasts). Plomp used the MTF concept to argue that, just like noise, multichannel amplitude compression with small time constants reduces the intensity contrasts and will thus lead to reduced intelligibility.

In a comment on Plomp's paper, Villchur (1989) objected to the above way of reasoning. Villchur argued that the reduction of the MTF in itself is not the reason for reduced intelligibility. He said (p. 425): "It does not follow that if noise and compression each reduce the MTF, and noise reduces intelligibility, compression must reduce intelligibility equally". More specifically, adding noise to the speech signal causes the weaker elements (consonants) to be masked, which is not the case with compression, where this information is preserved.

The question underlying this discussion is whether the MTF concept holds for dynamic compression. In the present paper, we will not restrict ourselves to the matter of compression per se. Using various types of envelope manipulations, we will evaluate among other things whether the implication that a reduced MTF leads to reduced intelligibility is generally valid. As will be shown in the next sections, this is not always the case. It is possible to create conditions which have equal MTFs, but yield quite different intelligibility scores (or vice versa).

For each of a range of frequency channels, two features of the speech signal will be investigated: the temporal envelope and the fine structure (carrier signal). Of these two, the envelope appears to be most important for intelligibility. This is particularly illustrated by channel vocoders, where the amplitude modulations are preserved (up to about 25 Hz), whereas the fine structure at the receiver side is provided by a pulse or noise generator (Flanagan, 1972; O'Shaughnessy, 1987). Intelligibility is worse if the fluctuations in the temporal envelope are not transferred adequately. From previous experiments with filtered temporal envelopes (chapters 2 and 3) we know that intelligibility at a critical signal-to-noise ratio (SNR) is virtually unaffected when amplitude modulations either above 16 Hz or below 4 Hz

are reduced. In the extreme case of complete suppression of the modulations (flat envelope) in 24 ¼-oct bands, the intelligibility score for sentences in quiet drops to about 5%, demonstrating that the fine structure alone supplies insufficient information.

Concerning the importance of modulations, a next question is whether troughs are equally important as peaks. Commonly, it is assumed that most information is conveyed in the peaks of the speech signal. This can be inferred from the idea that additional noise acts as a sort of 'fence' where only the peaks of the speech can rise above. The higher the noise level, the less speech peaks can be perceived, until eventually the entire speech signal is masked. The peaks in a narrow frequency band are about 9-12 dB above the long-term average level (Pavlovic, 1987). So, on a dB-scale, they are relatively unaffected for SNRs down to at least 0 dB (see Festen *et al.*, 1990, Fig. 2, for an example in an octave band), whereas the modulations in the troughs have disappeared. In an attempt to study the relative contribution of weaker speech components without using noise, Plomp and Van Beek (1990) eliminated energy below a certain level (3-6 dB above the long-term rms level, L_{eq}). By presenting only the spectro-temporal peaks above that level, intelligible speech could be obtained. However, no formal listening tests were carried out.

In summary, two important factors that might affect speech reception have been described: envelope versus fine structure and peaks versus troughs. The aim of this study is to determine how intelligibility depends on the preservation of peaks and/or troughs in the temporal envelope with and without affecting the fine structure. Adding noise to the speech signal reduces the amount of temporal modulations by filling the troughs and, at the same time, changes the fine structure. However, to merely measure the effect of modulation reduction on intelligibility, the fine structure should remain intact. Therefore, a processing scheme was used that operates directly upon the temporal envelope. The present experiments were run to investigate the effects of (1) a noisy envelope alone, (2) removing modulations in the troughs, and (3) removing modulations in the peaks. Answers to (2) and (3) can give an estimate of the relative contribution of peaks and troughs. This gives a method of finding a critical level, subdividing the envelope into equally important peak and trough parts. In addition, while preserving the speech envelope, the effect of a noise fine structure was examined.

4.2 METHOD

4.2.1 Speech processing and experimental design

The basics of the analysis-resynthesis algorithm described in the previous chapters was used. The same 130 sentences from a female speaker (Plomp and Mimpen, 1979), sampled at 15,625 Hz with 16 bits resolution, served as speech material.

The wideband speech signal was split up into 24 ¼-oct bands, using a linear-phase FIR filter bank with slopes of at least 80 dB/oct, covering the range 100-6400 Hz. From the output of each channel the Hilbert envelope was computed. Each envelope was modified (see below) and the new narrowband signal was obtained by multiplying each sample of the fine structure by the ratio of the modified and the original envelope. In this way all original amplitude modulations were eliminated. Finally, all modified channels were added and the level of the new wideband signal was adjusted to have the same rms value as the input signal.

Four methods for envelope modification were investigated. They will be referred to by the acronyms FT, FP, BLK, and SN, respectively. Before explaining their meaning, one feature should be mentioned first: the *target level* within each ¼-oct band. For processing purposes, one may think of an imaginary horizontal line through the temporal envelope. It will be expressed in dB *re* long-term rms value (which is based on the 130 sentences). Thus, when it coincides with the long-term-rms level (L_{eq}), we speak of a target level of 0 dB (one level per ¼-oct band accounts for all sentences); $-x$ dB means moving the target level downwards; $+x$ dB moving it upwards. Table 4.1 gives an overview of all processing methods described below. Figure 4.1 shows schematic diagrams and examples of the FT, FP, and BLK envelope processing methods.

TABLE 4.1 Survey of the six processing strategies

Processing	Fine structure	Temporal envelope
REF	speech + noise	speech + noise
SN	speech	speech + noise
FT	speech	speech in peaks, flat in troughs
FP	speech	speech in troughs, flat in peaks
BLK	speech	zero in troughs, flat in peaks
NFS	noise	speech

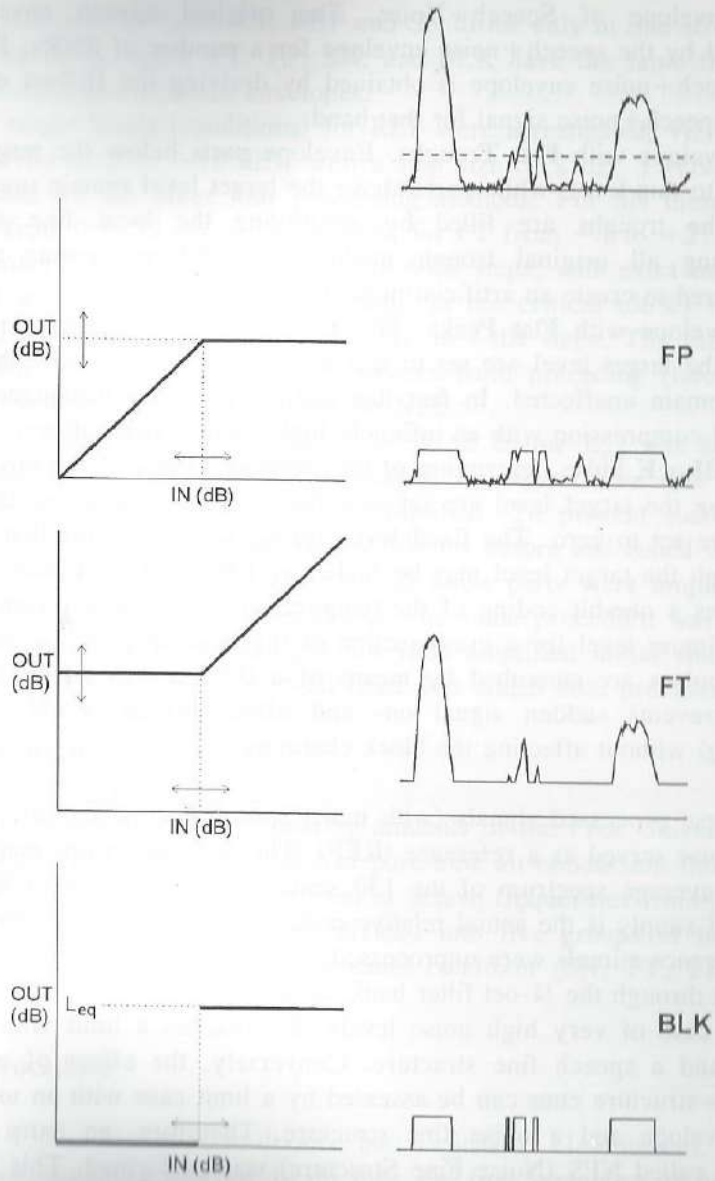

FIG. 4.1. From top to bottom: diagrams and examples of the modified temporal amplitude envelopes for FP, FT, and BLK, respectively. Top right is (a part of) an input envelope, with the target level drawn as a dashed line.

The four methods for envelope modification are

- SN: envelope of Speech+Noise. The original speech envelope is replaced by the speech+noise envelope for a number of SNRs. Each ¼-oct speech+noise envelope is obtained by deriving the Hilbert envelope of the speech+noise signal for that band.
- FT: envelope with Flat Troughs. Envelope parts below the target level are set to this level, while parts above the target level remain unaffected. Thus the troughs are filled by amplifying the local fine structure (removing all original trough modulations). FT processing may be considered to create an artificial noise-floor envelope.
- FP: envelope with Flat Peaks. The complement of FT. Envelope parts above the target level are set to this level, while parts below the target level remain unaffected. In fact this method performs instantaneous 24-channel compression with an infinitely high compression ratio.
- BLK: BLocK pulse description of the envelope. Parts of the envelope at or above the target level are set to a fixed level, parts below the target level are set to zero. The fixed level always equals L_{eq} for that channel (although the target level may be higher or lower). This method actually performs a one-bit coding of the temporal envelope, as a means to find the optimum level for a cross-section of the envelope. The edges of the block pulses are smoothed by means of a 0.5 ms half cosine window. This prevents sudden signal on- and offset (which would result in clicking) without affecting the block character.

Besides these processed signals (with intact speech fine structure), ordinary speech+noise served as a reference (REF). The noise spectrum matched the long-term average spectrum of the 130 sentences. In the case of REF, the target level simply is the actual relative noise level (i.e., the negative SNR). These reference stimuli were unprocessed, except that both speech and noise had passed through the ¼-oct filter bank.

In the case of very high noise levels, SN reaches a limit with a noise envelope and a speech fine structure. Conversely, the effect of a loss of speech fine-structure cues can be assessed by a limit case with an unaffected speech envelope and a noise fine structure. Therefore, an extra type of processing called NFS (Noise Fine Structure) was performed. This involved the combination of a random (noise) fine structure with a speech envelope. For this purpose, there was a parallel analysis (band filtering, envelope detection) of the separate speech and noise signals. For each ¼-oct band, the noise fine structure was multiplied by the ratio of the speech envelope and the noise envelope, sample by sample. In fact, NFS is a simple vocoder design (for a whispering voice).

As indicated in Table 4.1, REF and SN differ only in fine structure, their envelopes being equal; FT, FP, SN, and BLK have the same fine structure (speech only) but different envelopes.

Six target levels (conditions) for REF were investigated, viz. 0 to 10 dB (i.e., SNRs of 0 to −10 dB), with a step size of 2 dB. Twelve conditions were used for the other four processing methods. For SN these conditions ranged from 0 to 22 dB, in 2-dB steps; for FT from −1 to +21 dB in 2-dB steps; for FP from −39 to −3 dB, in 4-dB steps, with extra measurements at −21 and −25 dB (for 'fine tuning' in the critical (50%) intelligibility region); for BLK from −35 to +9 dB, in 4-dB steps. The ranges for the conditions were based on the experiences from preceding (informal) tests. Clearly, only one condition existed for NFS.

For FT and SN, a noise floor at 35 dB below L_{eq} was added to the wideband speech signal before processing, to ensure there was just enough signal in the (silent) troughs to be amplified. To prevent sudden on- and offset of a sentence, the noise started 500 ms before and lasted until 500 ms after the speech. These initial and final noise parts were amplified by the processing. For the sake of similarity, the same procedure was undertaken for FP (where the processing did not yield amplified initial and final noise parts); for BLK and NFS, −35-dB noise was added after processing.

4.2.2 Subjects

Subjects were 60 normal-hearing students of the Free University, whose ages ranged from 18 to 30. All had pure-tone air-conduction thresholds less than 15 dB HL in their preferred ear at octave frequencies from 125 to 4000 Hz and at 6000 Hz. They were divided into five groups of twelve. Each group was assigned to an experimental condition (SN, FT, FP, BLK, or REF + NFS, respectively).

4.2.3 Procedure

The 130 sentences were divided into 12 lists of 11 sentences. Since there were two sentences too few to do this, the first sentences of lists 1 and 2 also served as the first sentences of lists 11 and 12. This did not matter, because only the last ten sentences in a list were used for the intelligibility scores. Six lists were used for REF, all twelve lists were used for FT, FP, SN, and BLK. Lists were presented in a fixed order. The sequence of the

conditions was varied according to a digram-balanced 6×6 (REF) or 12×12 Latin square. So, each sequence was presented to one subject in the FT, FP, SN, and BLK tests, and to two subjects in the REF test. For NFS, four lists of 11 sentences pronounced by a male speaker were used; each of these lists was presented to three subjects.

All stimuli were presented at a level of approximately 65 dB(A). Every sentence was presented once, after which a subject had to reproduce it as accurately as possible. A response was scored as correct only if the entire sentence was reproduced correctly. In the REF test, the level of the masking noise was fixed at 65 dB(A) and the level of the sentences varied according to the condition. The noise started 500 ms before and ended 500 ms after the sentence.

The sentences were presented monaurally through a headphone (Sony MDR-CD999) at the subject's ear of preference and in a soundproof room. Before the actual tests (except for NFS, which followed directly after REF), a list of 11 sentences pronounced by a male speaker in a representative condition was presented, in order to familiarize the subjects with the procedure. For each test and each condition the scores for sentences 2-11 were counted.

4.3 RESULTS AND DISCUSSION

In presenting and discussing the results throughout the rest of the chapter, the following abbreviation will be used for the different processing methods and experimental conditions: The condition (target level) will be written in parentheses after the acronym of the processing algorithm; e.g., REF(8), FT(10), FP(−15), SN(0), BLK(−7), etc.

The mean scores (percentages) as a function of condition for each of five processing algorithms are given in Table 4.2. The figures in Table 4.2 are the arithmetical averages of the raw scores of 12 subjects. The mean score for NFS is 98.3%, which clearly shows that if the envelope is intact, the fine structure is of minor importance for intelligibility. For further statistical analysis of the data, arcsine transformed scores were used (Studebaker, 1985). The mean and standard error of the transformed scores was calculated for each condition; these means and standard errors were transformed back into percentages and are shown in Figs. 4.2 and 4.3.[1] All statistics consisted of a repeated-measures analysis of variance (ANOVA) with processing method as between-subjects factor and target level as within-subjects factor. In case of significant interactions, tests for simple effects (Kirk, 1968) were carried out.

TABLE 4.2. Mean raw scores (percentages) as a function of the condition (dB target level) for each of the five processing algorithms

	Processing				Processing	
Condition	FT	FP	BLK	Condition	SN	REF
21	28.3			22	17.5	
19	30.8			20	16.7	
17	23.3			18	16.7	
15	32.5			16	17.5	
13	40.0			14	20.0	
11	55.0			12	18.3	
9	65.8		10.0	10	24.2	0.0
7	84.2			8	33.3	6.7
5	90.0		70.0	6	55.8	40.8
3	95.0			4	85.8	69.2
1	99.2		98.3	2	94.2	83.3
-1	96.7			0	98.3	92.5
-3		99.2	96.7			
-7		100.0	97.5			
-11		96.7	98.3			
-15		93.3	95.8			
-19		75.8	94.2			
-21		73.3				
-23		65.0	85.0			
-25		50.8				
-27		35.8	74.2			
-31		18.3	51.7			
-35		8.3	36.7			
-39		8.3				

4.3.1 Speech+noise envelope, artificial noise-floor envelope

Figure 4.2 displays the scores for REF, FT, and SN. The results for REF display the well-known intelligibility curve (cf. Plomp and Mimpen, 1979), with a speech-reception threshold (SRT, level for 50%) at a target level of about 5.5 dB. The scores for REF are never better than for SN and FT. In terms of the SRT, the latter yield about 6.5 and 12 dB, respectively. The effects of processing, condition and the interaction between them are highly significant ($p<0.001$). *Post hoc* tests (Scheffé) on the simple effects

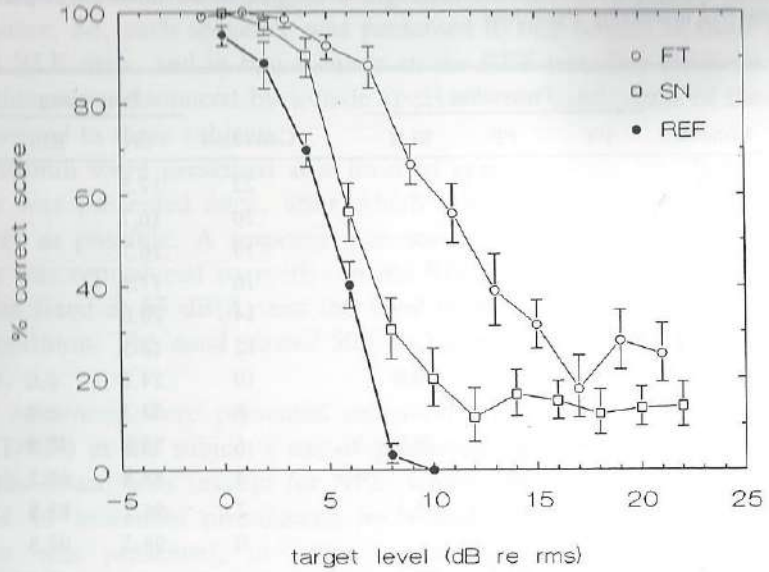

FIG. 4.2. Mean score and standard error (vertical bars) of FT, SN, and REF as a function of target level.

showed significantly lower scores for REF than for SN for target levels above 2-3 dB ($p<0.01$; $p<0.05$ at 5 dB). From 4 to 10 dB the scores for REF and SN decrease rapidly. Their functions show almost the same slope; the slope for FT is flatter. The scores for FT and SN differ significantly over the range 5 to 15 dB ($p<0.01$; for 15 dB $p<0.05$).

In the light of the processing methods (see Table 4.1), the difference in scores between REF and SN must be attributed to the unaffected fine structure in SN. Preservation of the fine structure if the original intensity contrasts are disturbed yields a 1-dB benefit of the SRT. In the case of extremely high target levels, when the envelope contains practically no speech information anymore, the fine structure seems to supply some minimal cues (0% for REF and 17% for SN). An explanation for the difference between the SN and FT scores must be based on the presence of non-relevant fluctuations, originating from the noise, in the temporal envelope of SN. Apparently, these fluctuations interfere with those of the speech signal and leave the listener with a 'sorting problem', i.e., he/she is unable to separate the relevant (speech) modulations from the non-relevant (noise) modulations.

In summary, flattening the troughs of the speech signal as a means to reduce the amount of temporal modulations (artificial noise-floor envelope) is less detrimental to intelligibility than the same modulation reduction brought about by the addition of noise. The benefit expressed in terms of the SRT is about 6.5 dB.

4.3.2. Removing envelope peaks and/or troughs

Figure 4.3 displays the scores for FP, BLK, and FT. The data for FP show that a considerable part of the temporal-envelope peaks (up to 15 dB below L_{eq}) can be 'chopped off' before a detrimental effect on intelligibility is noticeable. The SRT for FP is around −25 dB. In contrast to the other processing methods, the curve for BLK is typically nonmonotonic. Apparently, complete intelligibility is reached for a wide range between −19 and +1 dB. At lower target levels, all frequency bands are filled more evenly, resulting in less temporal cues for the listener. At higher target levels, BLK only codes the upper peak parts; sentences in these conditions consist of only a few short-duration block pulses per ¼-oct band. The limit cases for low and high target levels are quite different, viz. severely

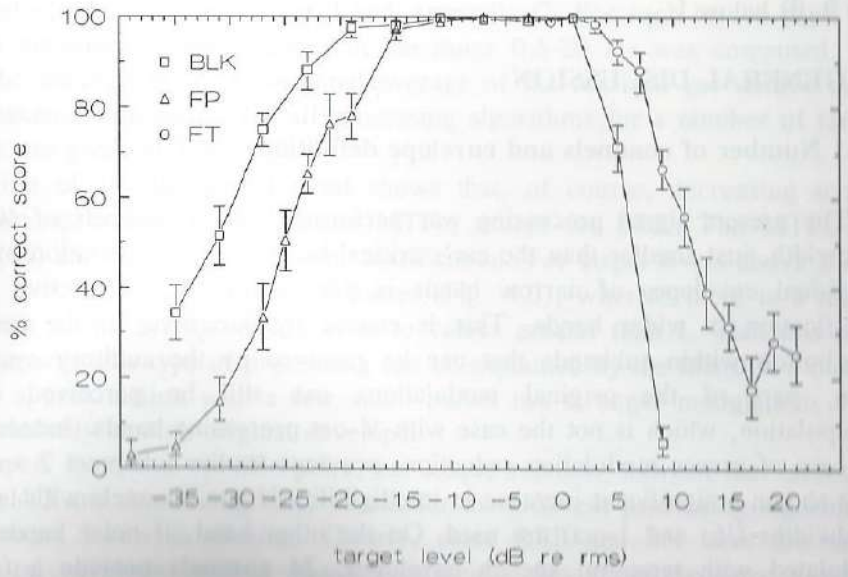

FIG. 4.3. Mean score and standard error (vertical bars) of BLK, FP, and FT as a function of target level.

compressed speech and complete silence, respectively. For high target levels both envelope and fine structure information is lost, so that intelligibility decreases rapidly (steep slope); for low target levels the loss of envelope information may be compensated for by relying more on the fine structure. The physical difference between FP and BLK may explain the discrepancy between their scores at low target levels. This difference consists of the presence (FP) or absence (BLK) of speech in the troughs. As for FP, at lower target levels the modulations in the troughs become relatively stronger. They may therefore disturb the perception of the peaks (or what is left from them) as separate entities. BLK speech does not have this problem; the peaks are well separated in time.

FP and FT have virtually the same slope (except for a different sign) from −31 to −15 dB and from 3 to 17 dB, respectively. The symmetry axis runs at a target level of about −6 dB. So, flattening the envelope x dB below (FT) or above (FP) that level results in virtually equal intelligibility (except for extreme deviations from this level). Simple one-bit coding of the envelope (BLK) retains perfectly intelligible speech when applied in a 20-dB range around L_{eq}−9 dB. FT versus FP on the one hand and BLK on the other are two approaches to estimate the relative contribution of peaks and troughs. Thus, the critical target level ('perceptual crossover level') for subdividing the envelope in two equally important parts is estimated at about 6 to 9 dB below L_{eq}.

4.4 GENERAL DISCUSSION

4.4.1 Number of channels and envelope definition

The present signal processing was performed on 24 channels of ¼-oct bandwidth, just smaller than the ear's critical bandwidth. Modification of the individual envelopes of narrow bands is perceptually more effective than modification in wider bands. This is caused by variations of the energy distribution within subbands that can be resolved by the auditory system. Thus, part of the original modulations can still be perceived after manipulation, which is not the case with ¼-oct processing bands. Indeed, in the case of severe modulation reduction, previous studies (chapters 2 and 3) have shown a significant increase in intelligibility if less channels with larger bandwidths (½- and 1-oct) are used. On the other hand, if noise bands are modulated with temporal speech envelopes, 24 channels provide a fairly detailed transmission of the spectrogram. The outcome of the NFS experiment may therefore not be very surprising. In a similar experiment

Shannon *et al.* (1994) reported nearly 100% sentence intelligibility with only four channels, indicating that little spectral detail is sufficient for speech recognition in quiet.

A second point is the definition of the temporal envelope. We adopted the Hilbert envelope, although many studies use a method of rectification and lowpass filtering. The Hilbert envelope has the clear advantage that it accurately follows all amplitude modulations of a frequency band, running smoothly over the actual waveform. As a consequence, any modification of the envelope can be performed without keeping unwanted (original) temporal modulations intact. In this way one can precisely control the envelope cues transmitted to the listeners.

4.4.2 Results in relation to the MTF

As mentioned in the Introduction, one aim of this study was to establish the relation between the intelligibility scores and (the prediction by) the MTF. For all processing methods, the MTF was measured for each of five octave bands with center frequencies at 0.25, 0.5, 1, 2, and 4 kHz. The measurements were based on the long-term envelope spectra (processed in ¼-oct bands versus unprocessed) of a 71-s speech fragment (30 concatenated sentences) and were performed according to the phase-locked MTF procedure described in chapter 3 and Appendix C. For each octave band the mean modulation reduction (m) in the range 0.5-20 Hz was computed. To get the average MTF, a weighted average of the mean m per octave band was taken.[2] The results for all processing algorithms for a number of target levels are given in Table 4.3.

First of all, the global trend shows that, of course, decreasing scores correspond to decreasing average MTFs, except for BLK. The MTFs for BLK(15) and BLK(10) need some clarification. For target levels above 2 dB, there is a sharp decrease in the scores (Fig. 4.3), whereas there is a major increase in the average MTF, even to values greater than 1. That this may happen with this type of processing can be explained by the fact that a block-pulse approximation with a few small pulses has stronger modulations than the relatively smooth original envelope.

Apart from this specific point, an unique relation between the values in Table 4.3 and the intelligibility scores for the different processing algorithms and target levels is missing. Clearly, since REF and SN have the same temporal envelope (with a ¼-oct resolution), they have the same average MTFs. These values are practically equal to the theoretical values.[3] However, the scores for REF and SN (Table 4.2, Fig. 4.2) are different for

TABLE 4.3. Average phase-locked MTFs of five processing algorithms for various target levels, based on a 71-s speech fragment.

Target level	Processing			
	REF/SN	FT	FP	BLK
20	0.01	0.07		
15	0.03	0.07		1.47
10	0.10	0.11		1.15
5	0.26	0.23		0.73
0	0.52	0.49	0.33	0.46
-5			0.24	0.31
-10			0.19	0.23
-15			0.16	0.18
-20			0.13	0.15
-25			0.11	0.13
-30			0.09	0.11

most target levels. For example, an average MTF of 0.10 at 10 dB corresponds to a score of either 0% (REF) or 24% (SN). The MTFs of REF(10) and FT(10) are almost equal, unlike their scores of 0% and 60%, respectively.

Comparison of FT(0) and FP(0) is even more illustrative of why one has to proceed with caution in applying the MTF as a direct measure for intelligibility. Scores for these conditions are high (nearly 100%), whereas their MTFs differ substantially, viz. 0.49 and 0.33. The MTF can even drop as low as 0.19 for FP(−10) with a corresponding score of about 97%.

These are just some examples; by further comparing the average MTF values with the measured intelligibility scores one can find more discrepancies. Let it be clear that the above is not meant to discredit the MTF concept in general. The MTF was mainly developed for practical purposes, and has proved to be very successful in studying the quality of speech transmission in the presence of disturbances like bandpass limiting, noise, reverberation, echoes, and (wideband) automatic gain control (Steeneken and Houtgast, 1980; Steeneken, 1992). But the experiments in this study have demonstrated that its use cannot simply be extended to any manipulation of the temporal speech envelope (see also Hohmann and Kollmeier, 1990; Verschuure *et al.*, 1993). Taking FP as an extreme example of fast multichannel compression and referring to the discussion mentioned in the Introduction, one must conclude that the relation between

MTF, compression, speech+noise, and intelligibility is rather obscure. Indeed, as pointed out by Plomp (1988), fast multichannel compression (even in a more moderate form) reduces the MTF, as does the addition of noise. But the results of these experiments suggest that Villchur (1989) was right in stating that the effects of noise and compression are different and that reduction of the MTF does not automatically reduce intelligibility.

4.4.3 Temporal envelope statistics

For a better insight in the amount of information that is (physically) present after processing, some data about the temporal envelope statistics may be helpful. Using the same 71-s fragment as for the MTF measurements, the percentage of the time that the amplitude envelope exceeded the target level was measured for six ¼-oct bands, distributed regularly over the spectrum. The results for a range of target levels are shown in Fig. 4.4. Not surprisingly, all curves are monotonically decreasing, with slightly different paths for the lower and higher frequency bands. The steeper the slope, the narrower the peaks in that frequency band. By relating these data to the intelligibility scores, one can see how sensitive listeners are to minor

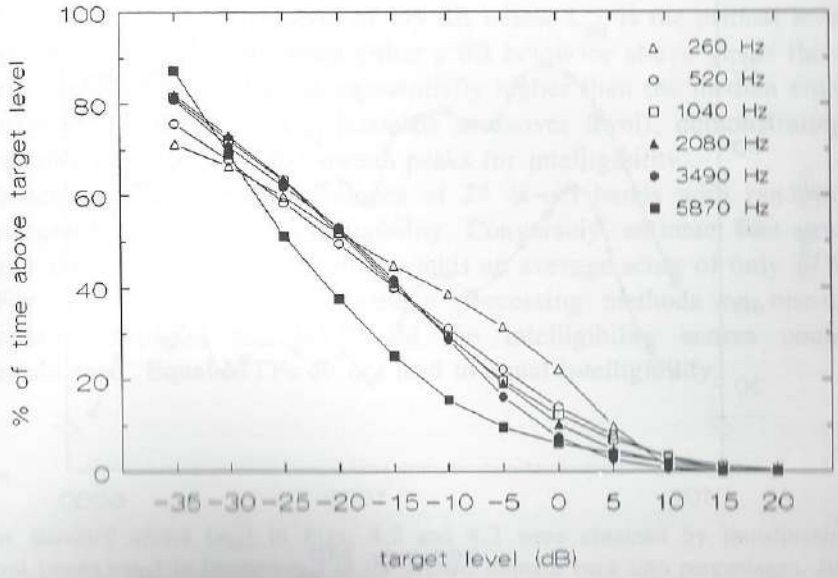

FIG. 4.4. Percentage of the time that the temporal envelopes in six ¼-oct bands of a 71-s speech fragment are above the target level.

differences in the temporal envelope. For example, at 10 dB, there are peaks for on average 2% of the time, the very amount that the envelopes contain after FT processing. Still, this results in a mean intelligibility score of 60%. At 15 dB there are on average 0.5% peaks, and the mean score for FT still is 33%. This sensitivity to a minimum peak remainder (even at 20 dB) may be the reason why the plateau for extreme FT conditions lies around a score of 25-30%. In chapters 2 and 3, using really flat envelopes, significantly lower intelligibility scores of 3-7% were found.

According to Fig. 4.4, the median envelope level, is about 18 dB below L_{eq}. Figure 4.5 displays L_{eq} and the median for all 24 ¼-oct bands of the 71-s fragment. For completeness, the dashed line shows L_{eq} for all 130 female sentences. The difference between L_{eq} and the median is relatively independent of the frequency band. The mean difference over the 24 bands is 18.4 dB with a standard deviation of 3.5 dB. This level is much lower than the crossover level of 9-9 dB below L_{eq} that was found for the intelligibility scores of FT versus FP and BLK. So, for the present sentence material, one can say that there are two different crossover levels: (1) an acoustic crossover level of on average 18 dB below L_{eq} that divides the temporal envelope into two equal peak and trough parts, and (2) a perceptual

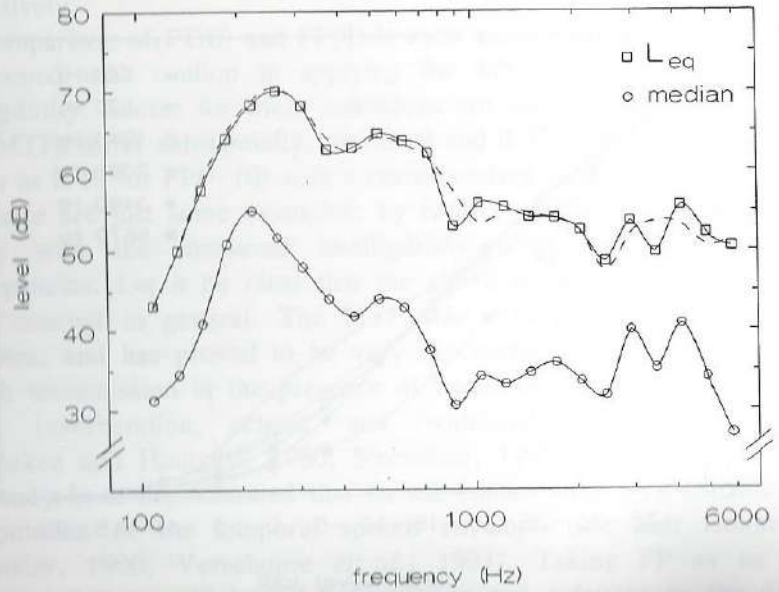

FIG. 4.5. L_{eq} and median for 24 ¼-oct bands, based on a 71-s speech fragment (30 concatenated female sentences). The dashed line reflects L_{eq} over all (130) sentences.

crossover level of 6-9 dB below L_{eq} that yields equal intelligibility scores when removing all modulations either x dB below or above that level. The fact that the perceptual level is 9-12 dB higher than the acoustic level suggests that the envelope peaks are more important for intelligibility than the troughs.

4.5 CONCLUSIONS

The main conclusions of this chapter are

(1) Reduction of the temporal modulations in speech by direct manipulation of the envelope is less detrimental to intelligibility than the same degree of reduction caused by adding noise. Preservation of the speech fine structure alone (speech+noise envelope) results in a 1-dB decrease of the SRT. If in addition an artificial noise-floor envelope is used, the SRT decreases by an extra 5-6 dB. This indicates that fine structure cues play a less important role than envelope cues and that noise introduces spurious modulations, disturbing the perception of the relevant speech modulations.

(2) The most important fraction of the dynamic range, providing essentially 100% sentence intelligibility, is between 19 dB below and 1 dB above L_{eq}.

(3) A perceptual crossover level of 6-9 dB below L_{eq} is the critical level for which removing modulations either x dB below or above yields the same intelligibility score. This is substantially higher than the median envelope level of 18 dB below L_{eq} (acoustic crossover level), demonstrating the relative importance of the speech peaks for intelligibility.

(4) Intact temporal speech envelopes of 24 ¼-oct bands with random fine structure retains perfect intelligibility. Conversely, an intact fine structure and a random temporal envelope yields an average score of only 17%.

(6) For the present type of envelope processing methods no one-to-one relation between the MTF and the intelligibility scores could be established. Equal MTFs do not lead to equal intelligibility.

Notes

[1] The standard errors (σ_M) in Figs. 4.2 and 4.3 were obtained by transforming the interval (mean$-\sigma_M$) to (mean$+\sigma_M$) in the arcsine domain back into percentages. Because of this (nonlinear) inverse transform, the mean is generally not in the middle of the interval.

[2] The weighting factors for the octave bands were derived from Steeneken and Houtgast (1980). Their original weighting factors account for a total of seven octave bands, viz. those used here and two with center frequencies of 125 Hz and 8 kHz. Because we omitted the latter two, the weighting factors for the remaining five bands were recalculated so that their sum equals 1: $W_{0.25}=0.196$; $W_{0.5}=W_1=0.157$; $W_2=0.255$; $W_4=0.235$.

[3] The theoretical values for the MTF in case of steady-state interfering noise are given by the formula $m=(1+10^{-SNR/10})^{-1}$, where m is independent of the modulation frequency (Houtgast and Steeneken, 1985).

CHAPTER 5

Speech intelligibility in noise: relative contribution of speech elements above and below the noise level*

Abstract

In the previous chapter the relative contribution of temporal modulations and fine structure to sentence intelligibility was investigated. This chapter reports additional listening experiments to assess in more detail the effect of masking noise on the peaks and troughs of the speech signal. For this purpose, the signal structure of each ¼-oct band in a 24-band filterbank (100-6400 Hz) was altered by manipulating the distribution of speech and noise over the sentences. Results for 12 normal-hearing subjects indicate that removing noise from the peaks has no effect on intelligbility; removing the speech signal from the noisy troughs, however, yields a 2-dB increase of the SRT. So, it appears that, even below the noise level, weak speech elements do contribute to intelligibility.

*Submitted as a Letter to the Editor in: J. Acoust. Soc. Am.

5.1 INTRODUCTION

When speech is presented against a noise background, the entire waveform is affected, but the reduction of intelligibility is generally attributed to masking of the weaker speech elements. Results from experiments in chapter 4 show that noise introduces spurious modulations, obscuring the relevant speech modulations (particularly in the troughs), rather than obscuring the speech fine structure. However, despite the masking effect of noise on low-level envelope and fine-structure speech components, there may still remain some cues that are beneficial to intelligibility. For example, with sinusoidally-modeled speech, which does not contain the noise-corrupted weaker speech elements, Kates (1994) found significantly reduced consonant recognition.

The experiments presented here can be regarded as an extension of some earlier work (chapter 4) in which the effect of removing modulations from the peaks and/or troughs of the temporal envelopes of narrow frequency bands (with preserved speech fine structure) on intelligibility in quiet was studied. The aim of the present experiments is to assess in more detail how the intelligbility of sentences in noise depends on the reduction of information in the peaks and troughs of the speech signal. For clean speech, Plomp and van Beek (1990) have put forward the idea of eliminating all spectrotemporal components that would fall below the noise level (troughs), in order to study the relative contribution of the weaker speech elements to intelligibility. This approach provides an unaffected transfer of the speech elements above the noise level (peaks), but the effect of spurious noise modulations in the troughs is not taken into account.

In the method presented in this paper, the distribution of speech and noise over the sentences was manipulated for a series of narrow frequency bands. Sentences were processed as to have speech only in the peaks and either noise only or speech+noise in the troughs. Compared to the ordinary speech+noise situation, the latter condition can be seen as removing noise from the peaks.

5.2 METHOD

5.2.1 Material, design

The speech material consisted of 130 everyday Dutch sentences of eight to nine syllables read by a female speaker (Plomp and Mimpen, 1979). The masking noise had the same spectrum as the long-term average spectrum of

the sentences. Sentences and noise were digitized at a sampling rate of 15,625 Hz and 16-bits resolution. Speech processing was done on an Olivetti M290S computer with an OROS-AU21 card with TMS320C25 signal processor.

In the analysis-resynthesis algorithm the speech signal is divided into 24 ¼-oct bands (100-6400 Hz), after which the Hilbert envelope of each band is computed. In each band the envelope triggers the type of signal that is transferred (speech, noise, or speech+noise). The envelope is used to compare the momentary speech level with the fixed average noise level for a certain speech-to-noise ratio (target level). The speech level can be above (peak) or below (trough) the target level of that band. If there is a peak, the original speech signal is transferred; if there is a trough, either only noise at the target level (TN, Troughs with Noise) or speech supplemented with noise at the target level (TSN, Troughs with Speech+Noise) is transferred. So, with both TN and TSN the peaks per ¼-oct band remain undisturbed, whereas the troughs contain noise and speech+noise, respectively. The intelligibility of TN and TSN was investigated for six speech-to-noise ratios (conditions), viz. −10, −8, −6, −4, −2, and 0 dB.

5.2.2 Subjects

Subjects were 12 normal-hearing students of the Free University, whose ages ranged from 19 to 25. All had pure-tone air-conduction thresholds less than 15 dB HL in their preferred ear at octave frequencies from 125 to 4000 Hz and at 6000 Hz.

5.2.3 Procedure

The 130 sentences were divided into 12 lists of 11 sentences, with copies of the first sentences of lists 1 and 2 for lists 11 and 12, respectively. Only the last ten sentences per list were used for the intelligibility scores. Lists were presented in a fixed order. For both TN and TSN, the conditions varied according to a digram-balanced 6×6 Latin square in which each sequence of conditions was presented to two subjects. Both the order of the tests (TN versus TSN) and the list series (lists 1-6 versus lists 7-12) were counterbalanced.

Stimuli were presented monoraully at a level of 65 dB(A) through a headphone (Sony MDR-CD999) at the subject's ear of preference. Every sentence was presented once, after which the subject had to repeat it. A

response was marked as correct only if the complete sentence was reproduced without a single error. To prevent sudden stimulus on- and offset, the noise started 500 ms before and ended 500 after the sentence. A TN introductory list by a male speaker in the −4-dB condition was presented in advance to familiarize the subjects with the procedure.

5.3 RESULTS

For comparison, the ordinary speech+noise conditions from chapter 4 are used as a reference (REF). The results for REF are based on scores of a different group of 12 young normal-hearing listeners in an identical experimental procedure.

Table 5.1 gives the mean raw scores for REF, TN, and TSN as a function of target level. For further statistical analysis, arcsine transformed scores were used (Studebaker, 1985). The mean scores and standard errors for TN, TNS, and REF - in percentages after inverse transform from the arcsine domain - are plotted in Fig. 5.1. The three curves show virtually the same slope, but the one for TN has shifted to the right. In terms of the SRT, TN and TSN yield values of −3.5 and −5 dB, respectively, as opposed to −5.5 dB for REF.

A repeated-measures ANOVA with processing method as between-subjects factor and condition as within-subjects factor showed significant effects for both factors ($p<0.001$) and for the interaction ($p<0.01$). Tests for simple main effects (Kirk, 1968) and *post hoc* tests (Tukey HSD) indicated that TSN is not significantly different from REF in any of the

TABLE 5.1. Untransformed mean scores (percentages) as a function of condition (S/N ratio in dB) for each of the three processing methods.

	Processing		
Condition	REF	TN	TSN
−10	0.0	0.0	0.0
−8	6.7	0.8	7.5
−6	40.8	13.3	25.0
−4	69.2	44.2	65.8
−2	83.3	75.8	91.7
0	92.5	90.8	94.2

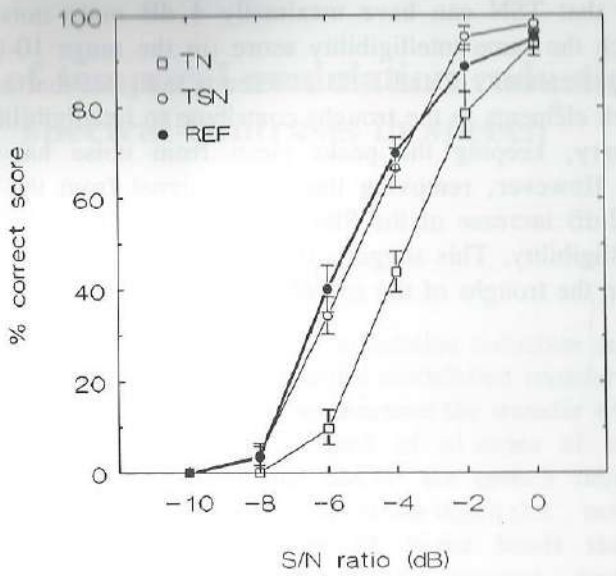

FIG. 5.1. Mean score and standard error of TN, TSN, and REF as a function of S/N ratio.

conditions. At −4 and −6 dB the scores for TN differ significantly ($p<0.01$) from REF (and from TSN); at −2 dB this only accounts for TN and TSN ($p<0.01$).

5.4 DISCUSSION AND CONCLUSIONS

The results indicate that it does not matter for intelligibility if the peaks remain 'clean', i.e., not affected by noise (TSN versus REF). This is not surprising, considering the fact that the peaks reach to about 10 dB above the long-term average speech level. In other words, even with REF the speech peaks remain virtually intact. If, on the other hand, the speech parts in the troughs are entirely replaced by noise, intelligibility at critical speech-to-noise ratios is significantly reduced (TN versus REF/TSN); the increase in terms of the SRT is about 2 dB.

Physically, TSN and TN only differ in the troughs, where the noise dominates. Despite this, it looks like one can listen 'through' the noise in the troughs. Given the fact that a steady-state noise with the long-term speech spectrum is used, the speech-to-noise ratio in every narrow (critical) band is

constant. In these cases the detection threshold for complex signals is about 4 dB below the noise level (Zwicker and Feldtkeller, 1967; Moore, 1982). This implies that TSN can have maximally 4 dB more noise than TN in order to reach the same intelligibility score (in the range 10-90%, see Fig. 5.1). The experimentally found 2-dB difference indicates that only part of the audible speech elements in the troughs contribute to intelligibility.

In summary, keeping the peaks clean from noise has no effect on intelligbility. However, removing the speech signal from the noisy troughs results in a 2-dB increase of the SRT, corresponding to a 25%-30% loss in sentence intelligibility. This suggests that even in noise listeners make use of information in the troughs of the speech signal.

CHAPTER 6

Effect of temporal modulation reduction on spectral contrasts in speech*

Abstract

In this chapter the effect of temporal modulation reduction on spectral contrasts is investigated. First, a spectral modulation transfer function (SMTF) is presented as a method to measure the transfer of spectral ripples (sinusoidal periods/oct) in each of a series of short-time spectral envelopes. Measuring the SMTF for speech subjected to uniform reduction of the temporal modulation depth (i.e., modulation-frequency-independent reduction) in 24 ¼-oct bands showed an almost equal uniform reduction of the spectral modulations. Furthermore, the SMTF was used to measure the reduction of spectral contrasts associated with lowpass and highpass temporal-envelope filtering [see chapters 2 and 3]. Comparison of the speech-reception threshold (SRT) for sentences in noise after direct reduction of spectral contrast (data from 10 normal-hearing subjects) with the previously obtained SRTs after temporal lowpass and highpass filtering revealed that the SRT-effect of highpass filtering can be attributed completely to the associated spectral reduction. However, for temporal lowpass filtering, the associated spectral modulation reduction appears to be only a secondary effect.

Paper submitted for publication in: J. Acoust. Soc. Am.

6.1 INTRODUCTION

Recent studies about the effects of spectral and temporal envelope filtering on speech reception have tried to assess the limits up to which modulations are needed for speech intelligbility in noise. Experiments by ter Keurs *et al.* (1992, 1993a) on spectral smearing (lowpass filtering of the spectral envelope) revealed that the speech-reception threshold (SRT) for sentences in noise does not increase for smearing bandwidths below ⅓ oct. In terms of ripple densities of the short-term spectral envelope this means that transfer of the lower spectral ripples up to about 1.5 periods/oct is sufficient. In chapters 2 and 3 we studied the effect of reducing specific temporal modulation frequencies by either lowpass or highpass filtering the temporal envelopes of various frequency bands. Limits for reducing low- and high-frequency modulations were found to be 4 Hz and 16 Hz, respectively, independent of the number of frequency channels.

In the above studies the temporal and spectral modulation domains have always been treated separately. However, manipulations in the temporal domain will have an effect in the spectral domain (and vice versa). For example, as indicated by Plomp (1988), reduction of temporal modulations as a result of multichannel amplitude compression may cause a loss of structure (intensity contrasts between formants) in the short-term spectral shapes. The question is how much and in which way spectral modulations are affected by modifications of the temporal envelopes of a series of frequency bands.

In this chapter we address the above question. First, a general method is proposed to quantify the reduction of spectral modulations. This method is applied on temporally processed speech, in order to find the amount of spectral modulation reduction associated with a given amount of temporal modulation reduction. Futhermore, by measuring the intelligibility of spectrally reduced speech, we wanted to find out whether the results of previous experiments on multichannel temporal-envelope filtering (chapters 2 and 3) could be explained by a mere reduction of spectral contrasts.

6.2 SPECTRAL MODULATION TRANSFER FUNCTION

In the concept of the modulation transfer function (MTF, Houtgast and Steeneken, 1985), temporal modulation reduction is defined as the extent to which the temporal intensity modulations in the input are preserved in the output. The amount of reduction is expressed in a factor (m) as a function of the temporal modulation frequency. Analogously, one can define spectral

modulation reduction as the extent to which spectral modulations are preserved in the short-term spectral envelope. To quantify the degree of reduction in the spectral domain, we need a method to estimate the value of the spectral modulation-reduction factor for a given input (original) and output (processed) speech signal. In terms of system analysis this amounts to finding the spectral modulation transfer function (SMTF), i.e., the reduction of spectral modulations for a range of spectral modulation frequencies. For the sake of clarity, the temporal and spectral modulation-reduction factors will be denoted as m_t and m_s, respectively.

Modulations in the temporal envelope of a particular frequency band are examined by considering how they are composed of elementary sinusoidal components, viz. the modulation frequencies (in Hz). In a similar way the spectral contrasts in a short-term speech spectrum can be regarded to consist of several sinusoidal modulations (ripples), so that spectral modulations can be described in terms of ripple densities. The ripple density is defined as the number of sinusiodal periods per octave in the short-term spectral envelope (van Veen and Houtgast, 1985). Thus, the short-term ripple spectrum can be defined as the Fourier transform of the short-term spectral envelope on a log-frequency scale.

The first step in computing the SMTF consists of the estimation of the short-term spectral envelope for each of a series of original and processed speech frames. This is done by subjecting the power spectrum to a cepstral analysis, including 'liftering' and adaptive level adjustment to run smoothly over the spectral peaks (ter Keurs *et al.*, 1992). The power spectrum is obtained from a fast Fourier transform (FFT) of non-overlapping Hamming-windowed segments of 16.4 ms duration. The spectral intensity envelope will on average yield relatively small values for the higher frequencies, since most energy is found in the low-frequency regions. To compensate for this spectral slope, each short-term spectral envelope is divided by the long-term intensity envelope. This method normalizes for the 'base-line' of modulations present in the long-term spectral envelope.[1] The resulting envelope may be regarded as the short-term deviation from the average. These differences in frame-to-frame deviation contain the relevant information for the identification of speech signals. As an example, Fig. 6.1(a) displays the spectral envelopes of an arbitrary original and processed speech frame, together with the long-term average speech envelope. Figure 6.1(b) shows the variations in spectral modulations (which are reduced in the processed signal) relative to the long-term average. The Fourier transform of the normalized intensity envelopes on a log-frequency scale yields normalized ripple spectra, which are used for the computation of the SMTF.

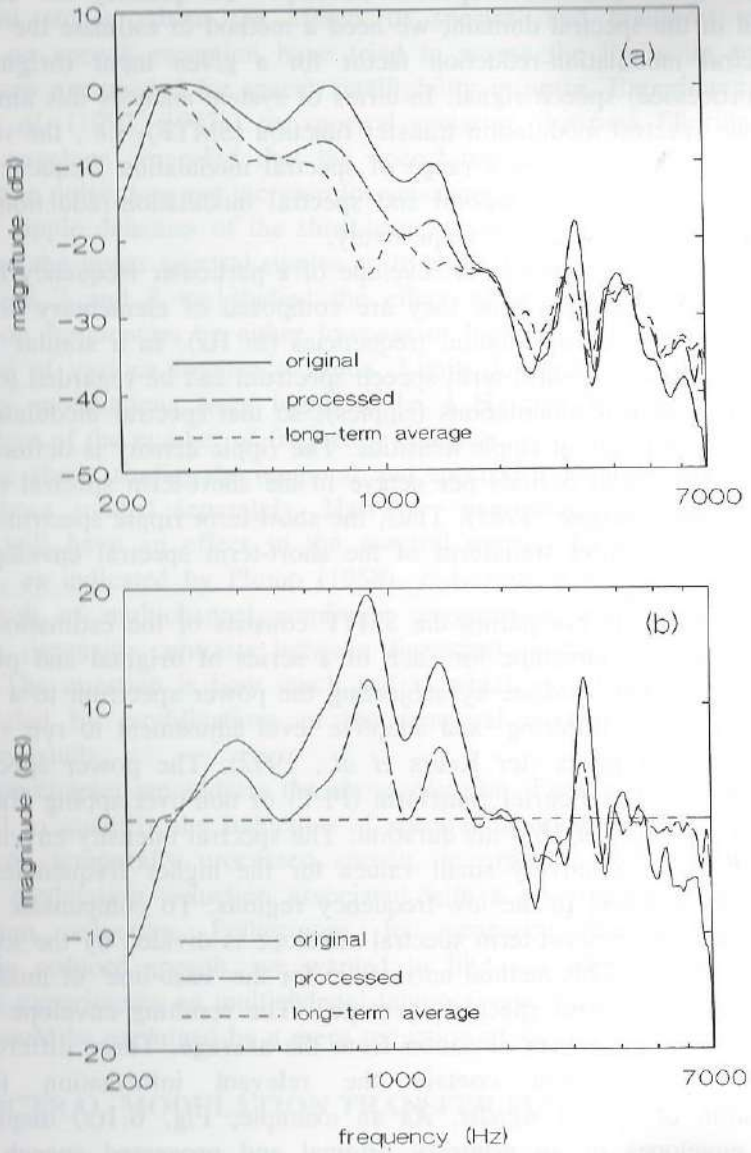

FIG. 6.1. Long-term average spectral envelope and short-time spectral envelopes of an arbitrary original and processed speech frame [panel (a)]. Panel (b) shows variations in spectral ripples relative to the long-term average.

The procedure for determining m_s for a range of ripple densities has been derived from the phase-locked approach for temporal modulations (see chapter 3 and Appendix C). That is, both intensity and phase of the ripples in the original and modified spectral envelopes are taken into account. The normalized ripple spectra were computed over a 4-oct range from 400 to 6400 Hz, for ripple densities up to 4 periods/oct, with ¼ periods/oct resolution.[2] The calculation of the SMTF was done on the average phase-locked ripple transfer of a few hundred corresponding short-term ripple spectra of the original and modified spectral envelopes.

6.3 SMTF AFTER UNIFORM TEMPORAL MODULATION REDUCTION

In order to asses the effect of temporal modulation reduction on the spectral modulations we will first focus on the base-line condition in which all temporal modulations are reduced equally. This means that a reduction is applied to the entire range of temporal modulation frequencies (uniform reduction), making it possible to express m_s as a function of an imposed m_t. One method to perform uniform reduction is by adding noise. However, noise reduces the modulation depth by filling the troughs of the temporal envelope. But on top of that, noise introduces additional non-relevant (disturbing) modulations. Therefore, we choose a method to reduce the fluctuations in the temporal envelope by proportionally raising the troughs and lowering the peaks.

All speech processing was done on an Olivetti M290S computer with an OROS-AU21 card with TMS320C25 signal processor. For the uniform reduction of temporal modulations, the wideband speech signal is split up into a 24 ¼-oct bands (linear-phase FIR filter bank, slopes of at least 80 dB/oct), covering the range 100-6400 Hz. The temporal intensity envelope of each band is obtained by squaring the Hilbert (amplitude) envelope. Subsequently, the fluctuations in the original intensity envelope (I_{org}) are reduced to get the modified (I_{mod}) envelope. The degree of reduction is controlled by the modulation-reduction factor m, which can take any value between 0 (all modulations suppressed) and 1 (all modulations intact). The relation beween I_{mod} and I_{org} at time t is given by

$$I_{mod}(t) = m \cdot I_{org}(t) + (1-m) \cdot \bar{I} \,, \qquad 0 \le m \le 1 \,, \qquad (1)$$

where $\bar{I}$ stands for the long-term average intensity of that band. Each sample of the original band signal is multiplied by the square root of the ratio of the

reduced and original intensity envelope, after which all modified bands are added. Note that the fine structure (carrier) in each channel is left untouched.

To estimate the reduction of spectral contrasts after the imposed uniform reduction of temporal modulations, the SMTF was measured for various values of m_t. For this purpose, we used a 71-s speech fragment (female talker), digitized at a sampling rate of 15 625 Hz and 16 bits resolution. This fragment typically consisted of 30 concatenated sentences, a subset of the material used in the intelligibility tests described in section 6.5.

The speech signal was temporally reduced, using m_t=0.1, 0.2, 0.4, 0.6, and 0.8 in each ¼-oct band. For each m_t, the spectral analysis yields values of m_s for a range of ripple densities. The results of the analysis are plotted in Fig. 6.2. As can be seen in the figure, m_s is virtually independent of the ripple density, and comes on average close to the value of m_t. A convenient single index for the amount of spectral modulation reduction can be obtained by taking the average for the lower range of ripple densities. In accordance with the perceptual significance of spectral modulations, we adopted an upper limit of 2 periods/oct (van Veen and Houtgast, 1985; ter Keurs *et al.*, 1993a). This mean modulation-reduction factor will be denoted by $\bar{m}_s$. Figure 6.3 shows $\bar{m}_s$ as a function of m_t. The data points can be fitted well by a straight line, which almost equals the main diagonal.

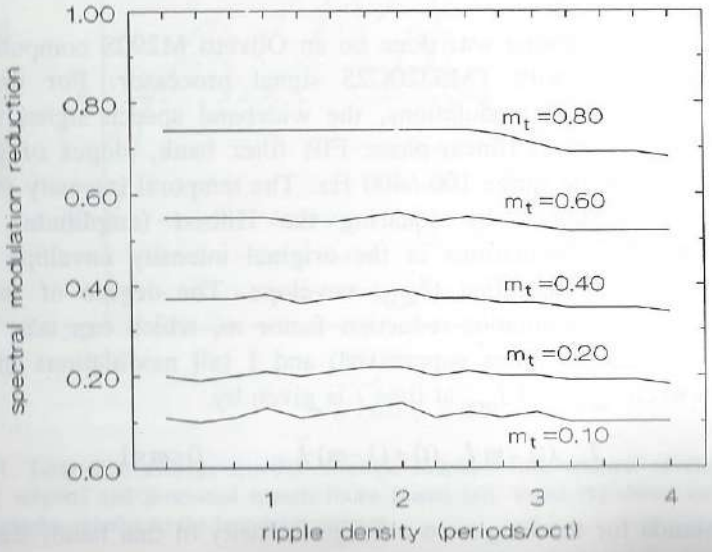

FIG. 6.2. SMTF resulting from imposing various degrees of uniform temporal modulation reduction (m_t).

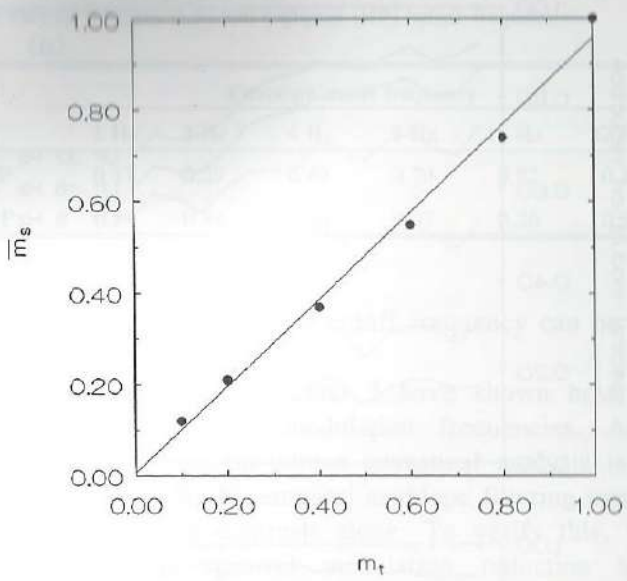

FIG. 6.3. Mean spectral modulation reduction as a function of the imposed uniform temporal modulation reduction.

6.4 SPECTRAL EFFECTS OF TEMPORAL-ENVELOPE FILTERING

The previous section pointed out a clear relation between temporal and spectral modulation reduction in the sense that uniform reduction in the temporal domain exhibits an essentially equal degree of uniform reduction in the spectral domain. A next question is what happens to the spectral contrasts if *specific* temporal modulation frequencies are reduced. This question is motivated by earlier studies (chapters 2 and 3) in which narrowband temporal envelopes were lowpass and highpass filtered. In this section we will examine the SMTF for various temporal envelope lowpass (LP) and highpass (HP) cutoff frequencies, processed in ¼-oct bands. Details about the temporal filtering methods can be found in the aforementioned chapters.

The same 71-s speech fragment was used for the temporal envelope filtering and subsequent spectral analysis. LP and HP filtering were performed on the amplitude envelope of each ¼-oct band, using cutoff frequencies of 1, 2, 4, 8, 16, and 32 Hz. Figure 6.4 shows the SMTF after LP [panel (a)] and HP [panel (b)] processing. The spectral contrasts appear to be reduced along with the reduction of temporal modulations, i.e., by a

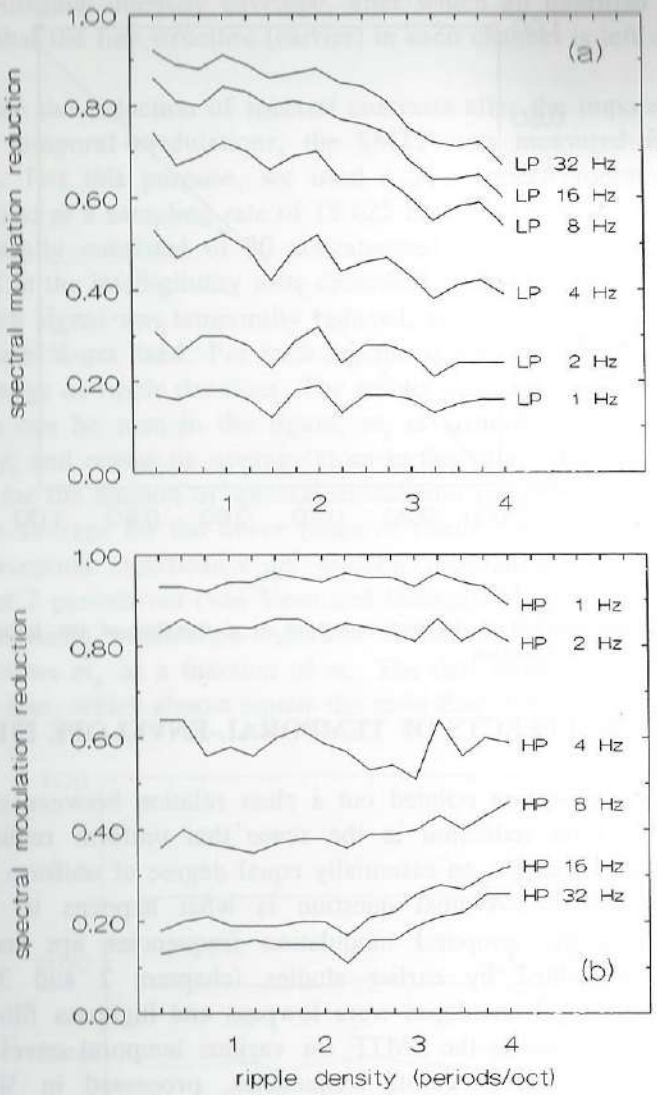

FIG. 6.4. SMTF resulting from imposing various degrees of lowpass [LP, panel (a)] and highpass [HP, panel (b)] temporal-envelope filtering. Parameters are the LP and HP envelope filter cutoff frequencies.

decrease of the LP cutoff frequency or an increase of the HP cutoff frequency. Spectral modulation reduction is virtually uniform for spectral densities up to about 2 periods/oct. The degree of spectral modulation

TABLE 6.1. Values of $\bar{m}_s$ (mean for 0.25-2 periods/oct) after temporal-envelope filtering for various lowpass (LP) and highpass (HP) cutoff frequencies.

	Envelope cutoff frequency					
	1 Hz	2 Hz	4 Hz	8 Hz	16 Hz	32 Hz
LP	0.17	0.29	0.49	0.70	0.82	0.89
HP	0.94	0.84	0.60	0.37	0.20	0.15

reduction as a function of LP and HP cutoff frequency can be expressed by the value of $\bar{m}_s$, as listed in Table 6.1.

The experiments in chapters 2 and 3 have shown how intelligibility suffers from reducing temporal modulation frequencies. An interesting question which follows from the above acoustical analysis is whether the reduction of intelligibility due to temporal envelope filtering can be explained by the reduction of spectral contrasts alone. To verify this, we measured how various degrees of spectral modulation reduction affect speech intelligibility.

6.5 PERCEPTUAL EVALUATION

6.5.1 Method

The major part of the procedure for reducing spectral modulations is based on an algorithm with short-term FFTs and overlapping additions (OLA) used by ter Keurs *et al.* (1992, 1993a). In fact, the only modification consists of reducing all spectral modulations equally instead of lowpass filtering (smearing) the spectral envelope. Since details of the FFT-OLA processing have been described in ter Keurs *et al.* (1992), we will restrict ourselves to the implementation of the spectral envelope reduction.

For each of a set of overlapping speech frames (16.4 ms duration, 4.1 ms shift), the short-term power spectrum is subjected to a cepstral analysis, resulting in the short-term spectral intensity envelope, $S_{org}(f)$. The modified spectral envelope, $S_{mod}(f)$, is obtained by reducing the modulations of $S_{org}(f)$ relative to $\bar{S}(f)$, the long-term average envelope adjusted to the over-all intensity of $S_{org}(f)$. That is,

$$S_{mod}(f) = m \cdot S_{org}(f) + (1-m) \cdot \bar{S}(f) \ , \qquad 0 \le m \le 1. \qquad (2)$$

The modified envelope is imposed on the original complex spectrum. The spectrum's original phase, harmonic structure, and total energy remain unchanged; only the spectral shape is modified. After inverse transformation into the time domain, the overlapping frames are added to reconstruct a continuous speech signal. Thus, in the extreme case of $m=0$, the output speech has a fixed spectral envelope-shape for the entire duration; phase, pitch, and local wideband energy are preserved.

6.5.2 Material, design

The speech material consisted of ten lists of 13 everyday Dutch sentences of eight to nine syllables read by a trained female speaker (Plomp and Mimpen, 1979). For the SRT measurements to be described below, a masking noise with the same spectrum as the long-term average of the 130 sentences was used. Both speech and noise were digitized at a sampling rate of 15 625 Hz with 16 bits resolution.

Ten conditions for spectral modulation reduction were investigated, with reduction factors running from 0.1 to 1.0 in steps of 0.1. The 1.0 (control) condition was obtained by analysis and resynthesis by the FFT-OLA system without applying any spectral reduction. A preliminary test had shown that for the 0.1 condition an intelligibility score in quiet of almost 90% was reached, so that a reliable SRT could be measured for all conditions. To ensure a constant (long-term) speech-to-noise ratio over the entire spectral range, the masking noise was processed separately for each condition in the same way as the sentences.

6.5.3 Subjects

Subjects were 10 normal-hearing students of the Free University, whose ages ranged from 18 to 27. All had pure-tone air-conduction thresholds less than 15 dB HL in their preferred ear at octave frequencies from 125 to 4000 Hz and at 6000 Hz.

6.5.4 Procedure

The ten sentence lists were presented in a fixed order. The sequence of the condidions was varied according to a digram-balanced 10×10 Latin

square, to avoid order an list effects. Details about the experimental procedure for the SRT test can be found in chapters 2 and 3.

The stimuli were presented monaurally through a headset (Sony MDR-CD999) to the subject's ear of preference in a soundproof room. Before the actual test, a list of 13 sentences pronounced by a male speaker in the 0.5 condition was presented to familiarize the subject with the procedure.

6.5.5 Results

Figure 6.5 shows the mean SRT for sentences in noise as a function of the spectral modulation reduction. A one-way analysis of variance for repeated measures revealed that the effect of reduction is highly significant ($p < 0.001$). *Post hoc* tests (Tukey HSD) showed that SRTs are significantly higher ($p < 0.01$) for reduction factors from 0.4 downward than for the unreduced (1.0) condition. The results demonstrate that up to 50% reduction of the original spectral contrasts has no effect on sentence intelligibility in a critical speech-to-noise situation.

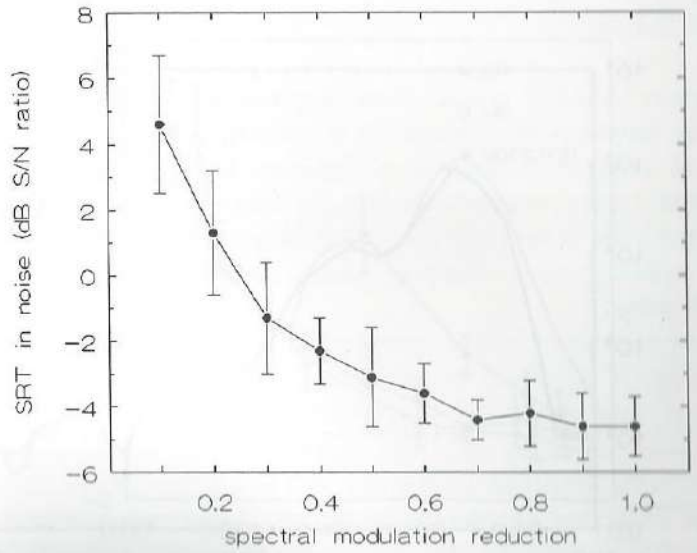

FIG. 6.5. Mean SRT for sentences in noise as a function of the imposed (uniform) spectral modulation reduction. Vertical bars represent the standard deviation for each condition.

6.6 DISCUSSION

The results of Section 6.3 (Fig. 6.3) show that uniform reduction of temporal modulations leads to uniform reduction of spectral modulations of (almost) equal magnitude. To recognize this finding, the nature of temporal and spectral modulations and the (physical) effect of reducing them may be thought of as follows. Temporal modulations are essentially defined as local deviations from the mean intensity per channel, measured over a relatively long speech fragment. The ensemble of mean intensities can be considered as the long-term average spectral envelope. Spectral modulations have been defined as deviations from this average in the short-term spectral envelope. Reducing temporal modulations in each channel to a certain degree means coming closer to the average. This holds for each moment in time. In other words, each short-time spectral envelope comes closer to the average spectral envelope. Temporal modulation reduction is thus basically similar to reduction towards the average bandfilter spectrum, which in our case (24 ¼-oct bands) lacks some detail in comparison to the average spectral envelope based on a Fourier analysis used for the SMTF. The slight difference

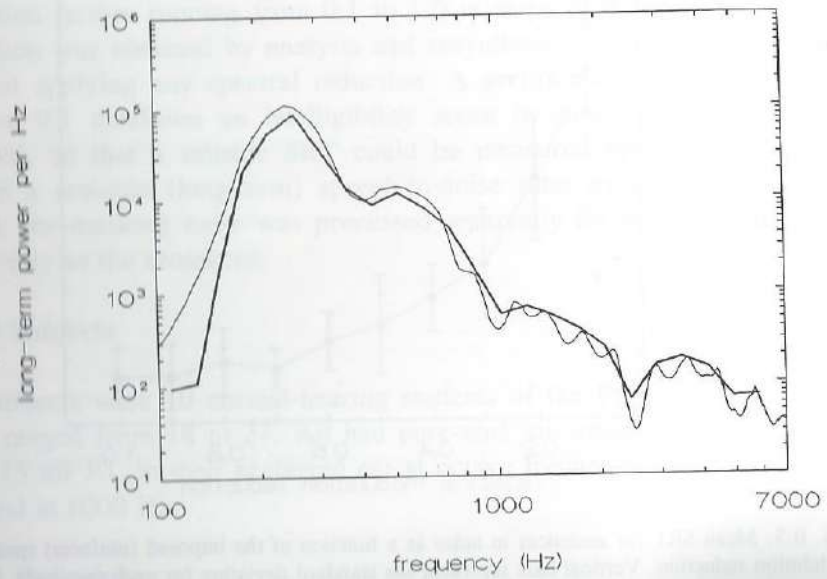

FIG. 6.6. Long-term average spectral envelopes based on the FFT power spectrum (normal line) and on the mean intensities of 24 ¼-oct bands (heavy line). For comparison both curves are expressed in power density.

between these two envelopes is most probably the reason why the data points in Fig. 6.3 do not lie exactly on the main diagonal. As an illustration, Fig. 6.6 displays the long-term average spectral envelopes, based on the FFT power spectrum (normal line) and the bandfilter output mean intensities (heavy line). Both curves are expressed in power density. The above implies that uniform reduction of temporal modulations is associated with an equal reduction of the spectral modulations, if a sufficiently large number of frequency bands is used.

The primary goal of the perceptual evaluation was to find out to what extent the previously found SRT-effect of lowpass and higpass temporal-envelope filtering may also be interpreted as reflecting the effect of the associated reduction in spectral contrasts. For this purpose we need to combine the data in Table 6.1 and Fig. 6.5. The relative SRTs (i.e., SRT relative to the unprocessed condition in the respective experiments) for the LP and HP cutoff frequencies after conversion to the corresponding spectral-modulation factors (Table 6.1) are plotted in Fig. 6.7, together with the data of the present experiment. Figure 6.7 clearly shows that the two curves for

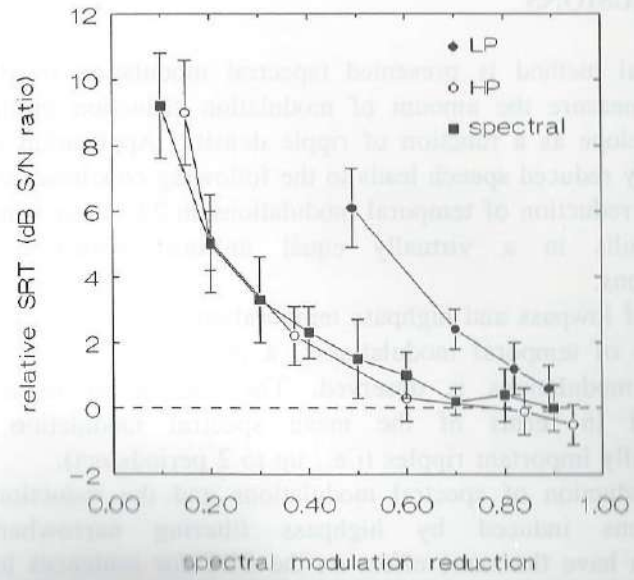

FIG. 6.7. Mean relative SRT and 95%-confidence limits as a function of spectral modulation reduction associated with lowpass (LP) and highpass (HP) temporal-envelope filtering, and for direct reduction of spectral contrasts (filled squares, cf. Fig. 6.5).

HP filtering and direct spectral reduction are very similar. In other words, the reduction of intelligibility due to temporal highpass filtering can be equally well interpreted in terms of its spectral effect. The curve for LP, however, rapidly diverges from the other two, except for high spectral modulation-reduction factors (>0.80). The detrimental effect of LP on the SRT is much larger than can be explained solely on the basis of the associated spectral modulation reduction. In the extreme conditions of 1- and 2-Hz LP cutoff frequency, intelligibility has dropped to such a degree that the SRT in noise cannot be measured anymore. Intelligibility scores in quiet are only 12% and 22%, respectively (cf. chapter 2, Fig. 2.3). In short, it looks like spectral modulation reduction is only a secondary effect in the case of temporal lowpass filtering.

The present study illustrates that manipulations of the (multichannel) *temporal* envelopes is associated with large effects on the (short-term) *spectral* envelopes. The reverse (not considerd here) is equally true. Therefore, in interpreting the subjective effects of such manipulations in one domain, one should be well aware of the associated effects in the other domain.

6.7 CONCLUSIONS

A general method is presented (spectral modulation transfer function, SMTF) to measure the amount of modulation reduction in the short-term spectral envelope as a function of ripple density. Application of the SMTF on temporally reduced speech leads to the following conclusions:

(1) Uniform reduction of temporal modulations in 24 ¼-oct bands (100-6400 Hz) results in a virtually equal uniform reduction of spectral modulations.

(2) In case of lowpass and highpass temporal-envelope filtering (non-uniform reduction of temporal modulations), a reasonably uniform reduction of spectral modulations is observed. The amount of reduction can be expressed in terms of the mean spectral modulation transfer of perceptually important ripples (i.e., up to 2 periods/oct).

(3) Direct reduction of spectral modulations and the reduction of spectral modulations induced by highpass filtering narrowband temporal envelopes have the same effect on the SRT for sentences in noise. This implies that a loss of intelligbility due to the reduction of low temporal modulation frequencies can be explained by the associated reduction of spectral modulations alone. This relationship is not demonstrated for lowpass filtering of the temporal envelope.

Notes

[1] A similar normalization is also applied in the calculation of the temporal modulation transfer function. That is, for each frequency band the reduction of temporal modulation frequencies is normalized with respect to the mean intensities of the temporal envelopes of both original and processed signals (Houtgast and Steeneken, 1985). For the ensemble of (narrow) frequency bands, this boils down to expressing the amount of temporal modulation reduction relative to the average intensity spectrum.

[2] The lower frequency of 400 Hz was chosen because we used a female voice in the applications of the SMTF throughout this paper. With a pitch varying around 200 Hz, the octave below 400 Hz would at times contain only one harmonic, yielding a less accurate estimate of the spectral envelope.

CHAPTER 7

Concluding remarks

In this study we investigated the significance of narrowband temporal speech envelopes for intelligibility. In chapter 1 we noted that a faithful transfer of temporal modulations is needed to ensure good intelligibility. The main outcome of this study is that we have quantified what 'faithful' means in this context. It is shown that narrowband temporal modulations can be reduced considerably, both in terms of modulation frequencies and of peak-to-trough difference, before a detrimental effect on intelligibility is noticeable. Rather substantial modifications in the temporal (and hence spectral) distribution of speech energy appear to be possible without losing essential information. This is illustrated by the finding that the two disjunct modulation-frequency regions above and below 8-10 Hz each contain sufficient information to yield 100% intelligibility in quiet. The redundancy of speech, for which this is an example in the modulation-frequency domain, implies that the selectivity of the normal ear is primarily attuned to understanding speech in difficult listening conditions. The more redundant cues are obscured or removed, the less speech will become resistant to external disturbances such as noise.

A question that remains is whether the limits for temporal modulations found for normal hearing apply to impaired hearing as well. One may assume that the hearing-impaired listener needs more of the speech information that is redundant for the normal-hearing listener. On the other hand, removing cues that are hardly or not perceived by the hearing impaired is likely to have no effect on their performance. Subjects with moderate sensorineural hearing losses have difficulties following higher temporal and spectral modulations (Plomp, 1984). Recent experiments have shown that their limited resolution of spectral contrasts is hardly related to their reduced speech intelligibility in noise (ter Keurs *et al.*, 1993b).

In both the temporal and the spectral domain, signal processing for the hearing-impaired has to concentrate on those features that are particularly important for intelligibility. As far as temporal modulations are concerned, this means, for instance, modulation frequencies below 16 Hz and fluctuations in the envelope peaks. The conclusion from chapter 3 that

modulations below 4 Hz do not contribute to intelligibility, even in noise, suggests that dynamic gain control as applied in hearing aids should be slower than 4 Hz in order to prevent a reduction of intelligibility. This is in agreement with the recommendation that the time constant for readjusting the amplitude-frequency response should preferably not be much shorter than 250 ms (van Dijkhuizen *et al.*, 1989; Plomp, 1994).

The results of chapter 4 impose certain restrictions on the general use of the MTF to predict speech intelligibility. For temporal smearing a sufficiently adequate prediction is found (chapter 2). However, other manipulations of the temporal envelope which lead to a reduction of the average modulation depth do not always yield a reduction of intelligibility. This is not only the case for speech presented in quiet as examined in chapter 4. The average MTF of the highpass filtered envelopes with 4-Hz cutoff frequency is substantially reduced compared to unprocessed speech (chapter 3, Fig. 3.8), whereas sentence intelligibility in noise is equal for both conditions. As mentioned before, one has to keep in mind that short everyday sentences, though appropriate to represent conversational speech, are highly redundant. As a consequence, their application often fails to reveal a difference in intelligibility, despite a reduction of the MTF (Steeneken, 1992). For example, at a speech-to-noise ratio of 0 dB, corresponding to an (average) MTF of 0.5, normal-hearing listeners reach an intelligibility of 100%. However, this only indicates that tests with redundant sentences are a poor means to explore MTF differences and does not derogate from the heart of the matter that conditions are found for which a one-to-one relation between MTF and intelligibility is lacking. We have demonstrated that the same average MTF resulting either from direct manipulation of narrowband temporal envelopes or from additive noise can lead to substantially different intelligibility scores.

In this respect the origin of modulation reduction is an important factor, and one should be careful in converting it into an apparent signal-to-noise ratio. It would be worthwhile to investigate whether a difference in intelligibility exists between the more deterministic approach of uniform temporal modulation reduction (as in chapter 6) and the use of additive noise. Anticipating that the former method has a smaller effect on intelligibility, due to the absence of spurious modulations introduced by the noise, the effect of masking speech modulations by other modulations on intelligibility should be further examined.

Finally, the relation between modulations in the temporal and spectral domain (chapter 6) requires more attention. This may be regarded as a plea for a two-dimensional approach of signal processing algorithms for speech enhancement. We have to focus on the continuity and mutual relations of the energy distribution over both time and frequency (possibly after separate one-dimensional temporal and/or spectral preprocessing). For example, in order to decide whether some local peak in the temporal envelope of a particular channel originates from either speech or noise, the coherence between adjacent channels has to be considered as well. Finding the spectro-temporal distribution of speech elements which yields a maximal transfer of information may lead to more advanced speech-processing schemes.

Summary

Speech can be described as a summation of a number of frequency channels with amplitude-modulated signals. Each channel consists of a fine structure (carrier signal) and a time-varying envelope. The variations in amplitude, as represented by this temporal envelope, contain the information that is essential for the identification of phonemes, syllables, words, and sentences. For disturbances that occur in practice, e.g., noise and reverberation, the deleterious effects on intelligibility appear to be due to a reduction of temporal modulation depth, as shown by studies on the modulation transfer function (MTF) and speech transmission index (STI). The aim of the present study is to systematically investigate how crucial temporal modulations are for intelligibility. In other words, what is the effect of reducing the degree to which temporal modulations are present in the speech signal?

For the signal processing, a basic analysis-resynthesis algorithm was developed, in which the temporal (Hilbert) envelopes of a number of consecutive frequency bands were modified. In this way the envelope could be manipulated without affecting the fine structure. The effects of these envelope modifications on sentence intelligibility were assessed for normal-hearing listeners, either by measuring the score in quiet or by measuring the speech-reception threshold (SRT) in speech-shaped noise.

After a general introduction in chapter 1, chapters 2 and 3 address the question of which frequencies in the temporal speech envelope are important to intelligibility. Chapter 2 describes a series of experiments in which the temporal envelope was lowpass filtered (smeared), resulting in a reduction of the higher modulation frequencies. The effect of smearing on sentence intelligibility was measured as a function of both lowpass cutoff frequency (0 to 64 Hz) and bandwidth of the channels (¼, ½, or 1 oct, covering the range 100-6400 Hz) on which envelope filtering took place. Results for 36 subjects show that, even in quiet, intelligibility is severely affected for very low cutoff frequencies (0-2 Hz), especially in narrow frequency bands. For cutoff frequencies of 4 Hz and above, the SRT in noise was measured, revealing no effect of processing bandwidth, and a progressive decrease of intelligibility up to a cutoff frequency of 16 Hz. The latter implies that, when lower modulation frequencies are present, modulations above 16 Hz only marginally contribute to the intelligibility of everyday sentences. The measured SRTs for cutoff frequencies above 4 Hz appear to correspond well to those predicted by the STI. For cutoff frequencies of 0-2 Hz the computed STI is rather low and does not account for the observed dependence of processing bandwidth.

In continuation of the smearing experiments (i.e., reducing the higher modulation frequencies), the effect of reducing the lower modulation frequencies was studied in a similar way in chapter 3. For this purpose, the temporal envelopes were highpass filtered, using cutoff frequencies between 1 and 128 Hz. The SRT for sentences in noise could be measured for highpass cutoff frequencies up to 32 Hz; for higher cutoff frequencies, sentence intelligibility in quiet was determined. Results for 42 subjects show that modulation frequencies below 4 Hz can be reduced without having a detrimental effect on intelligbility. A further increase of the cutoff frequency yields a gradual reduction of intelligibility, depending on the processing bandwidth (¼-, ½-, or 1-oct) for cutoff frequencies of 16 Hz and above. For a number of experimental conditions, the amount of modulation reduction was measured by means the MTF and compared to data on multichannel amplitude compression. Results suggest that compression leads to an insignificant loss of intelligibility (about 1-dB increase of the masked SRT). Comparision of the results for low- and highpass envelope filtering indicates that the crossover modulation frequency, which divides the temporal modulation spectrum into two perceptually equal parts, is about 8-10 Hz for all three processing bandwidths. This means that, in a critical speech-to-noise condition, reduction of temporal modulations either below of above 8-10 Hz results in the same (reduced) sentence intelligibility.

In chapters 2 and 3 the effect of envelope filtering on individual phonemes was also examined. Recognition scores in quiet for vowels and consonants in nonsense CVC and VCV syllables were obtained for various lowpass and highpass cutoff frequencies, using ¼-oct-band processing. As expected, recognition decreases for all phonemes if the lowpass cutoff frequency is lowered or if the highpass cutoff frequency is raised. On average, consonants suffer more from envelope filtering than vowels. Errors in vowel identification are mainly characterized by reduced recognition of diphthongs and by confusions between long and short vowels. The consonants can be divided into three categories, viz. stops, fricatives, and vowel-likes. The majority of the erroneous consonant identifications are found to be confusions within a category, i.e., information of manner of articulation is more or less preserved. As an exception, smeared stops show fricative and vowel-like responses, typically depending on the position in the syllable. Confusions that may be expected to follow from smearing (e.g., /p/-/f/, /t/-/s/, /k/-/χ/) are primarily found in final position.

The results of the above experiments suggest that there must be a minimum amount of variation in the temporal envelope, governed by certain modulation frequencies, to ensure sufficient modulation depth, i.e., the difference between the envelope peaks and troughs. Other possible ways to

reduce the modulation depth include compression and the addition of noise. Roughly, the former lowers the peaks of the temporal envelope, whereas the latter raises the troughs. However, adding noise to the speech signal also changes the fine structure and masks the weaker speech elements. Chapter 4 reports a number of experiments to investigate the relative contributions of temporal envelope and fine structure to intelligibility. For this purpose either the envelope or the fine structure for each of 24 ¼-oct bands was manipulated. The importance of envelope cues was evaluated by preserving the speech fine structure, while the envelope was modified using peak clipping, trough clipping, and infinite clipping (simple block pulse approximation). The importance of fine structure cues was evaluated using noise fine structure with intact speech envelopes and intact speech fine structure with noisy envelopes. In addition, ordinary speech+noise stimuli served as a reference. Comparison of sentence intelligibility scores in quiet for a total of 60 subjects shows that (1) fine structure cues play a much less important role than envelope cues; (2) envelope peaks are perceptually more important than troughs; (3) reduction of modulations caused by the addition of noise produces lower intelligibility scores than the same degree of reduction brought about by direct manipulation of the envelope (difference in terms of the SRT is almost 7 dB). The latter result is inconsistent with the MTF concept. That is, an MTF-based model of speech recognition would predict that equal reduction of modulation depth would have the same effect on intelligibility. It appears that noise introduces spurious modulations, disturbing the perception of the relevant speech modulations. In general, for the types of envelope processing used, no one-to-one relation between the MTF and the intelligibility scores could be established.

In chapter 5 we take a closer look at the detrimental effect of noise on intelligibility in terms of the relative contribution of speech parts above (peaks) and below (troughs) the noise level. For each of 24 ¼-oct bands, the distribution of speech and noise over sentences was manipulated by presenting speech in the peaks and either noise only or speech+noise in the troughs. Sentence intelligbility was measured with 12 subjects for various speech-to-noise ratios. The results indicate that, compared to ordinary speech+noise stimuli (data from chapter 4), removing noise from the peaks has no effect on intelligibility. Removing the speech signal from the noisy troughs, however, yields a 2-dB increase of the SRT. So, it appears that speech elements below the noise level partly contribute to intelligibility.

All of the experimental results in chapters 2-4 have been discussed in terms of (a reduction of) temporal modulations. But manipulations performed on a series of narrowband temporal envelopes will have an effect in the spectral domain as well. Chapter 6 addresses the question of how much and

in which way spectral contrasts are affected by the reduction of temporal modulations. As a counterpart of the MTF for measuring the transfer of temporal modulations, a spectral modulation transfer function (SMTF) is presented to determine the extent to which modulations (ripples per octave) in the short-term spectral envelope are preserved. Uniform reduction (i.e., independent of modulation frequency) of the temporal modulations in ¼-oct bands appears to result in a virtually equal degree of reduction of the spectral modulations. Measurements of the SMTF after temporal-envelope filtering (chapters 2 and 3) were carried out for a number of lowpass and highpass cutoff frequencies (1-32 Hz). When the effect of this type of temporal processing is expressed in terms of the associated reduction of spectral contrasts, an interesting question is whether spectral reduction alone can explain the deleterious effect on intelligibility. Sentences were processed with an analysis-resynthesis algorithm consisting of short-term fast Fourier transforms, 10%-90% contrast reduction in the spectral envelope, inverse transforms, and overlapping additions to reconstruct a continuous signal. The SRT in noise for the spectrally reduced sentences was measured with 10 subjects. Comparison of these SRTs with those obtained in chapters 2 and 3 reveals that the loss of intelligbility due to temporal highpass filtering can be attributed to the associated spectral modulation reduction. This result is not found for temporal lowpass filtering, suggesting that in that case spectral reduction is only a secondary effect.

Samenvatting

Spraak kan worden opgevat als een verzameling opeenvolgende (smalle) frequentiebanden met amplitude-gemoduleerde signalen. Iedere frequentieband bestaat uit een fijnstructuur (draaggolf) en een in de tijd variërende omhullende. Deze temporele omhullende geeft dus voor een beperkt frequentiegebied de variaties in geluidsterkte weer. Daarin ligt de informatie besloten die essentieel is voor het identificeren van fonemen, lettergrepen, woorden en zinnen. Voor in de praktijk voorkomende verstoringen van de spraak, zoals ruis en nagalm, blijken de nadelige effecten op het verstaan een gevolg te zijn van een vermindering van de temporele modulatiediepte. De mate waarin de temporele omhullende van het spraaksignaal behouden blijft is uitvoerig bestudeerd aan de hand van de modulatie-overdrachtsfunctie (MTF). Deze succesvolle benadering heeft geleid tot de ontwikkeling van een fysische maat voor de kwaliteit van spraakoverdracht, de spraak-transmissie index (STI). Het doel van dit onderzoek is op een systematische manier na te gaan in hoeverre temporele modulaties bepalend zijn voor de verstaanbaarheid. Anders gezegd, wat is het effect van vermindering van de temporele modulaties in het spraak-signaal? Het onderzoek is primair fundamenteel van aard, maar heeft ook een meer toegepaste kant. Kennis over de mate waarin de perceptie van spraak bestand is tegen degradatie levert gegevens op die bijvoorbeeld kunnen worden toegepast in signaal-bewerkingen t.b.v. slechthorenden.

Voor de signaalbewerking is een algemeen analyse-resynthese algoritme ontwikkeld, waarmee de temporele (Hilbert) omhullende van een aantal opeenvolgende frequentiebanden kan worden gewijzigd. Op die manier is het mogelijk de omhullende te manipuleren zonder de fijnstructuur aan te tasten. Het effect van deze manipulaties op de verstaanbaarheid van eenvoudige zinnen is gemeten bij jonge normaalhorende luisteraars. Afhankelijk van de experimentele conditie werd het percentage correct verstane zinnen in stilte bepaald of werd de zogenaamde spraakverstaanvaardigheidsdrempel (SRT) tegen een achtergrond van ruis gemeten. De SRT is gedefinieerd als die spraak-ruisverhouding in dB waarbij nog 50% van de zinnen foutloos kan worden gereproduceerd. De ruis heeft daarbij hetzelfde gemiddelde frequentiespectrum als de zinnen.

Na een algemene inleiding in hoofdstuk 1 richten de hoofdstukken 2 en 3 zich op de vraag welke omhullende-frequenties van belang zijn voor de verstaanbaarheid. Hoofdstuk 2 beschrijft een serie experimenten waarin de temporele omhullende laagdoorlaat wordt gefilterd (versmeerd), waardoor de hogere modulatiefrequenties worden gereduceerd. Het effect van versmering op de verstaanbaarheid werd gemeten als functie van zowel de afsnij-

frequentie van het laagdoorlaatfilter (0-64 Hz) als de breedte van de frequentiebanden (¼, ½, of 1 octaaf) waarop de bewerking plaatsvond. De resultaten bij 36 luisteraars tonen aan dat de verstaanbaarheid voor zeer lage afsnijfrequenties (0-2 Hz) al in stilte ernstig vermindert, vooral voor smalle frequentiebanden. Voor afsnijfrequenties van 4 Hz en hoger werd de SRT in ruis gemeten. Daaruit blijkt dat de verstaanbaarheid progressief toeneemt tot een afsnijfrequentie van 16 Hz, onafhankelijk van de bandbreedte. Met andere woorden, modulatiefrequenties boven 16 Hz leveren geen wezenlijke bijdrage aan de verstaanbaarheid van alledaagse zinnen (vooropgesteld dat lagere modulatiefrequenties wel in het signaal aanwezig zijn). De gemeten SRT-waarden voor afsnijfrequenties vanaf 4 Hz blijken goed overeen te komen met voorspellingen op basis van de STI. Voor lagere afsnijfrequenties is de berekende STI tamelijk laag en houdt geen rekening met de waargenomen afhankelijkheid van de bandbreedte.

Als vervolg op de temporele versmering wordt in hoofdstuk 3 het effect van reductie van lage modulatiefrequenties op eenzelfde wijze bestudeerd. Daartoe werd de temporele omhullende hoogdoorlaat gefilterd, met afsnijfrequenties tussen 1 en 128 Hz. De SRT voor zinnen in ruis kon worden gemeten voor afsnijfrequenties tot en met 32 Hz; daarboven werd de zinsverstaanbaarheid in stilte bepaald. De resultaten met 42 luisteraars geven aan dat de verstaanbaarheid voor afsnijfrequenties tot en met 4 Hz gelijk blijft aan die van onbewerkte spraak. Verhoging van de afsnijfrequentie geeft een geleidelijke afname van de verstaanbaarheid, welke vanaf 16 Hz afhankelijk is van de bandbreedte (¼, ½, of 1 octaaf). Voor een aantal experimentele condities werd de modulatiereductie gemeten aan de hand van de MTF en vergeleken met data over meerkanalige amplitude-compressie. De resultaten daarvan doen vermoeden dat compressie, althans bij normaalhorenden, een beperkte achteruitgang van de verstaanbaarheid geeft (ca. 1 dB verhoging van de SRT). Vergelijken we de resultaten van deze experimenten met die uit het voorgaande hoofdstuk, dan blijkt de 'cross-over' modulatiefrequentie—de frequentie die het temporele modulatiespectrum in twee perceptief even belangrijke stukken verdeelt—voor alle drie de bandbreedten bij 8 à 10 Hz te liggen. Dit betekent dat reductie van temporele modulaties òf beneden òf boven 8-10 Hz bij een kritische spraak-ruisverhouding tot dezelfde (verminderde) verstaanbaarheid leidt.

Naast de verstaanbaarheid van zinnen is in de hoofdstukken 2 en 3 ook gekeken naar het effect van filteren van de omhullende op de herkenning van individuele fonemen. Hiertoe werden identificatiescores in stilte voor klinkers en medeklinkers in nonsens CVC- en VCV-syllaben gemeten bij verschillende afsnijfrequenties van de laag- en hoogdoorlaatfilters (bewerking in ¼-octaaf banden). Als verwacht neemt de herkenning af naarmate de

afsnijfrequentie van laag- en hoogdoorlaatfilter lager resp. hoger is. Gemiddeld hebben medeklinkers meer te lijden dan klinkers. Fouten in de identificatie van klinkers worden voornamelijk veroorzaakt door slechtere herkenning van diftongen en door kort-lang/lang-kort verwarringen. De medeklinkers kunnen verdeeld worden in drie categorieën, nl. plosieven, fricatieven en klinkerachtigen. De meerderheid van de foutieve medeklinkerbenoemingen blijken verwarringen binnen een categorie te zijn, d.w.z. informatie over de wijze van articuleren blijft min of meer behouden. Uitzondering hierop vormen de versmeerde plosieven, die, afhankelijk van de positie in de syllabe, vaak als fricatief of klinkerachtig benoemd worden. Verwarringen die verwacht mogen worden als gevolg van versmering (b.v. /p/-/f/, /t/-/s/, /k/-/χ/) treden vooral op bij finale plosieven.

De resultaten van bovengenoemde experimenten geven aan dat er een minimum aan variatie in de temporele omhullende nodig is die zorgt voor voldoende modulatiediepte, d.w.z. voldoende verschil tussen pieken en dalen. Andere mogelijkheden om de modulatiediepte te reduceren zijn o.a. compressie en het toevoegen van ruis. Ruwweg kunnen we zeggen dat compressie de pieken verlaagt terwijl ruis de dalen verhoogt. Ruis brengt echter ook een verandering van de fijnstructuur en maskering van de zwakke spraakdelen teweeg. In hoofdstuk 4 wordt een aantal experimenten beschreven om de relatieve bijdrage van temporele modulaties en fijnstructuur aan de verstaanbaarheid te onderzoeken. Daartoe werden omhullende en fijnstructuur per ¼-octaaf band afzonderlijk gewijzigd. Zonder de spraakfijnstructuur aan te tasten werd het belang van de omhullende bestudeerd door de laatste te onderwerpen aan een drietal bewerkingen: piek-clippen, dal-clippen en 1-bit codering (blokgolfbenadering). Het belang van de fijnstructuur werd geëvalueerd door gebruik te maken van een ruisfijnstructuur met onaangetaste spraakomhullende en spraakfijnstructuur met een door ruis aangetaste omhullende. Gewone spraak+ruis diende als referentie. Vergelijking van zinsverstaanbaarheidsscores in stilte bij 60 proefpersonen toonde het volgende aan: (1) de fijnstructuur speelt een veel minder belangrijke rol dan de omhullende; (2) de pieken in de omhullende zijn perceptief belangrijker dan de dalen; (3) reductie van de spraakmodulaties door het toevoegen van ruis is nadeliger voor de verstaanbaarheid dan dezelfde mate van reductie als gevolg van directe manipulatie van de omhullende (verschil in termen van de SRT is bijna 7 dB). Dit laatste resultaat is in strijd met het MTF-concept; een verstaanbaarheidsmodel op basis van de MTF zou voorspellen dat als de modulatiereductie gelijk is, het effect op de verstaanbaarheid ook gelijk is. Het blijkt dat ruis storende modulaties introduceert die de perceptie van de relevante spraakmodulaties bemoeilijken. Voor de in de experimenten

gebruikte bewerkingen geldt in het algemeen dat er geen één-op-één relatie tussen de MTF en de verstaanbaarheidsscores kan worden aangetoond.

In hoofdstuk 5 kijken we nader naar het effect van ruis op de verstaanbaarheid in termen van de relatieve bijdrage van spraakdelen boven (pieken) en onder (dalen) de ruis. In 24 ¼-octaaf banden werd de verdeling van spraak en ruis over de zinnen gemanipuleerd door spraak in de pieken en ofwel ruis alleen ofwel spraak+ruis in de dalen aan te bieden. De zinsverstaanbaarheid werd voor verschillende spraak-ruisverhoudingen gemeten bij 12 luisteraars. De resultaten geven aan dat, vergeleken met gewone spraak+ruis stimuli (data uit hoofdstuk 4), het weghalen van ruis uit de pieken geen effect heeft op de verstaanbaarheid. Het weghalen van het spraaksignaal uit door ruis gedomineerde dalen geeft echter een verhoging van de SRT van 2 dB. Het blijkt dus dat spraakdelen onder het ruisniveau voor een deel bijdragen aan de verstaanbaarheid.

Alle experimentele resultaten uit hoofdstukken 2-4 zijn besproken in termen van (een reductie van) temporele modulaties. Maar manipulaties op een reeks smalbandige temporele omhullenden zorgen ook voor zekere spectrale veranderingen. In hoofdstuk 6 wordt onderzocht in hoeverre spectrale contrasten worden aangetast door de reductie van temporele modulaties. Als tegenhanger van de MTF voor het meten van de overdracht van temporele modulaties wordt een spectrale modulatie-overdrachtsfunctie (SMTF) beschreven om te bepalen in hoeverre modulaties in de spectrale omhullende (gemeten in perioden per octaaf) behouden blijven. Uniforme reductie (d.w.z. onafhankelijk van de modulatiefrequentie) van de temporele modulaties in ¼-octaaf banden blijkt een vrijwel gelijke mate van spectrale reductie te geven. SMTF-metingen na filteren van de temporele omhullende (hoofdstukken 2 en 3) werden uitgevoerd voor een aantal laag- en hoogdoorlaat-afsnijfrequenties (1-32 Hz). Wanneer we het effect van deze temporele bewerkingen uitdrukken in termen van spectrale contrastreductie, is het interessant te onderzoeken of spectrale reductie op zich een verklaring is voor het gemeten verlies in verstaanbaarheid. Daarom werden zinnen bewerkt met een analyse-resynthesesysteem bestaande uit Fourier transformaties op een reeks korte segementen, 10-90% reductie van de contrasten in de spectrale omhullende, inverse transformatie en overlappende optelling van de bewerkte segementen om weer een continu signaal te krijgen. De SRT in ruis voor de op die manier spectraal gereduceerde zinnen werd gemeten bij 10 luisteraars. Vergelijking van deze SRTs met die uit hoofdstuk 2 en 3 geeft aan dat spectrale reductie hetzelfde effect op de verstaanbaarheid heeft als hoogdoorlaat filteren van de temporele omhullende. Dit geldt echter niet voor laagdoorlaat filteren, wat betekent dat spectrale reductie in dat geval slechts een bij-effect is.

References

Bendat, J.S. and Piersol. A.G. (**1980**). *Engineering Applications of Correlation and Spectral Analysis* (Wiley-Interscience, New York).

Behrens, S., and Blumstein, S.E. (**1988**). "On the role of the amplitude of the fricative noise in the perception of place of articulation in voiceless fricative consonants," J. Acoust. Soc. Am. **84**, 861-867.

Bosman, A.J. (**1989**). "Speech perception by the hearing impaired," Ph.D. dissertation, University of Utrecht.

Carter, G.C., Knapp, C.H., and Nuttall, A.H. (**1973**). "Estimation of the magnitude-squared coherence function via overlapped fast fourier transform processing," IEEE Trans. Audio Electroacoust. **21**, 337-344.

van Dijkhuizen, J.N., Festen, J.M., and Plomp, R. (**1989**). "The effect of varying the amplitude-frequency response on the masked speech-reception threshold of sentences for hearing-impaired listeners," J. Acoust. Soc. Am. **86**, 621-628.

Drullman, R., Festen, J.M., and Plomp, R. (**1994a**). "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95**, 1053-1064.

Drullman, R., Festen, J.M., and Plomp, R. (**1994b**). "Effect of reducing slow temporal modulations on speech reception," J. Acoust. Soc. Am. **95**, 2670-2680.

Drullman, R. (**1994c**). "Temporal envelope and fine structure cues for speech intelligibility," accepted for publication in J. Acoust. Soc. Am.

Drullman, R. "Speech intelligibility in noise: relative contribution of speech elements below and above the noise level", submitted as a Letter to the Editor in J. Acoust. Soc. Am.

Drullman, R., Festen, J.M., and Houtgast, T. "Effect of temporal modulation reduction on spectral contrasts in speech," submitted for publication in J. Acoust. Soc. Am.

Duquesnoy, A.J., and Plomp, R. (**1980**). "Effect of reverberation and noise on the intelligibility of sentences in cases of presbyacusis," J. Acoust. Soc. Am. **68**, 537-544.

Festen, J.M., and Plomp, R. (**1981**). "Relations between auditory functions in normal hearing," J. Acoust. Soc. Am. **70**, 356-369.

Festen, J.M, van Dijkhuizen, J.N., and Plomp, R. (**1990**). "Considerations on adaptive gain and frequency response in hearing aids," Acta Otolaryngol. Suppl. **469**, 196-201.

Flanagan, J.L. (**1972**). *Speech Analysis, Synthesis, and Perception* (Springer-Verlag, Berlin), 2nd ed., Chap. 8, 323-330.

Freyman, R.L., Nerbonne, G.P., and Cote, H.A. (**1991**). "Effect of consonant-vowel ratio modification on amplitude envelope cues for consonant recognition," J. Speech Hear. Res. **34**, 415-426.

Gelfand, S.A., and Silman, S. (**1979**). "Effects of small room reverberation upon the recognition of some consonant features," J. Acoust. Soc. Am. **66**, 22-29.

Hohmann, V., and Kollmeier, B. (**1990**). "Sprachverständlichkeit bei Dynamik-kompression," in *Fortschritte der Akustik - DAGA '90* (DPG-Kongress-GmbH, Bad Honeff), 1115-1118.

Houtgast, T., and Steeneken, H.J.M. (**1973**). "The modulation transfer function in room acoustics as a predictor of speech intelligibility," Acustica **28**, 66-37.

Houtgast, T., and Steeneken, H.J.M. (**1985**). "A review of the MFT concept in room acoustics and its use for estimating speech intelligibility in auditoria," J. Acoust. Soc. Am. **77**, 1069-1077.

Houtgast, T., and Verhave, J.A. (**1991**). "A physical approach to speech quality assessment: correlation patterns in the speech spectrogram," in *Proceedings of the 3rd European Conference on Speech Communication and Technology*, edited by G. Pirani, Genova, September 1991 (IIC, Genova, Italy), Vol. 1, 285-288.

ter Keurs, M. (**1992**). "Intelligibility of spectrally smeared speech," Ph.D. dissertation, Free University of Amsterdam.

ter Keurs, M., Festen, J.M., and Plomp, R. (**1992**). "Effects of spectral smearing on speech reception. I," J. Acoust. Soc. Am. **91**, 2872-2880.

ter Keurs, M., Festen, J.M., and Plomp, R. (**1993a**). "Effects of spectral smearing on speech reception. II," J. Acoust. Soc. Am. **93**, 1547-1552.

ter Keurs, M., Festen, J.M., and Plomp, R. (**1993b**). "Limited resolution of spectral contrasts and hearing loss for speech in noise," J. Acoust. Soc. Am. **94**, 1307-1314.

Kirk, R.E. (**1968**). *Experimental Design: Procedures for the Behavioral Sciences* (Brooks/Cole, Belmont, CA), 1st ed., Chap. 8, 263-270.

Lahiri, A., Gewirth, L., and Blumstein, S.E. (**1984**). "A reconsideration of acoustic invariance for place of articulation in diffuse stop consonants: Evidence from a cross-language study," J. Acoust. Soc. Am. **76**, 391-404.

Ludvigsen, C., Elberling, C., Keidser, G., and Poulsen, T. (**1990**). "Predic-tion of intelligibility of non-linearly processed speech," Acta Otolaryngol. Suppl. **469**, 190-195.

Moore, B.C.J. (**1982**). *An Introduction to the Psychology of Hearing* (Academic, London, UK), Chap. 3, 82-83.

Nearey, T.M., and Assman, P.F. (**1986**). "Modeling the role of inherent spectral change in vowel indentification," J. Acoust. Soc. Am. **80**, 1297-1308.

Nittrouer, S., and Studdert-Kennedy, M. (**1986**). "The stop-glide distinction: Acoustic analysis and perceptual effect of variation in syllable amplitude envelope for initial /b/ and /w/," J. Acoust. Soc. Am. **80**, 1026-1029.

Ohde, R.N., and Stevens, K.N. (**1983**). "Effect of burst amplitude on the perception of stop consonant place of articulation," J. Acoust. Soc. Am. **74**, 706-714.

O'Shaughnessy, D. (**1987**). *Speech Communication* (Addison-Wesley, Reading MA), Chap. 7, 305-309.

Pavlovic, C.V. (**1987**). "Derivation of primary parameters and procedures for use in speech intelligibility predictions," J. Acoust. Soc. Am. **82**, 413-422.

Plomp, R. (**1984**). "Perception of speech as a modulated signal," in *Proceed-ings of the 10th International Congress of Phonetic Sciences*, edited by A. Cohen and M.P.R. van de Broecke (Foris, Dordrecht), 29-40.

Plomp, R. (**1986**). "A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired," J. Speech Hear. Res. **29**, 146-154.

Plomp, R. (**1988**). "The negative effect of amplitude compression in multichannel hearing aids in the light of the modulation-transfer function," J. Acoust Soc. Am. **83**, 2322-2327.

Plomp, R. (**1989**). "Reply to 'Comments on 'The negative effect of ampli-tude compression in multichannel hearing aids in the light of the modulation-transfer function' [J. Acoust. Soc. Am. **83**, 2322-2327 (1988)]'," J. Acoust. Soc. Am. **86**, 428.

Plomp, R., and Mimpen, A.M. (**1979**). "Improving the reliability of testing the Speech Reception Threshold for sentences," Audiology **18**, 43-52.

Plomp, R., and van Beek, J.H.M. (**1990**). "The spectrogram as an aid in studying speech intelligibility at low S/N ratios," J. Acoust. Soc. Am. Suppl. 1 **87**, S118(A).

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

Rabiner, L.R., and Gold, B. (**1975**). *Theory and Application of Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ), Chap. 2, 70-72.

Rodenburg, M. (1977). "Investigation of temporal effects with amplitude modulated signals," in *Psychophysics and Psychology of Hearing*, edited by E.F. Evans and J.P. Wilson (Academic, London, UK), 429-437.

Shannon, R.V., Zeng, F.-G., Wygonski, J., Kamath, V., and Ekelid, M. (**1994**). "Speech recognition with minimal spectral cues," J. Acoust. Soc. Am. **95**, 2876(A).

Steeneken, H.J.M. (**1992**). "On measuring and predicting speech intel-ligibility," Ph.D. dissertation, University of Amsterdam.

Steeneken, H.J.M. and Houtgast, T. (**1980**). "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**, 318-326.

Studebaker, G.A. (**1985**). "A 'rationalized' arcsine transform," J. Speech Hear. Res. **28**, 455-462.

Studebaker, G.A., Pavlovic, C.V., and Sherbecoe, R.L. (**1987**). "A frequen-cy importance function for continuous discourse," J. Acoust. Soc. Am. **81**, 1130-1138.

Van Tasell, D.J., Soli, S.D., Kirby, V.M., and Widin, G.P. (**1987**). "Speech waveform envelope cues for consonant recognition," J. Acoust. Soc. Am. **82**, 1152-1161.

van Veen, T.M., and Houtgast, T. (**1985**). "Spectral sharpness and vowel dissimilarity," J. Acoust. Soc. Am. **77**, 628-634.

Verschuure, J., Dreschler, W.A., de Haan, E.H., van Cappellen, M., Hammerschlag, R., Maré, M.J., Maas, A.J.J., and Hijmans, A.C. (**1993**). "Syllabic compression and speech intelligibility in hearing impaired listeners," Scand. Audiol. Suppl. **38**, 92-100.

Viemeister, N.F. (**1979**). "Temporal modulation transfer functions based upon modulation thresholds," J. Acoust. Soc. Am. **66**, 1364-1380.

Villchur, E. (**1989**). "Comments on 'The negative effect of amplitude compression in multichannel hearing aids in the light of the modulation-transfer function' [J. Acoust. Soc. Am. **83**, 2322-2327 (1988)]," J. Acoust. Soc. Am. **86**, 425-427.

Zwicker, E., and Feldtkeller, R. (**1967**). *Das Ohr als Nachrichtenempfänger* (Hirzel Verlag, Stuttgart, Germany), Chap. 6, 70-73.

APPENDIX A: SUMMED CONFUSION MATRICES FROM THE PHONEME IDENTIFICATION EXPERIMENTS IN CHAPTER 2

TABLE A1. Summed confusion matrices for 12 subjects in the six conditions: Initial consonants.

Stimulus/response LP 0 Hz

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	7	12	2	3	1	.	.	2	3	2	3	.	.	.	1	.	12	48
k	2	17	2	1	1	.	.	1	3	.	1	.	.	.	1	.	19	48
b	4	2	.	9	2	.	.	1	.	3	1	.	.	1	11	1	13	48
d	1	2	.	3	3	.	.	2	2	5	.	1	.	7	9	1	12	48
v	2	1	1	1	1	.	.	2	6	7	4	.	.	1	7	.	15	48
z	1	2	.	2	.	.	.	1	3	26	.	.	.	.	1	.	12	48
χ	1	5	1	3	.	2	2	7	6	4	.	2	.	.	5	.	10	48
n	5	3	.	3	2	.	.	3	.	.	3	6	.	4	3	3	13	48
l	1	3	.	3	.	.	.	1	.	3	1	1	.	13	5	2	15	48
w	1	.	.	2	4	.	.	5	3	.	1	.	.	1	9	.	22	48
j	4	1	.	4	2	.	.	2	3	1	2	1	.	2	6	2	18	48
h	.	4	1	4	1	.	1	2	2	.	1	1	.	.	8	1	22	48
sum	29	52	7	38	17	2	3	29	31	51	17	12	0	29	66	10	183	576

Stimulus/response LP 2 Hz

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	3	20	5	3	.	2	.	6	2	.	.	.	.	.	.	.	7	48
k	1	20	2	2	.	.	2	3	.	.	.	1	.	1	.	2	14	48
b	.	.	.	12	.	.	.	1	4	.	.	.	.	.	23	1	7	48
d	.	1	.	7	7	1	.	1	4	2	1	.	.	2	14	1	7	48
v	.	.	1	1	.	.	.	.	33	.	.	.	.	.	9	.	4	48
z	.	.	.	.	.	1	3	.	.	42	.	2	.	.	.	.	.	48
χ	6	4	.	.	.	2	.	35	.	1	.	.	.	.	.	.	.	48
n	.	.	.	1	.	.	1	.	1	2	1	15	.	4	14	1	8	48
l	.	.	.	.	.	.	.	.	1	2	1	.	.	28	11	.	5	48
w	.	.	.	5	.	.	.	1	3	.	.	.	.	.	32	.	7	48
j	.	.	.	3	.	.	.	2	.	.	.	.	.	.	20	11	12	48
h	.	.	.	.	1	.	.	2	.	.	.	.	.	.	2	1	42	48
sum	10	45	8	34	8	6	6	51	48	49	3	18	0	35	125	17	113	576

Stimulus/response LP 4 Hz

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	24	15	6	.	.	1	.	.	.	.	.	.	.	1	.	.	1	48
k	.	35	3	3	.	.	.	.	1	.	.	.	.	.	1	.	5	48
b	.	.	.	6	.	.	.	.	.	.	.	.	.	.	41	1	.	48
d	.	2	.	5	3	.	.	.	1	.	.	.	.	3	30	4	.	48
v	.	.	.	.	.	4	.	.	33	2	.	.	.	.	9	.	.	48
z	.	.	.	.	.	.	2	.	.	46	.	.	.	.	.	.	.	48
χ	.	.	.	.	.	.	2	45	.	1	.	.	.	.	.	.	.	48
n	.	.	.	.	.	.	.	.	.	.	.	42	.	2	3	1	.	48
l	.	.	.	.	.	.	1	.	.	2	.	.	.	38	6	1	.	48
w	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48	.	.	48
j	.	.	.	.	.	.	.	.	1	.	.	.	.	1	5	41	.	48
h	.	.	.	.	.	.	.	1	1	.	.	.	.	.	.	.	46	48
sum	24	52	9	14	3	5	5	46	37	51	0	42	0	45	143	48	52	576

TABLE A1. Continued.

Stimulus/response LP 8 Hz

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	36	5	3	.	1	.	.	.	2	1	.	.	.	.	.	.	.	48
k	1	34	7	.	.	.	.	3	.	.	.	1	.	.	.	.	2	48
p	.	.	.	20	.	.	.	.	.	.	.	.	.	.	28	.	.	48
d	.	1	.	1	29	.	.	.	.	.	.	.	.	.	16	1	.	48
v	.	.	.	.	.	10	.	.	37	.	.	.	.	.	1	.	.	48
z	.	.	.	.	.	.	3	.	.	45	.	.	.	.	.	.	.	48
χ	.	.	.	.	.	.	.	47	.	.	1	.	.	.	.	.	.	48
n	.	.	.	.	.	.	.	.	1	.	.	46	.	1	.	.	.	48
l	.	.	.	2	.	.	.	.	.	.	.	.	.	45	.	1	.	48
w	.	.	.	.	.	.	.	.	.	.	1	.	.	.	47	.	.	48
j	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48	.	48
h	.	.	1	.	.	.	.	1	.	.	.	.	.	.	1	.	45	48
sum	37	40	11	23	30	10	3	51	40	46	2	47	0	46	93	50	47	576

Stimulus/response LP 16 Hz

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	48	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48
k	2	37	3	.	.	1	.	4	1	.	.	.	.	.	.	.	.	48
b	.	.	.-	34	.	.	.	.	.	.	.	.	.	.	.	14	.	.
48																		
d	.	.	.	2	42	.	.	1	.	.	.	.	.	.	2	1	.	48
v	.	.	.	.	.	4	1	.	39	1	.	.	.	.	3	.	.	48
z	.	.	.	.	.	.	3	.	.	45	.	.	.	.	.	.	.	48
χ	.	.	.	.	.	.	.	47	.	.	.	.	.	1	.	.	.	48
n	.	.	.	.	.	.	.	.	.	.	.	48	.	.	.	.	.	48
l	.	.	.	.	.	.	.	.	1	.	.	.	.	47	.	.	.	48
w	.	.	.	.	.	.	.	.	.	.	1	.	.	.	47	.	.	48
j	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48	.	48
h	.	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	47	48
sum	50	37	3	36	42	5	4	52	41	46	1	49	0	48	66	49	47	576

Stimulus/response control

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	48	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48
k	1	47	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48
b	.	.	.	47	.	.	.	.	.	.	.	.	.	.	1	.	.	48
d	.	.	.	.	48	.	.	.	.	.	.	.	.	.	.	.	.	48
v	.	.	.	.	.	2	.	.	43	.	.	.	.	.	3	.	.	48
z	.	.	.	.	.	.	2	.	.	46	.	.	.	.	.	.	.	48
χ	.	1	.	.	.	.	.	47	.	.	.	.	.	.	.	.	.	48
n	.	.	.	.	.	.	.	.	.	.	.	48	.	.	.	.	.	48
l	.	.	1	.	.	.	.	.	.	.	.	.	.	47	.	.	.	48
w	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48	.	.	48
j	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48	.	48
h	.	.	.	.	.	.	.	.	.	.	1	.	.	.	.	.	47	48
sum	49	48	1	47	48	2	2	47	43	46	1	48	0	47	52	48	47	576

TABLE A2. Summed confusion matrices for 12 subjects in the six conditions: Final consonants.

Stimulus/response LP 0 Hz

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	.	.	.	.	.	10	10	23	.	.	1	.	.	1	.	.	3	48
k	1	.	.	.	.	13	9	21	.	.	1	1	.	.	.	2	.	48
p	1	.	.	.	.	11	9	22	.	.	1	.	.	2	.	1	1	48
f	.	.	.	.	.	15	12	15	1	.	.	.	.	3	.	.	2	48
s	2	.	.	.	.	9	26	9	.	.	.	1	.	.	.	.	1	48
χ	1	.	.	.	.	15	2	29	.	.	.	1	.	.	.	.	.	48
m	2	1	.	.	.	2	3	9	1	.	3	5	5	6	4	1	6	48
n	.	.	.	.	.	6	.	8	.	.	1	9	3	16	.	.	5	48
ŋ	4	.	.	.	.	4	3	8	.	.	4	8	4	7	1	1	4	48
l	2	.	.	.	.	5	3	7	.	.	3	2	1	14	3	1	7	48
w	.	.	.	.	.	7	2	11	.	.	4	6	1	6	6	.	5	48
j	.	.	.	.	.	5	2	13	.	.	1	5	2	6	3	2	9	48
sum	13	1	0	0	0	102	81	175	2	0	19	38	16	61	17	8	43	576

Stimulus/response LP 2 Hz

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	6	.	.	.	.	8	33	1	.	.	.	.	.	.	.	.	.	48
k	.	4	.	.	.	9	.	35	.	.	.	.	.	.	.	.	.	48
p	.	.	1	.	.	38	.	8	.	.	.	.	.	.	1	.	.	48
f	.	.	.	.	.	47	1	.	.	.	.	.	.	.	.	.	.	48
s	.	.	.	.	.	.	48	.	.	.	.	.	.	.	.	.	.	48
χ	.	.	.	.	.	1	1	46	.	.	.	.	.	.	.	.	.	48
m	.	.	.	.	.	.	.	.	.	.	25	15	3	.	5	.	.	48
n	.	.	.	.	.	1	.	.	.	.	6	25	9	3	3	.	1	48
ŋ	.	.	.	.	.	1	.	.	.	.	4	13	27	1	.	1	1	48
l	.	.	.	.	.	.	.	.	.	.	1	.	.	39	8	.	.	48
w	.	.	.	.	.	.	.	.	.	.	.	.	.	10	37	.	1	48
j	.	.	.	.	.	.	.	.	.	.	.	2	2	4	3	37	.	48
sum	6	4	1	0	0	105	83	90	0	0	36	55	41	57	57	38	3	576

Stimulus/response LP 4 Hz

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	36	.	.	.	.	1	10	.	.	.	.	.	.	1	.	.	.	48
k	1	26	3	.	.	8	1	9	.	.	.	.	.	.	.	.	.	48
p	2	.	28	.	.	17	.	1	.	.	.	.	.	.	.	.	.	48
f	.	.	.	.	.	47	1	.	.	.	.	.	.	.	.	.	.	48
s	.	.	.	.	.	.	48	.	.	.	.	.	.	.	.	.	.	48
χ	.	.	.	.	.	1	.	47	.	.	.	.	.	.	.	.	.	48
m	.	.	.	.	.	.	.	.	.	.	36	10	1	.	1	.	.	48
n	.	.	.	.	.	1	.	.	.	.	3	37	7	.	.	.	.	48
ŋ	.	.	.	.	.	.	.	.	.	.	3	5	40	.	.	.	.	48
l	.	.	.	.	.	.	.	.	.	.	.	.	.	42	5	1	.	48
w	.	.	.	.	.	.	.	.	.	.	.	.	.	4	44	.	.	48
j	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48	.	48
sum	39	26	31	0	0	75	60	57	0	0	42	52	48	47	50	49	0	576

TABLE A2. Continued.

Stimulus/response LP 8 Hz

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	48	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48
k	1	45	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48
p	.	.	48	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48
f	.	.	.	.	.	46	1	1	.	.	.	.	.	.	.	.	.	48
s	.	.	.	.	.	.	47	.	.	.	.	.	.	1	.	.	.	48
χ	.	.	.	.	.	.	1	46	.	.	.	.	.	1	.	.	.	48
m	.	.	.	.	.	.	.	.	.	.	43	3	1	.	.	1	.	48
n	.	.	.	.	.	.	.	.	.	.	.	44	4	.	.	.	.	48
ŋ	.	.	.	.	.	.	.	.	.	.	.	.	48	.	.	.	.	48
l	.	.	.	.	.	.	.	.	.	.	.	.	.	42	6	.	.	48
w	.	.	.	.	.	.	.	.	.	.	.	.	.	4	44	.	.	48
j	.	.	.	.	.	.	.	.	.	.	.	.	.	.	1	47	.	48
sum	49	45	50	0	0	46	49	47	0	0	43	47	53	47	52	48	0	576

Stimulus/response LP 16 Hz

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	47	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48
k	.	48	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48
p	.	.	47	.	.	.	1	.	.	.	.	.	.	.	.	.	.	48
f	.	.	.	.	.	48	.	.	.	.	.	.	.	.	.	.	.	48
s	.	.	.	.	.	.	48	.	.	.	.	.	.	.	.	.	.	48
χ	.	.	.	.	.	.	.	48	.	.	.	.	.	.	.	.	.	48
m	.	.	.	.	.	.	.	.	.	.	46	2	.	.	.	.	.	48
n	.	.	.	.	.	.	.	.	.	.	2	46	.	.	.	.	.	48
ŋ	.	.	.	.	.	.	.	.	.	.	1	.	47	.	.	.	.	48
l	.	.	.	.	.	.	.	.	.	.	.	.	.	45	3	.	.	48
w	.	.	.	.	.	.	.	.	.	.	.	.	.	3	45	.	.	48
j	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48	.	48
sum	47	49	47	0	0	48	49	48	0	0	49	48	47	48	48	48	0	576

Stimulus/response control

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	48	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48
k	.	48	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48
p	.	.	48	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48
f	.	.	.	.	.	47	.	.	1	.	.	.	.	.	.	.	.	48
s	.	.	.	.	.	.	48	.	.	.	.	.	.	.	.	.	.	48
χ	.	.	.	.	.	.	.	48	.	.	.	.	.	.	.	.	.	48
m	.	.	.	.	.	.	.	.	.	.	46	2	.	.	.	.	.	48
n	.	.	.	.	.	.	.	.	.	.	.	47	1	.	.	.	.	48
ŋ	.	.	.	.	.	.	.	.	.	.	.	.	48	.	.	.	.	48
l	.	.	.	.	.	.	.	.	.	.	.	.	.	42	6	.	.	48
w	.	.	.	.	.	.	.	.	.	.	.	.	.	1	47	.	.	48
j	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48	.	48
sum	48	48	48	0	0	47	48	48	1	0	46	49	49	43	53	48	0	576

TABLE A3. Summed confusion matrices for 12 subjects in the six conditions: Medial consonants.

Stimulus/response LP 0 Hz

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	.	1	.	.	.	10	1	20	7	1	.	.	.	.	.	.	8	48
k	.	.	.	.	.	4	1	32	5	.	.	.	.	.	.	.	6	48
p	.	.	1	.	.	7	3	26	6	.	.	.	.	.	.	.	5	48
b	.	.	.	10	1	1	1	3	19	1	.	.	.	.	1	.	11	48
d	.	.	1	4	9	1	1	1	13	2	.	.	.	2	1	.	13	48
f	.	.	.	.	.	12	3	19	11	.	.	.	.	.	.	.	3	48
s	1	.	.	.	.	7	16	11	6	2	.	.	.	.	.	.	5	48
χ	.	.	.	.	.	9	1	21	5	1	.	.	.	.	.	1	10	48
v	.	.	.	1	.	1	.	2	18	2	.	.	.	.	.	.	24	48
z	.	.	.	1	2	.	2	1	11	11	.	.	.	2	1	1	16	48
m	.	.	.	.	.	.	.	.	.	.	5	2	.	29	2	.	10	48
n	.	.	.	.	.	.	.	.	1	.	4	4	.	32	.	1	6	48
l	.	.	.	.	.	.	.	.	.	.	2	.	.	36	.	.	10	48
w	.	.	.	3	1	1	.	1	18	.	.	.	.	.	10	.	14	48
j	.	.	.	1	.	.	.	.	10	.	.	.	.	5	10	9	13	48
h	.	.	.	.	.	.	.	5	8	2	.	.	.	.	1	1	31	48
sum	1	1	2	20	13	53	29	142	138	22	11	6	0	106	26	13	185	768

Stimulus/response LP 2 Hz

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	10	2	2	.	.	6	5	4	8	4	.	.	.	.	.	.	7	48
k	.	12	.	.	.	4	.	21	4	.	.	.	.	.	1	.	6	48
p	.	2	5	.	.	5	.	17	8	.	.	.	.	.	.	1	10	48
b	.	1	.	12	4	2	.	6	14	.	.	.	.	.	8	.	1	48
d	.	.	.	.	20	.	2	5	11	1	.	.	.	2	.	1	6	48
f	.	.	.	.	.	30	.	6	12	.	.	.	.	.	.	.	.	48
s	.	.	.	.	.	.	43	1	.	4	.	.	.	.	.	.	.	48
χ	.	.	.	.	.	3	.	33	8	.	.	.	.	.	.	.	4	48
v	.	.	.	.	.	6	.	.	38	.	.	.	.	.	.	.	4	48
z	.	.	.	.	.	.	7	.	.	41	.	.	.	.	.	.	.	48
m	.	.	.	.	.	.	.	.	1	.	25	12	.	2	8	.	.	48
n	.	.	.	.	.	.	.	.	.	.	2	35	3	7	.	.	1	48
l	.	.	.	.	.	.	.	.	.	.	.	1	.	46	1	.	.	48
w	.	.	.	2	.	1	.	1	24	.	.	.	.	.	17	.	3	48
j	.	.	.	.	.	.	.	.	1	.	.	4	.	1	1	41	.	48
h	.	.	.	3	.	1	.	7	14	.	.	.	.	.	3	.	20	48
sum	10	17	7	17	24	58	57	101	143	50	27	52	3	58	39	43	62	768

TABLE A3. Continued.

Stimulus/response LP 4 Hz

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	20	6	1	.	2	1	5	2	3	2	.	1	.	1	.	1	3	48
k	4	19	2	.	.	.	.	15	.	.	.	.	.	.	.	.	8	48
p	1	9	18	1	1	6	.	3	2	.	.	.	.	.	1	.	6	48
b	.	1	.	23	5	.	.	6	7	.	.	.	.	.	6	.	.	48
d	.	.	.	.	38	.	.	3	1	3	.	.	1	1	1	.	.	48
f	.	.	.	.	.	42	.	.	6	.	.	.	.	.	.	.	.	48
s	.	.	.	.	.	.	42	.	.	6	.	.	.	.	.	.	.	48
χ	.	.	.	.	.	.	.	46	2	.	.	.	.	.	.	.	.	48
v	.	.	.	.	.	6	.	.	42	.	.	.	.	.	.	.	.	48
z	.	.	.	.	.	.	9	.	.	39	.	.	.	.	.	.	.	48
m	.	.	.	.	.	.	.	.	.	.	42	6	.	.	.	.	.	48
n	.	.	.	.	.	.	.	.	.	.	2	40	1	5	.	.	.	48
l	.	.	.	.	.	.	.	.	.	.	.	.	.	48	.	.	.	48
w	.	.	.	1	.	1	.	.	26	.	.	.	.	.	19	.	1	48
j	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48	.	48
h	.	.	.	.	.	.	.	6	17	.	.	.	.	.	4	4	17	48
sum	25	35	21	25	46	56	56	81	106	50	44	47	2	55	31	53	35	768

Stimulus/response LP 8 Hz

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	36	8	.	.	2	.	1	.	.	1	.	.	.	.	.	.	.	48
k	4	36	.	.	.	.	.	3	.	.	.	.	.	.	.	.	5	48
p	.	3	39	.	.	.	.	.	3	.	.	.	.	.	.	.	3	48
b	.	.	.	48	.	.	.	.	.	.	.	.	.	.	.	.	.	48
d	.	.	.	1	46	.	.	1	.	.	.	.	.	.	.	.	.	48
f	.	.	.	.	.	38	.	.	10	.	.	.	.	.	.	.	.	48
s	.	.	.	.	.	.	45	.	.	3	.	.	.	.	.	.	.	48
χ	.	.	.	.	.	.	.	48	.	.	.	.	.	.	.	.	.	48
v	.	.	.	.	.	9	.	.	39	.	.	.	.	.	.	.	.	48
z	.	.	.	.	.	.	9	.	.	39	.	.	.	.	.	.	.	48
m	.	.	.	.	.	.	.	.	.	.	47	1	.	.	.	.	.	48
n	.	.	.	.	.	.	.	.	.	.	.	43	.	5	.	.	.	48
l	.	.	.	.	.	.	.	.	.	.	.	.	.	48	.	.	.	48
w	.	.	.	3	.	2	.	.	22	.	.	.	.	.	21	.	.	48
j	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48	.	48
h	.	.	.	.	.	2	.	6	15	.	.	.	.	.	2	2	21	48
sum	40	47	39	52	48	51	55	58	89	43	47	44	0	53	23	50	29	768

TABLE A3. Continued.

Stimulus/response LP 16 Hz

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	40	8	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48
k	6	39	1	.	.	.	.	1	.	.	.	.	.	.	.	.	1	48
p	.	.	48	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48
b	.	.	.	48	.	.	.	.	.	.	.	.	.	.	.	.	.	48
d	.	.	.	1	47	.	.	.	.	.	.	.	.	.	.	.	.	48
f	.	.	.	.	.	38	.	.	10	.	.	.	.	.	.	.	.	48
s	.	.	.	.	.	.	45	.	.	3	.	.	.	.	.	.	.	48
χ	.	.	.	.	.	.	.	48	.	.	.	.	.	.	.	.	.	48
v	.	.	.	.	.	7	.	.	41	.	.	.	.	.	.	.	.	48
z	.	.	.	.	.	.	10	.	.	38	.	.	.	.	.	.	.	48
m	.	.	.	.	.	.	.	.	.	.	48	.	.	.	.	.	.	48
n	.	.	.	.	.	.	.	.	.	.	1	43	1	3	.	.	.	48
l	.	.	.	.	.	.	.	.	.	.	.	.	.	48	.	.	.	48
w	.	.	.	3	.	1	.	.	21	.	.	.	.	.	23	.	.	48
j	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48	.	48
h	.	.	.	.	.	2	.	4	15	.	.	.	.	.	2	4	21	48
sum	46	47	49	52	47	48	55	53	87	41	49	43	0	51	25	52	22	768

Stimulus/response control

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	43	1	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	44
k	.	44	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	44
p	.	.	44	.	.	.	.	.	.	.	.	.	.	.	.	.	.	44
b	.	.	.	44	.	.	.	.	.	.	.	.	.	.	.	.	.	44
d	.	.	.	.	44	.	.	.	.	.	.	.	.	.	.	.	.	44
f	.	.	.	.	.	37	1	.	6	.	.	.	.	.	.	.	.	44
s	.	.	.	.	.	.	41	.	.	3	.	.	.	.	.	.	.	44
χ	.	.	.	.	.	.	.	44	.	.	.	.	.	.	.	.	.	44
v	.	.	.	.	.	4	.	.	40	.	.	.	.	.	.	.	.	44
z	.	.	.	.	.	.	10	.	.	34	.	.	.	.	.	.	.	44
m	.	.	.	.	.	.	.	.	.	.	44	.	.	.	.	.	.	44
n	.	.	.	.	.	.	.	.	.	.	.	42	1	1	.	.	.	44
l	.	.	.	.	.	.	.	.	.	.	.	.	.	44	.	.	.	44
w	.	.	.	2	.	1	.	.	21	.	.	.	.	.	20	.	.	44
j	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	44	.	44
h	.	.	.	.	.	.	.	4	6	.	.	.	.	.	.	1	33	44
sum	43	45	44	46	44	42	52	48	73	37	44	42	1	45	20	45	33	704

TABLE A4. Summed confusion matrices for 12 subjects in six the conditions: Vowels.

Stimulus/response LP 0 Hz

	ɑ	a	ɑu	ɛ	e	ɛi	ø	I	i	ɔ	o	u	sum
ɑ	25	20	.	.	.	.	.	.	.	3	.	.	48
a	3	42	.	.	.	1	1	.	.	1	.	.	48
ɑu	27	11	3	1	.	.	4	.	.	2	.	.	48
ɛ	1	.	.	40	.	5	1	.	.	1	.	.	48
e	.	.	.	6	27	.	.	15	.	.	.	.	48
ɛi	.	.	.	40	.	7	.	.	.	.	1	.	48
ø	.	.	.	.	.	.	47	.	.	1	.	.	48
I	.	.	.	4	6	.	10	20	4	2	2	.	48
i	.	.	.	.	.	.	9	4	35	.	.	.	48
ɔ	1	1	.	.	.	.	5	2	.	22	11	6	48
o	.	.	.	.	.	.	4	.	.	17	26	1	48
u	3	2	.	1	1	.	4	1	4	2	1	29	48
sum	60	76	3	92	34	13	85	42	43	51	41	36	576

Stimulus/response LP 2 Hz

	ɑ	a	ɑu	ɛ	e	ɛi	ø	I	i	ɔ	o	u	sum
ɑ	35	13	.	.	.	.	.	.	.	.	.	.	48
a	1	46	1	.	.	.	.	.	.	.	.	.	48
ɑu	19	3	24	.	.	.	2	.	.	.	.	.	48
ɛ	.	.	.	43	1	2	.	1	.	.	1	.	48
e	.	.	.	.	42	.	.	4	1	.	.	1	48
ɛi	.	.	.	23	.	24	.	.	.	.	1	.	48
ø	.	.	.	.	.	.	48	.	.	.	.	.	48
I	.	.	.	.	2	.	1	44	1	.	.	.	48
i	.	.	.	1	.	.	.	2	45	.	.	.	48
ɔ	1	.	.	.	.	.	2	1	.	24	15	5	48
o	.	.	.	.	.	.	.	.	.	7	41	.	48
u	.	.	.	.	.	.	1	.	.	.	.	47	48
sum	56	62	25	67	45	26	54	52	47	31	58	53	576

Stimulus/response LP 4 Hz

	ɑ	a	ɑu	ɛ	e	ɛi	ø	I	i	ɔ	o	u	sum
ɑ	43	3	1	.	.	1	.	.	.	.	.	.	48
a	.	47	1	.	.	.	.	.	.	.	.	.	48
ɑu	4	.	43	.	.	.	.	1	.	.	.	.	48
ɛ	.	.	.	47	.	1	.	.	.	.	.	.	48
e	.	.	.	.	44	.	1	2	1	.	.	.	48
ɛi	.	.	.	4	.	44	.	.	.	.	.	.	48
ø	.	.	.	.	.	.	48	.	.	.	.	.	48
I	.	.	.	.	4	.	.	44	.	.	.	.	48
i	.	.	.	.	.	.	.	2	46	.	.	.	48
ɔ	1	.	.	.	.	.	.	.	.	38	7	2	48
o	.	.	.	.	.	.	.	.	.	3	45	.	48
u	.	.	.	.	.	.	.	.	.	.	.	48	48
sum	48	50	45	51	48	46	49	49	47	41	52	50	576

TABLE A4. Continued.

Stimulus/response LP 8 Hz

	ɑ	a	ɑu	ɛ	e	ɛi	ø	I	i	ɔ	o	u	sum
ɑ	47	1	.	.	.	.	.	.	.	.	.	.	48
a	1	47	.	.	.	.	.	.	.	.	.	.	48
ɑu	3	.	44	.	.	.	.	.	.	.	1	.	48
ɛ	.	.	.	48	.	.	.	.	.	.	.	.	48
e	.	.	.	.	46	.	2	.	.	.	.	.	48
ɛi	.	.	.	.	.	48	.	.	.	.	.	.	48
ø	.	.	.	.	.	.	48	.	.	.	.	.	48
I	.	.	.	.	2	.	.	46	.	.	.	.	48
i	.	.	.	.	.	.	.	.	48	.	.	.	48
ɔ	.	.	.	.	.	.	.	.	.	44	3	1	48
o	.	.	.	.	.	.	.	.	.	1	47	.	48
u	.	.	.	.	.	.	.	.	.	.	.	48	48
sum	51	48	44	48	48	48	50	46	48	45	51	49	576

Stimulus/response LP 16 Hz

	ɑ	a	ɑu	ɛ	e	ɛi	ø	I	i	ɔ	o	u	sum
ɑ	44	.	.	.	.	.	.	.	.	.	.	.	44
a	.	44	.	.	.	.	.	.	.	.	.	.	44
ɑu	2	.	41	.	.	.	1	.	.	.	.	.	44
ɛ	.	.	.	44	.	.	.	.	.	.	.	.	44
e	.	.	.	.	44	.	.	.	.	.	.	.	44
ɛi	.	.	.	.	.	44	.	.	.	.	.	.	44
ø	.	.	.	.	.	.	44	.	.	.	.	.	44
I	.	.	.	.	2	.	.	42	.	.	.	.	44
i	.	.	.	.	.	.	.	1	43	.	.	.	44
ɔ	.	.	.	.	.	.	.	.	.	37	7	.	44
o	.	.	.	.	.	.	.	.	.	.	44	.	44
u	.	.	.	.	.	.	.	.	.	.	.	44	44
sum	46	44	41	44	46	43	45	43	43	38	51	44	528

Stimulus/response control

	ɑ	a	ɑu	ɛ	e	ɛi	ø	I	i	ɔ	o	u	sum
ɑ	40	.	.	.	.	.	.	.	.	.	.	.	40
a	1	39	.	.	.	.	.	.	.	.	.	.	40
ɑu	.	.	40	.	.	.	.	.	.	.	.	.	40
ɛ	.	.	.	40	.	.	.	.	.	.	.	.	40
e	.	.	.	1	39	.	.	.	.	.	.	.	40
ɛi	.	.	.	.	.	40	.	.	.	.	.	.	40
ø	.	.	.	.	.	.	39	.	.	.	1	.	40
I	.	.	.	.	1	.	.	39	.	.	.	.	40
i	.	.	.	.	.	.	.	1	39	.	.	.	40
ɔ	.	.	.	.	.	.	.	.	.	35	5	.	40
o	.	.	.	.	.	.	.	.	.	.	40	.	40
u	.	.	.	.	.	.	.	.	.	.	.	40	40
sum	41	39	40	41	40	40	39	40	39	35	46	40	480

APPENDIX B: SUMMED CONFUSION MATRICES FROM THE PHONEME IDENTIFICATION EXPERIMENTS IN CHAPTER 3

TABLE B1. Summed confusion matrices for 12 subjects in the six conditions: Consonants.

stimulus/response HP 0 Hz

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	48	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48
k	3	45	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48
p	1	.	47	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48
b	.	.	.	48	.	.	.	.	.	.	.	.	.	.	.	.	.	48
d	.	.	.	.	48	.	.	.	.	.	.	.	.	.	.	.	.	48
f	.	.	.	.	.	38	.	.	9	.	1	.	.	.	.	.	.	48
s	.	1	.	.	.	.	41	.	.	6	.	.	.	.	.	.	.	48
χ	.	.	.	.	.	.	.	48	.	.	.	.	.	.	.	.	.	48
v	.	.	.	.	.	5	.	.	43	.	.	.	.	.	.	.	.	48
z	.	.	.	.	.	.	11	.	.	37	.	.	.	.	.	.	.	48
m	.	.	.	.	.	.	.	.	.	.	48	.	.	.	.	.	.	48
n	.	.	.	.	.	.	.	.	.	.	2	46	.	.	.	.	.	48
l	.	.	.	.	.	.	.	.	.	.	.	.	48	.	.	.	.	48
w	.	.	.	.	.	1	.	.	4	.	1	.	.	.	42	.	.	48
j	.	.	.	.	.	.	.	1	.	.	.	.	.	.	.	47	.	48
h	.	.	.	.	.	.	.	2	3	.	.	.	.	.	1	1	41	48
sum	52	46	47	48	48	44	52	51	59	43	52	46	0	48	43	48	41	768

stimulus/response HP 2 Hz

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	46	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48
k	2	41	4	.	.	.	.	1	.	.	.	.	.	.	.	.	.	48
p	4	.	43	.	.	1	.	.	.	.	.	.	.	.	.	.	.	48
b	.	.	.	47	1	.	.	.	.	.	.	.	.	.	.	.	.	48
d	.	.	.	.	48	.	.	.	.	.	.	.	.	.	.	.	.	48
f	.	.	.	.	.	29	3	2	14	.	.	.	.	.	.	.	.	48
s	.	.	.	.	.	.	36	.	.	12	.	.	.	.	.	.	.	48
χ	.	.	.	.	.	.	.	48	.	.	.	.	.	.	.	.	.	48
v	.	.	.	.	.	7	.	.	39	1	.	.	.	.	1	.	.	48
z	.	.	.	.	.	.	9	.	.	39	.	.	.	.	.	.	.	48
m	.	.	.	.	.	.	.	.	.	.	46	.	.	2	.	.	.	48
n	.	.	.	.	.	.	.	.	.	.	3	39	.	6	.	.	.	48
l	.	.	.	.	.	.	.	.	.	.	.	.	47	1	.	.	.	48
w	.	.	.	.	.	.	.	.	5	.	.	.	.	1	42	.	.	48
j	.	.	.	.	.	.	.	2	.	.	.	.	.	.	2	42	2	48
h	.	.	.	.	.	.	.	3	6	.	.	.	.	.	1	.	38	48
sum	52	43	47	47	49	37	48	56	64	52	49	39	0	56	47	42	40	768

TABLE B1. Continued.

stimulus/response HP 8 Hz

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	46	2	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	48
k	2	41	4	.	.	.	.	1	.	.	.	.	.	.	.	.	.	48
p	4	.	43	.	.	1	.	.	.	.	.	.	.	.	.	.	.	48
b	.	.	.	47	1	.	.	.	.	.	.	.	.	.	.	.	.	48
d	.	.	.	.	48	.	.	.	.	.	.	.	.	.	.	.	.	48
f	.	.	.	.	.	29	3	2	14	.	.	.	.	.	.	.	.	48
s	.	.	.	.	.	.	36	.	.	12	.	.	.	.	.	.	.	48
χ	.	.	.	.	.	.	.	48	.	.	.	.	.	.	.	.	.	48
v	.	.	.	.	.	7	.	.	39	1	.	.	.	.	1	.	.	48
z	.	.	.	.	.	.	9	.	.	39	.	.	.	.	.	.	.	48
m	.	.	.	.	.	.	.	.	.	.	46	.	.	2	.	.	.	48
n	.	.	.	.	.	.	.	.	.	.	3	39	.	6	.	.	.	48
l	.	.	.	.	.	.	.	.	.	.	.	.	47	1	.	.	.	48
w	.	.	.	.	.	.	.	.	5	.	.	.	.	1	42	.	.	48
j	.	.	.	.	.	.	.	2	.	.	.	.	.	.	2	42	2	48
h	.	.	.	.	.	.	.	3	6	.	.	.	.	.	1	.	38	48
sum	52	43	47	47	49	37	48	56	64	52	49	39	0	56	47	42	40	768

stimulus/response HP 32 Hz

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	35	7	1	1	.	1	1	2	.	.	.	.	.	.	.	.	.	48
k	3	21	5	.	.	2	.	13	1	.	.	.	.	.	.	.	3	48
p	.	1	37	1	.	3	.	4	1	.	.	.	.	.	.	.	1	48
b	.	.	.	43	.	.	.	1	3	.	.	.	.	.	1	.	.	48
d	.	1	.	2	41	.	.	1	1	.	.	.	.	1	.	.	1	48
f	.	.	4	.	.	11	1	20	11	.	.	.	.	.	.	.	1	48
s	.	.	.	.	.	1	20	13	1	11	.	.	.	.	.	.	2	48
χ	.	.	.	.	.	.	1	44	3	.	.	.	.	.	.	.	.	48
v	.	.	.	.	1	3	.	2	23	2	.	.	.	.	5	.	12	48
z	.	.	.	.	1	.	8	.	3	34	.	.	.	.	.	.	2	48
m	.	.	.	.	.	.	.	.	.	.	38	2	.	7	.	.	1	48
n	.	.	.	1	.	.	.	.	.	.	8	28	1	9	.	.	1	48
l	.	.	.	.	.	.	.	.	.	.	5	10	1	29	.	.	3	48
w	.	.	.	5	.	1	.	.	4	.	.	.	.	.	23	1	14	48
j	.	.	.	.	1	.	.	4	6	1	.	3	1	1	10	15	6	48
h	.	.	.	.	.	1	.	4	5	1	.	.	.	.	7	.	30	48
sum	38	30	47	53	44	23	31	108	62	49	51	43	3	47	46	16	77	768

TABLE B1. Continued.

stimulus/response HP 128 Hz

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	31	4	1	.	.	2	1	4	.	.	.	.	.	.	.	.	5	48
k	2	11	5	.	4	2	2	17	1	.	.	.	.	1	.	.	3	48
p	3	1	24	1	.	2	.	11	.	.	.	.	.	.	.	.	6	48
b	.	.	.	36	4	.	.	2	2	.	.	.	.	.	1	.	3	48
d	1	.	.	4	36	1	1	2	.	1	.	.	.	.	.	.	2	48
f	.	1	3	.	.	12	5	14	9	.	.	.	.	.	1	.	3	48
s	.	1	.	.	.	2	22	7	1	12	.	.	.	.	.	.	3	48
χ	.	.	1	1	.	5	2	28	2	3	.	.	.	.	.	.	6	48
v	.	.	.	.	.	1	2	2	20	1	.	.	.	.	4	1	17	48
z	.	.	.	.	.	.	6	.	3	32	.	.	.	.	1	2	4	48
m	.	.	.	.	.	.	.	.	.	.	19	5	.	14	3	.	7	48
n	.	.	.	.	.	1	.	1	.	.	5	18	1	17	1	1	3	48
l	.	.	.	.	.	.	.	.	.	.	.	7	.	32	.	.	9	48
w	.	.	.	3	2	1	1	.	3	.	.	.	1	.	23	1	13	48
j	.	.	.	.	2	1	.	3	5	1	.	3	.	1	7	13	12	48
h	.	.	.	2	1	.	.	1	3	1	.	1	.	.	7	2	30	48
sum	37	18	34	47	49	30	42	92	49	51	24	34	2	65	48	20	126	768

stimulus/response control

	t	k	p	b	d	f	s	χ	v	z	m	n	ŋ	l	w	j	h	sum
t	7	4	6	.	.	5	2	12	4	1	.	.	.	1	.	.	6	48
k	1	8	8	.	1	4	1	13	3	.	.	.	.	.	1	.	8	48
p	3	2	11	.	.	5	1	14	2	1	.	.	.	.	.	.	9	48
b	.	.	.	22	5	.	2	2	7	.	.	.	.	.	1	.	9	48
d	.	.	.	8	19	.	2	4	5	.	.	.	.	.	.	.	10	48
f	2	.	5	.	1	10	3	16	6	.	.	.	.	.	1	.	4	48
s	7	1	1	.	1	5	14	9	2	5	.	.	.	.	.	.	3	48
χ	3	.	4	.	.	4	3	23	3	.	.	.	.	.	.	.	8	48
v	.	.	.	7	2	.	1	2	10	1	.	.	.	.	7	1	17	48
z	.	.	.	2	3	.	5	.	2	17	.	.	.	2	6	.	11	48
m	.	.	.	.	.	.	.	.	2	.	18	1	.	20	1	.	6	48
n	.	.	.	.	.	.	.	1	1	.	6	9	.	20	1	4	6	48
l	.	.	.	.	.	.	1	.	.	.	4	4	.	26	2	.	11	48
w	.	.	.	4	2	1	.	.	4	.	.	.	.	1	16	.	20	48
j	.	.	.	.	.	.	.	.	5	1	.	2	.	1	12	13	14	48
h	.	.	.	.	.	1	1	2	4	1	.	.	.	.	6	1	32	48
sum	23	15	35	43	34	35	36	98	60	27	28	16	0	71	54	19	174	768

TABLE B2. Summed confusion matrices for 12 subjects in the six conditions: Vowels.

stimulus/response HP 0 Hz

	ɑ	a	ɑu	ɛ	e	ɛi	ø	I	i	ɔ	o	u	sum
ɑ	46	1	.	.	.	.	.	1	.	.	.	.	48
a	3	43	2	.	.	.	.	.	.	.	.	.	48
ɑu	2	.	46	.	.	.	.	.	.	.	.	.	48
ɛ	.	.	.	47	.	1	.	.	.	.	.	.	48
e	.	.	.	.	46	.	1	1	.	.	.	.	48
ɛi	.	.	.	.	.	48	.	.	.	.	.	.	48
ø	.	.	.	.	.	.	48	.	.	.	.	.	48
I	.	.	.	1	2	.	.	45	.	.	.	.	48
i	.	.	.	.	.	.	.	6	41	.	.	1	48
ɔ	.	.	.	.	.	.	.	.	.	39	8	1	48
o	.	.	.	.	.	.	.	.	.	2	46	.	48
u	.	.	.	.	.	.	.	.	.	.	1	47	48
sum	51	44	48	48	48	49	49	53	41	41	55	49	576

stimulus/response HP 2 Hz

	ɑ	a	ɑu	ɛ	e	ɛi	ø	I	i	ɔ	o	u	sum
ɑ	44	2	1	.	.	.	.	.	.	1	.	.	48
a	8	35	4	.	.	1	.	.	.	.	.	.	48
ɑu	1	.	46	.	.	.	.	.	.	1	.	.	48
ɛ	.	.	.	47	.	.	.	.	1	.	.	.	48
e	.	.	.	.	42	.	3	3	.	.	.	.	48
ɛi	.	.	.	.	.	48	.	.	.	.	.	.	48
ø	.	.	.	.	.	.	48	.	.	.	.	.	48
I	.	.	.	2	1	.	.	44	1	.	.	.	48
i	.	.	.	.	2	.	.	3	43	.	.	.	48
ɔ	.	.	.	.	.	.	.	.	.	39	9	.	48
o	.	.	.	.	.	.	.	.	.	3	45	.	48
u	.	.	.	.	.	.	.	.	.	.	2	46	48
sum	53	37	51	49	45	49	51	50	45	44	56	46	576

stimulus/response HP 8 Hz

	ɑ	a	ɑu	ɛ	e	ɛi	ø	I	i	ɔ	o	u	sum
ɑ	46	2	.	.	.	.	.	.	.	.	.	.	48
a	6	42	.	.	.	.	.	.	.	.	.	.	48
ɑu	18	1	28	.	.	.	.	.	.	.	.	1	48
ɛ	.	.	.	45	.	2	1	.	.	.	.	.	48
e	.	.	.	1	31	1	3	12	.	.	.	.	48
ɛi	.	.	1	14	.	33	.	.	.	.	.	.	48
ø	.	.	.	.	.	.	48	.	.	.	.	.	48
I	.	.	.	1	.	.	1	45	1	.	.	.	48
i	.	.	.	.	1	.	.	11	35	.	.	1	48
ɔ	1	.	.	.	.	.	.	.	.	30	9	8	48
o	.	.	.	.	.	.	.	.	.	11	35	2	48
u	.	.	.	.	.	.	.	.	.	.	.	48	48
sum	71	45	29	61	32	36	53	68	36	41	44	60	576

TABLE B2. Continued.

stimulus/response HP 32 Hz

	ɑ	a	ɑu	ɛ	e	ɛi	ø	I	i	ɔ	o	u	sum
ɑ	32	13	1	.	.	.	.	.	.	1	.	1	48
a	4	40	3	1	.	.	.	.	.	.	.	.	48
ɑu	20	7	18	.	.	.	.	.	.	1	2	.	48
ɛ	1	.	.	44	.	.	.	2	.	.	1	.	48
e	.	.	.	1	33	.	1	10	3	.	.	.	48
ɛi	.	.	.	30	1	17	.	.	.	.	.	.	48
ø	.	.	.	.	.	.	48	.	.	.	.	.	48
I	.	.	.	3	1	1	.	39	4	.	.	.	48
i	.	.	.	1	.	.	.	3	43	.	.	1	48
ɔ	.	.	.	.	.	.	.	.	.	18	14	16	48
o	.	.	.	.	.	.	.	.	.	9	37	2	48
u	.	.	.	.	.	.	1	.	.	2	.	45	48
sum	57	60	22	80	35	18	50	54	50	31	54	65	576

stimulus/response HP 128 Hz

	ɑ	a	ɑu	ɛ	e	ɛi	ø	I	i	ɔ	o	u	sum
ɑ	33	9	2	.	.	.	.	.	1	3	.	.	48
a	4	40	1	2	.	.	1	.	.	.	.	.	48
ɑu	22	7	12	.	.	.	1	.	.	3	3	.	48
ɛ	1	1	.	37	1	1	1	3	.	1	1	1	48
e	.	.	.	5	24	.	2	15	2	.	.	.	48
ɛi	.	.	.	33	1	13	.	.	.	1	.	.	48
ø	.	.	.	3	.	.	43	1	.	1	.	.	48
I	.	.	.	2	1	.	7	29	7	.	.	2	48
i	.	.	.	.	.	.	4	8	34	.	.	2	48
ɔ	.	.	2	.	1	.	5	1	.	20	11	8	48
o	.	.	1	.	1	.	2	.	.	11	31	2	48
u	1	1	.	.	.	.	4	1	3	2	.	36	48
sum	61	58	18	82	29	14	70	58	47	42	46	51	576

stimulus/response control

	ɑ	a	ɑu	ɛ	e	ɛi	ø	I	i	ɔ	o	u	sum
ɑ	26	14	3	.	.	.	.	.	.	4	.	1	48
a	5	40	1	2	.	.	.	.	.	.	.	.	48
ɑu	21	7	14	.	.	.	2	.	.	.	4	.	48
ɛ	.	.	.	39	1	1	.	2	.	.	.	5	48
e	.	.	.	1	21	.	3	22	1	.	.	.	48
ɛi	1	.	.	37	1	9	.	.	.	.	.	.	48
ø	.	.	.	3	1	.	44	.	.	.	.	.	48
I	.	.	.	3	1	.	13	23	6	2	.	.	48
i	.	.	.	.	.	.	2	11	35	.	.	.	48
ɔ	.	1	.	.	.	.	3	.	.	16	14	14	48
o	.	.	.	.	.	.	2	.	.	12	33	1	48
u	.	.	.	1	.	.	6	2	1	1	1	36	48
sum	53	62	18	86	25	10	75	60	43	35	52	57	576

APPENDIX C: RATIONALE OF THE PHASE-LOCKED MTF

Let x(t) and y(t) be the temporal intensity envelopes for input and output, respectively, with spectra X(f) and Y(f). The MTF is generally defined as the amplitude spectrum of the frequency response function H(f) of a transmission channel, that is

$$\mathrm{MTF} = |H(f)| = \alpha \frac{|Y(f)|}{|X(f)|} = \alpha \left(\frac{S_{yy}(f)}{S_{xx}(f)}\right)^{\frac{1}{2}}, \qquad \text{(C1)}$$

where $S_{xx}(f)$ and $S_{yy}(f)$ denote the autospectra (autospectral density functions) of x(t) and y(t) and α a normalization factor based on the mean intensities of x(t) and y(t). In this approach the phase information of the modulation components is not included. As a consequence, modulation frequencies introduced by the processing that are not in phase with the original envelope are wrongly considered to be transferred, thus underestimating the amount of reduction. In order to measure the transfer of only the original (reduced) intensity modulations a different frequency response $H_1(f)$ should be computed, using the cross-spectrum (cross-spectral density function) $S_{xy}(f)$ between x(t) and y(t) (Bendat and Piersol, 1980),

$$H_1(f) = \alpha \frac{S_{xy}(f)}{S_{xx}(f)} = \frac{C_{xy}(f) - jQ_{xy}(f)}{S_{xx}(f)}. \qquad \text{(C2)}$$

Equation (C2) gives the optimum linear frequency response, i.e., the frequency response which minimizes the nonlinearities in y(t) with respect to x(t). On the right hand side of Eq. (C2) $C_{xy}(f)$ denotes the cospectrum, which is the part where X(f) and Y(f) are in-phase, whereas the imaginary part $Q_{xy}(f)$ (quadspectrum) is the part where X(f) and Y(f) are 90° out of phase. So, assuming that only in-phase modulations are relevant for speech intelligibil-ity, the optimum estimate of the modulation transfer will be given by the so-called phase-locked modulation transfer function MTF_{pl}, defined as

$$\mathrm{MTF}_{pl} = \alpha \frac{C_{xy}(f)}{S_{xx}(f)}. \qquad \text{(C3)}$$

For the actual computation of both the normal MTF and the phase-locked MTF_{pl} (cf. Fig. 3.7 and 3.8), the long-term spectra of the (downsampled) squared amplitude envelopes of an octave-band filtered 71-s speech fragment (viz. 30 concatenated sentences of experiment 1 in Chapter 3) were used. The spectra were obtained by computing the average of 1024-point fast Fourier transforms (0.24-Hz resolution), using 50% overlap and Hanning weighting (Carter *et al.*, 1973).