Perception of Speech as a Modulated Signal

Reinier Plomp
Soesterberg and Amsterdam, the Netherlands

1. Introduction

In this contribution on the significance of hearing for speech research I would first like to demarcate what will be discussed. It is not my purpose to present a review of the state of the art of psychoacoustics as far as it is relevant to speech perception. Instead of this the attention will be focused on a comparison of the physical properties of the speech signal with the limits of the ear's capacity to handle this signal. To avoid misunderstandings, the auditory processes to be considered do not include the way in which the signal is decoded in phonetic terms. Therefore, the controversy of whether or not there is a phonetic mode will not bother us (for a recent review, see Repp, 1982).

The auditory system constitutes an important link in the transfer of information from the original speech signal produced by the speaker to the understanding of its message by the listener. Since speech intelligibility in general is the main criterion of the successfulness of this transfer, it will play an important role in our discussion. This means that the present approach does not consider the intelligibility, or any other property, of the individual speech sounds.

As will be shown, it is worthwhile to include in our considerations also the transfer of the speech signal from the mouth of the speaker to the ear of the listener, as far as the reverberation is concerned. This link in the chain is too often neglected in phonetics.

2. Analysis of Speech in Terms of Modulation Frequency

It is common in phonetics to describe the speech signal in terms related to its production, such as formant frequency and place of articulation. This results in a description specific for speech, but not generally applicable to all types of sound relevant in everyday listening. It will be clear that we cannot use such a production- oriented description in studying the transfer of the speech signal on its way from the speaker to the listener; the measures adequate in room acoustics (e.g. reverberation time, sound-absorption coefficient, volume, distance) are general, physical, quantities.

A similar argument holds for the auditory system in the strict sense of the term. Without going into details (see Plomp, 1976), we may compare this system with a set of (overlapping) band-pass filters tuned to different frequencies from low to high. The three main perceptual attributes of sounds: pitch, loudness, and timbre, are strongly related to the physical properties of fundamental frequency, intensity, and frequency spectrum, respectively. (For experimental evidence that sounds with different fundamental frequencies but equal absolute frequency spectra are very similar in timbre, see Plomp and Steeneken, 1971). Together, these three physical quantities given an adequate description of a speech vowel as a periodic vibration (fundamental frequency F_0) with a specific amplitude pattern of the harmonics. This approach also holds reasonably well for the voiced consonants (sonorants); in the nonsonorants the periodic vibration is replaced by a noise-like source. Thus, the speech signal can be interpreted as a carrier (periodic vibration or noise) with a frequency spectrum modulated continuously in time by the vocal tract. Although the temporal variations of F_0 contribute to speech intelligibility, this contribution is relatively small and will be left out of consideration here. This means that we will confine ourselves to the frequency spectrum and how it varies in time.

The significance of this perception-oriented approach can be demonstrated by means of the spectrogram. In a spectrogram the frequency spectrum measured with a set of band-pass filters is plotted as a function of time. It does not represent the fine structure of the signal (instantaneous amplitude) but gives the temporal intensity envelope for the different frequency bands. The important role the spectrogram has played in phonetic research during the last decades may be regarded as a demonstration that the spectrogram is an excellent representation of the information-bearing characteristic of speech.

This conclusion suggests that it makes sense to analyze the speech-signal envelope, reproduced in the spectrogram, in terms of sinusoidal components, as is usual in systems analysis. This analysis should be distinguished from the traditional frequency analysis in terms of audio frequencies. The spectrogram gives the intensity envelope both in time (horizontal) and in frequency (vertical) and it is these two envelopes that should be analyzed. The resulting frequencies are not audio frequencies but modulation frequencies describing the temporal and spectral variations. (In order to avoid confusion between these two types of frequencies, the prefixes 'audio' and 'modulation' will be frequently used.)

As a further illustration of what is important in the transfer of the speech signal, let us consider the spectrogram as an optical image to be transferred. It will be clear that its finer details, quite essential for identifying specific phonemes, are lost if the spectrogram is reproduced on a TV screen or as a newspaper picture with large dots; the medium, including the eye, should be able to preserve the relevant details. In recent years it has become common to quantify the quality of the image transfer by means of the spectral modulation transfer function, representing the faithfulness with which spatial sinusoidal brightness variations are preserved as a function of modulation frequency. Similarly, the transfer of speech as a sound signal can be quantified by means of temporal (in Hz) and spectral (for example, in periods/octave) modulation frequencies. Hence it makes sense to study the speech signal as radiated from the mouth in terms of modulation frequencies.

Since on temporal modulation rather more data are available than on spectral modulation, I will start by discussing what sinusoidal modulations in time are present in speech signals, the effect of reverberation on the transfer of these modulations from the speaker to the listener, and the way in which the limited capacity of the ear to perceive modulations can be expressed, too, by a modulation transfer function. Subsequently, the same points will be considered for the spectral modulations.

3. Temporal Modulation

3.1. The temporal envelope of the speech signal

The intensity of a speech signal as a function of time can be measured with the aid of a squaring circuit followed by a low-pass filter with a cut-off frequency of, say, 50 Hz. In this way a signal is obtained only determined by the fluctuating envelope, not by the fine structure (viz. the audio frequencies) of the speech signal. Figure 1 illustrates such an intensity envelope for a speech fragment of 10 sec; the dashed line represents the average intensity, $\overline{I}$, of this signal.

By means of a frequency analysis of the intensity envelope function of the speech signal the relative importance of different modulation frequencies can be determined. Steeneken and Houtgast (1983) analyzed one-minute speech

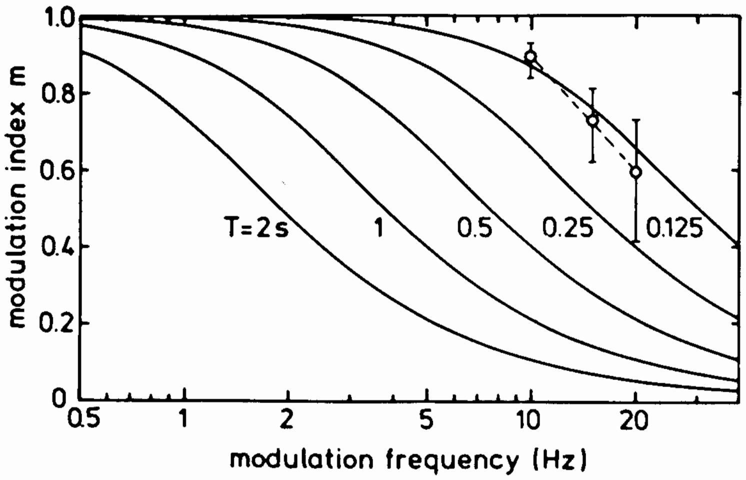

Figure 1 Intensity envelope for a 10-sec speech fragment.

fragments of connected discourse from ten male and ten female speakers who read the same text. By speeding up the envelope function by a factor of 400, they obtained a signal fluctuating in the normal audi-frequency range. This signal was analyzed with a set of one-third octave band-pass filters. Since the modulation index, m, is an appropriate measure for specifying modulation transfer functions, as we will see below, this measure will also be used for the speech signal. It is defined as the average peak value of the filter output amplitude divided by the average value of the unfiltered signal, $\overline{I}$; in Fig. 1 the sinusoids represent the case m=1.

In Fig. 2 the modulation index, averaged over ten male speakers, is plotted as a function of modulation frequency, F (centre frequency of the 1/3-oct band-pass filters). The 1-min speech segments were first analyzed in terms of audio frequencies by means of octave band-pass filters; then for each octave band the temporal intensity envelope was analyzed in terms of modulation frequencies. We see that the resulting curves are rather similar, except for their vertical positions. For all audio frequencies the most important modulation frequencies are 3-4 Hz, related to the number of words/syllables pronounced per sec. With ten female speakers, as well as with other texts, almost the same results were found.

Adopting as a criterion the modulation frequency for which the modulation index is reduced to half its peak value, we see that modulation frequencies are present in speech up to about 15 Hz.

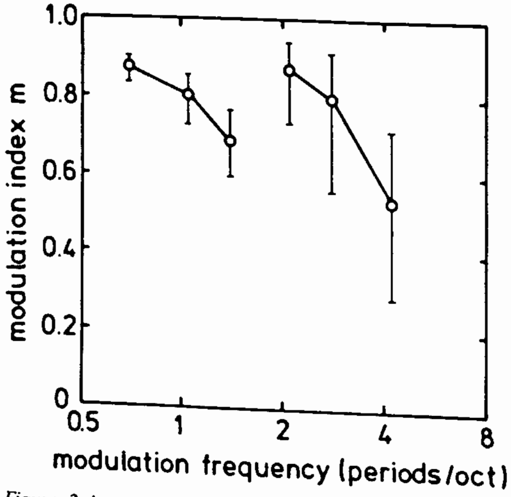

Figure 2 Average temporal envelope spectra in terms of the modulation index for one-min connected discourse from ten male speakers. The number of sentence units, stressed syllables etc. per sec is indicated. The parameter is the centre frequency of the one-oct audio-frequency band.

3.2. The speaker-to-listener modulation transfer function in rooms.

The transfer of the speech signal from the speaker to the listener depends in an enclosed space on the reverberation characteristics of the room. The sound travels from the mouth of the speaker to the ears of the listener via a great many different transmission paths. At a short distance the direct path may be the most important one, but for most positions the sound level will be determined by the indirect paths, each including a number of reflections. Since the time of arrival is given by the total lentgh of the path, the differences in path length result in a blurring of fine temporal details of the speech signal's envelope, comparable with the optical effect of a lens system that is out of focus. In terms of modulation, it means that slow sinusoidal modulations are well preserved, whereas fast modulations are attenuated.

This(modulation)frequency-dependent behaviour of an enclosure can be expressed in the so-called temporal modulation transfer function (TMTF). For an input signal (band of noise) with a 100-% sinusoidally varying intensity, $\overline{I}_i\,(1 + \cos 2\pi Ft)$, at the position of the speaker, the output signal at the listener's ear is given by

$$\overline{I}_o\ [1+m\cos(2F\pi t-\varphi)] \qquad (1)$$

where F= modulation frequency and m as a function of F is defined as the TMTF.

Since the TMTF was first introduced in room acoustics by Houtgast and Steeneken (1973) as a predictor of speech intelligibility, various ways of deriving m(F) from the room parameters have been explored. In the simplest case the sound field is considered as the result of a statistical process of sound reflections without any directional preference, the direct path being excluded. For such a diffuse sound field m(F) is given by

$$m(F) = (1+0.207F^2T^2)^{-1/2} \qquad (2)$$

where T = reverberation time in sec, the time in which a sound decays by 60 dB (Houtgast, Steeneken and Plomp, 1980; Schroeder, 1981. In Fig.3 this equation is plotted as a function of modulation frequency, with T as the parameter. The figure illustrates the blurring effect of high modulation frequencies and the significant role of reverberation time. The reverberation times in everyday life are largely restricted to the range from T= .5 sec (typical living room) to T=2 sec (good concert hall).

For a more accurate determination of the TMTF the statistical approach has to be replaced by a geometrical approach based on the exact dimensions of the room, the positions of the speaker and the listener, and the sound absorption properties of the different boundary surfaces. For a rectangular room the algorithm has been given by Plomp, Steeneken and Houtgast (1980), for a room with oblique walls shaped as a trucated pyramid by

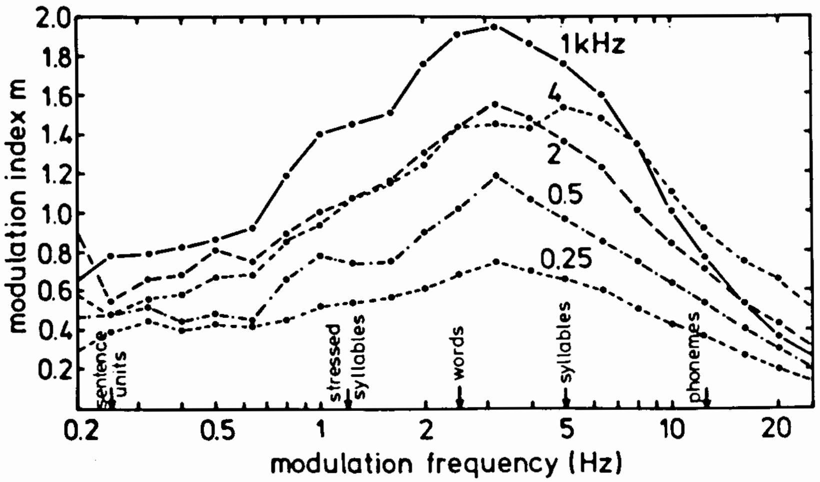

Figure 3 Temporal modulation transfer function in the diffuse sound field with the reverberation time of the room as the parameter. The data points (mean value and standard deviation for 50 subjects) represent the TMTF of the auditory system for 1000 Hz.

Wattel et al. (1982). Whereas those algorithms were based on mirror images, the more general approach by means of ray-tracing has been presented by van Rietschote, Houtgast and Steeneken (1981). In all these models the presence of noise was also taken into account. It has been shown (Houtgast et al. 1980) that from the TMTF over the modulation frequency range 0.4-40 Hz a single measure can be derived, the Speech Transmission Index (STI), which is an excellent predictor of speech intelligibility; as has been verified, this holds generally for Western languages (Houtgast and Steeneken, 1983). This means that the algorithms for computing the TMTF from the room parameters are important tools in designing acoustically good classrooms, conference rooms, auditoria, etc. The TMTF concept has been successfully extended to include band-limiting, nonlinear distortion, and other disturbances of the speech signal which may be present in communication channels (Steeneken and Houtgast, 1980).

4. The temporal modulation transfer function of the ear

In the same way as an enclosure the ear may be regarded as a link in the speech transmission chain that is able to transfer faithfully slow variations of sound intensity but is unable to follow fast modulations. If we could be sure that the ear is linear for modulations, we could derive the TMTF from the just-noticeable intensity modulation as a function of modulation frequency (Viemeister, 1979), but this assumption is not justified. Since we cannot measure directly at the output of the auditory system, we have to use an external test sound to investigate the ear's modulation transfer function. This can be done by measuring the detection thresholds for a very short test sound

coinciding in time with the peaks of the sinusoidally intensity-modulated stimulus and with the valleys of this signal, respectively. The peak-to-valley difference, ΔL, in dB, can be translated in the modulation index m.

$$\Delta L = 10\log\frac{1+m}{1-m},\ \text{or}\ m = \frac{10^{\Delta L/10}-1}{10^{\Delta L/10}+1}. \qquad (3)$$

A condition for this approach is that detection exclusively depends on signal-to-noise ratio which is true over a large intensity range.

In the literature only few data on peak-to-valley differences are reported (Rodenburg, 1977; Festen et al., 1977; Festen and Plomp, 1981). In the experiment by Festen and Plomp, the sinusoidally modulated sound was white noise low-pass filtered with a cutoff frequency of 4000 Hz. In a two-alternative forced-choice procedure the detection threshold of a 0.4-msec click, octave-filtered around 1000 Hz, was measured for modulation frequencies of 10, 15, and 20 Hz. The results, averaged over 50 normal-hearg subjects, are plotted in Fig. 3.

Taking the frequency for which m=0.5 as a measure of the limit up to which the ear is able to follow temporal modulations, we arrive at a value of about 25 Hz. It is of interest that in the first channel vocoders, already more than 50 years ago, intensity fluctuations up to 25 Hz were considered to be important in speech perception (Flanagan 1965, p. 246).

This experiment leaves unanswered the question of the degree to which the ear's TMTF may depend on audio frequency. The scarce data on the threshold for just-noticeable modulations strongly suggest that for frequencies as low as 250 Hz the TMTF shifts (maybe by as much as a factor of two) to lower frequencies, with the reverse holding for frequencies as high as 4000 Hz (Viemeister, 1979).

Figure 3 allows us to express the ear's sensitivity to modulations in reverberation time, resulting in an estimate of T=0.12-0.15 sec. Since for rooms the reverberation time is almost always at least 0.4 sec, it is clear that in everyday situations the room rather than the ear is the limiting factor in our ability to perceive temporal intensity fluctuations of sounds.

In this derivation of the TMTF it has been taken for granted that the role of phase in the transfer of temporal modulations is negligible. If a room modified the phase relation between the various modulation components, this should affect speech intelligibility. For the diffuse sound field underlying the curves of Fig. 3, the phase shift, relative to F=0Hz, is increased to only 45° at the modulation frequency for which m=0.5, with an asymptote of 90° at high modulation frequencies. Experiments by Viemeister (1977) indicate that for the auditory system, too, the phase shift may be neglected for the range of modulation frequencies relevant in speech perception.

5. Spectral Modulation

5.1. The spectral envelope of the speech signal

Analogously to the case of temporal modulations, we would like to analyze spectral speech envelopes in terms of modulation frequencies. Since no data for a speech fragment of which the frequency spectrum is measured periodically in time are available, we have to estimate the upper limit of spectral modulation frequencies from audio-frequency spectra of individual speech sounds. Both in view of their temporal prominence in speech and their peakedness, vowels are most appropriate for investigating this upper limit.

The spectral envelopes of vowels are characterized by a series of formants, of which the lower three are the most important ones. In addition to their frequency and level, these formants are described by their bandwidth and their interdividual spread in frequency.

Experiments by Dunn (1961) have shown that, in the mid-frequency range (800-2500 Hz), formant bandwidth is, roughly, about 6%. Assuming triangular spectral formant shapes, this implies that two formants have to differ about 15% in order to be separated by a spectral valley of 4.77 dB, corresponding to m=0.5 (equation 3). This peak distance of 15% determines the upper modulation frequency present in speech spectra, equal to about 5 periods/octave.

This value should be considered as an upper estimate, excluding the interindividual spread in formant frequencies. Since speech recognition is based on the absolute rather than the relative position of the spectrum along the frequency scale, it is reasonable to take the interindividual spread into account. For male vowel spectra the standard deviation of formant frequencies is about 10% (Pols, Tromp and Plomp, 1970) which means that 68% of the peaks are within a range of 20% around the average frequency for that particular formant. Interpreting this 20% as a bandwidth to be combined with the 6% of the formant bandwidth, we arrive at a lower estimate of the limit of modulation frequencies present in speech of about 1 period/octave. On the basis of 1/3- octave vowel spectra, the same lower estimate has been obtained (Plomp, 1983).

6. The transfer from the speaker to the listener in a room

The fact that sounds reach the ear via a great many different transmission paths does not only influence the temporal modulations present in the speech signal at the listener's position, but also its spectrum. For steady-state pure tones the sound pressure level at a large distance from the speaker has a theoretical uncertainty with a standard deviation of 5.57 dB (Schroeder, 1954). Measurements at a great many locations in a concert hall have confirmed this value (Plomp and Steeneken, 1973). This uncertainty is a consequence of the vectorial addition of sound waves with random phases; it

is inherent in a diffuse sound field and cannot be reduced by acoustical measures.

In order to get some insight into the effect of this 'noisy' character of the transfer of sound in a room on the speech signal, we can compare it with the interindividual spread in vowel spectra. Using data from Klein, Plomp and Pols (1970) it was found that the spectral variance due to reverberation is about as large as the interindividual spectral variance for male speakers pronouncing the same vowel (Plomp and Steeneken, 1973).

7. The spectral modulation transfer function of the ear

The sum of white noise and its replica by τ sec results in so-called comb-filtered noise with an intensity varying sinusoidally along the frequency scale at a frequency of 1/τ Hz (thus a delay of 5 msec gives noise with peaks at distances of 200 Hz). Similarly as temporally modulated noise, this signal can be used for measuring the peak- to-valley difference for a test tone at a fixed frequency.

With the same group of 50 normal-hearing subjects for which the TMTF was measured (Fig. 3), the spectral modulation transfer function (SMTF) with comb-filtered noise was also investigated. Short 1000-Hz test tones (duration 15 msec) were presented either during or immediately after 500-msec noise bursts. These two conditions, simultaneous masking, were chosen because they result in different values of the ear's SMTF, as the results in Fig. 4 show. This difference is attributed to a sharpening mechanism (lateral suppression, comparable with Mach bands in vision) not effective in simultaneous masking (for more details see Houtgast, 1974; Plomp, 1976).

According to Fig. 4, the ear is able to detect spectral modulations up to a

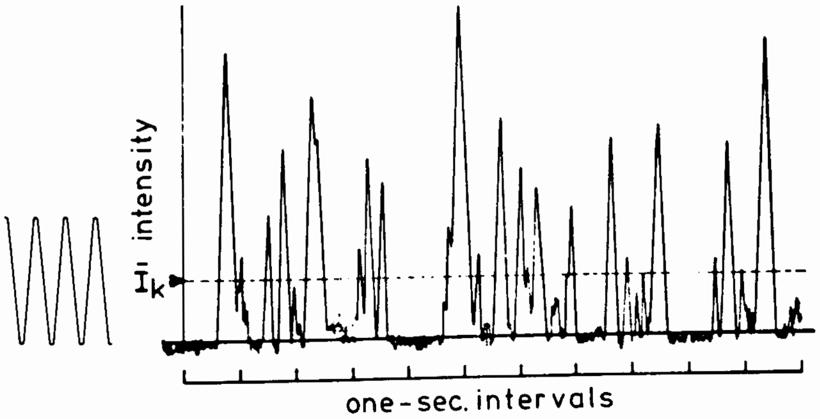

Figure 4. Spectral modulation transfer function (mean value and standard deviation for 50 subjects) of the auditory system for 1000 Hz. The left-hand curve holds for simultaneous masking, the right-hand curve for forward masking.

limit of, on the average, about 4 periods/octave, with lateral suppression included, and of about half this value for the condition without the effect of lateral suppression. From experiments by Houtgast (1974) we may conclude that these figures, measured for 1000 Hz, hold generally over the speech-frequency range. The diagram indicates that for the first-mentioned condition the interindividual differences are rather large; this confirms the finding in other experiments that subjects differ considerably in the effectiveness of the ear's sharpening mechanism.

8. Effect of Hearing Impairment

In the previous sections we have discussed values for the upper limits of temporal and spectral modulation frequencies present in the speech signal as well as the upper limit of the ear's capacity to detect modulations. These data can be summarized as follows:

Type of modulation	Speech signal	Hearing	
		Normal	Impaired
Temporal (Hz)	15	25	10
Spectral (per/oct)	1 to 5	2 to 4	1.4 to 1.6

From this table we may conclude that speech and normal hearing are well matched; the modulations present in speech cover the frequency range over which the ear is able to follow these modulations.

This agreement is disturbed in the case of hearing impairment. The right most column of the table gives average values for a group of 22 hearing-impaired subjects with moderate hearing losses (30 to 60 dB for 1000 Hz) of sensorineural origin. These recent data from Festen and Plomp (1983) were gathered with the same experimental procedures as used in the case of normal hearing.

We see that hearing-impaired subjects are, on the average, unable to hear the fast temporal intensity variations present in speech. According to the TMTF curves of Fig. 3, their hearing handicap in a reflection-free room is , in this respect, comparable with the situation for normal-hearing listeners in a room with a reverberation time of about 0.4 sec.

The table shows that not only the ear's TMTF but also its SMTF is reduced by the hearing impairment. This reduction is much stronger for the nonsimultaneous condition than for the case of simultaneous masking resulting in almost equal values in periods/octave. This suggests that lateral suppression is rather susceptible to hearing loss.

It is well-known that hearing-impaired subjects have special difficulties in understanding speech in noisy and reverberant environments (cf. Plomp and Mimpen, 1979; Plomp and Duquesnoy, 1980; Duquesnoy, 1982). We may conclude that, at least partly, these difficulties are due to the reduction in the ear's upper limits for the detection of temporal and spectral modulations. There are indications that by compensating for the ear's decrease in sensitivity to temporal modulations, speech intelligibility can be enhanced (Lewien, 1982; see Schroeder's contribution to this congress).

References

Dunn, H.K. (1961). Methods of measuring vowel formant bandwidths. *Journal of the Acoustical Society of America,* **33,** , 1737-1746.

Festen, J.M. et al. (1977). Relations between interindividual differences of auditory functions. In: *Psychophysics and Physiology of Hearing.* (E.F. Evans and J.P. Wilson Eds.) London: Academic Press, 311-319.

Festen, J.M. and Plomp, R. (1981). Relations between auditory functions in normal hearing. *Journal of the Acoustical Society of America,* **70,** 356-369.

Festen, J.M. and Plomp, R. (1983). Relations between auditory functions in impaired hearing. *Journal of the Acoustical Society of America,* **73,** 652-662.

Flanagan, J.L. (1965). Speech Analysis, Synthesis and Perception. Berlin: Springer Verlag.

Houtgast, T. (1974). *Lateral Suppression in Hearing.* Doctoral Dissertation, Free University Amsterdam.

Houtgast, T. and Steeneken, H.J.M. (1973). The Modulation Transfer Function in room acoustics as a predictor of speech intelligibility. *Acustica,* **28,** 66-73.

Houtgast, T. and Steeneken, H.J.M. (1983). A multi-language evaluation of the RASTI method for estimating speech intelligibility in auditoria. *Acustica,* in press.

Houtgast, T., Steeneken, H.J.M. and Plomp, R. (1980). Predicting speech intelligibility in rooms from the Modulation Transfer Function, I. General room acoustics. *Acustica,* **46,** 60-72.

Klein, W., Plomp, R. and Pols, L.C.W. (1970). Vowel spectra, vowel spaces, and vowel identification. *Journal of the Acoustical Society of America,* **48,** 999-1009.

Lewien, T. (1982). *Filterung von Spracheinhüllenden zur Verständlichkeitsverbesserung bei Innenohr-schwerhörigkeit.* Doctoral Dissertation, Georg-August-University, Göttingen.

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

Plomp, R. (1983). The role of modulation in hearing. In: *Hearing-Physiological Bases and Psychophysics.* (R. Klinke and R. Hartmann, Eds.) Berlin: Springer Verlag, 270-275.

Plomp, R. and Steeneken, H.J.M. (1971). Pitch versus timbre. In: *Proceedings Seventh International Congress on Acoustics, Budapest, Vol.* **3,** 377-380.

Plomp R. and Steeneken, H.J.M. (1973). Place dependence of timbre in reverberant sound fields. *Acustica,* **28,** 50-59.

Plomp, R., Steeneken, H.J.M. and Houtgast, T. (1980). Predicting speech intelligibility in rooms from the Modulation Transfer Function. II. Mirror image computer model applied to rectangular rooms. *Acustica,* **46,** 73-81.

Pols, L.C.W., Tromp, H.R.C., and Plomp, R. (1970). Frequency analysis of Dutch vowels from 50 male speakers. *Journal of the Acoustical Society of America,* **53,** 1093-1101.

Repp, B.H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin,* **92,** 81-110.

Rietschote, H.F. van, Houtgast, T. and Steeneken, H.J.M. (1981). Predicting speech intelligibility in rooms from the Modulation Transfer Function. IV. A ray-tracing computer model. *Acustica,* **49,** 245-252.

Rodenburg M. (1977). Investigation of temporal effects with amplitude modulated signals. In: *Psychophysics and Physiology of Hearing.* (E.F. Evans and J.P. Wilson, Eds.). London: Academic Press. 429-437.

Schroeder, M. (1954). Die Statistischen Parameter der Frequenzkurven von grossen Räumen. *Acustica,* **4** , 594-600.

Schroeder, M.R. (1981). Modulation transfer functions: Definition and measurement. *Acustica,* **49,** 179-182.

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

Steeneken, H.J.M. and Houtgast, T. (1983). The temporal envelope spectrum of speech and its significance in room acoustics. In: *Proceedings of the Eleventh International Congress on Acoustics*, Paris, **7**, 85-88.

Viemeister, N.F. (1977). Temporal factors in audition: A system analysis approach. In: *Psychophysics and Physiology of Hearing.* (E.F. Evans and J.P. Wilson, Eds.). London: Academic Press, 419-427.

Viemeister, N.F. (1979). Temporal modulation transfer functions based upon modulation thresholds. *Journal of the Acoustical Society of America,* **66,** 1364-1380.

Wattel, E., Plomp, R., Rietschote, H.F. van, and Steeneken, H.J.M. (1981). Predicting speech intelligibility in rooms from the Modulation Transfer Function. III. Mirror image computer model applied to pyramidal rooms. *Acustica,* **48,** 320-324.