on measuring and predicting speech intelligibility

H. J. M. STEENEKEN

On measuring and predicting speech intelligibility

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam,
op gezag van de Rector Magnificus
prof. dr. P.W.M. de Meijer,
in het openbaar te verdedigen in de Aula der Universiteit
(Oude Lutherse Kerk, ingang Singel 411, hoek Spui),
op donderdag 11 juni 1992 te 15.00 uur

door

Herman Jacobus Marie Steeneken

geboren te Delft

Faculteit der Letteren

Promotiecommissie:

promotor:	prof.dr.ir. L.C.W. Pols
co-promotor:	dr.ir. T. Houtgast
overige leden:	prof.dr.ir. F.A. Bilsen
	prof.dr. E. de Boer
	prof.dr.ir. R. Plomp
	dr. M.E.H. Schouten
	prof.dr. G.F. Smoorenburg

De uitgave van dit proefschrift is ondersteund door het Instituut voor Zintuigfysiologie-TNO te Soesterberg.

Omslagontwerp: Ilse Houtgast
Druk: Bariet B.V. Ruinen
ISBN: 90-6743-209-1

Ter nagedachtenis aan mijn ouders.
Aan mijn (geduldig) gezin.

VOORWOORD

Het onderzoek dat in dit proefschrift wordt beschreven is uitgevoerd op het Instituut voor Zintuigfysiologie TNO, te Soesterberg. Dit betekent veel. Onderzoeken en adviseren kan men niet alleen; elk produkt (publikatie, rapport en advies) is het werk van meerdere mensen. Ook het hier beschreven onderzoek, door mij neergelegd in deze wetenschappelijke verantwoording, vertegenwoordigt de bijdrage van een onderzoeksgroep die met een zeer beperkt aantal medewerkers op velerlei gebied actief is en internationale erkenning geniet.

Met Tammo Houtgast, Louis Pols, Johan Riemersma en Guido Smoorenburg heb ik vele discussies mogen voeren die als "fertilizer" voor de rijping van het onderzoek kunnen worden beschouwd. Ik meen dat zowel de wetenschappelijke opvoeding door Reinier Plomp als de ruimte die hij mij gaf tot zelfontplooiing mede tot dit proefschrift hebben geleid.

Zeer direct waren Evert Agterhuis en Hans van Raaij bij dit onderzoek betrokken. Evert heeft met groot enthousiasme alle metingen met proefpersonen uitgevoerd. Hans heeft het realiseren van het digitale gedeelte van de meetopstelling voor zijn rekening genomen. Frank Geurtsen heeft al het spraakmateriaal opgenomen en nauwkeurig geijkt.

Moderne tekstverwerking en grafische programma's maken het mogelijk van een proefschrift een vrijwel professioneel drukwerk te maken. Ik ben Koos Wolf erkentelijk voor de bijpassende professionele bijdrage aan de vormgeving van de figuren, Leny v.d. Boon voor de (zeer consequente) opmaak van de tekst en Ilse Houtgast voor het ontwerpen van de omslag. Melvyn Hunt improved the English of an earlier version of this thesis.

Ik denk dat geen mens kan gedijen zonder een thuisbasis en rustpunt, waar op het juiste moment tot activiteit of juist tot "gas terugnemen" wordt gemaand. In dit verband heeft mijn gezin een grote (en meestal onzichtbare) bijdrage aan het tot stand komen van dit proefschrift geleverd.

ON MEASURING AND PREDICTING SPEECH INTELLIGIBILITY

CONTENTS

1 INTRODUCTION

Speech is considered to be the major means of communication between people. In many situations, however, the speech signal we are listening to is degraded, and only a limited transfer of information is obtained. This may be due to factors related to the speaker, the listener, and the type of speech, but in most situations it is due to the transmission of the speech signal from the speaker to the listener. The purpose of this study is to quantify these limitations and to identify the physical aspects of a communication channel that are primarily related to the intelligibility[1] of the speech signal passed through such a channel.

Three distinct components are recognized: production properties (speech and speaker related), transmission aspects, and perceptual aspects (related to hearing aspects). Each component includes many variables which may interact with each other.

On the production side of the speech signal we can identify: the class of speaker (male, female, child), speaker-specific properties (speaking rate, speaking style, speech disorders). During transmission, a degradation may occur that results in a decrease of the information content[2] of the speech signal such as: limitations of the frequency range, the dynamic range, and distortion components. On the listener side, perceptual aspects, including frequency resolution, thresholds, and auditory masking, have to be considered.

All these aspects have been studied in the literature during the past half century. This has resulted in design criteria for transmission channels and in the development of speech quality measures, speech intelligibility tests, articulation tests, and a few diagnostic and objective assessment methods. We will consider some of these aspects, mainly in an attempt to improve the *prediction* of intelligibility for various types of distortion (band-pass limiting and noise masking). Three methods of assessment can generally be distinguished:
(a) subjective measures making use of speakers and listeners,
(b) predictive measures based on physical properties, and
(c) objective measures obtained by measurements with specific test signals.

---

[1] According to the ASA (1960) standard on monosyllabic-word intelligibility, the term "intelligibility" is used whenever units of speech material that are used for testing consist of complete and meaningful words, phrases, or sentences; the term "articulation" is used when units of speech material for testing consist of meaningless syllables or fragments. We will use the term intelligibility.

[2] Information content: properties of a speech signal that contribute to identification of a speech item (phoneme, word, or sentence).

- (a) Subjective tests make use of various types of speech material. All these tests have their specific advantages and limitations mostly related to the speech items tested. Frequently used speech elements for testing are phonemes, words (digits, alphabet, short words), sentences, and sometimes a free conversation. The scoring methods used with subjective tests are: the recall of the items presented, rating on a subjective scale, or rank ordering of the test conditions. The recall procedure can be based on a given limited set of responses or on an open response design in which all possible alternatives are allowed as a response.

- (b) Predictive measures based on physical and perceptual properties quantify the effect on the speech signal and the related loss of intelligibility due to for instance: a limited frequency transfer, masking noises with various spectra, reverberation and echoes, and a nonlinear transfer resulting from peak clipping, quantization, or interruptions.

On the perceptual (listener) side, knowledge about hearing aspects, such as frequency resolution, auditory masking, and reception thresholds, can be used to predict the intelligibility for a given condition.

One of the first descriptions of a model to predict the effect of a transmission path on the intelligibility of speech was presented by French and Steinberg (1947) and later evaluated by Beranek (1947). This work formed the basis for the so-called Articulation Index (AI), which was described, evaluated and made accessible by Kryter (1962a). The AI consists of the sum of the information content of 20 equally contributing frequency bands. These frequency bands cover the speech frequency range. The information content within each band depends on the signal-to-noise ratio, where the speech signal is described by the long-term frequency spectrum.

The effect of auditory spread-of-masking is modelled by the application of a specific correction to the shape of the noise spectrum.

The effect of peak clipping and reverberation is separately modelled by the application of specific correction factors.

The AI is frequently used for the design and the prediction of the quality of systems for which a limited frequency transfer and masking noise are the major distortions.

Peutz (1971) developed an algorithm to predict intelligibility in auditoria. This measure is based on reverberation time, distance between speaker and listener, and signal-to-noise ratio.

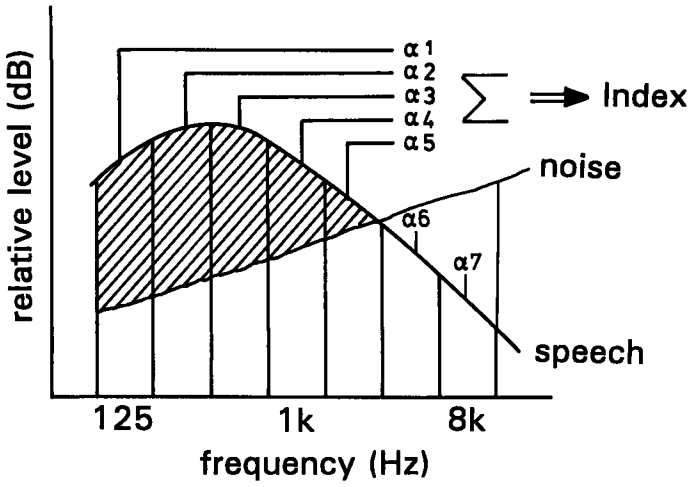

Fig. 1.1 Illustration of the long-term spectrum of a speech signal masked by noise, and the weighted summation to an objective intelligibility prediction.

- (c) An objective method to measure the speech transmission quality of an existing communication channel was developed by Houtgast and Steeneken (1971), and Steeneken and Houtgast (1980). This method is based on the application of a specific test signal. The transmission quality is derived from an analysis of the received test signal, and is expressed by an index, the Speech Transmission Index (STI). Similar to the AI concept, the STI is based on a weighted contribution of frequency bands. Where the original AI concept makes use of 20 frequency bands of equal importance (each with a different bandwidth), the STI uses a fixed bandwidth (octave bands) with a weighted contribution (weighting factor α_k) as indicated in Fig. 1.1. The STI value is obtained from measurements on the transmission channel in operation, whereas the AI is based on a calculation scheme making use of a physical description of the transmission channel. The STI measurement requires a special test signal from which the effective signal-to-noise ratio in each octave band at the receiving side is determined and used for the calculation of the STI. The specific features of this approach are that the test signal design allows an adequate interpretation of other degradations than just a limited frequency transfer and masking noise, for example nonlinear distortion and distortion in the time domain. Hence almost all types of distortion and their combinations that may occur on an analogue or digital (wave-form-based) transmission path are

accounted for. However distortions such as frequency shifts and voiced/unvoiced decision errors that may occur with certain types of vocoders, are not included in this concept. Application of the STI for telecommunication channel evaluation makes use of a specific measuring device (Steeneken and Agterhuis, 1982). Over the years, twenty-five of these devices have been built and have been distributed to many laboratories all over the world. A full description of the present STI-measuring method is provided in appendix A1.

Based on the STI concept, the RASTI method (Room Acoustical Speech Transmission Index) was developed in 1979 (Steeneken and Houtgast, 1979; Houtgast and Steeneken, 1984). This simplified method was especially developed as a screening device for applications in room acoustics. A device based on this method was commercialized by Brüel & Kjær (instrument no. 3361) and is presently being used by acoustic consultants.

Both the AI and the STI were validated with word tests of Phonetically Balanced[3] CVC-type nonsense words[4]. The calculation schemes of the AI and the test signal of the STI are based on the average long-term speech spectrum. This means that the predictions made by the AI and STI are related to the mean global transmission quality and do not reflect detailed phonetic aspects ("what is the intelligibility of fricatives") or specific speaker variations ("are female voices well perceived"). In many situations, such additional diagnostic information would be very useful to improve the performance of a transmission channel.

Ten years of our own experience with the development and the application of the STI has shown the need for further improvements, for instance when applied to conditions with a limited or non-contiguous frequency transfer. Also, effects of speaker variation, the sex of the speaker, and the individual relation with consonant and vowel recognition need further attention.

We will critically consider the model upon which the AI and the STI are based, as well as the speaker characteristics, and the speech material used for the evaluation.

In chapter 2 the validity and the limitations of the current model, in which each frequency band gives an independent contribution to intelligibility, is

---

[3] Phonetically Balanced (PB) indicates that the frequency distribution of the phonemes used for the test are representative of the language.

[4] CVC-nonsense words are words of the type Consonant-Vowel-Consonant, obtained from a random choice of initial consonants, vowels, and final consonants. This leads to a mixture of meaningless (nonwords) and meaningful words.

studied and some erroneous results are analyzed. Subsequently an extension of this model is introduced in order to improve the predictions of the model. Besides the independent contribution of each frequency band we have introduced some factors to account for the correlation between adjacent frequency bands. Also, the relation between signal-to-noise ratio and the information content of the various frequency bands is studied.

The relation between the scores for different speech items at the phoneme level is studied in chapter 3. This leads to a classification of four groups of phonemes with an identical degradation within each group, at various transmission conditions. Also the relation between sentence intelligibility and phoneme and word scores is studied.

Chapter 4 is concerned with the question how the information content within different frequency bands depends on the speech items studied. Different frequency-weighting functions were found for the different groups of phonemes and for short words. This can be related to the results of earlier studies (French and Steinberg, 1947; Steeneken and Houtgast, 1980; Studebaker et al., 1987).

In chapter 5 the extended model for predicting intelligibility is validated experimentally, for male and female speech, and various types of transmission channels.

2 MUTUAL DEPENDENCE OF OCTAVE-BAND-SPECIFIC CONTRIBUTIONS TO SPEECH INTELLIGIBILITY

Summary

Currently used objective measures for predicting the intelligibility of speech assume that this can be modelled as a simple addition of the contributions of individual frequency bands. The Articulation Index (AI, French and Steinberg, 1947) and Speech Transmission Index (STI, Steeneken and Houtgast, 1980) are based on this assumption. There is evidence that the underlying assumption of mutually independent frequency bands is not valid and may lead to erroneous prediction for conditions with a limited frequency transfer or with gaps or selective masking in the frequency domain.

We designed an experiment where the contribution of individual frequency bands is studied. For this purpose the speech signal is subdivided into seven octave bands with centre frequencies ranging from 125 Hz to 8 kHz. For 26 different combinations of three or more octave bands the CVC-word score (Consonant-Vowel-Consonant, nonsense words) was obtained at three signal-to-noise ratios.

For predicting the observed intelligibility, a revised model is proposed which accounts for mutual dependency between adjacent octave bands by the introduction of a so-called redundancy correction. Consequences for the existing objective measures are discussed.

2.1 Introduction

The assumption that the intelligibility of a speech signal is based on the sum of the contributions of individual frequency bands was proposed between 1925 and 1930 by Fletcher and modelled by French and Steinberg (1947). They found that this frequency-specific information content of a speech signal is not equally distributed along the frequency range of the signal. The procedure they used for estimating the relative contribution of different frequency regions was based on the syllable-articulation score, measured as a function of the cut-off frequency of a high-pass filter or a low-pass filter for various speech levels. From these results twenty contiguous frequency bands were obtained of which the bandwidths were adjusted to provide an equal contribution to a defined index, the so-called Articulation Index (AI).

Given the twenty equally contributing frequency bands, the actual contribution of each band is determined by a factor W. Based on the dynamic range of the fluctuations among speech spectra of different speech items as described by Dunn and White (1940) and the threshold of hearing, the factor W

was defined proportional to "the fraction of the intervals of speech in a band that can be heard".

The total information content of a signal is the sum of the W values for all frequency bands. The factor W will be reduced if degradation of the signal due to noise masking and peak limitation occurs. If we neglect other effects such as auditory masking, time-domain distortions and nonlinear distortion, a simple equation defines the AI:

$$AI = \frac{1}{20} \sum_{k=1}^{20} W_k \qquad (2.1.1)$$

The factor W is obtained by a linear transformation of the speech-to-noise ratio in dB within each contributing frequency band to a factor between 0 and 1. This is done in such a way that a signal-to-noise ratio of 18 dB or higher corresponds with a value "1" of factor W and a signal-to-noise ratio of -12 dB or lower corresponds with a value "0".

This approach assumes that the information content within a frequency band is simply defined by the signal-to-noise ratio and not by the absolute signal level in relation to adjacent frequency bands. This assumption was indirectly verified by the validation of AI and STI (Kryter, 1962b; Steeneken and Houtgast, 1980), where different signal-to-noise ratios at various levels were combined, and recently also by Van Dijkhuizen et al. (1987). They found that within a range of spectral tilt of more than 17 dB/oct at a constant signal-to-noise ratio the intelligibility did not change.

The AI concept was made more accessible by Kryter (1962a). This was achieved by the introduction of worksheets and tables with correction factors that also took into account auditory spread-of-masking in the estimation of the factor W. Also, for practical reasons a conversion to a 1/3-octave band and a 1/1-octave band approach was made by introducing a frequency-weighting factor (α_k) for each contributing frequency band (k), rather than by having equally contributing frequency bands with a different bandwidth. For the 1/1-octave-band concept, five octave bands with centre frequencies from 250 Hz up to 4000 Hz are used. The AI is then defined by:

$$AI = \sum_{k=1}^{5} (\alpha_k \cdot W_k), \quad \text{with} \quad \sum_{k=1}^{5} \alpha_k = 1.0 \qquad (2.1.2)$$

Hence a prediction of the AI is based on physical properties of the communication channel considered (mainly frequency transfer and noise spectrum).

The Speech Transmission Index (STI) is very similar to the AI concept, but is extended to distortion in the time domain (Houtgast and Steeneken, 1972) and to nonlinear distortion (Steeneken and Houtgast, 1980). The STI was designed for the measurement of speech transmission quality with the use of artificial test signals. As described in appendix A1, the approach is based on the Modulation Transfer Function (MTF), and does account for a wide variety of degradations, including reverberation and echoes.

In the study described in this chapter we will concentrate on degradation in the frequency domain and therefore no distortion in the time domain is considered (echoes, reverberation, and automatic gain control). For this purpose the method can be simplified to a (modulation-frequency-independent) factor TI (Transmission Index). The TI is similar to the factor W of the AI approach, indicating the contribution of each octave band.

For practical reasons, the frequency resolution of the STI concept is based on seven octave bands with centre frequencies from 125 Hz to 8 kHz. The STI becomes:

$$STI = \alpha_1 \cdot TI_1 + \alpha_2 \cdot TI_2 + \ldots \quad \ldots + \alpha_7 \cdot TI_7 \qquad (2.1.3)$$

where $\sum_{k=1}^{7} \alpha_k = 1$.

The optimization of the frequency-weighting factors, the calculation of the TI's, and the design of the test signal parameters were performed by optimizing the prediction by the STI of the observed CVC-word score with an iterative procedure. For this study, 167 different transmission channels were used, intended to be representative of most types of distortion. The optimal relation (Steeneken and Houtgast, 1980) between STI and the subjective data according to this concept is given in Fig. 2.1.1. The vertical spread, expressed by the standard deviation, is s = 5.6%. This vertical spread is a measure of the prediction accuracy of the CVC-word score by the STI. It was shown that this vertical spread is due, in about equal proportions, to the variance introduced by the limitation of the STI concept and to the variance within the observed subjective data, introduced with the limited number of speakers (4) and listeners (5).

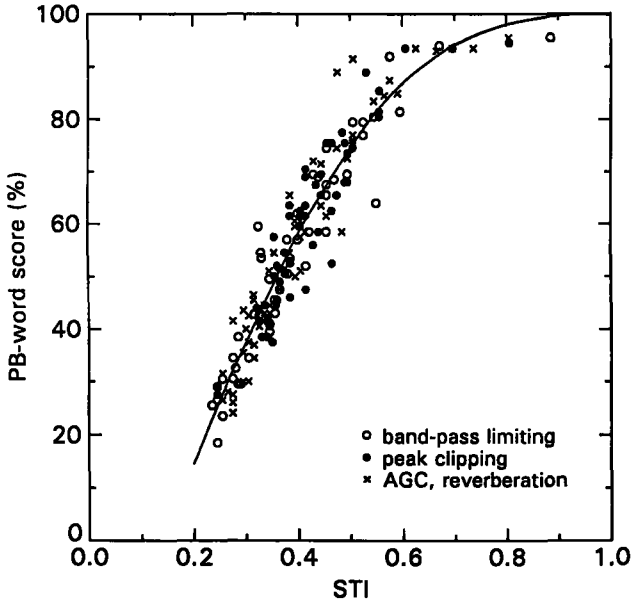

Fig. 2.1.1 Relation between STI and the score of phonetically-balanced CVC words for 167 transmission conditions including representative types of distortion. Adopted from Steeneken and Houtgast (1980). The standard deviation, representing the vertical spread around the best-fitting third-order polynomial, is s = 5.6%.

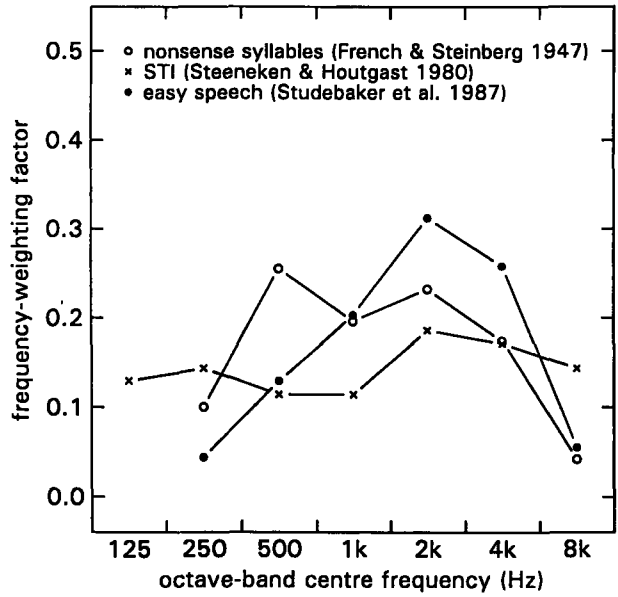

Fig. 2.1.2 Frequency-weighting functions α_k for the octave-band contribution to the AI and STI adopted from French and Steinberg (1947), Steeneken and Houtgast (1980), and Studebaker et al. (1987).

Although the type of speech material for both the AI and the STI was similar, quite different frequency-weighting factors α_k were found. Pavlovic (1987) compared these frequency-weighting factors, together with the results of a study of Studebaker et al. (1987). In this latter study "easy" speech was used from which the sentence intelligibility was estimated for a range of high- or low-pass filter conditions, similar to the Fletcher and Steinberg approach (1929). A comparison between the frequency-weighting factors of these studies is given in Fig. 2.1.2 (the sum of each series of frequency-weighting factors is normalized to "1"). Various aspects, such as the frequency range of the contributing frequency bands along the frequency scale, the speech items used for the validation (words, connected discourse), and the speech spectrum from which the frequency-band-specific signal-to-noise ratios are calculated, may account for the dissimilarity between the frequency-weighting functions.

The STI approach discussed so far is a simple additive model (Eq. 2.1.3). Presently, we are better aware of the fact that the energy contents in (narrow) adjacent frequency bands for different speech sounds may be correlated. Hence the levels in these bands show a high degree of co-variation and the information provided by such bands may be considered redundant to some degree. This correlation was recently studied for continuous speech (Ter Keurs and Houtgast, 1987; Houtgast and Verhave, 1991). The additive models of AI and STI do not account for redundancy associated with correlation between adjacent frequency bands and assume statistical independence between frequency bands. The evaluation of the original model was restricted to band-pass limited speech signals within contiguous pass-bands, and the possible effect of correlated bands on the total score could therefore not be investigated properly with this material.

Kryter (1960) actually found some discrepancy between subjective intelligibility scores for conditions based on contiguous frequency-band channels and conditions based on channels with gaps in the frequency transfer. It was concluded in that study that "the best single contiguous band-pass system requires twice the effective bandwidth compared to the best three-band system (500 Hz bandwidth each) to achieve equal performance as measured by speech intelligibility". It was found that the missing contributions from the frequency regions of the gaps did not result in a lower intelligibility as was predicted by the AI. The redundancy effect might provide a plausible explanation (see also Grant and Braida, 1991).

The studies described above also do not discuss the relation between optimal frequency weighting and other parameters, such as signal-to-noise ratio, type of speech material, and the sex of the speaker.

Therefore an experiment was designed where these relations could be studied separately. For this purpose a balanced set of frequency-transfer conditions (including nonadjacent frequency bands) was used in order to estimate the relations between adjacent frequency bands. This set of conditions was applied at three different speech-to-noise ratios, where the masking noise had a frequency spectrum equal to the long-term speech spectrum. This speech-like noise spectrum was different for the male and female speech conditions. The procedure results in equal masking of all frequency bands at the same signal-to-noise ratio and allows estimation of the frequency-weighting factors for each signal-to-noise ratio independently.

As the experiments were performed for male and female speech separately, a comparison of the results for both types of speech signals can be made.

2.2 Experimental design

The goal of this study is to define individual parameters in the AI and STI models for the optimal prediction of intelligibility between speaker and listener. For this purpose a set of conditions was defined in terms of the octave bands included in the frequency transfer and the signal-to-noise ratio. The experimental design was focused on the prediction of the information content, and was based on frequency bands according to the STI-concept: seven octave bands with centre frequencies from 125 Hz to 8 kHz.

2.2.1 *Description of the measuring conditions*

In appendix A2 an overview is given of the various frequency-transfer conditions which are based on combinations of seven octave bands. Each octave band can be included independently in the frequency transfer. Selection of adjacent octave bands results in a contiguous frequency transfer, while selection of nonadjacent frequency bands results in gaps in the frequency transfer. Four main groups of frequency-band selections are used and are described in detail in the appendix. These groups are:

- a frequency transfer with a "rippled" envelope. According to the STI concept (see Fig. 2.1.2), the joint contribution of the lower four octave bands is about equal to that of the upper three octave bands to the STI. Conditions no 1 and no 2 represent this opposition. The other conditions up to no 8 show an increased ripple density of the frequency transfer, with alternation of the successive octave bands as the upper limit.

- a second group consists of all possible adjacent triplets of octave bands (five conditions).
- a third group consists of triplets of only nonadjacent octave bands. In this set of 11 possible combinations, the lower and upper frequency bands occur six times while the octave band in the centre of the frequency range occurs only three times. In order to obtain a balanced design a limited number of conditions was selected where each octave band is used three times.
- the fourth group is composed of a contiguous frequency transfer from four adjacent octave bands up to the full range of seven octave bands.

Some groups share identical configurations of octave bands. These conditions were selected only once, and are indicated in the appendix by referring to the condition number used first.

The difference between the speech of male and female speakers as used for this experiment is that the female speech essentially has no energy (information) in the octave band with centre frequency 125 Hz. Therefore, the conditions including this frequency band were not used in combination with the female speech. The design described above results in 26 different frequency-transfer conditions for male speech and 17 different conditions for female speech.

Another variable in this experiment is the signal-to-noise ratio. By using a masking noise with a frequency spectrum equivalent to the long-term speech spectrum, the signal-to-noise ratio is equal for all selected frequency bands. The long-term frequency spectrum is different for male and female speech. Based on the average speech spectrum from four male and four female speakers two noise spectra were selected for use in the experiments. The speakers were the same as used for the subjective tests in this experiment. The signal-to-noise ratios used in the experiments were 15 dB, 7.5 dB, and 0 dB.

Because of the severe limitation of the frequency transfer in some of the conditions relative to some wide-band conditions, the full intelligibility range from bad-to-excellent was covered.

As predicted from the existing STI-model, a signal-to-noise ratio of 15 dB or higher will not affect intelligibility. Therefore, the highest signal-to-noise ratio was limited to 15 dB. Hence a noise floor at -15 dB (corresponding to this signal-to-noise ratio) was introduced to mask gaps in the frequency transfer and to avoid information transfer for frequency ranges covered by the slopes of the selected filters.

The 26 frequency-transfer conditions of the male speakers and the 17 frequency-transfer conditions of the female speakers combined with three signal-

to-noise ratios result in 78 and 51 conditions for the male and female speakers respectively.

2.2.2 *Subjective intelligibility measurement*

In order to obtain the relation between STI values as defined by the signal-to-noise ratio and intelligibility scores, subjective intelligibility measurements were performed. Four male and four female speakers were used for this subjective evaluation. The speech material consisted of an equally balanced set of 51 different CVC words embedded in a carrier phrase. Hence all possible initial consonants[1] of Dutch (a total of 17) are represented three times in each list. As the number of vowels (a total of 15) and the number of final consonants (a total of 11) is smaller than the number of initial consonants some of the phonemes were used more than three times in each list. A complete description of this test material is given in chapter 3, in which the subjective evaluation is discussed.

The use of CVC words not only provides us with a "syllable articulation" score similar to the method as used by Fletcher and Steinberg, but also results in individual phoneme scores. The four male and four female speakers were split into two pairs of male and two pairs of female speakers, each pair of speakers was then linked with one of two groups of four female listeners. If the four male and four female speakers are labelled M1-4 and F1-4 and the two listener groups of four listeners LG1 and LG2, the following combinations were used:

	Male speakers				Female speakers			
	M1	M2	M3	M4	F1	F2	F3	F4
LG1	x	x			x	x		
LG2			x	x			x	x

The male and female speaker conditions are considered separately, resulting in 16 speaker-listener pairs for the male conditions and 16 speaker-listener pairs for the female conditions.

---

[1] Phonemes with an occurrence below approximately 1% were omitted (see section 3.3.1).

2.2.3 *Experimental set-up and calibration*

The experimental set-up was established by using a (real-time) digital signal processor system (Van Raaij and Steeneken, 1991). This system applied a FIR filter to the speech signal in order to obtain the required frequency-transfer condition. The FIR filter consisted of 767 taps. A sampling time of 40 μs was used with an effective resolution of the analogue-to-digital and digital-to-analogue conversion of 14 bit. This resulted in a filter transfer characteristic with a maximum slope of 15 dB per 1/12 octave for frequencies below 500 Hz and a maximum slope of 25 dB per 1/12 octave for frequencies above 3000 Hz. The system also performed the addition of noise to the filtered speech at the defined signal-to-noise ratio.

The signal-to-noise ratio was determined by the application of a speech level measure according to Steeneken and Houtgast (1986) - for more details see appendix A7. The speech level measure is based on the (A-weighted) RMS value with a threshold applied to cancel the effect of silent periods between the speech utterances.

2.3 Experimental results

2.3.1 *Evaluation of experimental results with the existing STI-model*

For all experimental conditions the STI value can easily be calculated with equation 2.1.3. Given the frequency-weighting factors, it is uniquely defined by the contributing octave bands and the defined signal-to-noise ratio.

The contributing frequency bands are defined in appendix A2. The transmission index (TI_k), for each frequency band (k) depends on the signal-to-noise ratio (SNR) and is given (according to the STI concept) by:

$$TI_k = \frac{SNR_k + 15}{30}, \text{ where } 0 \le Ti_k \le 1 \qquad (2.3.1)$$

This means that the TI_k for the selected signal-to-noise ratios of 15, 7.5, and 0 dB becomes 1.0, 0.75, and 0.5 respectively. The calculation of the TI_k and the values of the octave-band-specific weighting factor α_k are according to Steeneken and Houtgast (1980) and are not optimized for this set of data. The mean CVC-word scores as resulting from the subjective listening experiments, for the male and female speakers separately, are given in appendix A2.

Based on the CVC-word scores and the corresponding STI values for each condition, scatter diagrams for the male and female speakers were made to represent the relation between the observed word score and the STI values. These are given in Figs 2.3.1 and 2.3.2. The data points for the four groups of frequency-transfer conditions as described in section 2.2.1 are marked differently. The optimal relation between the two variables is described by means of a best-fitting third-order polynomial[2], and is quantified by the standard deviation (s) along the CVC-word score axis.

A standard deviation s = 12.8% is obtained for the 78 male conditions and s = 8.8% for the 51 female conditions. This vertical spread is much higher than the earlier mentioned value obtained with the original 1980 experiments where s = 5.2% for male speakers and for 49 conditions with a contiguous frequency transfer and four different types of noise. The unusual frequency-transfer conditions used here, especially those conditions with a rippled frequency transfer and with gaps in the frequency transfer, show a systematically different relation between the observed word score and the STI value. Conditions with triplets of adjacent frequency bands and part of the conditions with a rippled spectral envelope (also the conditions based on adjacent frequency bands) show a relatively high STI value in relation to the CVC-word score. The conditions based on triplets of nonadjacent frequency bands and part of the conditions with the rippled spectral envelope (those with gaps in the frequency transfer) show a relatively low STI value in relation to the corresponding CVC-word score.

As a first step to improve the relation between the STI and the scores of the present data, we used an iterative procedure to calculate the optimum frequency-weighting factors α_k. The criterion was to minimize s. For a given set of α_k-values, the s-value was calculated as a function of the value of the frequency-weighting factor for each contributing octave band individually. After this procedure all frequency-weighting factors were adjusted. For each band a new value was selected between the original weighting factor and the value corresponding to the minimum of s at 1/3 of the range between the two values. This procedure was chosen to avoid overshoot in the iteration procedure. With the new set of α_k-values the procedure was repeated. We found that a stable minimum for the final set of α_k factors was obtained after approximately 5

---

[2] According to Fletcher and Galt (1950) the relation between AI and word score is exponential, therefore they recommend an exponential fit. For the data in this study a third-order polynomial gives a slightly better fit. The best fitting third-order polynomial can be calculated directly, the best fitting exponential function is obtained in an iterative procedure.

iterations. It was verified that different starting configurations all converged to the same set of frequency-weighting factors.

It was found that for the optimal set of α_k-values, the relation between the STI value and the CVC-word score for this set of data gives s = 6.8% for the male speakers and s = 6.0% for the female speakers. The frequency-weighting factors for this optimal relation are given in Fig. 2.3.3. These factors are different from those of Fig. 2.1.4 and suggest that these optimal frequency-weighting factors depend on the conditions used for the evaluation. This may be due to the additive nature of the STI concept. The model does not account for any correlation between adjacent frequency bands. Overestimation of the total information content for conditions without gaps in the frequency transfer may occur. This effect and the relation between the signal-to-noise ratio and the octave-band-specific transmission index, TI_k, will be discussed in the next sections.

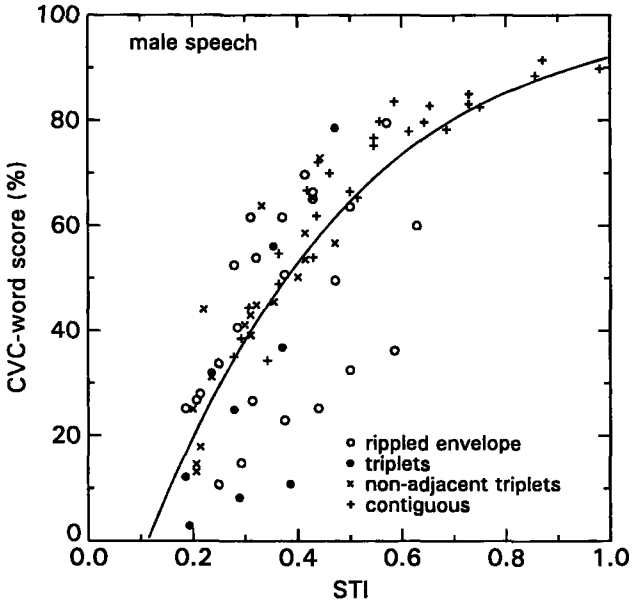

Fig. 2.3.1 Relation between the STI and the CVC-word score for the conditions involving MALE speech, band pass limiting and noise. The parameters used for the STI calculation were adopted from the procedure described previously by Steeneken and Houtgast (1980). The standard deviation, representing the vertical spread around the best-fitting third-order polynomial is s = 12.8%.

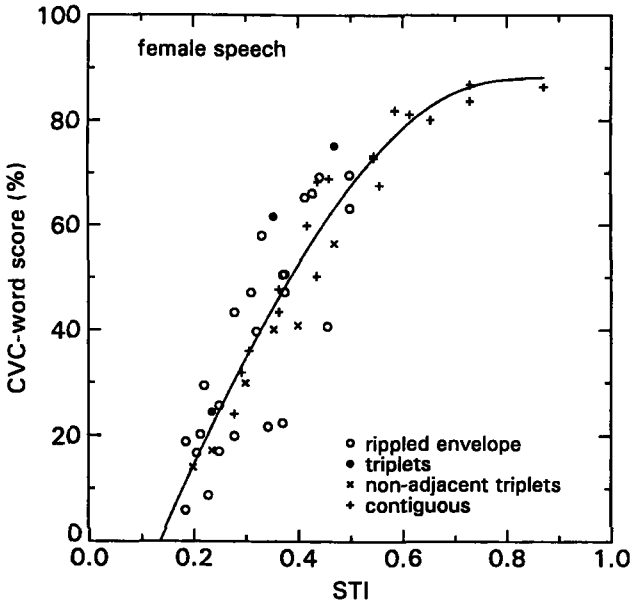

Fig. 2.3.2 Relation between the STI and the CVC-word score for the conditions involving FEMALE speech, band-pass limiting and noise. The parameters used for the STI calculation were adopted from the procedure described previously by Steeneken and Houtgast (1980). The standard deviation, representing the vertical spread around the best-fitting third-order polynomial is s = 8.8%.

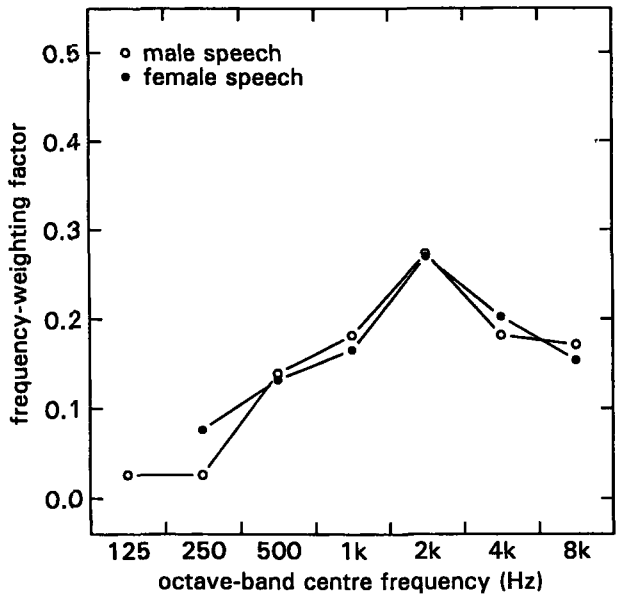

Fig. 2.3.3 Frequency-weighting factors α_k for the optimal octave-band contribution to STI for the male and female conditions of this experiment.

2.3.2 *Extension of the model with a frequency-dependent redundancy factor*

As mentioned before, a correlation between the fluctuations within adjacent frequency bands was found, among others, by Houtgast and Verhave (1991). This correlation suggests that the assumption of independent contributions of frequency bands to the AI or STI is, in principle, not valid.

If the information content of two adjacent frequency bands is redundant, a simultaneous contribution of these bands will result in an overestimation of the effective information content. Therefore a reduction should be applied for conditions of such a simultaneous contribution. This can be done in various ways. One simple way is the addition of correction terms to the original equation (2.1.3). The resulting equation is given by:

$$\text{Index} = \alpha_1 \cdot TI_1 - \beta_1 \cdot \sqrt{(TI_1 \cdot TI_2)} + \alpha_2 \cdot TI_2 - \beta_2 \cdot \sqrt{(TI_2 \cdot TI_3)} + \ldots \ldots + \alpha_n \cdot TI_n \qquad (2.3.2)$$

where $$\sum_{k=1}^{n} \alpha_k - \sum_{k=1}^{n-1} \beta_k = 1.$$

In this equation the factors α_k and TI_k are comparable to those in equation 2.1.3. The redundancy correction term depends on the redundancy factor β_k and the simultaneous contribution of two adjacent frequency bands given by the two TI_k factors. Taking the root of the TI_k factors is not essential but makes the dimensionality of the terms in the expression uniform. The effect of redundancy is modelled in this equation for the six pairs of adjacent frequency bands only. A simple extension can be made for relations between nonadjacent frequency bands as well, but it has been avoided in order to limit the number of parameters.

We applied this extended model to the data set described before. In the iteration procedure the factors α_k and β_k were also optimized simultaneously. The results for the male and female speakers are given in Figs 2.3.4 and 2.3.5. The vertical spread for these relations is s = 4.7% for the male speech and s = 4.2% for the female speech. The frequency-weighting factors α_k and redundancy factors β_k obtained with an individual optimization of the male and female results are given in Fig. 2.3.6. The frequency-weighting factors are almost identical for both speech types except for the octave band with centre frequency 125 Hz, which is not relevant for female speech. The new index is referred to as STI_r (revised, redundancy).

The redundancy correction increases for frequencies below 1 kHz and above 4 kHz and is negligible for frequencies around 2 kHz. The relatively high

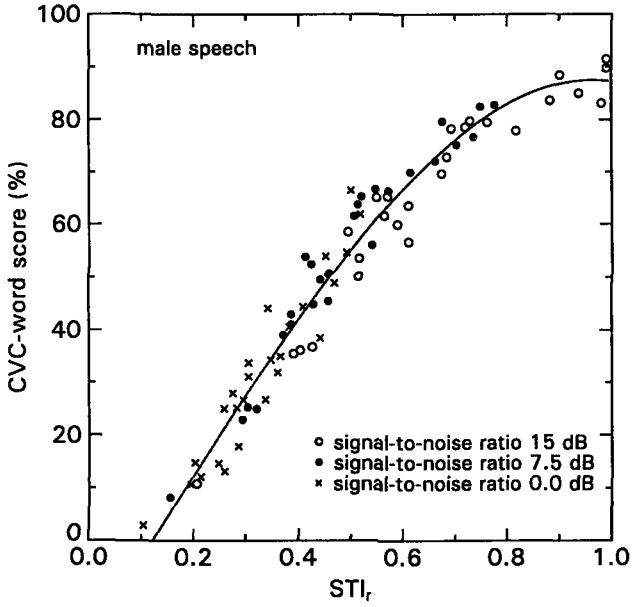

Fig. 2.3.4 Relation between the STI_r and the CVC-word score for the conditions involving MALE speech, band-pass limiting and noise. The standard deviation, representing the vertical spread around the best-fitting third-order polynomial is s = 4.7%.

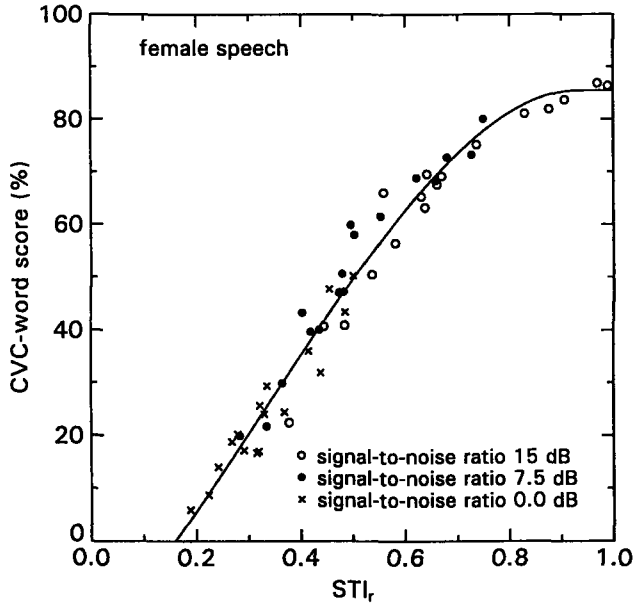

Fig. 2.3.5 Relation between the STI_r and the CVC-word score for the conditions involving FEMALE speech, band-pass limiting and noise. The standard deviation, representing the vertical spread around the best-fitting third-order polynomial is s = 4.2%.

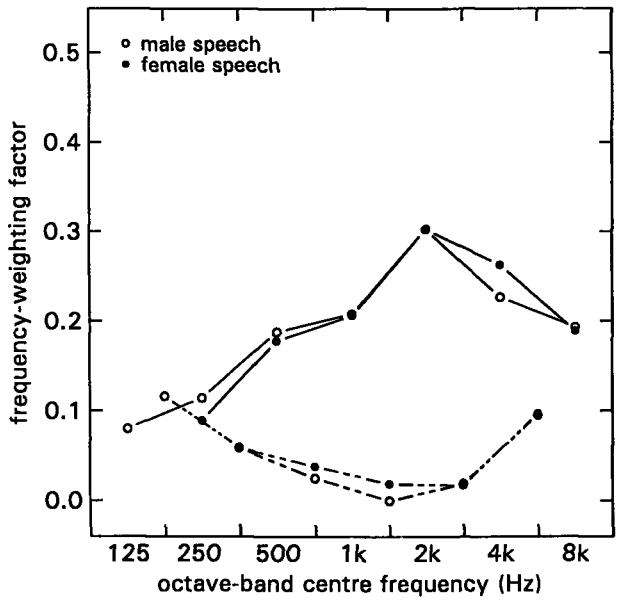

Fig. 2.3.6 Frequency-weighting factors for the octave-band contribution α_k (solid line) and the redundancy factors β_k (dashed line) for the male and female conditions.

frequency-weighting factor α_5 for the 2 kHz octave band, indicates that this frequency band should perhaps be divided into two (or more) frequency bands in order to obtain a smoother frequency-weighting factor curve. For the same reason, the frequency bands below 500 Hz could probably be combined. As a consequence, the difference between male and female results would completely disappear, which is a reduction of the diagnostic information.

For the optimization, the transmission index TI_k was obtained according to the original STI concept as given by equation 2.3.1. In the next section, the relation between TI_k and signal-to-noise ratio will be studied in more detail.

2.3.3 *Information content and optimal weighting as a function of the signal-to-noise ratio*

The measuring conditions of the experiments described in section 2.2 can be divided into three subsets, each with a different signal-to-noise ratio (15, 7.5, and 0 dB). The noise signals used had a frequency spectrum equal to the long-term speech spectrum of the four male speakers and of the four female speakers used in the experiments. This implies a constant (long-term) signal-to-noise ratio

for all frequency bands. A separate optimization can therefore be accomplished for each set of the conditions at a fixed signal-to-noise ratio (26 for the male speakers and 17 for the female speakers). As the signal-to-noise ratio does not change among conditions and among the contributing frequency bands, the octave-band-specific transmission index, TI_k, is also fixed. For the calculation of the new index we will, for a moment, keep the value of TI_k equal to unity for the contributing frequency bands. In this way, the STI for the set of conditions with a signal-to-noise ratio of 15 dB will be identical as before but the STI values for the two sets of conditions with the lower signal-to-noise ratios will be overestimated (the decrease due to the lower signal-to-noise ratio is not reflected in the TI_k value according to Eq. 2.3.1) The amount of overestimation provides a measure for deriving the actual TI_k at the two lower signal-to-noise ratios.

For each of the three limited sets of conditions with $TI_k = 1.0$ the relation between the observed CVC-word score and the STI value (with α_k and β_k from Fig. 2.3.6) was calculated and is given for the male and female conditions in Figs 2.3.7 and 2.3.8. The corresponding standard deviations for the male speakers and the three signal-to-noise ratios are s = 3.4%, s = 4.0%, and s = 5.9% respectively. For the female speakers these corresponding standard deviations are s = 2.4%, s = 4.3%, and s = 4.4%. The standard deviation increases for conditions with a low signal-to-noise ratio. This can be explained by the variation among the long-term speech spectra of the individual speakers.

The horizontal distance between the three curves is related to the reduction of the information content due to noise masking. Normally the effect of noise is accounted for by an appropriate reduction of the transmission index, TI_k, for the conditions with a lower signal-to-noise ratio. We obtained this reduction for three levels of the intelligibility score (CVC-word scores of 70%, 50% and 30% respectively) by calculating the relative horizontal shift of the curves. This is given in Table 2.3.1 together with the reduction factor for the given signal-to-noise ratio according to the original STI concept (Eq. 2.3.1). The results show that the reduction of the information content, related to the signal-to noise ratio, is largely independent of the intelligibility range, and is almost identical for male and female speech, and fairly well predicted by the original STI concept. We verified the effect of the slightly increased TI_k values in comparison with the original values (0.8 versus 0.75, and 0.51 versus 0.50) for the male conditions and did not obtain a significant improvement of the prediction accuracy.

Table 2.3.1 Reduction of the transmission index TI for male and female speech due to a reduction of the signal-to-noise ratio from 15 dB to 7.5 dB, and from 15 dB to 0 dB. The values are derived from Fig. 2.3.7 for the male speech and Fig. 2.3.8 for the female speech, at three levels of the CVC-word score. The reduction according to the original STI concept is also given.

CVC-word score	SNR 15/7.5 male	female	STI	SNR 15/0 male	female	STI
70%	0.80	0.80	0.75	-	-	0.50
50%	0.80	0.80	0.75	0.51	0.62	0.50
30%	0.81	0.85	0.75	0.51	0.58	0.50

In a separate analysis, the factors α_k and β_k were optimized for the three sets of data individually. The result is presented in Figs 2.3.9 and 2.3.10; the octave-weighting factors and the redundancy factors are given for the three signal-to-noise ratios and for male and female speech. The figures show that these factors, separately obtained for each signal-to-noise ratio and speech type, are quite similar. This indicates that the model is robust for these parameters.

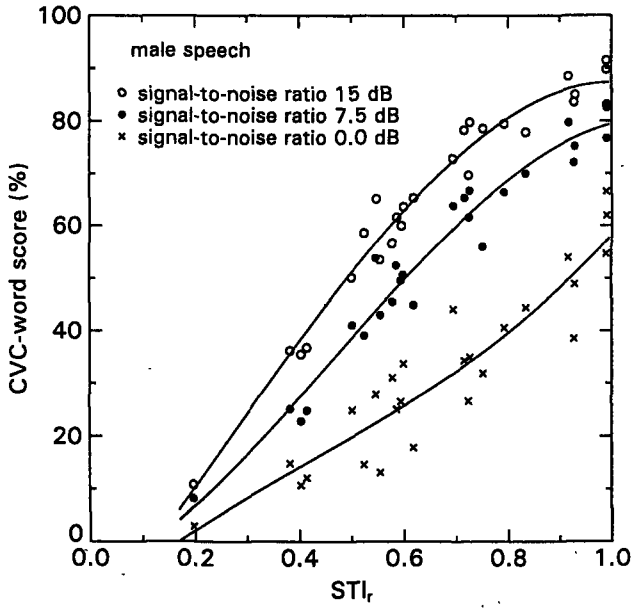

Fig. 2.3.7 Relation between the STI_r (with $TI_k=1.0$) and the CVC-word score for the conditions involving MALE speech, band-pass limiting at three signal-to-noise ratios. The parameters used for the Index calculation were optimized for each signal-to-noise ratio separately.

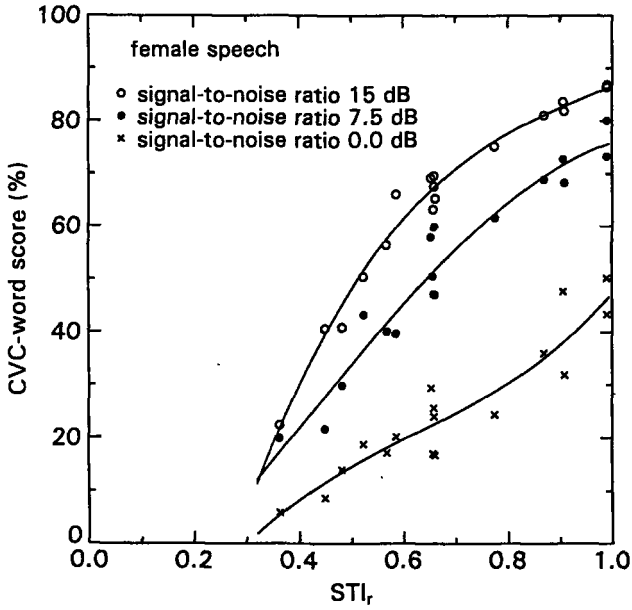

Fig. 2.3.8 Relation between the STI_r (with $TI_k = 1.0$) and the CVC-word score for the conditions involving FEMALE speech, band-pass limiting at three signal-to-noise ratios. The parameters used for the Index calculation were optimized for each signal-to-noise ratio separately.

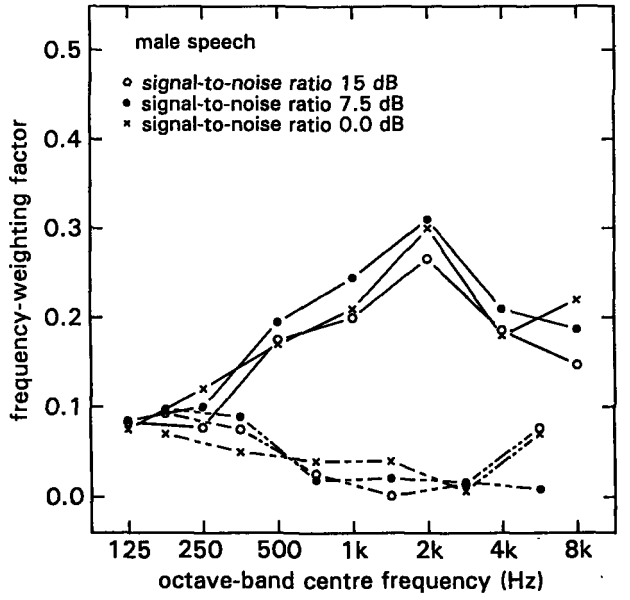

Fig. 2.3.9 Frequency-weighting factors for the octave band contribution α_k (solid line) and redundancy factors β_k (dashed line) for the MALE conditions at three signal-to-noise ratios.

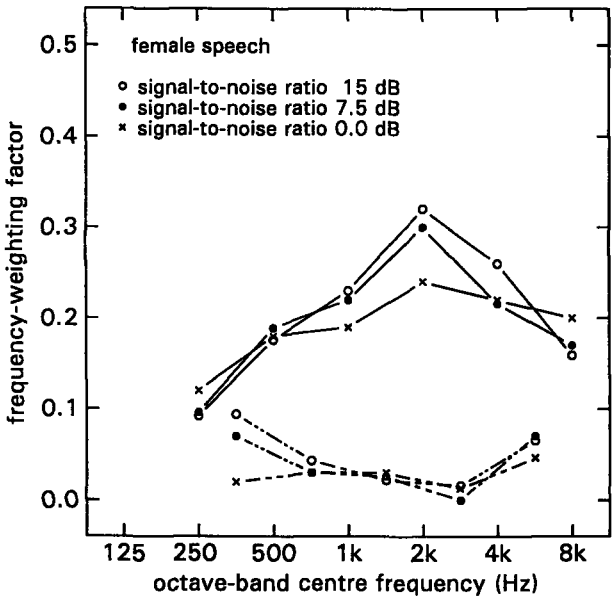

Fig. 2.3.10 Frequency-weighting factors for the octave band contribution α_k (solid line) and redundancy factors β_k (dashed line) for the FEMALE conditions at three signal-to-noise ratios.

2.3.4 *The frequency-band-specific contributions for consonants and vowels*

The long-term speech spectrum used for the masking noise of the experiments in this study is determined mainly by the contribution of the vowels and vowel-like consonants as these phonemes appear frequently and have a higher energy than the other types of phonemes. Hence the actual signal-to-noise ratio is dependent on the type of phoneme. Also, the distribution of the information content in the frequency domain might be phoneme-type-specific. As all CVC responses by the listener were typed into a computer we could, apart from word scores, also obtain the scores for the initial consonants, the vowels, and the final consonants. It was found that the scores for the initial consonants and final consonants are highly correlated for the different transfer conditions. However, the relation between the consonant scores and the vowel scores depends on the frequency transfer. The poor relation between the two major phoneme types is given in Fig. 2.3.11 for the male speakers. A similar relation (not shown) is obtained for the female speakers. This suggests that a specific set of frequency-weighting factors is required for an optimal prediction of the score for each of these groups of phonemes. We optimized the frequency-weighting

factors and the redundancy factors separately for the initial-consonant scores, the final-consonant scores, and the vowel scores. The optimal relations between the predicted intelligibility with the new index and the observed initial-consonant scores and vowel scores for male speech only are given in Figs 2.3.12 and 2.3.13. Similar results were obtained for the female speech conditions.

In Figs 2.3.14 and 2.3.15 the frequency-weighting factors α_k and redundancy factors β_k are given for the three phoneme groups and the male and female speech. These figures indicate that the optimal set of frequency-weighting factors is similar for initial and final consonants and for male and female speech but not similar to that obtained for vowels. As the results in this study are based on the calculation of the index by an equal signal-to-noise ratio concept in all frequency bands, no correction for individual phoneme spectra was made. This could be performed by correcting the signal-to-noise ratio within each frequency band in accordance with the spectral shape of the speech signals of the phoneme (group) considered, or by measuring the signal-to-noise ratio with a test signal whose frequency spectrum is adapted to the spectrum of the phoneme (group) considered. This will be discussed in chapter 4.

The relation between phoneme groups and frequency-weighting functions may be essential for the variety of frequency-weighting functions found by various studies (see Fig. 2.1.2) and will be studied separately in chapter 4.

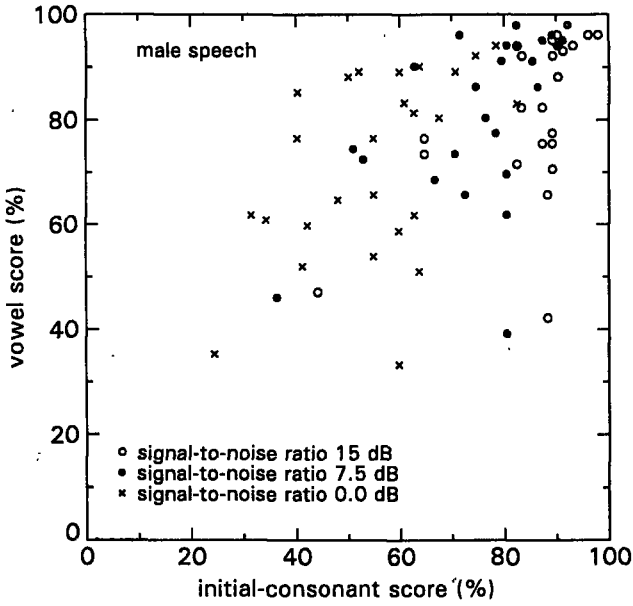

Fig. 2.3.11 Relation between initial-consonant scores and vowel scores for 78 transfer conditions for male speech.

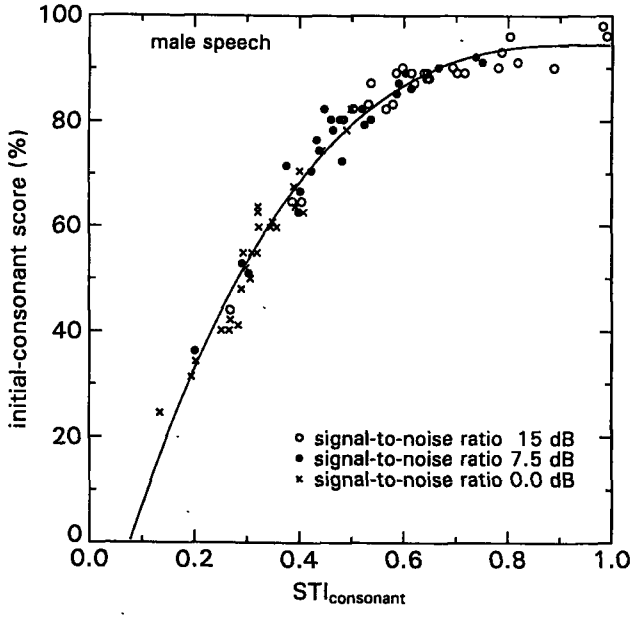

Fig. 2.3.12 Relation between the $STI_{consonant}$ and the observed INITIAL-CONSONANT score for the conditions involving male speech, band-pass limiting and three signal-to-noise ratios. The parameters for the STI_r calculation were obtained using an iterative optimization procedure.

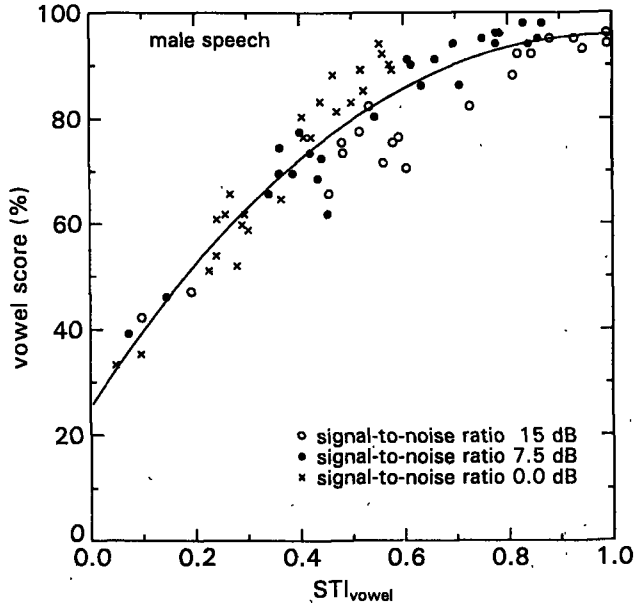

Fig. 2.3.13 Relation between the STI_{vowel} and the observed VOWEL scores for the conditions involving male speech, band-pass limiting and three signal-to-noise ratios. The parameters for the STI_r calculation were obtained using an iterative optimization procedure.

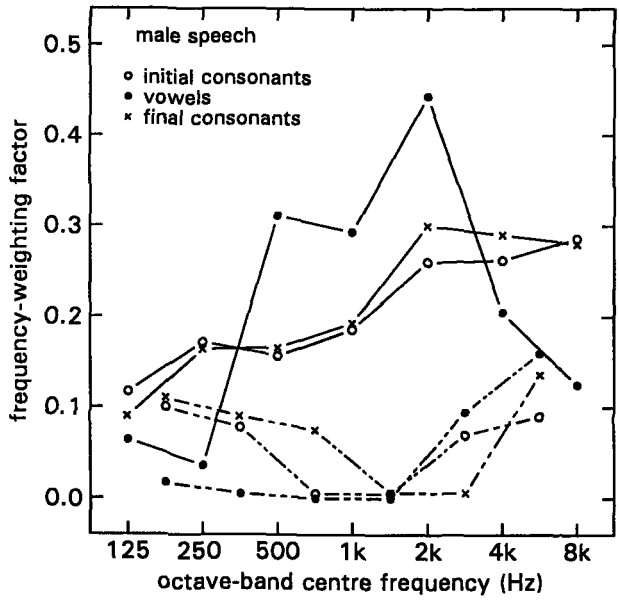

Fig. 2.3.14 Frequency-weighting factors for the octave band contribution α_k (solid line) and redundancy β_k (dashed line) for the MALE conditions and three phoneme groups.

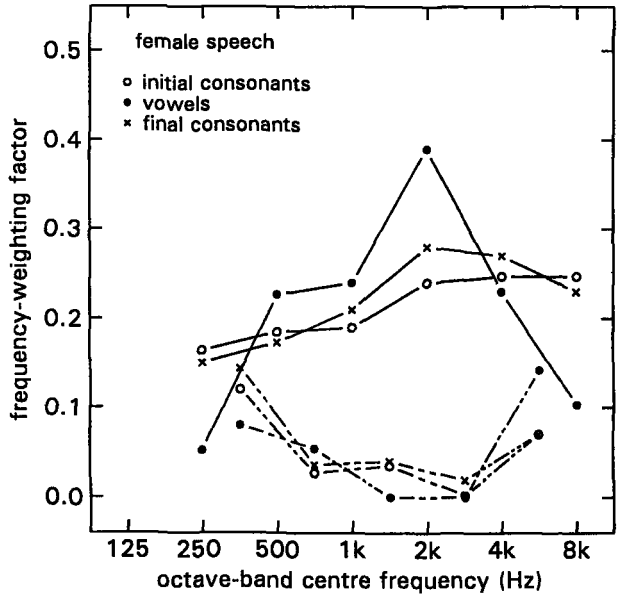

Fig. 2.3.15 Frequency-weighting factors for the octave band contribution α_k (solid line) and redundancy β_k (dashed line) for the FEMALE conditions and three phoneme groups.

2.4 Application of the model to earlier studies

In order to verify the new model for independent sets of data, the results of two earlier experiments could be used: The data from the study upon which the original STI model was based (Steeneken and Houtgast, 1980), and results from a study by Kryter (1960).

The new model with the redundancy correction was applied to the 1980 experimental results where band-pass limiting was combined with noise. In this study three representative contiguous frequency bands and four different types of noise were used. These parameters were combined to give 49 transmission conditions. The vertical spread after optimization of the α_k frequency-weighting values was s = 5.2% (Steeneken and Houtgast, 1980, Fig. 4). We could reproduce these results with the present automatic optimization program. The application of the new (redundancy) model to the original data set improved the relation between the observed scores (Phonetically Balanced CVC words) and the objective index. The vertical spread found was s = 4.4%. This improvement is not very large, and indicates that the existing STI method, not including the redundancy correction, is valid for the type of transmission conditions for which it was designed, e.g contiguous frequency-transfer conditions.

More extreme frequency-transfer conditions were used by Kryter (1960) in a study to reduce bandwidth with a minimum loss of the information content of speech. He performed a systematic study on transmission in relation to a frequency transfer consisting of combinations of three 500 Hz wide frequency bands. These bands were positioned along the frequency scale at 250 Hz intervals. In this way 51 combinations were made in a frequency range between 250 Hz and 3250 Hz. Unfortunately no frequencies above 3250 Hz were used. We divided the frequency range into 12 frequency bands in 250 Hz steps. Kryter used two signal-to-noise ratios for his experiments: 20 dB, and 10 dB. Since the masking noise was equal to the long-term speech spectrum, we could use a similar optimization between the subjective results and the index as described before in section 2.3.2. An important difference with the experiments described in this study is the extension to twelve frequency bands with equal bandwidth rather than seven octave bands. For the subjective evaluation Kryter used the "standard Harvard" Phonetically Balanced (PB) word test. The experiments were performed with only one speaker and six listeners.

The optimal relation between the PB-word score and the Index based on 12 contributing frequency bands is given in Fig. 2.4.1. The corresponding standard deviation becomes s = 3.9%. The relation between the subjective results and the predictions is similar to the results found in this study.

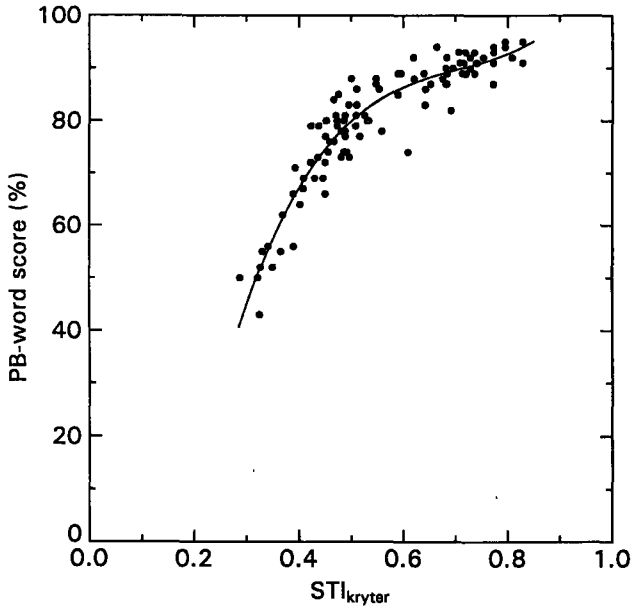

Fig. 2.4.1 Relation between the STI_{kryter} and the PB-word score for the conditions adopted from Kryter (1960) involving one male speaker, band-pass limiting and noise. The vertical spread around the best fitting third-order polynomial is s = 3.9%.

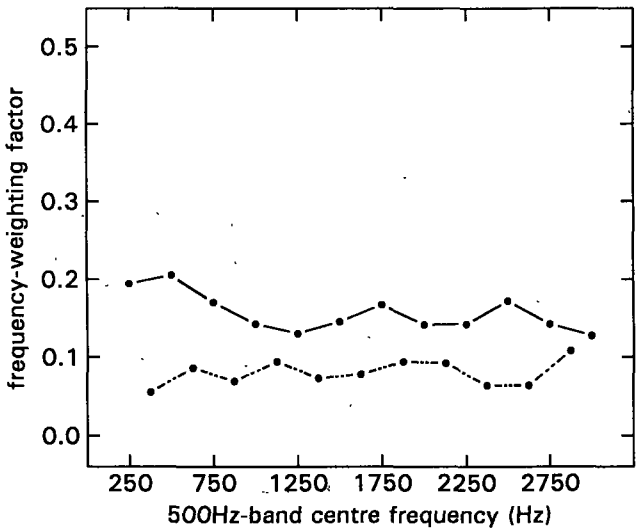

Fig. 2.4.2 Frequency-weighting factors for the 250 Hz band contributions α_k (solid line) and redundancy correction β_k (dashed line).

An optimization of the Kryter data without the application of a redundancy correction resulted in a standard deviation s = 4.7%.

In Fig. 2.4.2 the frequency-weighting factors and redundancy factors are given as resulting from the optimization procedure. A fairly flat function of these factors is obtained along the frequency axis. This may indicate that the linear frequency scale (in this restricted frequency region) results in a more efficient distribution of the contributing frequency bands.

2.5 Discussion and conclusions

In this chapter the relation between speech intelligibility and various physical parameters was examined. It was found that extension of the existing additive model (Eq. 2.1.3), with a correction factor accounting for the correlation of adjacent frequency bands, results in a more accurate prediction of the intelligibility. Such a redundancy correction β_k is included in:

$$STI_r = \alpha_1 \cdot TI_1 - \beta_1 \cdot \sqrt{(TI_1 \cdot TI_2)} + \alpha_2 \cdot TI_2 - \beta_2 \cdot \sqrt{(TI_2 \cdot TI_3)} + .. \quad .. + \alpha_n \cdot TI_n \qquad (2.5.1)$$

where $$\sum_{k=1}^{n} \alpha_k - \sum_{k=1}^{n-1} \beta_k = 1 \ .$$

Actually we used two versions of the redundancy correction ($\beta_k \sqrt{(TI_k \cdot TI_{k+1})}$, $\beta_k \cdot TI_k \cdot TI_{k+1}$), and $\beta_k = 0$ (no redundancy correction). For all versions we calculated the optimal frequency-weighting α_k and redundancy β_k. As described previously, this optimization was performed in an iterative procedure based on the prediction accuracy of the observed intelligibility (CVC-word score) by the STI. This prediction accuracy is given by the vertical spread (standard deviation, s) around the best fitting third-order polynomial between these two variables.

In Table 2.5.1 the standard deviations are given for the two redundancy correction methods and for no redundancy correction, including all measuring conditions (26 frequency-transfer conditions for male speakers and 17 frequency-transfer conditions for female speakers at three signal-to-noise ratios), and for each signal-to-noise ratio separately. For the conditions with a fixed signal-to-noise ratio no difference is obtained by the application of a root in the equation as the TI values are either 1 or 0, indicating that a frequency band is or is not contributing. As can be seen in the table, a substantial improvement in

prediction accuracy is obtained by adding a redundancy correction. The application of the root in equation 2.5.1 does not have an important effect.

A part of the standard deviation (s) might be caused by statistical inaccuracy with respect to the values of the observed CVC-word scores: the standard error of the score. An indication of this standard error was obtained by splitting each of the CVC-word scores obtained at the various transfer conditions (16 CVC-word scores from 4 speakers and 4 listeners), into two values based on 8 CVC-word scores (2 speakers and 4 listeners). The standard deviation of these differences was found to be s_{score} = 5.26% for the male and s_{score} = 5.73% for the female speech. It can be shown that the corrected standard deviation (s_{corr}), reflecting only the error in the objective STI method, can be calculated by:

$$s_{corr} = \sqrt{(s^2 - \frac{1}{4} s^2_{score})} \qquad (2.5.2)$$

These corrected standard deviations for the calculations based on all 78/51 conditions in Table 2.5.1, are given between brackets.

Table 2.5.1 Prediction accuracy expressed by the standard deviation (%) around the best fitting third-order polynomial of the CVC-word score for three versions of equation 2.5.1. The calculations include a redundancy correction according to: $\beta_k.\sqrt{(TI_k.TI_{k+1})}$, $\beta_k.TI_k.TI_{k+1}$, and $\beta_k = 0$ (no redundancy correction). The calculations are based on the 78 male speech conditions and 51 female speech conditions for the three signal-to-noise ratios of 15 dB, 7.5 dB, and 0 dB. The values between brackets represent the corrected standard deviation.

Male speech Redundancy correction:	$\beta_k.\sqrt{(TI_k.TI_{k+1})}$	$\beta_k.TI_k.TI_{k+1}$	$\beta = 0$
all	4.73 (3.93)	4.49 (3.64)	6.76 (6.23)
SNR=15dB	3.37	3.37	6.62
SNR=7.5dB	3.69	3.69	6.41
SNR=0dB	3.19	3.19	4.25
Female speech **Redundancy correction:**	$\beta_k.\sqrt{(TI_k.TI_{k+1})}$	$\beta_k.TI_k.TI_{k+1}$	$\beta = 0$
all	4.21 (3.08)	4.60 (3.60)	5.99 (5.26)
SNR=15dB	2.38	2.38	4.98
SNR=7.5dB	2.79	2.79	4.48
SNR=0dB	2.47	2.47	3.85

The deviation from the additivity assumption has led to an extension of the original model (Eq. 2.1.3) as used previously by French and Steinberg (1947) for the AI, and by Steeneken and Houtgast (1980) for the STI. A redundancy correction was added (Eq. 2.5.1) for several reasons: the difference in intelligibility scores for conditions with adjacent or nonadjacent frequency bands and the finding of a high correlation between adjacent octave bands. The deviation from the additivity assumption for different adjacent and nonadjacent frequency-band combinations has been reported earlier by Pollack (1948), Licklider (1959), Kryter (1960), and recently by Grant and Braida (1991). Grant and Braida summarized four possible reasons to explain the deviation from the additivity assumption: (1) overlap of the slopes of the filters used to obtain gaps in the frequency transfer, (2) self-masking of the speech signal for closely-spaced frequency bands, (3) the effect of spectrum sampling (a frequency spectrum is better described with a given number of observations distributed all over the frequency band rather than concentrated in a smaller range), (4) a high degree of correlation between adjacent frequency bands. This last factor agrees best with our extension of the additivity model, although other solutions could be possible as well.

We applied an *arcsine* transformation to the subjective data as proposed by Studebaker et al. (1987). Such an *arcsine* transformation rescales data which cluster at a maximum or a minimum value on the scale. For example, such a clustering effect is obtained with sentence intelligibility tests (see chapter 3, Fig. 3.2.1), where a saturation at 100% is already obtained at poor-to-fair transmission conditions. As the subjective data obtained in the experiments described above are fairly equally distributed along the scale, no different frequency-weighting factors were obtained by applying this transformation (see section 3.3.4, Fig. 3.3.1).

The optimal frequency-weighting function was found to be rather independent of several parameters such as signal-to-noise ratio, and whether male or female speakers were used. This frequency weighting, including the redundancy correction, is given in Figs 2.3.6, 2.3.9, and 2.3.10. All these figures show a high contribution for the octave-band with centre frequency 2 kHz and a low redundancy correction for this octave-band with adjacent frequency bands. The opposite behaviour was found for the lower and upper frequency bands which show a high redundancy with adjacent frequency bands. This may be due to: (a) a high correlation within speech spectra at the lower and the upper end of the spectrum, and (b) the selection of octave-bands. A selection of two or three smaller frequency bands around 2 kHz would have led to a lower

frequency weighting (for each frequency band individually), and an increased redundancy correction. This was also found for the calculations performed with the Kryter data (see section 2.4). The extreme situation of $\alpha \approx \beta$ in a given frequency region would imply a total dependency of adjacent frequency bands in that region. In this case a redefinition of the distribution of the frequency bands along the frequency scale should be considered. This cannot be evaluated with the experiment described above as the selection of the different frequency-transfer conditions is related to octave bands.

Remaining questions

The frequency-weighting factors, as obtained independently for consonants and vowels, are quite dissimilar (Fig. 2.3.15). As shown in Fig. 2.3.11 the percentage correct for both groups of phonemes, obtained for the various frequency-transfer conditions, are also very dissimilar. However, even separating phonemes simply into consonants and vowels is rather arbitrary; other groupings of phonemes are more likely. As Miller and Nicely indicated for some distorted speech conditions, several independent grouping parameters (speech production features) can be distinguished (voicing, nasality, etc.). Further analysis of the observed results is required to identify phonemes or groups of phonemes with a similar response at various transmission conditions. This will be described in chapter 3.

The experimental design of the study described above is based on many different frequency-transfer conditions in combination with a masking noise with a spectrum equal to the long-term speech spectrum. This allows a prediction of the physical parameters of the different transfer conditions, without the need of performing actual measurements of the signal-to-noise ratios. If speech items other than the CVC words are studied, such as phoneme groups (each with a different long-term spectrum), different signal-to-noise ratios will be obtained. It is then required to correct the signal-to-noise ratio with respect to the difference of the long-term spectrum of the speech signals of the phoneme group considered and the long-term spectrum of speech (connected discourse) as used for the experiments. Another possibility is to measure these signal-to-noise ratios with an adequate test signal rather than to predict them. A study focused on the relation between optimal frequency weighting for different types of speech items is described in chapter 4.

Conclusions

- The addition of a redundancy correction between adjacent frequency bands for estimation of intelligibility by an additive model leads to a more accurate prediction (lower standard deviation).

- Optimal frequency-weighting factors do not depend on the sex of speakers or the signal-to-noise ratio.

- Optimal frequency-weighting factors depend on type of phoneme; e.g. different frequency-weighting factors were obtained for consonants and vowels. This effect requires further attention.

- A better prediction of intelligibility than found by the original authors was obtained by using a redundancy correction for two earlier studies (Kryter, 1960; Steeneken and Houtgast, 1980).

- A robust relation between signal-to-noise ratio and information content was found for male and female speech at various intelligibility levels.

3 SUBJECTIVE PHONEME, WORD, AND SENTENCE INTELLIGIBILITY MEASURES

Summary

A frequently used measuring method for obtaining word and phoneme scores is based on nonsense syllables of the CVC-word type (Consonant-Vowel-Consonant). This type of test word was used for the studies on band-pass limiting and noise masking (chapters 2 and 4), and on communication channels (chapter 5).

The effect of several parameters on the scores, such as speakers, listeners, and speaker sex was analyzed. Also, the relation between the CVC-word scores and the individual phoneme type (initial consonants, vowels, and final consonants), and phoneme-group scores (fricatives, plosives, vowel-like consonants, and vowels) was analyzed. For phonetic grouping, principal-component analysis and multi-dimensional scaling were used. It was found that at the phoneme level, four groups can be identified with a fairly similar response for various transmission conditions. The individual scores for phoneme groups can be used as a diagnostic tool for improving communication systems and to improve predictive intelligibility measurements.

The relation between sentence intelligibility and the CVC-word score was also studied. For this study the Speech Reception Threshold (SRT) was used. This method tunes a condition to a 50% sentence intelligibility level by adding noise at a required signal-to-noise ratio. Beside the usual 50% sentence intelligibility level, the 25% and the 75% sentence intelligibility level could also be obtained from our data. A good relation could be found with the CVC-word score, as well as with combinations of some phoneme-group scores.

3.1 Introduction

Subjective intelligibility tests and speech quality tests are used in many disciplines, but mainly in speech audiometry, evaluation of speech transmission systems, evaluation of room acoustics, communication in noise, and research in speech perception. One of the first applications of intelligibility testing was in speech audiometry. Reviews on this subject are given by Bosman (1989) and by Feldmann (1960). The evaluation of speech transmission systems was initiated by Fletcher and Steinberg (1929) who developed several intelligibility tests for the evaluation of telephone systems. Bolt and MacDonald (1949) and Beranek (1954) describe intelligibility assessment in auditoria for speech degraded by noise and reverberation.

Miller and Nicely (1955) focused on speech material and made a major contribution to the characterization of the different phonemes spoken over voice communication systems. On the basis of their results they suggested five articulatory features: voicing, nasality, affrication, duration, and place of articulation.

In this thesis we deal with specific transfer conditions for a variety of speech transmission channels. The subjective evaluation is performed with a CVC-word test (Consonant-Vowel-Consonant) based on nonsense words embedded in a carrier phrase. The word score is considered to be a predictor of speech intelligibility.

Various studies (Kryter, 1962b; Steeneken and Houtgast, 1980) have shown that the intelligibility of speech signals degraded by transmission through a communication channel can be predicted from various physical properties of that channel. This prediction is based on a weighted contribution of a number of relevant frequency bands and is robust for various parameters, for example male/female speech, signal-to-noise ratio, etc. However, as we showed in chapter 2, a separate validation for consonant and vowel scores resulted in a different set of optimal frequency-weighting factors for the different frequency bands. Therefore, rather than the arbitrary optimization with consonants and vowels, a more advanced analysis of the data is required to investigate the perception of phonemes in degraded speech signals.

In the next section various tests for the assessment of speech communication systems are discussed and the arguments for the CVC-word test are presented. The experiments on band-pass limiting and communication channels as described in chapters 2 and 5 provided a useful data set for the validation of the subjective test method, which are presented in section 3.4. The relation between the test results of the CVC-word test and sentence intelligibility is described in section 3.5.

3.2 Overview of some subjective intelligibility and speech quality tests

A number of subjective intelligibility tests have been developed for the evaluation of speech communication channels (Fletcher and Steinberg, 1929; Egan, 1944; Miller and Nicely, 1955; House et al., 1965; Voiers, 1977a). In general, the choice of the test is related to the purpose of the study: are we comparing and rank-ordering systems, are we evaluating a system for a *specific application*, are we supporting the *development* of a system, or are we studying speech *perception*? For each type of application a different test may be appropriate. An overview of assessment methods for speech synthesis systems is given by Pols (1991), and an overview focused on the assessment of speech communication systems is given by Steeneken (1992).

Subjective intelligibility tests can be largely categorised by the speech items tested and by the response procedure used. The smallest items tested are at the segmental level, i.e phonemes. Other test items are CV, VC, and CVC combinations, nonsense words, meaningful words, and sentences.

Besides intelligibility scores, speech quality can also be determined by questionnaires or scaling methods, using one or more subjective scales such as: overall impression, naturalness, noisiness, clarity, etc. Speech quality assessment is normally used for communications with a high intelligibility, for which most tests based on intelligibility scores cannot be applied because of ceiling effects.

The overview given below describes representative tests from this segmental level up to sentence level, as well as tests giving a general impression of transmission or speech quality.

Tests at phoneme and word level

A frequently used test for determining phoneme scores is the rhyme test. A rhyme test is a forced-choice test in which a listener, after each word that is presented, has to select his response from a small group of visually presented alternatives. In general, the alternatives only differ with respect to the phoneme at one particular position in the test word. For example, for the Dutch language and for a test with a plosive in the initial consonant position, the possible alternatives might be: Bam, Dam, Pam, Tam, Kam. A rhyme test is easy to apply and does not require much training of the listeners. Frequently used rhyme tests are the Modified Rhyme Test (MRT, testing consonants and vowels) and the Diagnostic Rhyme Test (DRT, testing initial consonants only).

The MRT is based on six alternatives (Fairbanks, 1958; House et al., 1965), the DRT is based on two alternatives (Voiers, 1977; Peckles and Rossi, 1973; Sotscheck 1982; Steeneken, 1982). For the DRT the alternatives are based on testing single articulatory features mainly according to the concept defined by Miller and Nicely (1955). Studies have shown that the DRT, because of the limited number of alternatives, is less sensitive and may force listeners to respond differently from their perceptual impression (i.e the phoneme actually heard by the listener might not be included in the two alternatives presented by the test, Steeneken, 1987b; Greenspan, 1989).

A more general approach is obtained with a test with an open response, such as with monosyllabic word tests (Fletcher, 1929; Egan, 1944). Open response tests make use of short nonsense or meaningful words of the CVC type. Sometimes VCV words, CV words, VC words, CCVC words, or CVCC words are used. This may depend on features of the particular language (for example

Italian has no closed syllables) or the wish to evaluate specific clusters such as consonant clusters or diphone clusters. With nonsense words and an open response, the listener can respond with any combination of phonemes corresponding to the type of word as defined beforehand. This procedure requires extensive training of the listeners.

The test results can be presented as phoneme scores and word scores but also as confusions between the initial consonants, vowels, and final consonants.

The confusion matrices obtained with open response tests provide useful (diagnostic) information for improving the performance of a system (Steeneken, 1986; Bos and Steeneken, 1991). Multidimensional scaling techniques may help to visualize the relations between the stimuli.

With word tests it is recommended to embed the words in a carrier phrase. Such a carrier phrase (which is neglected in many studies) will cause representative echoes and reverberation in conditions with a distortion in the time domain. Also automatic gain control (AGC) settling will be established by the carrier phrase. An important aspect of using a carrier phrase is also that it stabilizes the vocal effort of the speaker during the pronunciation and that it reduces the vocal stress on the test words. Finally it can function as a cue to the listener that the next test word is going to be presented.

Tests at sentence level

Sentence intelligibility is sometimes measured by asking the subjects to *estimate* the percentage of words correctly heard on a 0-100% scale. This scoring method tends to give a wide spread among listeners. Sentence intelligibility saturates to 100% at poor signal-to-noise ratios, the effective range is small (see Fig. 3.2.1).

The speech reception threshold (SRT) measures word or sentence intelligibility against a level of masking noise. The listener has to recognize a word or sentence presented at a fixed level and masked by noise at a variable level. After a correct response the noise level is increased, while after a false response the noise level is decreased. This procedure leads to an estimation of the noise level where a 50% correct identification of the words or sentences is obtained (Plomp and Mimpen, 1979). The quality of the speech (and/or of the listener) is related to the amount of noise required for masking. The procedure has the advantage that it can be performed with naive listeners and gives very reproducible results. The standard deviation of the masking noise level for repeated tests with the same speaker and listener is close to 1.5 dB.

Recently the use of anomalous sentences has been getting a great deal of attention in combination with the assessment of speech synthesis systems. These

syntactically correct but semantically anomalous sentences consist of approximately seven words (Benoit, 1990). The words are taken from sets of common monosyllabic words from which a virtually unlimited number of sentences can be generated randomly according to some predefined grammatical structures. The robustness of this test has not yet been shown, but it is being studied in a current European Esprit programme on assessment measures.

Quality rating

Quality rating is a more general method, used to evaluate the user's acceptance of a transmission channel or speech output system. The claim of some investigators (Goodman and Nash, 1984) is that a quality rating reflects the total auditory impression of speech by a listener and can be used to discriminate between a number of intervals ranging from excellent to bad. For quality ratings, normal test sentences or a free conversation are used to obtain the listener's impression. The listener is asked to rate his impression on a subjective scale such as the five-point scale: bad, poor, fair, good, and excellent. Different types of scales are used, including: intelligibility, quality, acceptability, naturalness etc. Quality rating or the so-called Mean Opinion Score (MOS) gives a wide variation among listener scores (Steeneken, 1987b). The MOS does not give an absolute measure since the scales used by the listeners are not calibrated. Therefore the MOS can be used only for rank-ordering conditions. For a more absolute evaluation, the use of reference conditions is required as an anchor.

Other methods (more or less related to the scaling methods used with the MOS) are paired comparison, categorial and magnitude estimations, semantic scaling, and the Diagnostic Acceptability Measure (DAM; Voiers, 1977b).

Relation between various measures

Fig. 3.2.1 gives, for five intelligibility measures, the score as a function of the signal-to-noise ratio of speech masked by noise (Steeneken, 1987b). This gives an impression of the effective range of each test. The given relation between intelligibility scores and the signal-to-noise ratio is valid only for noise with a frequency spectrum similar to the long-term speech spectrum, which makes the signal-to-noise ratio the same for each frequency band. This is for instance the case with voice-babble. A signal-to-noise ratio of 0 dB then means that speech and noise have an equal spectral density.

As can be seen from the figure, the CVC-nonsense words discriminate over a wide range, while meaningful test words[1] have a slightly smaller range

[1] Meaningful test words are normally phonetically balanced (PB), hence the frequency distribution of the phonemes is representative for the language used.

(Anderson and Kalb, 1987). The digits and the alphabet give a saturation at a signal-to-noise ratio of -5 dB. This is due to: (a) the limited number of test words and (b) the fact that recognition of these words is controlled mainly by the vowels rather than by the consonants. Vowels have an average level approximately 5 dB above the average level of consonants, and are therefore more resistant to noise. On the other hand nonlinear distortions, such as clipping, will have a greater impact on vowels than on consonants. Therefore the use of the digits and the alphabet, for which recognition is based mainly on vowels, may lead to misleading results. This is indicated in Fig. 3.2.2. In this figure the initial-consonant score is given versus the vowel score as obtained from CVC-word tests for 78 different transmission conditions. The graph shows that a high vowel score and a low consonant score can be obtained for one type of channel (e.g. band-pass limiting) while conversely a low vowel score combined with a high consonant score can be obtained for another type of channel (e.g. peak clipping). This indicates that the exclusive use of either consonants or vowels in a subjective test may lead to an incorrect evaluation of the transmission quality. A combination of consonants and vowels, as with CV or CVC words, is required.

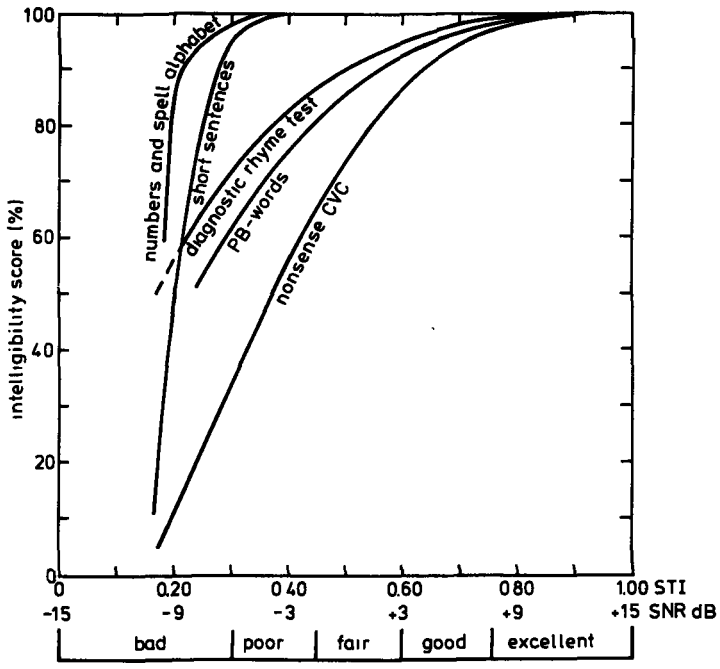

Fig. 3.2.1 Qualification of some intelligibility measures and their relation with signal-to-noise ratio for noise with a spectrum shaped according the long-term speech spectrum.

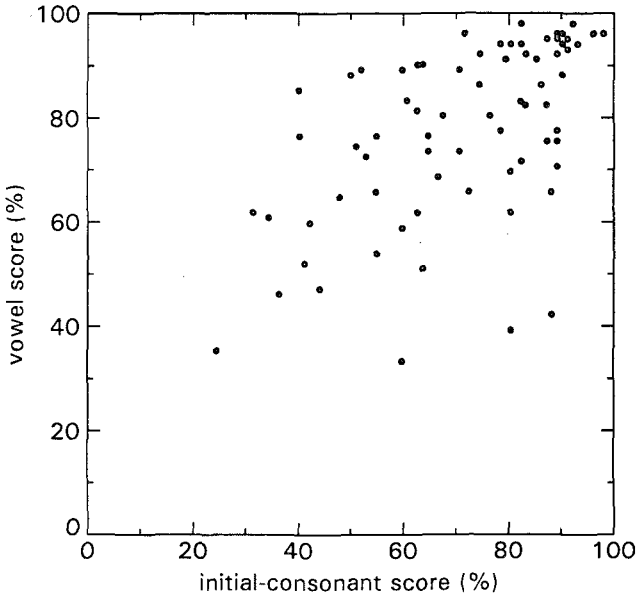

Fig. 3.2.2 Initial-consonant score versus vowel score obtained from CVC words for 78 transmission conditions with various combinations of bandwidth, noise, and signal-to-noise ratio.

The reproducibility of a test strongly depends on the number of speakers and listeners used for the experiments. When a session in which N conditions are measured is repeated in a second session, the test-retest reproducibility can be quantified by an index such as Cronbach's α (Cronbach, 1951).

$$\alpha = \left(1 - \frac{S_B^2}{S_W^2} \right) \frac{N}{N-1} \qquad (3.2.1)$$

The index α reflects the ratio of the total variance of the intelligibility scores among all test conditions (within-session variances, S^2_W), and the variance between replications of the test conditions (between-session variances, S^2_B), of the speaker-listener scores. Perfect reproducibility corresponds to $\alpha \approx 1$, and α tends to zero as reproducibility gets worse.

Fig. 3.2.3 shows the value of α as a function of the number of speaker-listener pairs for some of the intelligibility tests discussed above (Steeneken, 1987b). The graph shows that the index increases as a function of the number of

speaker-listener pairs. In general the more complex a test (from simple vowels to nonsense words), the more reproducible the test results are (close to $\alpha = 1$). This may be due to the effective range of each test method (Fig. 3.2.1). A CVC test with four speaker-listener pairs gives the same α-index as a DRT with nine speaker-listener pairs.

For our present subjective measurements we mainly used the CVC test. This test was selected for various reasons but mainly because of the variation in the speech items tested (word scores and phoneme scores are obtained), the ease of administering the test, and the reproducibility. Nevertheless, nonsense words are artificial and not representative of connected discourse. For this reason the relation between CVC-word tests and sentence intelligibility was studied separately (see section 3.5). The speech reception threshold was used as a measure of sentence intelligibility.

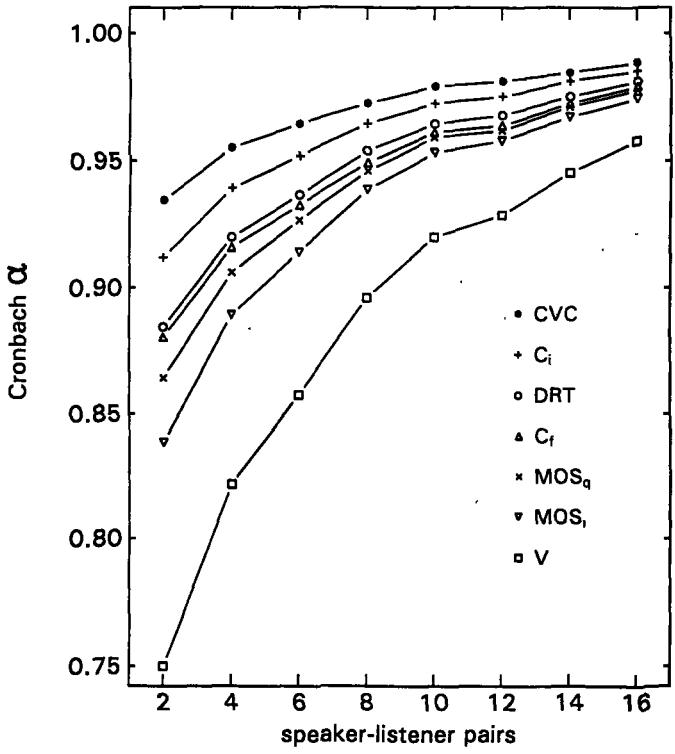

Fig. 3.2.3 Test-retest index, α, as a function of the number of speaker-listener pairs for some intelligibility measures (vowels, mean opinion scores based on quality and intelligibility, initial and final consonants, diagnostic rhyme test, and CVC-word scores).

3.3 Description of the CVC-word test and the scoring method

3.3.1 *Speech material*

The use of CVC-nonsense words permits experiments with an open response task and also discriminates between conditions with good and excellent quality where other tests are saturated (Steeneken 1987b). For an effective use of the open response design a word list with all possible phonemes: initial consonants, vowels, and final consonants, is required. However, to keep the database manageable, phonemes with a frequency of occurrence less than approximately 1% in spoken language are omitted (v.d. Broecke et al., 1987). For the Dutch language this results in 17 initial consonants (C_i), 15 vowels (V), and 11 final consonants (C_f). Each word list consists of 51 words, hence each initial consonant is present three times in each list. For the vowels and final consonants additional repetitions of some of the phonemes are used in order to get multiples of 17. This arrangement provides an approximation of equal phonetic balancing.

The phonemes in orthographic, SAMPA, and IPA notation (SAMPA, 1989) are listed below. The additional phonemes (to reach a total of 17), are shown in brackets.

C_i:	p, t, k, b, d, g, f, s, h, v, z, m, n, l, r, j, w	(orthographic)
	p, t, k, b, d, x, f, s, h, v, z, m, n, l, R, j, w	(SAMPA)
	p, t, k, b, d, χ, f, s, h, v, z, m, n, l, r, j, ʋ	(IPA)
V:	ie, ee, i, e, u, aa, a, uu, eu, o, oo, oe, ei, ui, au (a,e)	(orthographic)
	i, e:, I, E, Y, a:, A, y, 2:, O, o:, u, Ei, 9y, Au	(SAMPA)
	i, e, I, ɛ, œ, a, ɑ, y, ø, ɔ, o, u, ɛi, œy, ɑu	(IPA)
C_f:	p, t, k, f, s, g, m, n, ng, l, r (l,t,r,s,n,m)	(orthographic)
	p, t, k, f, s, x, m, n, N, l, R	(SAMPA)
	p, t, k, f, s, χ, m, n, ŋ, l, r	(IPA)

All C_iV combinations occur in the Dutch language. However, some VC_f combinations do not exist. These restrictions are:

ijr, uir, aur, (diphtong - r)
ieng, oeng, uung, ijng, uing, aang, aung, oong, eung, eeng (diphtong or long vowel - ŋ).

With these restrictions, the test words within a list of 51 words are random combinations of CVC's, resulting in both nonsense as well as meaningful words. Some undesirable words ("dirty" words) were replaced. We also arranged that successive test words never had two or more phonemes in common.

An example of a CVC-word list in orthographic notation:

dies	fijs	zek	lal	van	waum
kel	joof	neng	nig	tum	suug
keun	huik	joer	beum	puit	beun
sop	fag	buim	mies	neer	ris
hung	rool	lok	gan	reem	zaul
soon	maar	giep	tuum	wet	par
faut	kijt	haar	puun	ges	dit
jol	vijp	doer	vef	zaf	wung
teel	moet	laas			

The CVC words are embedded in five different carrier phrases. The main reasons for the use of a carrier phrase were stated above. In particular, some types of distortions cannot be studied adequately with isolated words. With reverberation or other distortions in the time domain for example, the masking introduced by the first part preceding the test word is essential, while the second part is essential to mask echoes of the test word. The effect of using a carrier phrase for speech signals distorted by reverberation is shown in a study by Houtgast and Steeneken (1984). This study was based on an international experiment from eleven laboratories using different test material for the evaluation of the same conditions consisting of combinations of reverberation and noise.

The five different carrier phrases used here consist of one or two words preceding the CVC word and one word following the CVC word. The phonemes just before and just after the CVC words have been selected to induce different masking effects in the case of echoes and reverberation.

Examples of test words in carrier phrases:

Attentie	*dies*	einde
En nu	*fijs*	over
En zo	*zek*	onder
Versta	*lal*	uit
Volgende	*van*	aan

3.3.2 *Speakers*

For CVC tests, 4-8 speakers and 4-8 listeners are generally used. It has been argued that the reproducibility of a test strongly depends on the number of speaker-listener pairs and that depending on the test method 8-16 of these pairs are required. The amount of variance among individual results is equal for speakers and listeners, so in a balanced experiment the number of speakers and listeners should be equal. Therefore a speech data-base of word lists was recorded for four male and four female speakers. The speakers were native Dutch between 25 and 50 years of age, and they did not have a noticeable accent. Each speaker read 50 different word lists of embedded CVC words as described above. This results in a total of 400 different lists. When recording a list the speaker was situated in an anechoic room (noise level below 20 dBA) and read the carrier phrase with the test word from a display. Each phrase was presented separately in a sequence of one phrase every 3 seconds. Hence, a list of 51 words and phrases lasts 153 seconds.

This data-base was used for both experiments described in this thesis (band-pass limiting and communication channels, chapters 2, 4, and 5). The data-base was recorded on digital audio tape. A complete description of these recordings and the calibration procedure is given by Steeneken, Geurtsen, and Agterhuis (1990).

3.3.3 *Listeners*

A full experimental design of the 78 male conditions on band-pass limiting (chapter 2) and the 68 communication channel conditions (chapter 5) yields for four speakers a total of 584 lists. For the 51 female conditions on band-pass limiting and the 68 communication channel conditions for four speakers a total of 476 lists is obtained. With this huge design it was not possible to use the required number of listeners (equal to the number of speakers) within a reasonable time period. We therefore used an incomplete design of two groups of four female listeners. Each group of four listeners listened to only two of the male and two of the female speakers. The four listeners of a group worked simultaneously in an anechoic room (room noise level below 20 dBA). For reasons of good acoustical reproduction, the stimuli were presented binaurally via headphones at a level of approximately 70 dB SL (sensation level). The listener was requested to type what she had heard on a silent keyboard. The responses of the listeners were collected and stored by a computer.

The listeners were students between 20 and 25 years of age, and they were paid for their participation. All listeners had normal hearing: no hearing loss of

more than 15 dB in the frequency range between 125 Hz and 8 kHz was found. Each listener group worked for half a day, five days a week, over almost four weeks. A few days were skipped because of system failure or absence of one or two listeners. Approximately 40 lists were presented during a working day. Lists were presented in sequences of five to ten lists with breaks in between. The listeners were trained for two half days before the actual data collection started.

3.3.4 *Scoring program*

A scoring program was used for the calculation of the word scores and the phoneme scores. This was performed for each of the 16 speaker-listener combinations for the male and female speech conditions. In addition to the mean word score and mean phoneme-group score, the median values were calculated. This median is more robust against any drop outs, i.e. extreme individual scores due to system errors or a temporary loss of attention of a listener. Separate confusion matrices were obtained for the phoneme groups at the three phoneme positions of the CVC words. An overview of the output of the scoring program is given in appendix A5.

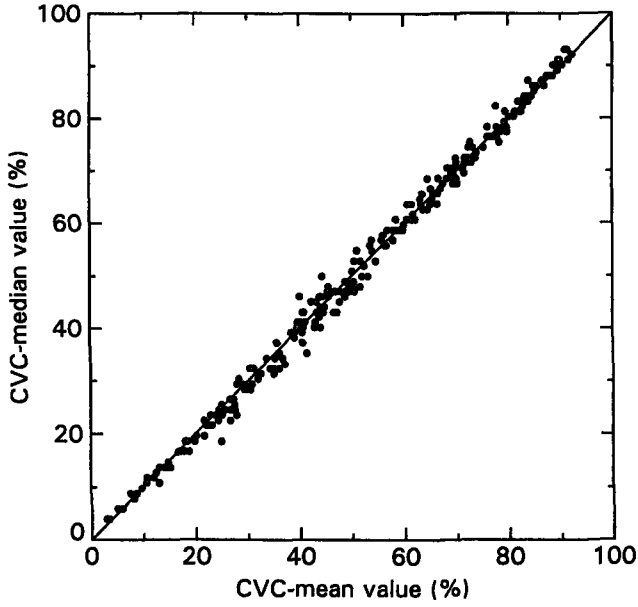

Fig. 3.3.1 Relation between the median values and the mean values of the 16 CVC-word scores (4 speakers, 4 listeners) for all 265 conditions of both experiments. The correlation coefficient between the two sets of data is r=0.998.

Unless otherwise noted, the validation of the word scores is based on the mean values of 16 speaker-listener pairs. The relation between these mean values and the median values of the CVC-word scores of all male and female conditions of both experiments (a total of 265 conditions) is shown in Fig. 3.3.1. The correlation coefficient between the two averaging methods is r = 0.998. The figure shows no extreme data points of the kind expected in case of severe drop outs. It should also be noted that a continuous distribution of the data-points along the scale is found, indicating an equal representation of all word intelligibility levels.

3.3.5 *Learning effects*

A small subset of all conditions was repeated several times during the total experiment. In this way the effect of learning (test-retest) could be quantified at several points in time during the course of the experiments. For listener group 1, 10 lists from the first, second, sixth, and eleventh listening day after the training days were presented again at the end of the experiment. Group 2 ran out of time and only a repetition of lists from the first, fifth and ninth day could be obtained. The days selected for the repetition were (except day 1 and 2) the first working days of a new week.

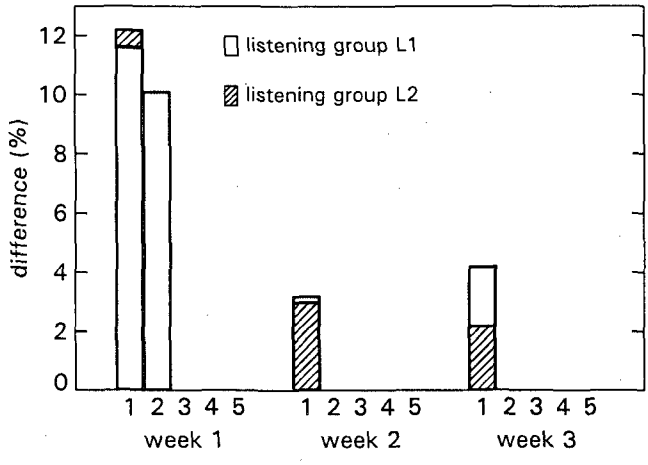

Fig. 3.3.2 Differences between the mean CVC-word scores of the working day indicated and the last day of the test, for a minimum of 10 of the same conditions, and the same speakers and word lists.

In Fig. 3.3.2 a bar diagram is given, representing the mean absolute differences between the CVC-word scores obtained for the working day indicated and the repetition at the end of the experiments for the same speaker-listener pairs, the same word lists, and the same transfer conditions. These difference scores are given for the two listener groups separately. Between the first two days and the end of the experiments a learning effect of approximately 10% is found, the effect of learning decreases to 3% between the second week and the end of the experiments. Similar results were found for the individual phoneme-group scores (not shown). A significance test indicated that, with the variances obtained with these experiments for 4 speaker-listener pairs, a difference greater than 1.2% is significant at a 1% significance level ($p = 0.01$).

The presentation sequence of the two experiments was balanced for the two listening groups, e.g. listener group 1 performed the experiment on band-pass limiting and noise first, listener group 2 performed the experiments on communication channels first. Within each of the two experiments the presentation of word lists was in random order with respect to speaker and condition. The measuring results of the first day after the training day were not used as they were repeated at the end of the experiments. Because of this design it may be expected that the results are not systematically affected by the significant effects of learning.

3.4 Evaluation of the CVC-word score and individual phoneme scores

3.4.1 *Variation among speakers and listeners*

An analysis of variance (ANOVA) was carried out on the mean word scores and the mean scores for the initial consonants, vowels, and final consonants. It was confirmed (by a procedure within the ANOVA program) that the data were normally distributed to fulfil the requirements of ANOVA. As the experimental design, due to practical reasons, was limited, not all combinations of speakers, listeners, and conditions could be investigated. This incomplete design results in an analysis of variance with nested variables. Nests were used for the speakers and the two groups of listeners.

We performed separate analyses for the two experiments on band-pass limiting and communication channels. As the experimental design of the experiment on band-pass limiting was different for the male (78 conditions) and female (51 conditions) speakers, the ANOVA was also separated.

Table 3.4.1 Analysis of variance (ANOVA) on CVC-word scores, initial-consonant scores, vowel scores, and final-consonant scores. The effect of speakers and listeners is given for a mixed analysis of male and female speakers, and for male and female speakers separately. The significance (p) and the mean scores (%) were calculated for the 68 communication channel conditions according to the description in appendix A3.

Variable	CVC	C_i	V	C_f
Speaker sex				
significance	0.07	0.33	< 0.01	0.08
male speakers	59.5	77.3	86.1	80.8
female speakers	50.4	74.3	77.7	74.3
Male speakers				
significance speakers	< 0.01	< 0.01	< 0.01	< 0.01
M1	68.2	83.9	90.7	84.5
M2	59.8	78.3	86.2	79.7
M3	54.7	71.1	85.5	80.3
M4	55.5	75.8	82.1	78.8
significance listeners	< 0.01	0.01	< 0.01	< 0.01
M1,2 - L1	67.0	83.1	90.3	82.6
M1,2 - L2	62.9	79.1	88.5	83.3
M1,2 - L3	63.1	81.7	87.4	80.5
M1,2 - L4	63.1	80.5	87.6	81.9
M3,4 - L5	54.9	73.9	83.1	77.6
M3,4 - L6	54.9	72.2	84.8	80.7
M3,4 - L7	58.9	76.0	86.4	80.7
M3,4 - L8	51.9	71.6	80.8	79.3
Female speakers				
significance speakers	< 0.01	< 0.01	< 0.01	< 0.01
F1	48.5	73.0	76.3	72.0
F2	57.6	79.6	81.0	79.8
F3	51.6	74.4	79.3	75.7
F4	43.9	70.3	74.1	69.7
significance listeners	0.05	< 0.01	0.02	< 0.01
F1,2 - L1	55.9	78.0	81.8	76.4
F1,2 - L2	52.6	75.0	78.5	77.2
F1,2 - L3	50.5	75.8	76.7	73.7
F1,2 - L4	53.4	76.3	77.7	76.4
F3,4 - L5	46.8	72.9	74.8	70.6
F3,4 - L6	48.1	71.7	77.9	73.9
F3,4 - L7	51.2	74.3	79.7	74.1
F3,4 - L8	44.9	70.5	74.5	72.2

Table 3.4.2 Analysis of variance (ANOVA) on CVC-word scores, initial-consonant scores, vowel scores, and final-consonant scores. The effect of speakers and listeners is given for an analysis of male and female speakers separately. The significance (p) and the mean scores (%) were calculated for the 78/51 conditions according to the description in appendix A2.

Variable	CVC	C_i	V	C_f
Male speakers				
significance speakers	< 0.01	< 0.01	< 0.01	<0.01
M1	54.9	77.3	78.6	79.0
M2	47.8	71.1	76.8	75.0
M3	50.0	68.5	79.3	80.1
M4	50.7	72.2	78.4	78.0
significance listeners	0.01	0.13	< 0.01	< 0.01
M1,2 - L1	53.3	75.8	79.9	76.7
M1,2 - L2	52.6	73.3	79.2	79.7
M1,2 - L3	49.6	74.7	76.0	75.3
M1,2 - L4	50.0	73.1	75.9	76.4
M3,4 - L5	47.0	69.7	75.6	76.0
M3,4 - L6	50.6	69.4	79.5	80.5
M3,4 - L7	53.4	72.3	81.5	80.0
M3,4 - L8	50.2	70.2	78.8	79.5
Female speakers				
significance speakers	< 0.01	< 0.01	< 0.01	< 0.01
F1	46.9	71.0	76.6	72.7
F2	49.7	75.2	73.9	77.6
F3	49.4	72.9	77.6	75.0
F4	43.9	71.1	72.3	70.7
significance listeners	< 0.01	< 0.01	< 0.01	< 0.01
F1,2 - L1	50.4	74.4	77.5	75.1
F1,2 - L2	50.0	72.7	76.8	77.0
F1,2 - L3	46.1	73.7	71.9	73.5
F1,2 - L4	46.8	71.6	74.8	74.9
F3,4 - L5	42.9	70.4	70.7	69.3
F3,4 - L6	46.7	71.6	75.3	73.9
F3,4 - L7	50.6	74.3	78.3	74.9
F3,4 - L8	46.4	71.8	75.4	73.3

The experiments on communication channels (68 conditions) could be combined for male and female speakers.

The levels of significance of the variables speaker sex, speaker identity, and listener identity are given in the Tables 3.4.1 and 3.4.2 for both experiments. The four male speakers are labelled M1-4, the four female speakers F1-4. The

listeners of group 1 are labelled L1-4 and of group 2 L5-8. The mean individual speaker scores and listener scores are also given. As speakers and listeners were nested a Newman-Keuls test was performed to analyse the significance within subgroups.

Table 3.4.1 indicates that for $p<0.08$ the difference between male and female speakers is significant for CVC words, vowels, and final consonants (mean male CVC-word score 59.5%, and mean female CVC-word score 50.4%). For the initial consonants $p=0.33$.

For most phoneme-type scores, the speakers (and also the listeners), were in most conditions significantly different, $p<0.01$. The Newman-Keuls test on individual speakers and individual listeners also indicated a significant difference between most of the individuals for all items tested. As the variation among speakers is roughly equal to the variation among listeners the assumption that we had to use an equal number of speakers and listeners is valid.

3.4.2 *Relations between CVC-words, phoneme groups, and phoneme types*

Representation of the scores and analysis of the results of a CVC-word test can be performed at various levels. The word score is often used as a measure of the intelligibility of a total communication channel, the phoneme-group scores (C_i, V, C_f) and the individual phoneme scores are more specific and may be used to convert the results to other intelligibility measures or for diagnostic purposes.

In this section we will make an analysis at these two levels namely: (1) the relation between the CVC-word scores and the group scores of the phonemes at the three phoneme positions in a CVC word, and (2) analysis at the phoneme level to identify phoneme clusters with correlated scores at various transmission conditions.

3.4.2.1 CVC-word scores and phoneme-group scores

The word score reflects the correct responses to the complete CVC word; hence, all three phonemes within a word must be correct. This word score is related to the scores of the three phoneme groups C_i, V, C_f. If these phonemes can be considered statistically independent, the word score should be equal to the product of the individual phoneme-group scores. We have verified this assumption. Both scores (the actual CVC-word scores and the scores predicted from the product of the probabilities for the three phoneme groups) are obtained from the scoring program (see appendix A5). We calculated the

correlation coefficient and the first-order regression between the two scores for both experiments and for the male and female speech. The correlation coefficient (r) and regression coefficients of the best fitting straight line (CVC_{word} = a . $CVC_{product}$ + b) are:

band-pass limiting	male speech	r = 0.998	n = 78
		a = 0.960	b = 4.21
	female speech	r = 0.998	n = 51
		a = 0.972	b = 3.86
communication channels	male speech	r = 0.997	n = 68
		a = 0.944	b = 5.66
	female speech	r = 0.998	n = 68
		a = 0.959	b = 4.67

Hence, a high correlation between the two measures is obtained and the (nonsense) word score can be predicted from the scores of the individual C_i, V, and C_f scores. However, a small systematic effect, suggesting some statistical dependency among the phoneme scores, was found for all four independent analyses. The regression coefficient "a" is slightly lower than 1.0 and an intercept of approximately 4% was found.

The calculation of the word score from the product of probabilities of the phoneme groups opens the possibility to obtain word scores based on other phoneme distributions than the "equally balanced" concept as used in this study. An example involving the PB-word score (Phonetically Balanced nonsense words) is given in appendix A5.

Consonant and vowel scores may, to some degree, be independent variables for different transmission conditions. Therefore, the correlation between these groups was calculated for the two experiments and for the male and female speech. The correlation coefficients are given in Table 3.4.3. The correlation coefficient between initial consonants and final consonants is r > 0.95, this indicates a close relation. The correlation coefficients between the vowels and the two consonant groups range from 0.567 to 0.738. This low correlation was already shown in Fig. 3.2.2 with a scatter plot of the initial consonants and vowels in male speech (r = 0.567). It indicates that a representative intelligibility test cannot be based on either consonants or vowels.

Table 3.4.3 Correlation coefficients between initial consonants C_i, final consonants C_f, and vowels V, for male and female speech and for the two experiments on band-pass limiting and communication channels.

Band-pass limiting

	Male speech			*Female speech*		
	C_i	C_f	V	C_i	C_f	V
C_i	-			-		
C_f	0.955	-		0.974	-	
V	0.567	0.653	-	0.682	0.722	-

Communication channels

	Male speech			*Female speech*		
	C_i	C_f	V	C_i	C_f	V
C_i	-			-		
C_f	0.951	-		0.967	-	
V	0.720	0.692	-	0.734	0.738	-

3.4.2.2 Relations between groups of phonemes

The separation in terms of consonants and vowels is quite arbitrary, as it relates only to the phoneme position in the CVC words. A robust grouping of phoneme types related to production phenomena such as voicing, duration, or an identical obstruction in the vocal tract might be of more interest. Therefore a detailed analysis at the phoneme level should be performed to identify phonemes or groups of phonemes with correlated scores for various transmission conditions. This is also of interest for the evaluation and comparison of transmission conditions and for the efficient design of robust intelligibility tests. Two kinds of information are available:

(a) the individual phoneme scores.

The scores of several phonemes or groups of phonemes may be correlated for various transmission conditions. This can be analyzed with a principal-component analysis.

(b) the confusions among phonemes (only at the same position in the CVC words).

Confusions among phonemes may indicate a close relation. For this purpose a multi-dimensional scaling technique can be used.

Both methods will be applied to the present data and lead to a selection of four groups of phonemes with quite unique properties at the various transmission conditions.

An analysis to identify phonemes or groups of phonemes with correlated scores at various transmission conditions can be performed by using a principal-component analysis on the individual phoneme scores. A total of 43 phonemes (17 initial consonants, 15 vowels, and 11 final consonants) as obtained from the CVC-word test were used for this principal-component analysis. The scores in both experiments (band-pass limiting, and communication channels), and for the male and female speech conditions were analyzed individually. These principal-component analyses were based on the correlation matrix of the input variables rather than on the co-variance matrix. This was done because of the wide variation of the mean scores and the standard deviations of the different phonemes as shown in appendix Table A4.5-A4.7.

In Fig. 3.4.1 the factor loadings for the first dimension from the original 43 input dimensions are given (from the experiment on band-pass limiting). A similar factor loading for different input variables (phonemes) means a similar scoring pattern and a high correlation between these variables. As can be seen in Fig. 3.4.1 the factor loadings for most phonemes are fairly similar for this dimension.

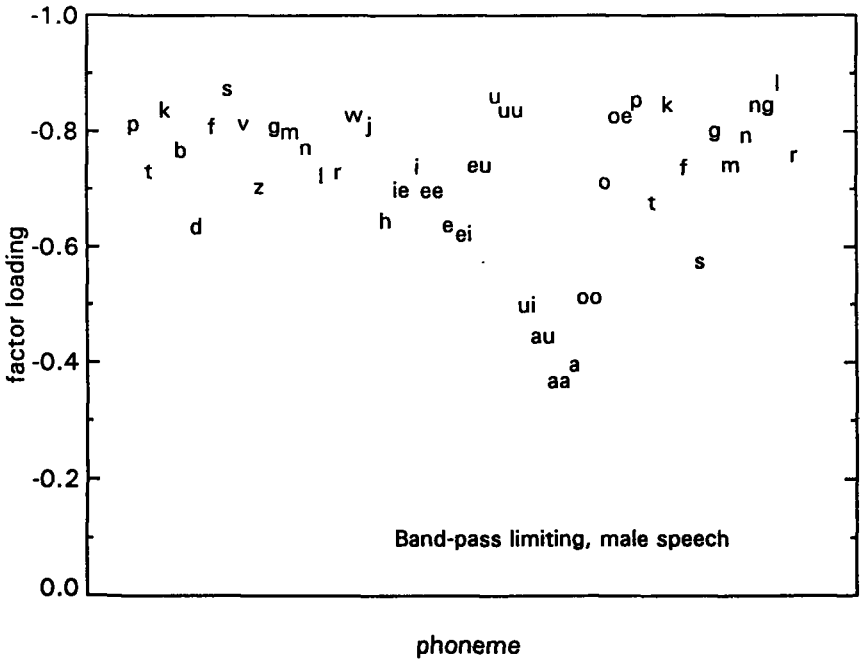

Fig. 3.4.1 Factor loading for the FIRST dimension of the 43 possible phoneme responses as obtained with the experiments on band-pass limiting and for male speech. The variance explained by this dimension is 54.5%. The phonemes are given by the orthographic notation.

It indicates that this dimension is related to an overall effect: the overall score. The variance explained by dimension 1 is 54.5%. A similar result for dimension 1 was obtained for the data set of the female speech and for the data sets of the communication channel conditions (not shown).

The factor pattern of dimension 2 is given in Fig. 3.4.2. It shows, in general, a separation between plosives and fricatives on one side (positive factor loading), and vowel-like consonants and vowels on the other side (negative factor loading). The amount of variance explained by dimension 2 is 15.3%. The variance explained by each higher dimension was below 7%, and the factor loadings did not show a systematic pattern. The factor loading of dimension 2 from the principal-component analysis applied to the data set of the female speech and to the data set of the communication channel experiment showed a similar pattern (not shown).

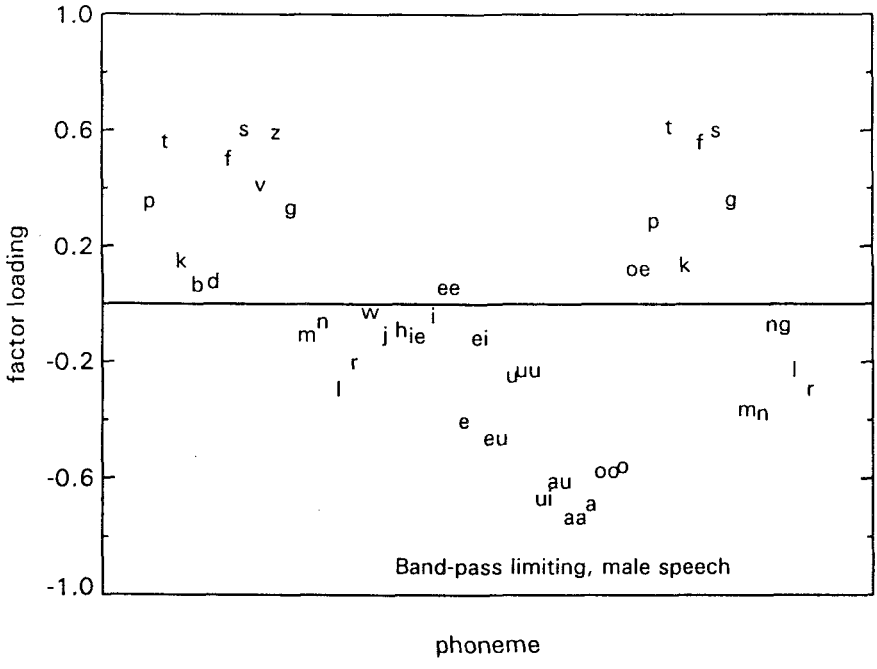

Fig. 3.4.2 Factor loading for the SECOND dimension of the 43 possible phoneme responses as obtained for male speech with the experiment on band-pass limiting. The variance explained by this dimension is 15.3%. The phonemes are given by the orthographic notation.

Our goal, a robust grouping of phoneme types from our scoring data, which show a relation with production phenomena such as voicing, duration, or an identical obstruction in the vocal tract, is not obtained with this analysis. This may be due to the nature of this analysis which is based only on the correlation between correct responses of phonemes over conditions.

Confusions between phonemes (i.e. phonemes which are perceptually similar) have so far not been taken into account. In Table 3.4.4 a confusion matrix for initial consonants, male speech, and 26 selected conditions of the band-pass limiting experiment is given.

Table 3.4.4 Confusion matrix for initial consonants, male speech, and 26 selected conditions of the experiment on band-pass limiting. Only the conditions without noise but with all combinations of band-pass limiting were used. The initial consonants are given by the SAMPA notation.

response	p	t	k	b	d	f	s	v	z	x	m	n	l	R	w	j	h
stimulus																	
p	1068	22	37	62	8	12	4	4	0	9	4	0	0	0	2	0	3
t	38	1099	51	0	29	6	3	3	1	3	0	0	0	0	0	0	2
k	52	58	1105	1	3	4	0	3	0	9	0	0	0	0	0	0	1
b	112	1	2	1002	41	0	0	0	0	0	11	7	2	0	50	3	3
d	8	113	16	49	1031	0	0	0	0	1	0	7	4	0	5	1	0
f	44	6	2	1	0	915	10	193	1	53	0	0	0	0	0	2	5
s	22	29	9	0	4	52	1037	13	41	14	0	0	0	0	0	1	1
v	6	3	1	4	1	337	11	739	35	34	0	0	0	2	43	11	8
z	2	5	0	1	4	6	161	27	934	3	0	0	0	7	24	44	18
x	9	2	4	0	0	26	0	11	0	1083	0	0	0	1	0	0	12
m	1	0	0	5	0	0	0	0	0	0	1068	113	25	1	6	2	15
n	0	0	0	0	0	0	0	0	0	0	111	1081	33	0	2	7	1
l	11	0	0	0	0	0	0	0	0	0	12	59	1112	12	7	25	4
R	1	1	0	2	0	0	0	2	0	15	0	2	9	1161	3	1	39
w	6	0	0	3	7	1	0	13	2	0	30	7	5	25	1065	27	17
j	0	0	0	0	0	0	0	2	5	0	2	11	13	6	21	1163	12
h	9	0	1	8	0	4	0	4	0	6	7	1	3	12	16	20	1145

The conditions with noise masking at a signal-to-noise ratio of 7.5 dB and 0 dB are omitted because of the lower scores and consequently the high number of random responses of the listeners. In the table the phonemes which show many mutual confusions are grouped together. This results in a clustering of the plosives (p, t, k, b, d), the fricatives (f, s, v, z, x) and the vowel-like consonants (m, n, l, R, w, j, h). Some confusions are found between phonemes belonging to different clusters: f,s → p,t; v,z → w,j,h, and b → w.

The confusions can be considered as a similarity measure in a multi-dimensional space, i.e. a high number of confusions means a close relation. Multidimensional scaling techniques such as INDSCAL (INDividual SCALing, Caroll and Chang, 1970; Riemersma, 1974) aim to use the similarities between stimuli to locate them in a multidimensional space. This type of analysis requires a symmetrical (dis)similarity matrix; that is, the distance from a to b equals that from b to a. The confusion matrices themselves are not symmetrical, and a procedure is required to transfer the confusion matrix to a symmetrical similarity matrix. Although several procedures can be used, it has been previously established that the procedure developed by Houtgast and first used by Klein et al. (1970) offers reliable results. This procedure is based on the concept that similarity between two stimuli is reflected by the correspondence between the distribution of all responses obtained for the two stimuli. Thus, for example, the dissimilarity between p and t is reflected by the (adequately normalized) difference between the response distributions as given by the two rows in Table 3.4.4.

The INDSCAL program (Riemersma, 1974) used for the analysis allows a multi-dimensional scaling on several similarity matrices resulting in a mean presentation of the stimuli based on these matrices (stimulus space) but also in an estimation of the fit of the individual similarity matrices into the stimulus space. This individual fit is presented in the condition space. We transformed the original 26 confusion matrices (underlying Table 3.4.4, one for each condition) into 26 symmetrical similarity matrices and applied the INDSCAL analysis. In Fig. 3.4.3, panel A, the two-dimensional representation of the stimulus space is given. Dimension 1 relates to voicing and explains 43% of the total variance, dimension 2 discriminates between plosives, fricatives, and the other phonemes. This dimension explains 44% of the total variance. With a three-dimensional analysis only a very small increase of the explained variance was obtained and therefore the two-dimensional analysis is presented.

In panel B of Fig. 3.4.3 the individual fit of the 26 conditions is given. For each condition the relative weight is plotted for dimensions 1 and 2. A perfect fit

with the stimulus space is obtained with an equal weight for both dimensions at a distance of "1" from the origin, hence at the position marked "x". All data points except 1, 9, 10 and 12 have a similar weighting for both dimensions and fit rather well with the total stimulus space of Fig. 3.4.3 A. The conditions 1, 9, 10, and 12 have a low weighting for dimension 2, meaning that the fricatives and plosives for those four conditions are more similar than represented in panel A. If we observe the off-diagonal confusions in Table 3.4.4, many confusions from f,s → p,t and from v,z → w,j,h are found. It was checked that all f,s → p,t confusions and the majority of the v,z → w,j,h confusions were due to the four conditions mentioned. If we examine the frequency range of these conditions according to the description given in appendix A2, all these conditions are low-pass filtered. Hence the high frequency information required for correct recognition of fricatives and plosives as shown for dimension 2 was not available.

In the stimulus space we may observe three clusters of phonemes: plosives (p, t, k, b, d), fricatives (f, s, z, v, x) and vowel-like consonants (m, n, l, j, h, w, R). A similar clustering was obtained with the analysis of the initial consonants of the female speech. Similar results were also obtained from the experiments with the communication channels (not shown).

The confusion matrix and the results of the INDSCAL analysis for the vowels do not show an ordering or clustering as obtained with the consonant analysis.

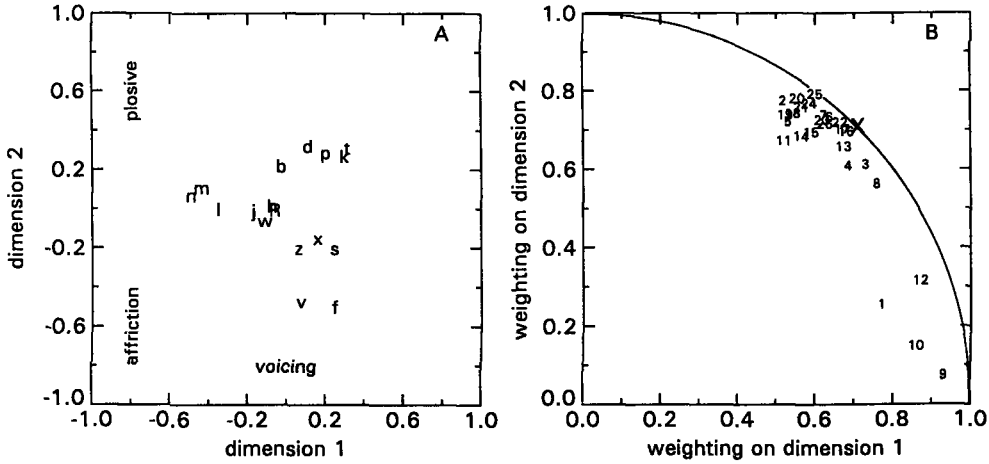

Fig. 3.4.3 Stimulus (A) and condition (B) space for the initial consonants. The optimal relation within the stimulus space for the 26 selected conditions is given in panel B. The explained variance for dimension 1 is 43% and for dimension 2 44%.

Based on the principal-component analysis of the phoneme scores and the multi-dimensional scaling of the confusions among consonants and vowels, four groups of phonemes can be recognized: vowels, fricatives, plosives, and vowel-like consonants. Phonemes within these groups have a quite similar response for the various transmission conditions. We verified this selection by considering correlation coefficients between the mean scores for these four groups. As no different results were found for both experiments (band-pass limiting and communication channels) we combined the results. In Table 3.4.5 these correlation coefficients for the male and female speech are given. The correlation between the initial consonants and final consonants for the three groups (fricatives, plosives, and vowel-like consonants) and for the male and female speech is $r > 0.92$. The consonant groups "fricatives" and "vowel-like" consonants have the lowest correlation, $r = 0.5$. The "plosives" have a correlation with both the fricatives and the vowel-like consonants of $r = 0.8$. The correlation of the vowels with fricatives and plosives is $r = 0.4$ to $r = 0.6$ respectively. The correlation between vowels and vowel-like consonants is $r = 0.7$ - 0.85. This means that for an intelligibility test vowel-like consonants do not mirror the results obtained for the vowels. Table 3.4.5 shows similar behaviour for the female speech.

Table 3.4.5 Correlation coefficients between three different groups of initial and final consonants and vowels, for male and female speech and combined for the two experiments on band-pass limiting and communication channels. For each phoneme group the mean score (m), the standard deviation (s), and the number of conditions (n) are also given.

Male speech

	$C_{i,fr}$	$C_{i,pl}$	$C_{i,vl}$	$C_{f,fr}$	$C_{f,pl}$	$C_{f,vl}$	V
$C_{i,fr}$	-						
$C_{i,pl}$	0.836	-					
$C_{i,vl}$	0.508	0.791	-				
$C_{f,fr}$	0.966	0.791	0.443	-			
$C_{f,pl}$	0.920	0.931	0.710	0.883	-		
$C_{f,vl}$	0.461	0.714	0.926	0.404	0.646	-	
V	0.411	0.567	0.696	0.356	0.514	0.808	-
m	71.6	75.3	76.3	82.2	79.8	75.3	82.3
s	20.3	19.3	17.9	22.7	20.6	13.5	14.8
n	146	146	146	146	146	146	146

Female speech	$C_{i,fr}$	$C_{i,pl}$	$C_{i,vl}$	$C_{f,fr}$	$C_{f,pl}$	$C_{f,vl}$	V
$C_{i,fr}$	-						
$C_{i,pl}$	0.888	-					
$C_{i,vl}$	0.512	0.719	-				
$C_{f,fr}$	0.957	0.862	0.474	-			
$C_{f,pl}$	0.908	0.950	0.728	0.889	-		
$C_{f,vl}$	0.509	0.696	0.946	0.477	0.703	-	
V	0.456	0.611	0.780	0.439	0.587	0.845	-
m	69.7	77.0	73.9	79.3	74.7	68.9	77.0
s	17.6	18.0	18.7	22.8	21.3	16.5	16.7
n	119	119	119	119	119	119	119

Additional to the correlation between the different phoneme groups, the mean score (m) and the standard deviation (s) were calculated over the 146 different conditions for the male speech and the 119 different conditions for the female speech. The phoneme-group scores for final consonants are higher than for initial consonants. This may be due to the different size of the set of phonemes of the two groups (17 initial consonants and 11 final consonants). Initial fricatives represent the lowest scores for both male and female speech. The highest scores with the lowest standard deviation are obtained for the vowels. This may be due to the relatively high speech level of vowels. The relative levels of the speech signals of these four phoneme groups and the average signal spectra are described in appendix A6.

3.5 Relation between Speech Reception Threshold and word or phoneme scores

Speech communication normally is based on connected discourse and therefore sentence intelligibility should be an appropriate measure to evaluate speech communication systems. However, an evaluation of a system with this type of speech only discriminates in a small range of transmission qualities (see Fig. 3.2.1). Therefore a phoneme or word test is normally used. These tests have been discussed previously (see section 3.2). In general the relations among the various tests are obtained for a specific communication system. The relation between sentence intelligibility and word scores or phoneme-group scores based on a variety of transmission conditions is, as far as we know, not available.

A method for estimating sentence intelligibility (Speech Reception Threshold, SRT) at a 50% level in relation to a masking noise, was evaluated by Plomp and Mimpen (1979). In this method a set of sentences pronounced by a speaker is masked by noise. Normally the frequency spectrum of the noise for this test is equal to the long-term speech spectrum of that particular speaker. With an iterative procedure the noise level for a 50% correct recall of the sentences is obtained. This method is highly reproducible and is frequently used in studies on speech perception of the hearing impaired (such as in relation to hearing aid adjustments). The restriction of the method is that the relation with the masking noise is obtained at a 50% sentence intelligibility level only, which corresponds to poor communication quality. This also makes the method less useful for the assessment of transmission conditions at a higher quality level (see Fig. 3.2.1). Therefore, other test methods with a better sensitivity at higher qualities (tests at the phoneme or word level) are used more frequently.

In this section we will study the validity of the CVC-word test in comparison with the sentence intelligibility obtained with the SRT. For this purpose a limited set of twelve representative conditions was selected from the conditions of the two experiments on band-pass limiting and communication channels. This selection is based on several criteria: various types of transmission conditions (band-pass limiting, noise, nonlinear distortion, and distortion in the time domain), and conditions for which the relations between the consonant scores and vowel scores differ substantially.

The conditions in the experiments on band-pass limiting and communication channels are grouped in triplets with a decreasing signal-to-noise ratio. For the experiments on band-pass limiting and noise three fixed signal-to-noise ratios were used (15 dB, 7.5 dB, and 0 dB), where the noise had a spectrum equal to the long-term spectrum of speech for the males or the females. For the experiment on communication channels, a condition without noise was used, together with two conditions with the same transfer parameters but with an additional noise signal. The SRT was measured for the noise signal derived from the transmission condition selected. In this way a direct relation with the phoneme scores (earlier obtained for the same conditions) can be obtained.

As will be described in the next section, the SRT method was extended to obtain the 25% and the 75% intelligibility level as well. This offers the possibility of using a more rigid statistical evaluation (such as a multiple regression) in order to find the optimal relation between sentence intelligibility and the scores of various phoneme groups.

3.5.1 *Experimental design*

3.5.1.1 Description of the SRT measuring method

In the SRT method successive sentences masked by noise are presented to a listener. If the listener recognizes each word of the sentence correctly, the noise level at the presentation of the next sentence is increased by 4 dB. This procedure is repeated until the listener incorrectly recognizes a sentence. The noise level is then decreased by 4 dB. In general, three sentence presentations are used to adjust the noise to a level where the listener recalls at least one sentence incorrectly. After this initial adjustment procedure, the level correction steps are set to 2 dB, and a presentation sequence of ten sentences is started. During these ten sentences the listener will respond (on the average) correctly to half of the sentences presented and incorrectly to the other half, hence tracking close to a 50% correct recognition level. The mean noise level during the sentence presentations represents the Speech Reception Threshold at 50% sentence intelligibility. An example of the noise levels used in such an experiment is given in Table 3.5.1. The mean noise level and the standard deviation of the ten presentations are given. In fact the four examples given in Table 3.5.1 refer to the same transmission condition but are related to four different male speakers.

Plomp and Mimpen (1979) calculated only the mean presentation level representing the 50% correct sentence recognition level. They used the method for one male and one female speaker. This resulted in a very steep psychometric curve (percent correct recognition versus noise level). The extension to more than one speaker results in a less steep psychometric curve, allowing the estimation of other "correct response" levels as well. In this study we introduced the use of the 25% and the 75% sentence intelligibility level. The advantage of having more than just one intelligibility level is the possibility of applying a statistical procedure to find an optimal relation with other measures, such as word score and combinations of phoneme groups. An example of a relation between correct sentence recognition and relative noise level, as given in Table 3.5.1, can be plotted as a response curve according to Fig. 3.5.1. From this stimulus-response curve the required noise levels at various sentence intelligibility levels can be derived.

In order to make a fair comparison, the same four male and four female speakers were used for the SRT measurements as for the CVC measurements. The sentences used were a subset of the sentences according to Plomp and Mimpen (1979).

Table 3.5.1 Example of noise level and sentence recognition (correct +, incorrect -) for ten successive presentations. The mean noise level, m, represents the SRT (dB), the variation around this mean is given by the standard deviation, s.

6	8	10	12	10	8	10	8	6	8	m = 8.6	s = 1.9
+	+	+	-	-	+	-	-	+			
4	6	4	6	8	6	4	2	4	6	m = 5.0	s = 1.7
+	-	+	+	-	-	-	+	+			
6	4	2	0	2	4	6	4	2	4	m = 3.4	s = 1.9
-	-	-	+	+	+	-	-	+			
2	4	2	4	6	8	6	4	2	4	m = 4.2	s = 2.0
+	-	+	+	+	-	-	-	+			

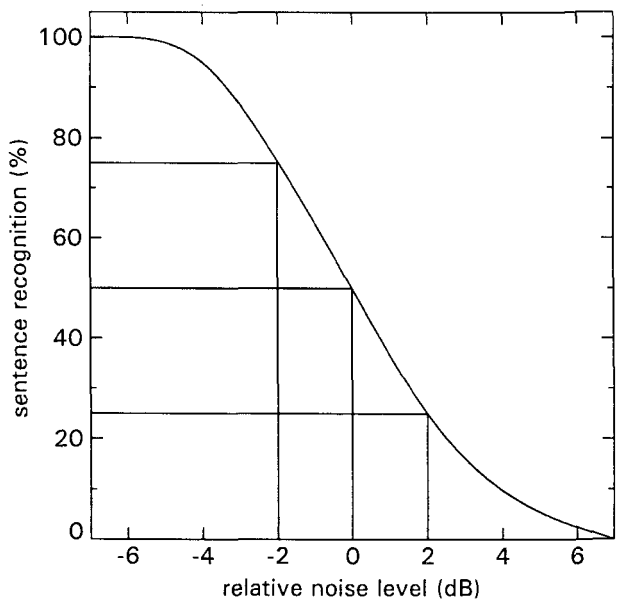

Fig. 3.5.1 Sentence correct score (%) as a function of the relative noise level (dB) during the presentation. The curve is obtained from Table 3.5.1 and is based on the same speech transfer condition, but on four different male speakers. The values for the 75%, 50%, and 25% sentence intelligibility levels are indicated.

3.5.1.2 Phoneme scores at a given SRT noise level

As indicated before, the conditions selected for the SRT evaluation were obtained from the experiments on band-pass limiting and communication channels in which groups of three different noise levels were evaluated. Hence, the phoneme score or CVC-word score corresponding to the SRT can be derived from the results of these experiments by interpolation at the required noise level. This is indicated in Fig. 3.5.2. The CVC score together with the scores for the initial consonants, final consonants, and vowels are given as a function of the signal-to-noise ratio. This example is according to the frequency transfer of condition 7 (signal-to-noise ratios 15, 7.5, and 0 dB) of the experiment on band-pass limiting and noise as described in appendix A2. The mean SRT for this condition and for the same male speakers as used for the CVC test is 5.4 dB. This is indicated in the graph. The speech level of the sentences was obtained according to the method described in appendix A7 and is identical to the method used for the CVC test.

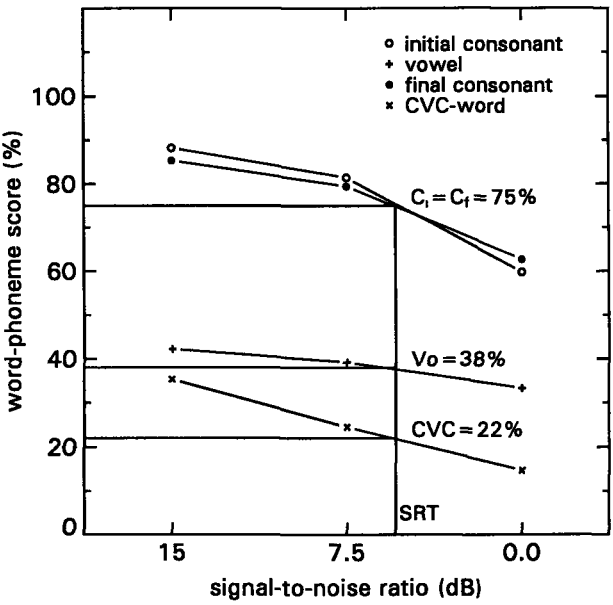

Fig. 3.5.2 Illustration of the interpolation procedure to obtain the C_i, C_f, V, and CVC score at a given SRT value. In this example a SRT of 5.4 dB (for 50% sentence intelligibility) is used. This example corresponds to condition 7, male speech. A similar procedure was used to obtain the score for the phoneme groups (fricatives, plosives, vowel-like consonants and vowels).

At the given SRT level the corresponding CVC and phoneme scores and the scores for the fricatives, plosives, and vowel-like consonants were obtained. In a similar way, the word and phoneme scores corresponding to 25% and 75% SRT were found. In some cases a correct interpolation was not possible as the given SRT value was not within the range of the curves. Hence, these data points could not be obtained.

3.5.1.3 Experimental conditions

As described earlier, twelve conditions for male speech and twelve conditions for female speech were selected. These conditions were based on various transmission channels and various relations between the scores of the different phoneme groups. The twelve male and female conditions are described in Table 3.5.2. Appendices A2 and A3 provide a more complete specification.

As four speakers were used for both groups of conditions, a total of 96 measuring conditions were obtained. Each speaker recorded 5 lists of 3+10 sentences. One list was used for training of the listener. The type of experiment allows a short training procedure in which the listeners are exposed to only one pilot condition. In order to avoid recognition of sentences presented earlier, a list can only be presented once to a listener, and therefore 24 (96:4) different listeners are required. The combinations of speaker, listener, and condition were balanced.

Table 3.5.2 Experimental conditions for male and female speech as used for the SRT experiments.

	Male speech		**Female speech**	
No.	from expm.	condition no.	from expm.	condition no.
1	comm. chan.	5	comm. chan.	5
2	„	12	„	12
3	„	14	„	14
4	„	56	„	56
5	„	23	„	23
6	„	28	„	28
7	band-pass	3	band-pass	3
8	„	4	„	4
9	„	6	„	9
10	„	12	„	10
11	„	14	„	11
12	„	15	„	12

The listeners were seated in an anechoic room and used an intercom system to repeat the perceived sentence to the experimenter. The experiment was computer controlled (sentence presentation, noise level adjustment, and response storage) except for the interpretation of the recognized sentence responses. This was done by the experimenter who responded to the system by hitting a "correct" or "false" key.

3.5.2 *Experimental results*

For all 12 conditions and the male and female speech, the SRT values corresponding to a sentence intelligibility of 50% were obtained by calculating the mean signal-to-noise ratios. The corresponding standard deviations were also calculated. The SRT values for the 25% and 75% sentence intelligibility levels were obtained from the psychometric response curves in the same way as in the example given in Fig. 3.5.1. The SRT values for the three sentence intelligibility levels are given in Table 3.5.3 for the male speech and Table 3.5.4 for the female speech. The corresponding initial-consonant scores, the final-consonant scores, the vowel scores, the word scores, and the consonant-group scores (fricatives, plosives, and vowel-like consonants) are also given. We calculated for each phoneme group the mean score and the corresponding standard deviation for the given set of conditions. Normally the CVC-word scores are used as a measure of word intelligibility. Therefore, these scores are also given in the tables, together with the C_iV and the C_iVC_f-word scores predicted by the product of the individual C_i, V, and C_f probabilities. If, due to interpolation restrictions, scores could not be obtained, this is shown in the table. As will be discussed later, the scores for condition 4 (based on echoes) were excluded.

The prediction of sentence intelligibility by (a combination of) phoneme scores

The mean scores for the various phoneme and word groups represent the relation with sentence intelligibility at three levels. The standard deviations in relation to the differences between these mean scores for the three sentence intelligibility levels express the reliability of this relation (accuracy of the prediction). Small differences of 3-5% (in % correct scores for 25%, 50%, and 75% sentence intelligibility) and high standard deviations of 18-23% are obtained for the fricatives. This means that the distributions for the three sentence intelligibility levels are not significantly different. For the CVC-word scores, differences between the mean scores at the three sentence intelligibility

levels of 6% and standard deviations of 4-6% are obtained. This illustrates a better prediction of sentence intelligibility.

To quantify these relations, we calculated the correlation coefficients between the phoneme-group scores and the sentence intelligibility. These correlation coefficients are given in the bottom row of Tables 3.5.3 and 3.5.4. For the fricatives a correlation coefficient r = 0.101 - 0.115 is obtained, and for the C_iV word a correlation coefficient of r = 0.77 - 0.78.

We calculated the optimal prediction of sentence intelligibility with a linear combination of consonant and vowel scores by using a multiple regression[2] technique. For this purpose, we used the STATPAC software (Smoorenburg, 1989).

As multiple regression is based on a weighted combination of the input variables and hence results in an additive model, we also applied a log transform[3] on the data, which results in a multiplicative model with the same multiple regression algorithm. The highest correlation coefficient (R = 0.861) was obtained for a combination of the fricative, plosive, vowel-like consonant, and vowel scores. These and other correlation coefficients together with the corresponding weighting coefficients are given in Table 3.5.5. In general a better prediction is obtained by using the scores of the four phoneme groups than by using the initial consonant, final consonant, and vowel scores. The contribution of the plosives is very small. It was tested that omission of the plosive scores resulted in a slightly lower correlation coefficient.

---

[2] With multiple regression the optimal weighting coefficients are obtained to fit a set of input variables with a given target variable. The linear prediction for C_i, V, and C_f with coefficients a,b,c,d is according to: $a.C_i + b.V + c.C_f + d$.

[3] Application of a log-transform on the input data results in a multiplicative model with the coefficients (a,b,c) applied as $C_i^a \cdot V^b \cdot C_f^c$. Correlation coefficients obtained with a multiple regression technique are normally indicated as "R".

Table 3.5.3 SRT and corresponding phoneme-group scores for twelve reference conditions, male speech, and three SRT levels.

75% sentence intelligibility:											
No.	SRT	s	C_i	C_f	V	CVC	C_{fr}	C_{pl}	C_{vl}	C_iV	C_iVC_f
1	4.6	–	50	67	67	29	40	50	61	33.5	22.4
2	0.5	–	44	54	76	21	18.5	42	62.5	33.4	17.1
3	6.1	–	50	60.5	80	28.5	40	48	63	41	24.2
4	6.4	–	77	85	96	65	72	76	80.5	73.9	62.8
5	2.9	–	66.5	73	64	34	83	72	50	42.6	31.1
6	3.5	–	63	68	64	31	73	75	58	40.3	27.4
7	7.4	–	80.5	79	39	24.5	83	89	71	31.4	24.8
8	2.5	–	50.0	63.5	89	35	47	60	55	44.5	28.3
9	4.7	–	66.5	75	75	42	78	68	58	49.9	37.4
10	5.5	–	56.0	65.0	86.0	36.0	53	56	62	48.2	31.3
11	4.2	–	58	69	64	29	68	57	48	37	25.6
12	2.2	–	67	70	56	30.5	78	67	51	37.5	26.3
		m	59.2	67.6	69.1	31.0	60.1	62.2	58.1	39.9	26.9
		s	10.6	6.9	14.3	5.7	21.6	13.7	6.8	6.1	5.3
		condition 4 not included									
50% sentence intelligibility:											
No.	SRT	s	C_i	C_f	V	CVC	C_{fr}	C_{pl}	C_{vl}	C_iV	C_iVC_f
1	2.6	2.5	43.5	63.5	54.5	21.5	34.0	43.5	53.0	30.0	15.0
2	−1.0	2.4	40.0	50.0	72.5	15.5	13.0	36.5	58.0	29.0	14.5
3	4.1	2.7	42.0	53.0	73.5	22.5	31.5	41.5	53.5	29.9	15.9
4	3.9	2.8	69.5	79	92.5	55	66	70.5	75	64.3	50.8
5	0.9	2.1	60	70	60	27	83	65	39.5	36.0	25.2
6	1.0	3.0	55	67	63	26	67.0	62.5	45.5	34.7	23.2
7	5.4	1.2	75	75	38	22	82	82	62	28.5	21.4
8	1.1	2.7	45.5	59	87	30	42	56.5	50.5	39.6	23.4
9	2.7	2.8	58	70.5	70.5	34	72.5	62	49	33.6	23.7
10	2.7	4.3	49	57	81.5	27	44	50	54	39.9	22.8
11	3.2	2.6	53	65	62	24	65	52	42	32.9	21.4
12	1.2	1.9	65	65	53	28	78	65	51	34.5	22.4
		m	53.3	63.2	65.0	25.2	55.6	56.0	50.7	33.5	20.8
		s	10.8	7.7	13.9	4.9	23.7	13.1	6.6	4.0	3.8
		condition 4 not included									
25% sentence intelligibility:											
No.	SRT	s	C_i	C_f	V	CVC	C_{fr}	C_{pl}	C_{vl}	C_iV	C_iVC_f
1	0.6	–	37.5	61.5	45	15	27.5	38	45.5	16.9	10.4
2	−3.5	–	out of range for accurate interpolation								
3	1.5	–	33.0	44.0	66.0	15.0	22.0	33.0	41.5	21.8	9.6
4	1.0	–	61.5	73	88	44	59.5	62	68	54.1	39.5
5	−1.1	–	53	65.5	54	19	83	58	30.5	28.6	18.7
6	−2.5	–	45	66	58	18	60	51.5	34	26.1	17.2
7	3.4	–	69	70	36	20	81	74	54	24.8	17.4
8	−1.9	–	32	45.0	83	13	31.5	47	40	26.6	12
9	0.0	–	48.0	65.0	65.0	23.5	65.0	55.0	37.5	31.2	20.3
10	−0.3	–	39	47	76	15	34.5	44	45	29.6	13.9
11	1.2	–	46	58	55	15.5	59	46	34	25.3	14.7
12	−1.5	–	62	60	48	22	76	58	44	29.8	17.9
		m	46.5	58.2	58.6	17.6	54.0	50.5	40.6	26.6	15.2
		s	12.1	9.5	14.3	3.5	23.2	11.7	6.9	4.3	3.7
		conditions 2, 4 not included									
	corr. coeff.	r	0.437	0.445	0.299	0.759	0.115	0.362	0.738	0.770	0.754

Table 3.5.4 SRT and phoneme-group scores for twelve reference conditions, female speech, and three SRT levels.

No.	SRT	s	C_i	C_f	V	CVC	C_{fr}	C_{pl}	C_{vl}	C_iV	C_iVC_f
75% sentence intelligibility:											
1	7.4	–	57	60	67.5	27.5	52	51	65	38.5	23.1
2	6.8	–	51.5	49	63.5	20.5	34.5	48	67.5	32.7	16
3	9.6	–	57.5	59.5	72	32.5	42.0	56	70	41.4	24.6
4	4.4	–	67	67	85	41.5	62	65	71	57	38.2
5	7.1	–	76	76	55	37	88	88	60.5	41.8	31.8
6	11.5	–	out of range for accurate interpolation								
7	10.6	–	81	75	47	32	80.5	89	76	38.1	28.6
8	3.0	–	53.0	59.0	82.0	29.0	44.0	60.5	56.0	43.5	25.6
9	1.0	–	53	59.5	73	30	57	53	50	38.7	23
10	8.3	–	73	72	57	30	74	85	53	41.6	30
11	4.1	–	64	65	60	33	67	70.5	52	38.4	25
12	3.0	–	62.0	64.0	84.5	38.0	49.0	63.0	70.0	52.4	33.5
		m	62.8	63.9	66.2	31.0	58.8	66.4	62.0	40.7	26.1
		s	10.5	8.4	12.0	5.0	17.8	15.8	9.0	5.1	5.1
		conditions 4, 6 not included									
50% sentence intelligibility:											
1	5.4	2.6	51	54	60	22	46	46	58.5	30.6	16.5
2	4.8	2.4	47.5	44	60.5	17	31	42.5	64	28.7	12.6
3	7.6	1.4	45.5	48	64.5	18	29	44.5	64.5	29.3	14.1
4	2.4	1.9	59	59	81	34	57	60	63.5	47.8	28.2
5	5.1	3.4	70	70	47	28	87	84	52	32.9	23.0
6	8.5	2.4	76	74	57	35	76	81	70	43.3	32.1
7	8.6	2.5	77	71	45	26.5	78.5	85	72	34.7	24.6
8	1.7	1.0	48	55.5	80	23.5	40	56.5	50	38.4	21.3
9	−1.0	2.2	44	48.5	67	15	50	43	37	29.5	14.3
10	5.3	3.7	65	65	52	25	72	75	46	33.8	22.0
11	3.1	1.4	60	62	58	28	67	67	49	34.8	21.6
12	−0.9	1.6	47	51	79	18	33	47	55.5	37.1	18.9
		m	57.4	58.5	60.9	23.3	55.4	61.0	56.2	33.9	20.1
		s	12.7	10.4	11.4	6.0	21.2	17.7	10.8	4.5	5.7
		condition 4 not included									
25% sentence intelligibility:											
1	3.4	–	44	47.5	52.5	16	38	42.5	53	23.1	11.0
2	2.8	–	44	40.5	51.5	13.5	28	38	61.5	22.7	9.2
3	5.6	–	41.5	41.5	57	14	26	40	55.5	23.7	9.8
4	0.4	–	53.5	53.5	77.5	27	50.5	55.5	55.5	41.5	22.2
5	3.1	–	67	68	43	24	84	78	47	28.8	19.6
6	4.0	–	62	61	53	23	68	70	53	32.9	20.0
7	6.6	–	72	66	42	21	76	78	65	30.2	20
8	−0.8	–	38	48	75	14	31	48.5	36.5	28.5	13.7
9	−3.0	–	out of range for accurate interpolation								
10	1.3	–	51.5	56	46	17	67	56	39	23.7	13.3
11	1.1	–	55	58	53	22	66.5	60	44	29.2	16.9
12	−2.5	–	out of range for accurate interpolation								
		m	52.8	54.6	52.6	18.3	53.8	56.8	50.5	27.0	14.8
		s	12.1	10.2	9.8	4.2	22.8	15.8	9.7	3.7	4.4
		conditions 4, 9, 12 not included									
	corr. coeff.	r	0.335	0.391	0.435	0.714	0.101	0.238	0.439	0.789	0.678

Table 3.5.5 Optimal correlation coefficients (R) and corresponding weighting factors for two sets of phoneme combinations to predict sentence intelligibility.

		coefficients		
Male speech	R	C_i	V	C_f
additive model	0.765	0.832	1.021	0.849
multiplicative model (log)	0.761	1.094	1.304	0.859
Female speech				
additive model	0.787	2.173	1.496	-1.124
multiplicative model (log)	0.830	2.992	2.139	-1.592

		coefficients			
Male speech	R	C_{fr}	C_{pl}	C_{vow}	V
additive model	0.861	0.754	-0.682	1.932	0.629
multiplicative model (log)	0.859	0.392	-0.066	1.766	0.785
Female speech					
additive model	0.793	0.815	-0.278	0.960	1.405
multiplicative model (log)	0.821	0.862	-0.255	1.149	1.864

For all relations between the sentence intelligibility scores and the phoneme-group scores the contribution of condition no. 4 was excluded. The main distortion used for this condition was a distortion in the time domain. This was achieved by addition of an echo with a delay of 200 ms to the speech signal, and resulted in a lower sentence intelligibility compared with the CVC-word score and the phoneme scores than for the other eleven transmission conditions. A careful observation of the speech for this condition revealed that the individual words were understood quite well, as long as the listener could focus his attention on a particular word. This is the situation for CVC words in a well known carrier phrase, and results in a relatively high phoneme and word score. However, during the presentation of a full sentence the listener is confused by the echoes and detects only some of the words correctly. Reproducing the complete sentence is then much more difficult. For this reason we excluded the results for this condition.

3.6 Discussion and conclusions

Evaluation at word level

A significant difference was found between the word scores ($p <= 0.07$), the vowel scores ($p < 0.01$), and the final consonant scores ($p <= 0.08$) for male and female speakers (for the male speakers a higher score was obtained, see Table

3.4.1). No significant difference between male and female speakers for the initial-consonant scores was found ($p=0.33$).

The variances among speakers and the variances among listeners are both significant (Tables 3.4.1 and 3.4.2) and of the same order of magnitude. It can be concluded that an equal number of speakers and listeners is required for a balanced experimental design.

Word scores were obtained directly, as a correct response to the complete CVC word, but word scores could also be predicted from the product of the individual phoneme-group scores of the initial consonants, vowels, and final consonants. This prediction is valid if the phoneme scores are statistically independent. A high correlation between these two types of word-scores was found for four different experimental conditions including male and female speech ($r > 0.99$). The regression coefficients show that the relation between the two scoring methods is almost linear. A small systematic difference was found. Calculation of the word score based on individual phoneme-group scores allows a different balancing of the phonemes.

Evaluation at phoneme level

Miller and Nicely (1955) studied the differences between phonemes for various transmission conditions, which were based on band-pass limiting and noise. They grouped the phonemes according to production features such as: voicing, nasality, affrication, and place of articulation (labial, dental, etc).

In this study two aspects of the listener responses are considered, the distribution of the correct responses for all 43 different phonemes, and the confusions among phonemes within a group (initial consonants, vowels, and final consonants).

A high correlation between the scores of initial and final consonants was found ($r = 0.96 - 0.97$), but a lower correlation between consonants and vowels ($r \approx 0.57 - 0.73$), see Table 3.4.3. In appendix A4, the individual phoneme scores for male and female speech are given. These results indicate that, in general, vowels are better understood than consonants. A systematic analysis among phonemes, indicating a similar response for the various transmission conditions, was based on correlation of all the phoneme scores as obtained with a principal-component analysis. The results of such a principal-component analysis are given in Figs 3.4.1 and 3.4.2, and indicate a close relation among some groups i.e. plosives and fricatives, vowel-like consonants, and vowels. Especially the results for the vowels are not consistent for the analysis conditions as presented in Fig. 3.4.2. This may be due to certain specific transmission conditions.

We also considered the confusions among phonemes. For this purpose we applied a multidimensional scaling technique. Due to the composition of the test material (CVC words) only confusions within the groups of initial consonants, vowels, and final consonants could be obtained. With the INDSCAL multidimensional scaling technique a two-dimensional representation was obtained for the initial consonants (see Fig. 3.4.3). This representation was obtained for the male speech and for the conditions with a signal-to-noise ratio of 15 dB of the experiments with band-pass limiting and noise. Similar results were obtained for female speech and the experiment on communication channels. In Fig. 3.4.3 a cluster of the plosives (p, t, k, b, d), a cluster of the vowel-like consonants (m, n, l, j, w, R, h), and a cluster of the fricatives (v, f, z, s, x) are obtained. Dimension 1 of this figure represents voicing, and dimension 2 is related to the obstruction in the vocal tract (plosive/affrication). This corresponds with the results reported by Steeneken (1987) based on DRT experiments where only the dimensions "labial" and "sustention" (both related to the same type of obstruction in the vocal tract) discriminated between the various transfer conditions.

Based on the analyses described in section 3.4, a grouping of plosives, fricatives, vowel-like consonants, and vowels is proposed. As a verification we calculated the correlation coefficients between scores of each pair of these groups (Table 3.4.5). The highest correlation coefficient for the four phoneme groups was found between fricatives and plosives ($r = 0.84$), however, only a few confusions between these groups were obtained (Table 3.4.4). The vowels had the lowest correlation with the other three groups.

The results indicate that vowels cannot be neglected in intelligibility testing. It also indicates that intelligibility tests cannot be based on vowels only. Hence tests with digits and the spell alphabet (alpha, bravo, ...), which are mainly based on vowel recognition, may produce invalid results.

Evaluation at sentence level

For many reasons, as described above, the assessment of speech communication channels is preferably performed with word tests. Voice communication however, is mainly based on connected discourse. Therefore we studied the relation between the intelligibility of words and sentences. Twelve representative conditions were used, being a selection (based on maximal discrimination between consonant and vowel scores) of the conditions of the experiments as described above. Sentence intelligibility was measured with the speech reception threshold method (SRT). This SRT was extended in order to obtain not just a sentence intelligibility level of 50% but also a level of 25% and 75%. The correlation between the sentence intelligibility scores and the

corresponding CVC-word scores was found to be r = 0.76 for the male speech and r = 0.71 for the female speech. The optimal relation between the sentence intelligibility scores and the scores of the phoneme groups was obtained with a multiple regression. The optimal correlation coefficients for male and female speech are R = 0.86 and R = 0.82 respectively.

With the limitation that sentence intelligibility ranges from 0-100% in a small range of CVC-word intelligibility (Fig. 3.2.1), it can be concluded that the CVC-word score (for the equally-balanced nonsense CVC words) is closely related with sentence intelligibility (of simple Dutch sentences as defined by Plomp and Mimpen, 1979).

Conclusions

- We have found four representative groups of phonemes, each with a specific transmission quality, for different transfer conditions. With the scores of these four groups a transmission condition can be characterized in a diagnostic manner. This concept can be used to extend the objective intelligibility measuring procedure from a prediction of word scores to this more diagnostic phoneme-group principle (this will be studied in chapter 4).

- A close relation was found between sentence intelligibility and the CVC-word score. An optimal relation was found with a multiplicative combination of the four phoneme-group scores.

4 FREQUENCY-WEIGHTING FUNCTIONS FOR PHONEME GROUPS AND CVC WORDS

Summary

In chapter 2 robust frequency-weighting factors and redundancy-correction factors were found with respect to male and female speech and for various signal-to-noise ratios. Different frequency-weighting factors were found for consonants and vowels. Based on the phoneme groups described in chapter 3, a phoneme-group-specific frequency weighting and redundancy correction is obtained in this chapter. These frequency-weighting functions depend on the speech items considered and may explain the differences found with other studies (French and Steinberg, 1947; Steeneken and Houtgast, 1980; Pavlovic, 1984; Studebaker et al., 1987). Also the test signal spectrum used for the measurements or for the calculations is of interest. This parameter has a small effect on the shape of the optimal frequency-weighting functions.

The variation among the scores of different speaker-listener combinations is also studied. A speaker-listener-dependent frequency-weighting function and redundancy correction is obtained.

4.1 Introduction

It was found in chapter 2 that the information content of a speech signal within different frequency bands (the frequency-weighting factors) depends on the speech items considered, but does not depend on various other parameters such as male-female speech and signal-to-noise ratio. Different frequency-weighting factors were found for different speech items. This corresponds with the results found in other studies (French and Steinberg, 1947; Steeneken and Houtgast, 1980; Studebaker et al., 1987).

A robust frequency-weighting function is of interest for an optimal design of speech communication channels, and for an explanation of the differences between the results of the various studies mentioned above.

We found different frequency-weighting functions for consonants and vowels. The separation between consonants and vowels, as described in chapter 2, was rather arbitrary and was related only to the word type (CVC words) used for the subjective evaluation. From the study described in chapter 3 we obtained a subdivision of the phonemes into four groups (fricatives, plosives, vowel-like consonants, and vowels). The phonemes within each group show a quite similar variation of the recognition rate for various types of degradation. Therefore we will study the relation between the intelligibility of each individual phoneme group and the phoneme-group-specific STI_s (specific).

Because of the experimental set-up used in chapter 2, where the prediction of the CVC-word score was performed by using pre-defined signal-to-noise ratios for a masking signal with a spectrum equal to the long-term speech spectrum, only a global prediction at the word level could be obtained. A specific prediction, based on optimal frequency-weighting factors for a particular phoneme group, could not be obtained, mainly because the frequency spectrum of the applied masking signal was not adapted to the frequency spectrum of a specific group of phonemes. A phoneme-group-specific correction of the signal-to-noise ratio is required. This can be obtained by a phoneme-group-spectrum dependent correction of the signal-to-noise ratio or, more pragmatically, by the measurement of the signal-to-noise ratio with an adapted test signal, of which the frequency spectrum is adjusted according to the phoneme group considered.

We will also study the variation of the STI_r in relation to the observed intelligibility scores of individual speaker-listener combinations.

4.2 Experimental design

As described in appendix A1, the STI is calculated from the signal-to-noise ratios obtained for seven octave bands. The test signal for which the signal-to-noise ratios are determined has a frequency spectrum equal to the long-term speech spectrum, hence the signal-to-noise ratios are representative of connected discourse. As found in chapter 3, four different groups of speech items can be considered: fricatives, plosives, vowel-like consonants, and vowels. All these groups of phonemes have their own specific frequency spectrum (see appendix A6). This results, in combination with a noise signal, into phoneme-group-specific signal-to-noise ratios. For linear systems and for a given test signal spectrum, the derived signal-to-noise ratio can be adjusted for any speech spectrum. However, this is not possible for nonlinear transfer conditions or for conditions with a limited dynamic range. In these cases, the test signal level and test signal spectrum are important variables and have to be applied in actual measurements. Therefore in the study described below, four different test signal spectra are used, which are representative of the mean spectra of the four phoneme groups.

Test signal spectrum

The speech spectra for the four phoneme groups were obtained from speech tokens which were annotated at the phoneme level. Hence the spectrum of each individual phoneme of these speech tokens could be used for the calculation of the mean spectrum of the corresponding phoneme group. The

mean spectra are based on the equivalent level (energetic summation). The data are derived from an earlier study (Steeneken and Van Velden, 1989). From this study speech tokens of ten male and ten female speakers (different from the speakers used for the intelligibility tests) were available. Each speech token was based on a sentence with a duration of five seconds. Since each speaker spoke the sentence twice a total of 100 seconds of male and female speech was available. The annotation was performed by hand with an interactive computer system. Only the beginning of each phoneme was labelled, the end label was placed at the beginning of the next phoneme. While calculating the mean spectra a threshold was applied to avoid the effect of silent intervals between phonemes. The mean spectra for the four phoneme groups and for male and female speech are given in appendix A6.

Experimental set-up

The experimental conditions were the same as those of the experiment described in chapter 2.

For the objective measurements the STIDAS-IID measuring device (Steeneken and Agterhuis, 1982) was used. The design of this device allows for simultaneous measurements in seven octave bands with centre frequencies from 125 Hz to 8 kHz. We used a measuring program with such a simultaneous measuring scheme and with one different modulation frequency for each frequency band (see appendix A1).

4.3 Experimental results

4.3.1 *Phoneme-group-specific predictions*

An optimization of the STI procedure was performed for each of the four groups of phonemes individually. We also included the CVC-word scores. As found before, different weighting factors for the octave-band-specific contributions and redundancy corrections can be expected (see section 2.3.4).

As before, the criterion for an optimal relation was to minimize the standard deviation around the best-fitting third-order polynomial between the predicting STI values and the subjective phoneme-group or word scores. This optimization was performed for two models of the STI calculation scheme: a redundancy correction according to $\beta\sqrt{(TI1.TI2)}$, and no redundancy correction ($\beta = 0$). In Table 4.3.1 these standard deviations (s in %) are given for the male speech conditions and for the female speech conditions. Exclusion of the redundancy correction results in a less accurate prediction of the fricative,

plosive, vowel, and CVC-word score by the corresponding, individually optimized, STI value. For vowel-like consonants (especially for female speech) no major improvement was found by the application of a redundancy correction.

The smallest standard deviation is found for the phoneme groups of vowels, and vowel-like consonants. This may also be related to the variation among the scores obtained for these phoneme groups individually (see chapter 3, Table 3.4.6).

Four different sets of frequency-weighting factors and redundancy correction factors were found for the optimal relation between the observed phoneme-group scores and the phoneme-group-specific STI. These frequency-weighting and redundancy factors for male and female speech separately, are given for each of the four phoneme groups in Figs 4.3.1 - 4.3.4, respectively. The frequency-weighting and redundancy factors obtained for the CVC words are given in Fig. 4.3.5.

Table 4.3.1 Prediction accuracy (%) expressed by the standard deviation around the best fitting third-order polynomial for fricatives, plosives, vowel-like consonants, vowels, and CVC words. The calculations include a redundancy correction according to: $\beta \surd(TI1.TI2)$, as well as without redundancy correction ($\beta = 0$). The calculations are based on the male speech and female speech conditions of the experiment on band-pass limiting and noise masking.

	Male speech		Female speech	
Redundancy	$\beta \surd(TI1.TI2)$	$\beta = 0$	$\beta \surd(TI1.TI2)$	$\beta = 0$
fricatives	3.91	4.44	4.31	4.54
plosives	5.57	6.78	5.84	6.51
vowel-like cons.	4.03	4.45	4.22	4.19
vowels	3.63	5.28	2.88	3.75
CVC words	4.63	6.89	4.49	6.41

In general a high contribution and a low redundancy correction are found for the octave bands with centre frequencies 1 kHz to 4 kHz. For fricatives and plosives an increasing contribution at higher frequencies is found, vowel-like consonants show a fairly flat weighting function and a low redundancy correction. This is also reflected in the small difference between the standard deviations given in Table 4.3.1 for the calculation with and without a redundancy correction. For vowels a high contribution and a low redundancy correction are found for the octave bands with centre frequencies of 500 Hz to 2 kHz.

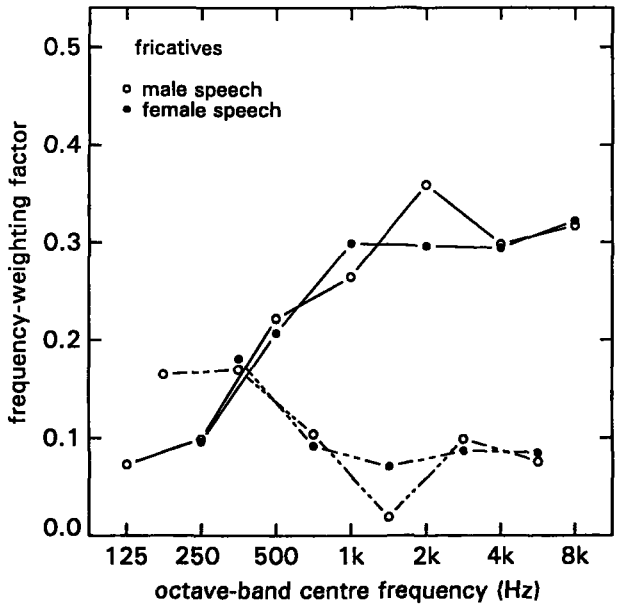

Fig. 4.3.1 Frequency-weighting factors for the octave-band contribution α_k (solid line) and redundancy correction β_k (dashed line) for the FRICATIVES and for the male and female speech.

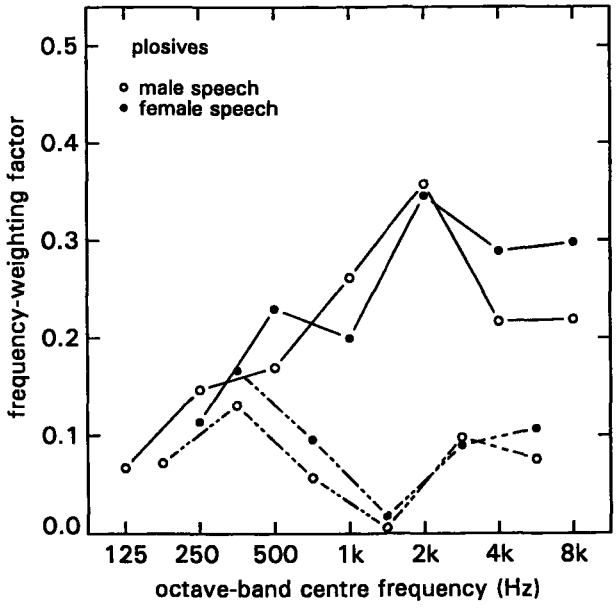

Fig. 4.3.2 Frequency-weighting factors for the octave-band contribution α_k (solid line) and redundancy correction β_k (dashed line) for the PLOSIVES and for the male and female speech.

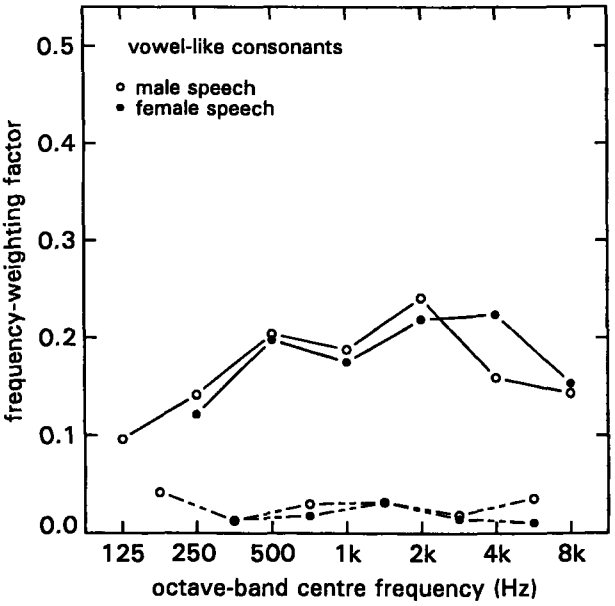

Fig. 4.3.3 Frequency-weighting factors for the octave-band contribution α_k (solid line) and redundancy correction β_k (dashed line) for the VOWEL-LIKE consonants and for the male and female speech.

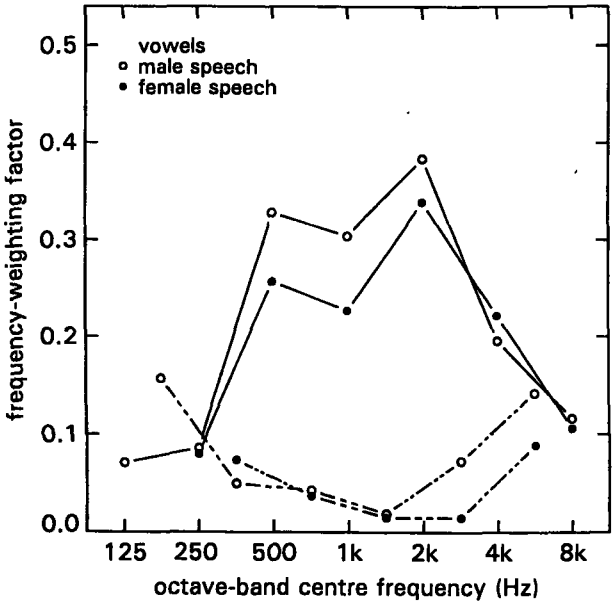

Fig. 4.3.4 Frequency-weighting factors for the octave-band contribution α_k (solid line) and redundancy correction β_k (dashed line) for the VOWELS and for the male and female speech.

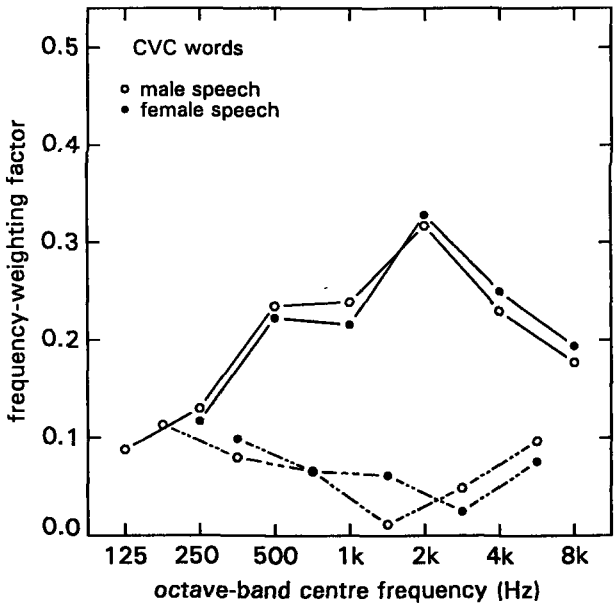

Fig. 4.3.5 Frequency-weighting factors for the octave-band contribution α_k (solid line) and redundancy correction β_k (dashed line) for the CVC words and for the male and female speech.

All these results are similar for male and female speech with the exception that for female speech the octave band with centre frequency 125 Hz is not included.

The combination of a relatively low octave-band contribution and a high redundancy correction may indicate that the information content in adjacent frequency bands is highly correlated and that these frequency bands may be combined. For example the octave bands with centre frequencies at 125 Hz and 250 Hz, for the male speech conditions and all phoneme groups, show such a small contribution and such a high redundancy correction that these two low-frequency-bands might be combined. On the other hand the contribution of the octave band with centre frequency 2 kHz is for all phoneme groups fairly high with a low redundancy correction between the lower adjacent octave band with centre frequency 1 kHz. Smaller frequency bands, giving a better resolution along this part of the frequency scale with respect to the contribution to the STI, might be considered.

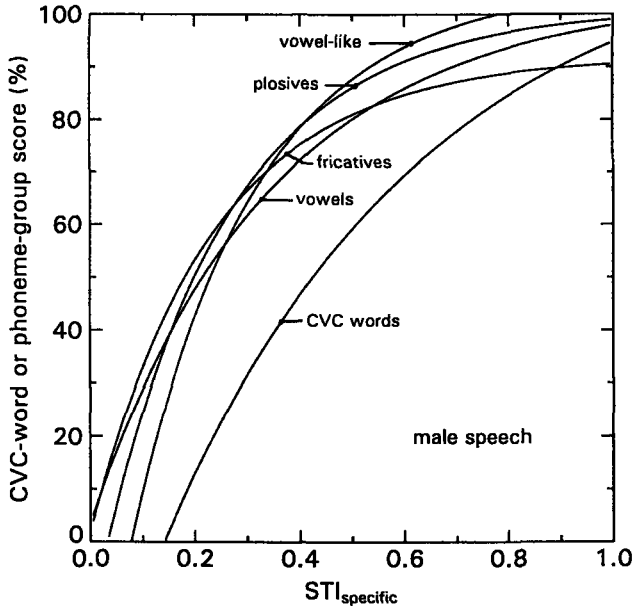

Fig. 4.3.6 Relation between predicted phoneme-group scores and the corresponding phoneme-group-specific STI_s for MALE speech. The relation for the CVC-word score is also given.

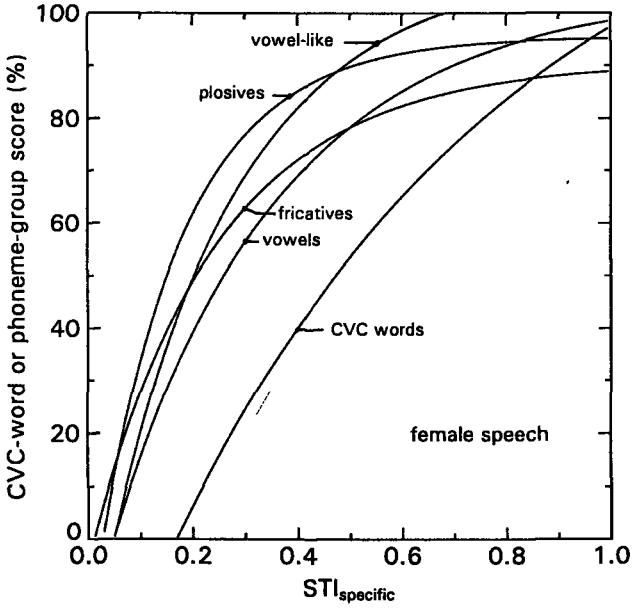

Fig. 4.3.7 Relation between predicted phoneme-group scores and the corresponding phoneme-group-specific STI_s for FEMALE speech. The relation for the CVC-word score is also given.

Different relations were found between the observed phoneme-group scores and the phoneme-group-specific STI_s, see Figs 4.3.6 and 4.3.7. Rather than giving the best-fitting third-order polynomial, the best-fitting exponential function[1] is given. The relation for the CVC-word score and the STI_r is also given and is virtually identical to the one found in chapter 2 (see Figs 2.3.4 and 2.3.5).

4.3.2 *The effect of the test signal spectrum on the frequency-weighting functions*

In section 4.3.1 we calculated the optimal prediction of the phoneme-group scores by the phoneme-group-specific STI_s, based on a test signal with the spectrum and the level of the speech signals of the corresponding phoneme group. A further point of interest is the relation between the spectrum of the applied test signal and the optimal frequency-weighting functions. To investigate this, we also optimized the phoneme-group-specific STI_s on the basis of measurements with nonmatched test signals, namely those specific of the *other* phoneme groups. This procedure results in an over- or under-estimation of the signal-to-noise ratios within frequency bands where the spectrum level does not coincide with the spectrum level of the test signal of the phoneme group considered. The accuracy of the prediction of the phoneme-group-specific STI_s, for each of these combinations was calculated. In Table 4.3.2 the vertical spread around the best fitting third-order polynomial is given for all combinations and for male and female speech separately.

As expected, the best prediction for each group is found for a combination with a test signal spectrum matched to the same phoneme group. These results are given on the diagonal of Table 4.3.2 and are underlined. Groups with a similar long-term speech spectrum (fricatives/plosives, and vowels/CVC word[2]) show a similar, or in a few cases a slightly better, prediction accuracy.

---

[1] Although a slightly better prediction between the STI_r and the observed intelligibility is obtained by using a third-order polynomial, an exponential curve was used instead. This was done because this function always presents an increment of the predicted score for increasing STI_r values. With the third-order polynomial an unpredictable slope is obtained outside the range of the data points. This sometimes results in a negative slope at higher STI_r values (for $STI_r > 0.95$).

[2] The long-term frequency spectrum of the CVC words is dominated by the vowel spectra.

Table 4.3.2 Prediction accuracy expressed by the standard deviation around the best fitting third-order polynomial (%) for a test signal spectrum according to the mean spectrum of fricatives, plosives, vowel-like consonants, vowels, and the long-term spectrum of phonetically balanced CVC words. The calculations include an optimization with each test signal spectrum and all groups of phonemes. The calculations are performed both for the male and for the female speech conditions of the experiment on band-pass limiting and noise masking.

Male speech test signal spectrum from	fricatives	plosives	Predicted scores for vowel-like consonants	vowels	CVC words
fricatives	3.91	6.05	6.51	6.22	5.82
plosives	3.98	5.57	4.56	5.43	5.34
vowel-like cons.	4.61	5.37	4.03	5.00	5.24
vowels	4.78	5.29	4.70	3.63	4.26
CVC words	4.37	5.37	4.32	3.64	4.63
Female speech test signal spectrum from	fricatives	plosives	Predicted scores for vowel-like consonants	vowels	CVC words
fricatives	4.31	6.86	8.38	7.22	6.84
plosives	4.12	5.84	4.73	4.44	4.30
vowel-like cons.	4.50	5.50	4.22	3.32	4.44
vowels	4.85	5.13	4.89	3.08	4.43
CVC words	4.15	4.38	5.24	2.88	4.49

It was no surprise to find that the optimal frequency-weighting function and redundancy correction also depend on the test signal spectrum used for the measurements. An overestimation of the test signal level within a specific octave band results in an overestimation of the signal-to-noise ratio and of the corresponding factor TI_k (see Eq. 2.3.2), which is compensated for by a lower frequency-weighting factor α_k.

A shift of the total test signal level (as is obtained for the speech spectrum of fricatives) results in a systematically lower STI value and is not reflected in the frequency-weighting function. In Figs 4.3.8 and 4.3.9 the frequency-weighting factors for fricatives and vowels and for three test signal spectra (vowel spectrum, fricative spectrum, and long-term speech spectrum) are given. The significant changes in the frequency-weighting functions and redundancy correction coincide with the spectral difference as given in appendix A6. What is reassuring however, is the fact that the frequency-weighting factors show a consistent structure despite substantial variation in the test spectra.

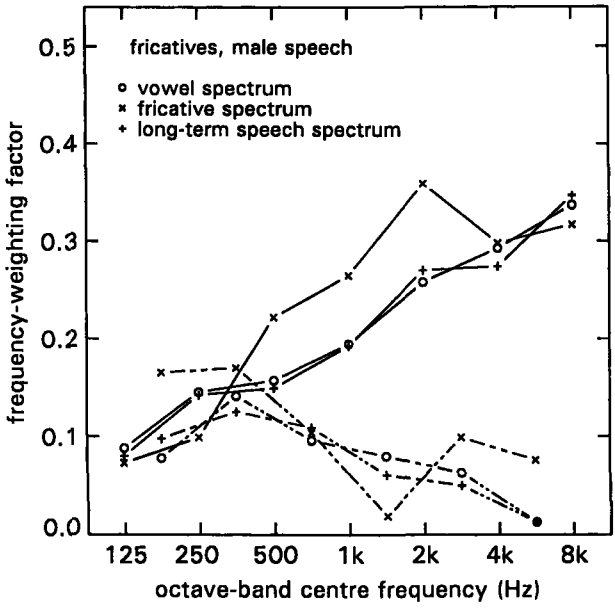

Fig. 4.3.8 Frequency-weighting factors for the octave-band contribution α_k (solid line) and redundancy correction β_k (dashed line) for the FRICATIVES obtained with three different test-signal spectra.

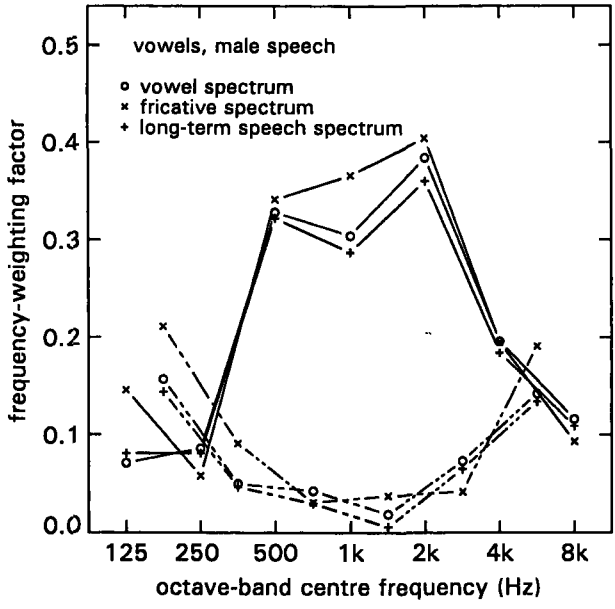

Fig. 4.3.9 Frequency-weighting factors for the octave-band contribution α_k (solid line) and redundancy correction β_k (dashed line) for the VOWELS obtained with three different test-signal spectra.

4.3.3 *Prediction of the CVC-word score from phoneme-group-specific STI's*

It was found that the prediction accuracy by the STI_r for the observed CVC-word score, with the use of a redundancy correction according to $\beta\sqrt{(TI_1.TI_2)}$, was 4.63% for the male speech and 4.49% for the female speech (see Table 4.3.1). The CVC-word score, however, can also be predicted by combining phoneme-group scores obtained from the STI values for the fricatives, plosives, vowel-like consonants, and vowels. This is performed in two steps: (1) calculation of the initial consonant and final consonant score, and (2) calculation of the CVC-word score from the product of the initial consonant, vowel, and final consonant probabilities. In Figs 4.3.10 and 4.3.11 the relation is given between the predicted CVC-word scores thus obtained, and the actual CVC-word scores for male and female speech. The prediction accuracy is s = 4.11% for the male speech and s = 3.63% for the female speech. Hence a small improvement is obtained.

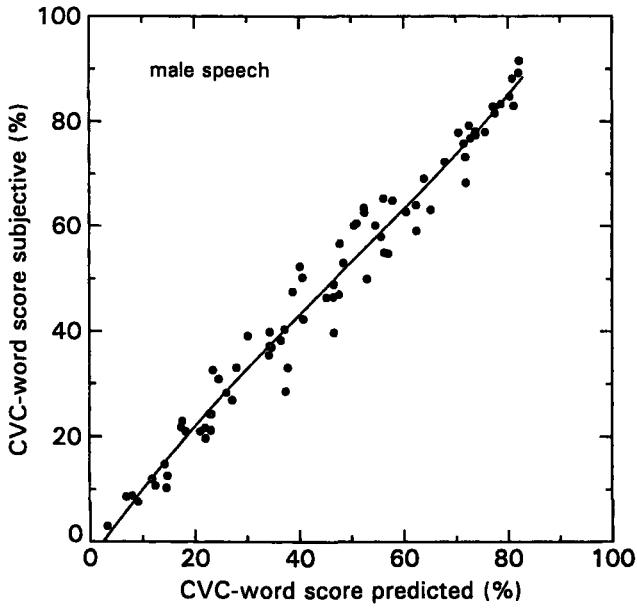

Fig. 4.3.10 Relation between predicted CVC-word scores and observed CVC-word scores for male speech. The predicted CVC-word scores are obtained from the product of the initial consonant, vowel, and final consonant scores for MALE speech.

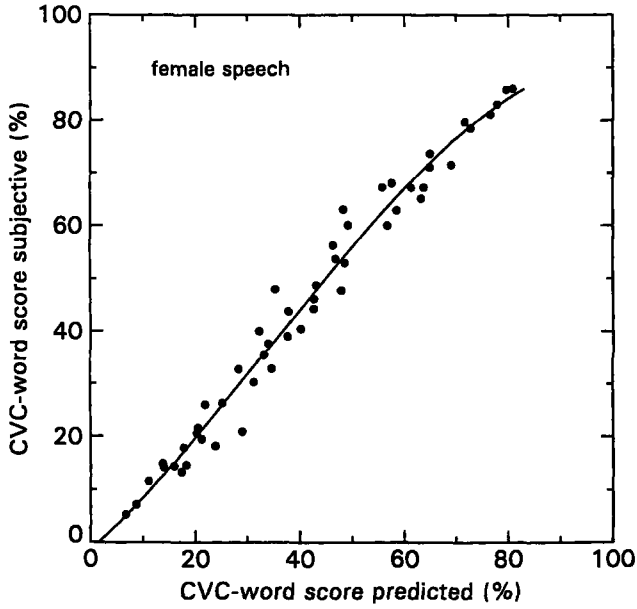

Fig. 4.3.11 Relation between predicted CVC-word scores and observed CVC-word scores for female speech. The predicted CVC-word scores are obtained from the product of the initial consonant, vowel, and final consonant scores for FEMALE speech.

The advantage of predicting the word score by a (weighted) combination of the predicted phoneme-groups scores is that it is not restricted to the example with the equally balanced CVC words as demonstrated above, but can also be used to predict the word score of PB-words or any other combination, including rhyme tests. The restriction is, however, that the product of the probabilities of the phoneme-group score should be highly correlated with the word score (this has been verified for nonsense words, see section 3.4.2.1). Bronkhorst et al. (1992) described a method to extend this prediction to meaningful words, hence the effect of context is included in this method.

4.4 The effect of speaker variation

Significant differences between CVC-word scores obtained with different speakers and listeners were found and described in section 3.4.1. Similar to the difference in speech intelligibility as obtained with male and female speech, the variation among speakers and listeners is of interest for the evaluation of com-

munication channels (especially channels with a fair-to-poor transmission quality). None of the existing objective measures takes this speaker-listener variation into account.

Related to the speakers, various specific speech-production parameters were identified, mainly described by physical parameters such as vocal effort (mainly related to the test signal level), fundamental frequency (not modelled), long-term speech spectrum (related to the test signal spectrum), and speaking rate (related to the frequency range of the modulation transfer function). More difficult to quantify is "speaker efficiency", related to intonation and word stress. Some of these speaker-listener parameters can be accounted for by the frequency-weighting factors and redundancy correction and by a specific relation between the STI_r and the CVC-word score.

We calculated for the four male and the four female speakers, combined with the two listener groups, the optimal "speaker"-specific frequency-weighting factors and redundancy-correction factors. These are given for the male and female speakers in Figs 4.4.1 and 4.4.2, respectively.

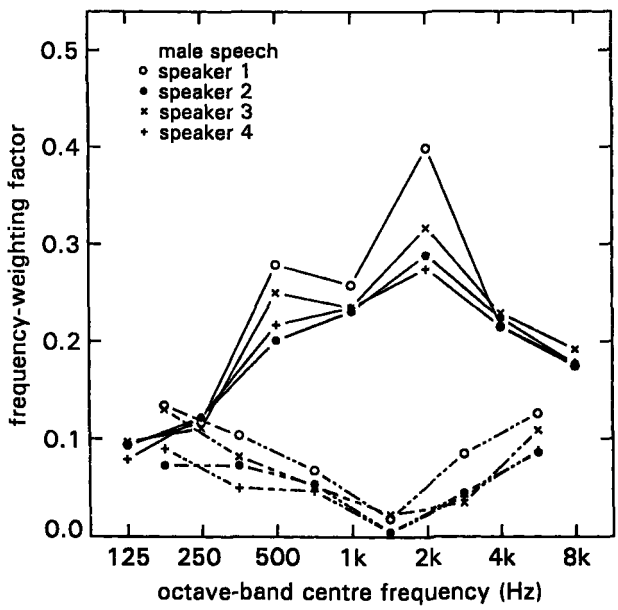

Fig. 4.4.1 Speaker-specific frequency-weighting factors for the octave-band contribution α_k (solid line) and redundancy correction β_k (dashed line) for CVC-word scores for four MALE speakers.

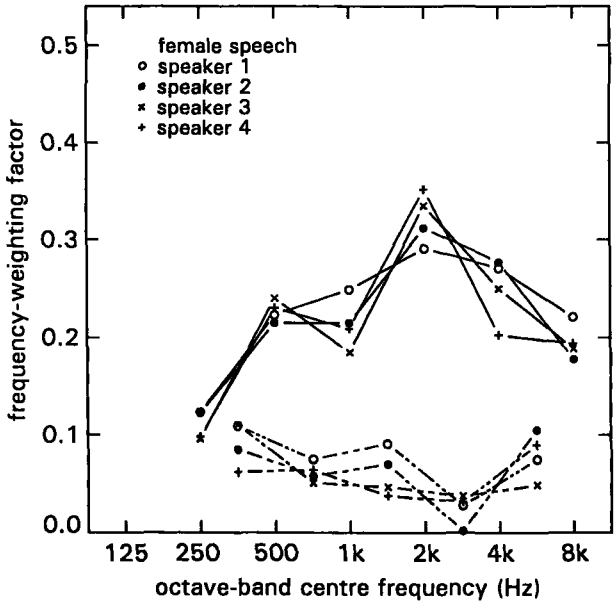

Fig. 4.4.2 Speaker-specific frequency-weighting factors for the octave-band contribution α_k (solid line) and redundancy correction β_k (dashed line) for CVC-word scores for four FEMALE speakers.

It was found with an analysis of variance (ANOVA) that the frequency-weighting and redundancy correction factors for the four speakers are significantly different. For both speaker groups a $p<0.01$ was obtained. The analysis was based on the frequency weighting and redundancy correction for each individual speaker-listener group combination. Replicas were obtained by dividing the conditions into two sets; for each group the optimal frequency-weighting function and redundancy correction were calculated separately.

4.5 Discussion and conclusions

In chapter 2 we have already indicated that in the literature several different frequency-weighting factors have been found, see Fig. 2.1.2. Pavlovic (1987) compared some of these functions and included more recent results. This comparison is one of the essential parts of the recently proposed revision of the existing "American national standard methods for the calculation of the articulation index" (ANSI, 1992). In the draft document four different speech tests are considered and these are related to four different types of speech: nonsense syllables, PB-words, initial consonants, and short passages. The

frequency-weighting functions (referred to in the document as "importance functions") are given for three different resolutions in the frequency domain, namely critical bands, 1/3-octave bands, and 1/1-octave bands. In Fig. 4.5.1. these functions are given for the octave-band resolution. The AI includes the octave bands with centre frequencies from 250 Hz to 8 kHz, hence the frequency band with centre frequency 125 Hz is not specified. The frequency-weighting function as obtained in the present study is also given. We corrected our own frequency-weighting factors for redundancy by subtracting half of the redundancy factor from both adjacent frequency bands.

The American speech material used by Pavlovic (1984) and the Dutch CVC words from this study are similar in structure. By comparing the two corresponding curves in Fig. 4.5.1 it can be seen that different frequency-weighting factors were obtained for the frequency bands with centre frequencies 125 Hz, 250 Hz, 4 kHz, and 8 kHz. This may be due to the different languages (American English versus Dutch), or to the experimental design[3].

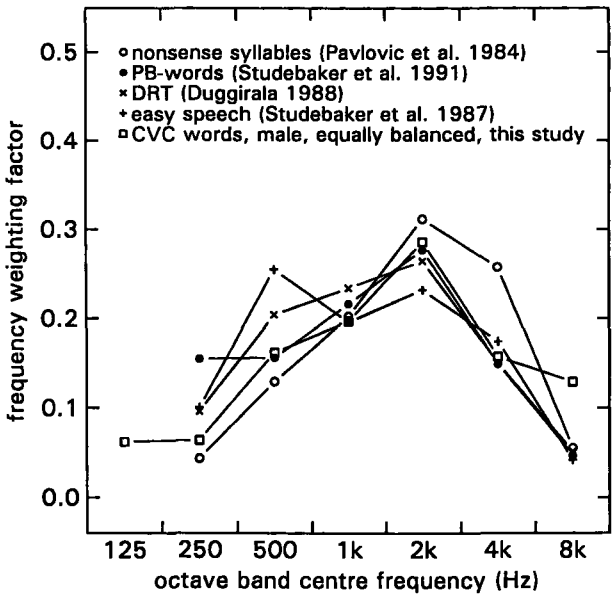

Fig. 4.5.1 Frequency-weighting functions for the octave-band contribution to the AI as proposed in the draft American National Standard (1992). The frequency-weighting function (corrected for redundancy) as obtained in this study for CVC words is also given.

[3] It should be noted that some of the frequency-weighting functions, included in the draft standard, are based on a rather limited set of data points. For instance the "importance function" for nonsense syllables is based on 13 different conditions, and speech of 2 speakers only (this resulted in a set of 17 frequency-weighting factors each specified within four decimals).

The frequency-weighting factors obtained for the DRT words (Duggirala, 1988), are similar to the frequency-weighting factors as obtained by us for initial consonants and final consonants (see Fig. 2.3.14). However, as discussed before a test based on the DRT concept may produce misleading results.

Given the restrictions of the various studies as described above we think that, if a standardization is required, this might be in terms of the phoneme-group-specific frequency-weighting functions, which can be used to extend the method to words consisting of any combination of phonemes.

As the frequency-weighting functions have been obtained for optimal prediction of the corresponding phoneme-group scores, these functions can be considered as the optimal filter characteristic for phoneme-oriented speech transmission and recognition. Even the redundancy correction can be used if the information within certain frequency bands is masked (constant level) or unreliable. Therefore it may be useful to apply this type of filter as a preprocessor for phoneme oriented speech recognizers.

Conclusions

- The frequency-weighting functions for the four phoneme groups used in this study are very different from each other and indicate that a different optimal frequency range is required for a correct recognition of the phonemes within each group at restricted transmission conditions.

- The frequency-weighting functions for male and female speech, obtained in independent studies, are similar. This indicates the robustness of the obtained results.

- The relation between the phoneme-group scores and the STI_r is also phoneme-group-dependent and saturates at different intelligibility levels.

- A small improvement of the prediction accuracy of CVC words was obtained by calculating the CVC-word score from the individual predicted phoneme-group scores. This procedure allows an extension of this concept to any type of word combination.

- The prediction accuracy of the observed intelligibility by the phoneme-group-specific STI_s varies for the different phoneme groups. The best prediction was obtained for vowels, vowel-like consonants, fricatives and also CVC words (s = 3.6 - 4.6%). A less accurate prediction was obtained for plosives (s = 5.6 - 5.8%). Such a prediction accuracy is, expressed in terms of signal-to-noise ratio, smaller than 2 dB.

- Speaker-specific relations between the STI_r and the observed intelligibility, and speaker-specific frequency-weighting factors and redundancy corrections were found. These can be used to predict speaker variation.

5 VALIDATION OF THE STI METHOD WITH THE REVISED MODEL

Summary

The revised model for the STI, and the corresponding frequency-weighting functions are validated with an independent set of 18 transfer conditions. It was found that a good prediction with this design can be obtained. The revised model was also applied to 50 other communication channels including nonlinear distortion, echoes, automatic gain control, and wave-form coding. For these types of distortion additional parameters of the test signal are of interest and can be tuned for an optimal fit with the observed intelligibility scores.

5.1 Introduction

In the preceding chapters a revised model of the STI algorithm and a redefinition of the parameter values were obtained. Improvements involved the application of a redundancy correction between adjacent frequency bands, the extension to female speech, and the extension to prediction of the intelligibility of specific groups of phonemes (chapters 2 and 4). For validation purposes we will apply this revised model to transfer conditions on communication channels. These conditions include band-pass limiting, noise, nonlinear distortion, and distortion in the time domain.

The conditions with band-pass limiting and noise are used for an *independent* verification of the parameters as obtained before.

In this chapter the STI_r values of the various transfer conditions are obtained by measurement. They represent an example of the practical applicability of this method. For comparison, we start this chapter with an overview of the existing objective intelligibility measuring systems (without discussing the concepts).

5.2 Objective measuring methods for predicting speech intelligibility

The first description of the use of a computational method for the prediction of the intelligibility of speech and its realization in an objective measuring device, was given by Licklider et al. (1959). They described a system which could measure the spectral correspondence between speech signals at the input and at the output of the transmission channel under test, the so-called Pattern Correspondence Index (PCI). This PCI shows a remarkable similarity with the AI

(Articulation Index), although the approach is quite different. A spectrally-weighted contribution of the similarity between temporal envelopes of the speech signals at the input and at the output of a transmission channel is used for the computation of the PCI. A total of 15 minutes of speech was required for this analysis. The paper reports that the results of a comparison between the PCI and human listener evaluation show a monotonic relation for conditions with an increasing effect of one type of distortion. Contributions of different types of distortion show a "sufficient agreement". The electronic design of the system was described by Schwarzlander (1959). Licklider proposed an improvement of the PCI by making use of synthetic signals, physically related to average speech, and with a duration of about one second for the total measurement of the PCI.

Five years later Kryter and Ball (1964) described a system called the Speech Communication Index Meter (SCIM), which was based on the AI as described by Kryter (1963). The measurements were mainly concentrated on deriving the signal-to-noise ratio within a frequency range of 100-7000 Hz, a dynamic range of 30 dB, and on auditory masking corrections according to the AI concept. An evaluation of the system was performed for several types of transmission conditions, including low-pass filtering, noise, frequency shifts, and clipping. This was done by comparing SCIM results with human listener results based on the MRT (Modified Rhyme Test). Since Kryter and Ball published all the original SCIM scores and MRT results, we were able to calculate the relation between the two results by using the same method as in our study described before (see section 2.4). We calculated the third-order regression and the vertical spread around the regression line. This vertical spread, expressed by the standard deviation, was s = 9.1%.

In 1970 Houtgast and Steeneken developed a system based on the use of an artificial test signal which was transmitted over the channel-to-be-tested and which was analyzed at the output. The test signal was an amplitude-modulated noise signal with a square-wave amplitude modulation. Hence the signal level alternated between two values. The difference between these two levels was 20 dB and the switching rate was 3 Hz. The noise carrier had a frequency spectrum corresponding to the long-term speech spectrum. It was the first approach in which speech-related phenomena, concerning dynamic variations and temporal variation, were included in an artificial test signal. The essential point of this approach was that the resulting level variation at the output of a communication system reflects the signal-to-noise ratio, providing a basis for subsequent calculations according to the AI concept. The method was based on measurements in five octave bands (centre frequencies 250 Hz - 4 kHz). The

effect of band-pass limiting, noise, peak-clipping, and reverberation on intelligibility was included in the test signal concept and in the evaluation procedure. This resulted in an index ranging from 0 - 1, the so-called Speech Transmission Index (STI). A measuring device was developed, based (at that time) on analogue circuits, which could determine the STI within 10 s. It should be noted that this method is different from the STI approach published later and described in appendix A1.

The relation between this objective STI and the subjective CVC-word scores, obtained with speakers and listeners, was studied for 50 different transmission conditions. It was found that the standard deviation around a third-order regression line between STI and the word score for phonetically-balanced nonsense words was 6.5% (Houtgast and Steeneken, 1970).

The next step was to use a test signal with various modulation frequencies instead of the fixed (3 Hz) square-wave modulation signal. This modulated test signal was based on the measurement of the fluctuations of the envelope of connected discourse (Houtgast and Steeneken, 1971). The envelope fluctuations were determined for separate frequency bands (octave bands). As the envelope function is unique for a certain combination of successive speech sounds, a more global quantification was used: the frequency spectrum of the envelope fluctuations, called the "envelope spectrum" (see Fig. 5.2.1). This envelope spectrum (with a frequency range from about 0.2 Hz to 25 Hz) was measured in 1/3-octave bands and normalized with respect to the mean level (intensity).

The transfer of these fluctuations of speech by a communication channel can be obtained by comparing the envelope spectra of the same speech signal at the input and at the output of the channel under test (Steeneken and Houtgast, 1973). Hence, natural speech is used as the test signal. The effect of noise on the envelope spectrum of speech is independent of the fluctuation frequency, however, this is not the case for distortions in the time domain. Reverberation will act as a low-pass filter for fluctuations and can be predicted for an exponential decay. There is a simple relation between the relative decrease of the fluctuations and the signal-to-noise ratio. Hence this method can be used to measure the effective signal-to-noise ratio as a function of fluctuation frequencies.

Next to the use of natural speech as a test signal, Houtgast and Steeneken (1972) also proposed the use of an artificial test signal, and the testing of each relevant fluctuation frequency separately. This resulted in the so-called Modulation Transfer Function (MTF). The MTF can be considered to act on the envelope of a signal between input and output of a transmission channel.

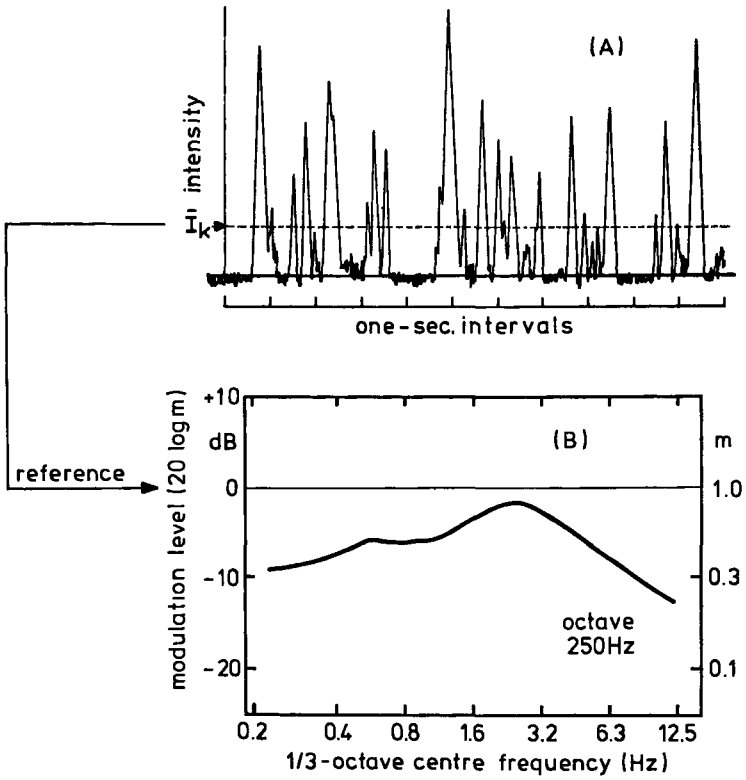

Fig. 5.2.1 Envelope function (panel A) of a 10s speech signal filtered for the octave band with centre frequency 250 Hz. The corresponding envelope spectrum (panel B) is normalized with respect to the mean signal intensity (I_k).

The method was extensively evaluated for conditions with noise, reverberation, and echoes. The analysis and the generation of the echo conditions, at that time, were performed with a digital (PDP-7) computer, a system with a 1.75 μs cycle time and 8K-words of memory!

Payne and McManamon (1973) introduced the Speech Quality Measure (SQM) for communication channels. This system was based on the AI concept. The authors mentioned limitations for digital encoding, fading, and nonlinear distortion. They remarked: "when using the system it should be checked to have none of these distortions present". The test signal was based on 20 tones with frequencies at the mid-point of the 20 frequency bands with "equal contribution to intelligibility" as used for the original AI concept. The paper also proposes the use of mini-computers to perform the analysis and to display the results. No validation was reported.

Steeneken and Houtgast (1980) extended the MTF approach (that had already been validated for channels with noise, echoes, and reverberation) to channels with distortions more specific for communication channels, namely band-pass limiting, noise, nonlinear distortion, quantization errors from digital coders and reverberation. This measuring procedure is described in appendix A1.

Quackenbush et al. (1988) gave an overview of "Objective measures of speech quality" especially applied to digital coders. They also evaluated some objective measures, which were mainly based on signal-to-noise ratios. As a subjective measure the Diagnostic Acceptability Measure (DAM; Voiers, 1977) was used. The DAM is based on the rating of aspects concerning the speech signal and the background by a listening panel. No robust evaluation of the DAM is reported.

5.3 Experimental design

A significant improvement in the prediction accuracy of the STI was obtained by extension of the additive model with a correction for the correlation between adjacent frequency bands. Also, an extension was made for the prediction of the intelligibility of female speech. In this section these improvements will be validated with an independent set of 68 communication channels. These transfer conditions are described in appendix A3 and were not used for the evaluation of the STI model as described in chapters 2 and 4.

5.3.1 *Description of the measuring conditions*

For the experiments on communication channels three main types of distortion were used: band-pass limiting, nonlinear distortion, and distortion in the time domain. More specific: 18 conditions were obtained by combinations of two types of band-pass limiting and four types of noise, 26 conditions based on nonlinear distortion were obtained with peak clipping, centre clipping and digital wave-form coders, and 24 conditions with distortion in the time domain were based on automatic gain control and echoes. These distortions were combined with different types of noise at various signal-to-noise ratios and, if applicable, with band-pass limiting. A complete description of the total set of 68 communication channels is given in appendix A3.

5.3.2 *Experimental set-up*

Similar to the experiments described before, a subjective evaluation was performed with four male and four female speakers and two groups of four listeners. These subjective measurements and the experimental set-up are identical to that used for the experiments as described in chapter 2. A complete description is given in sections 2.2.2, 2.2.3 and 3.3.

The STI_r values were obtained by measurement of the 68 transmission conditions. For the conditions without distortion in the time domain a simplified measuring method was used (restricted to 3 modulation frequencies for each octave band). The conditions with echoes and automatic gain control require a measurement of the full modulation transfer function including 14 modulation frequencies, as described in appendix A1.

5.4 **Experimental results**

The validation of the STI_r was performed for CVC words. We studied the fit of the STI_r value and the corresponding CVC-word score with the relation as obtained before for male and female speech separately (see Figs 4.3.1 and 4.3.2). This study was performed separately for the 18 conditions including band-pass limiting and noise, the 26 conditions including nonlinear distortion, and the 24 conditions including distortion in the time domain.

5.4.1 *Communication channels with band-pass limiting and noise*

The conditions labelled 1-18, as described in appendix A3, consist of two band-pass limitations, four types of noise, combined at three different signal-to-noise ratios. One band-pass limitation corresponds with the so-called telephone bandwidth (300-3400 Hz). The four noise spectra include white noise, pink noise, low-frequency noise, and noise with a frequency spectrum equal to the long-term speech spectrum.

The corresponding CVC-word scores for these conditions (1-18) are included in appendix A4, Tables A4.2 and A4.4 for the male and female speech respectively.

The STI_r was measured according to the description in appendix A1, and the calculation of the STI_r value included the redundancy correction. As no distortion in the time domain is considered, the modulation transfer will be independent of the modulation frequency.

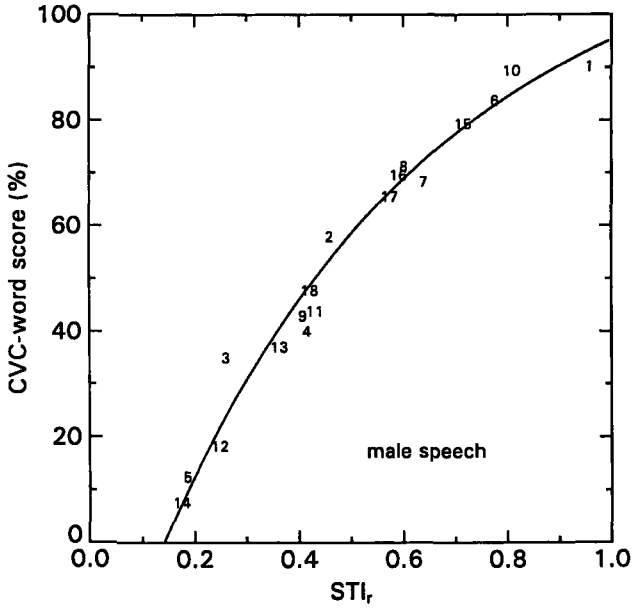

Fig. 5.4.1 Relation between the STI_r and the CVC-word score for the 18 transfer conditions including band-pass limiting and noise for MALE speech. The standard deviation, representing the vertical spread around the polynomial (cf. section 4.3.2.1) is s = 4.4%.

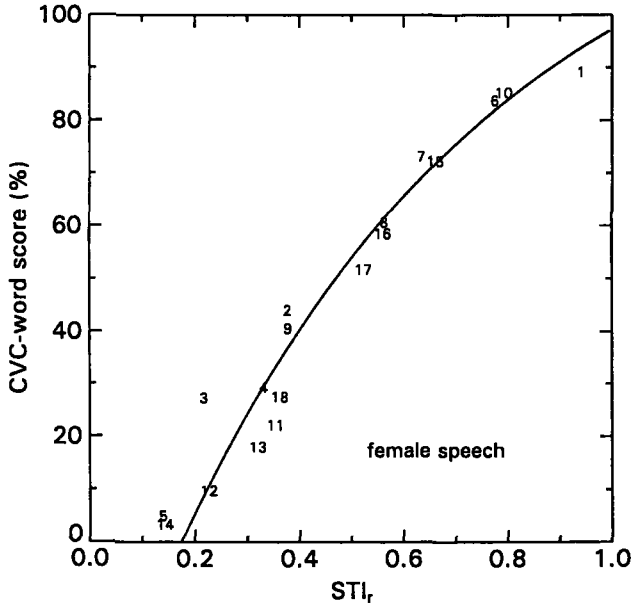

Fig. 5.4.2 Relation between the STI_r and the CVC-word score for the 18 transfer conditions including band-pass limiting and noise for FEMALE speech. The standard deviation, representing the vertical spread around the polynomial (cf. section 4.3.2.1) is s = 6.6%.

The relation between the STI_r value and the observed CVC-word score is given in Fig. 5.4.1 for the male speech conditions and in Fig. 5.4.2 for the female speech conditions. The standard deviation around the *predefined* relation (the exponential function was obtained from section 4.3.1 and is no optimization of the present data) is s = 4.4% for the male speech, and s = 6.6% for the female speech.

The goodness of the fit for these two independent sets of conditions indicates that the revised model of the STI concept and the corresponding frequency-weighting function and redundancy correction used for the calculations are robust, and offer for this type of communication channels an accurate prediction of the observed intelligibility score.

5.4.2 *Communication channels with nonlinear distortion*

For the verification with a subset of 26 communication channels with nonlinear distortion we applied peak clipping (as introduced by overloaded systems), centre clipping (as introduced by carbon microphones), and quantization errors (as introduced by wave-form coding, pulse-code modulation and delta-modulation). The effect of nonlinear distortion is different for each condition and varies in the amplitude domain. We combined these conditions with band-pass limiting, noise, and (for wave-form coders only) with bit errors. These combinations are representative of communication channels in adverse conditions. We obtained 26 different conditions and measured the STI_r in the way described in section 5.4.1. However, channels with a nonlinear transfer introduce harmonics and intermodulation components into other octave bands, and therefore require a more sophisticated test signal. It should be noted that the structure of this more sophisticated test signal should have no influence on the test results in the preceding paragraph. The most representative test signal would be running speech, in which only the speech signal in the octave band being tested is replaced by the sine-modulated test signal. In this way representative disturbing components are introduced into the octave band being tested, uncorrelated with the sinusoidal intensity modulation of the test signal within that octave band. For practical reasons the running speech signal is replaced by an artificial fluctuating signal with fluctuations similar to running speech. The relative level of this randomly fluctuating signal is an important parameter since it defines the amount of degradation. For instance, if the level of the random fluctuations of the test signal is too low, a relatively small amount of distortion components will be introduced into the other octave bands and the resulting effective signal-to-noise ratio will be too high. Hence the corresponding

STI_r will also be too high. In an iterative procedure we tuned this level for an optimal fit with the existing relation between STI_r and the observed CVC-word score, similar to the method as described by Steeneken and Houtgast (1980). In Figs 5.4.3 and 5.4.4 the data points around the existing relation between STI_r and observed CVC-word score are given.

As can be seen from the graphs the data points 29-32, corresponding with the centre clipping conditions, do not coincide with the other data points. Centre clipping appears to have a dramatic effect on the intelligibility, not reflected in the STI_r.

The standard deviation representing the vertical spread around the predefined polynomial is s = 6.46% for the male speech and s = 7.83% for the female speech. The centre clipping conditions are omitted in this standard deviation.

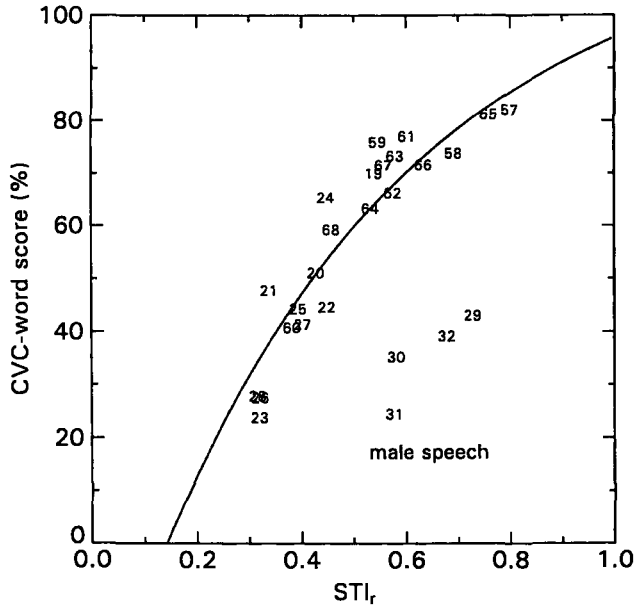

Fig. 5.4.3 Relation between the STI_r and the CVC-word score for the 26 communication channel conditions including nonlinear distortion for MALE speech. The standard deviation, representing the vertical spread around the polynomial (cf. section 4.3.1) for the conditions excluding centre clipping (29-32) is s = 6.46%.

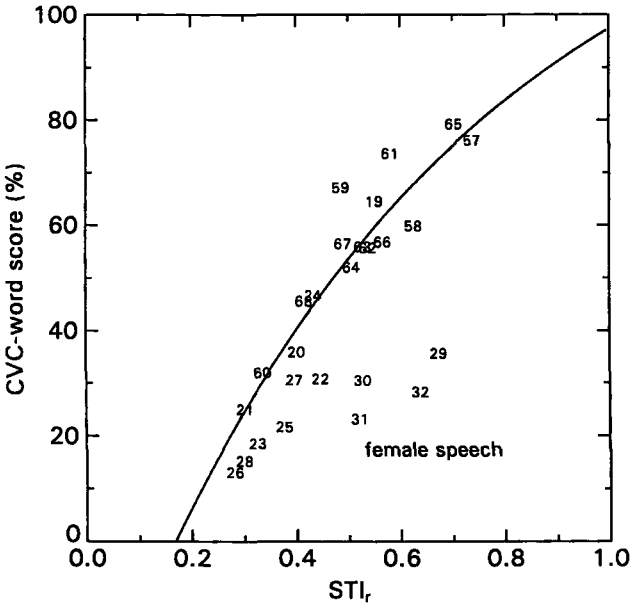

Fig. 5.4.4 Relation between the STIr and the CVC-word score for the 26 communication channel conditions including nonlinear distortion for FEMALE speech. The standard deviation, representing the vertical spread around the polynomial (cf. section 4.3.1) for the conditions excluding centre clipping (29-32) is s = 7.83%.

5.4.3 *Communication channels with distortion in the time domain*

One of the major advantages of the MTF concept, as presented by Houtgast and Steeneken (1973), was to describe distortions in the time domain by the transfer of individual modulation frequencies (see section 5.1.1). The relation between the frequency range of the fluctuations in connected discourse and the range of the modulation frequencies of the MTF is obvious (Steeneken and Houtgast, 1980). In general a frequency range of 0.63 Hz to 12.5 Hz is considered.

For transmission conditions with reverberation the fast fluctuations of the speech signal will be reduced, hence the higher modulation frequencies of the MTF. On the other hand, transmission conditions including automatic gain control may reduce the slow fluctuations in speech, related to the lower modulation frequencies. Echoes introduce a ripple onto the modulation transfer function related to the time delay of the echo. A suppression (anti-phase) or amplification (in phase) is obtained for certain modulation frequencies (comb-filter effect).

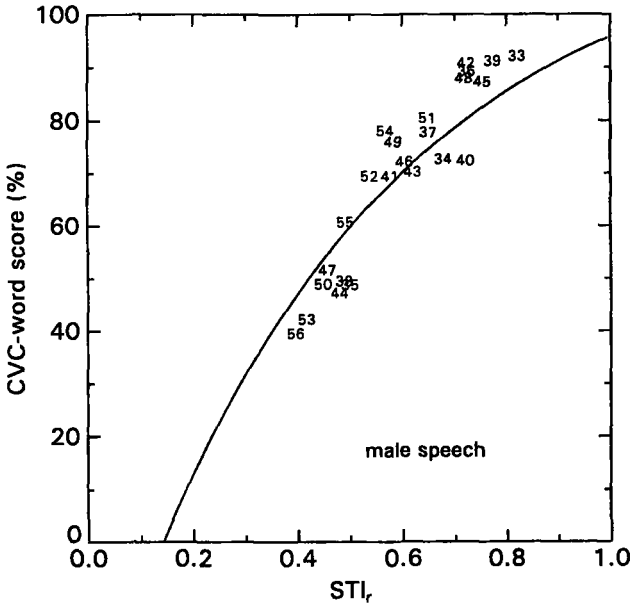

Fig. 5.4.5 Relation between the STI_r and the CVC-word score for the 24 communication channel conditions including distortion in the time domain for MALE speech. The standard deviation, representing the vertical spread around the polynomial (cf. section 4.3.1) is s = 6.93%.

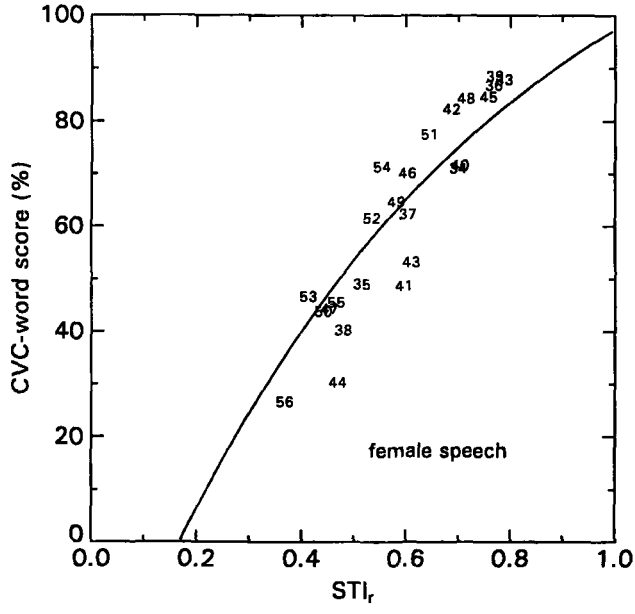

Fig. 5.4.6 Relation between the STI_r and the CVC-word score for the communication channel conditions including distortion in the time domain for FEMALE speech. The standard deviation, representing the vertical spread around the polynomial (cf. section 4.3.1) is s = 8.21%.

A careful selection of the modulation frequencies included in the MTF is required to model the effect of distortion in the time domain as reflected in the STI_r in a similar way as was performed for noise masking. Hence for conditions with the same effect on intelligibility but based on either noise masking or distortion in the time domain, identical STI_r values should be obtained.

We verified the revised STI-model for 24 different conditions including automatic gain control and echoes. Unfortunately we could not obtain conditions with reverberation. In Figs 5.4.5 and 5.4.6 the data points around the existing relation between STI_r and the observed CVC-word score are given for male and female speech.

5.5 Discussion and conclusions

The calculation scheme and the corresponding parameters, defined in the previous chapters of this study, were validated with an independent set of 18 transmission conditions including band-pass limiting and noise. It was found that the relation between the STI_r and the observed CVC-word score for the present set of data corresponds well with the relation found before. This was expressed by the vertical spread of the data points around the existing functions ($s = 4.4\%$ for the male speech and $s = 6.6\%$ for the female speech).

A similar verification was performed for two other sets of transmission conditions, one including nonlinear distortion and a second including distortion in the time domain. The fit of these data points is similar to the experimental results as found before by Steeneken and Houtgast (1980). This indicates that the STI_r does not introduce systematic differences into the relation with the observed intelligibility for various types of degradation.

Conclusions

- 18 Transmission conditions (with band-pass limitations and noise) were used to verify (in an independent way) the revised STI_r model and the redefined frequency-weighting functions. The results showed a standard deviation of $s = 4.4\%$ for the male and $s = 6.6\%$ for the female speech conditions.

- Application of the model to nonlinear channels and channels with a distortion in the time domain produced somewhat lower results than obtained with linear channels and showed the general applicability of the method.

6 RECAPITULATION AND MAIN CONCLUSIONS

We all know many examples of communicative situations with a poor intelligibility, such as public address systems at railway stations, auditoria, telephone communication in a noisy environment, portable transceivers, secure-voice systems, speech-synthesis systems, etc. Although many intelligibility tests and speech-quality tests have been developed during the past 60 years, only few designers of speech-communication systems evaluate their design with tests using speech or speech-related signals.

In this book we describe various intelligibility measures based on subjective methods (making use of speakers and listeners), on predictive methods (making use of a description of the physical properties of a system), and on objective measures (making use of artificial speech-like test signals). The main emphasis of the present study is on a further improvement of the relation between some of the subjective and objective measures. We also investigated the relation between phoneme and word scores and sentence intelligibility.

6.1 Subjective intelligibility measures

Sentence intelligibility

Speech communication is normally based on connected discourse and therefore the intelligibility of sentences is an appropriate measure to quantify the quality of a communication system. However, for various reasons the use of sentences for the assessment of communication systems is limited (one reason is that it discriminates only between bad and poor conditions, see section 3.2). Therefore, intelligibility measures based on words or phonemes are used more frequently. We have studied the relation between the intelligibility of specific simple sentences and phoneme and word scores. We obtained this relation for three levels of sentence intelligibility (75%, 50%, and 25%), which turned out to correspond, for male speech, to an average CVC-word score of 31%, 25%, and 18%, respectively (Table 3.5.3). The correlation coefficient between the two measures is $r = 0.76$ for male speech and $r = 0.71$ for female speech. By making use of a multiple regression technique, an optimal relation between sentence intelligibility and a combination of consonants and vowel scores could also be obtained. The corresponding correlation coefficients between sentence intelligibility and a weighted combination of specific phoneme-group scores are: $R = 0.86$ for male speech and $R = 0.82$ for female speech.

There is a good correspondence (typically within 1 dB signal-to-noise ratio) with the results described here and the relations obtained before and given in Fig. 3.2.1 (based on simple military communications). The relation found in this study is valid for simple (Dutch) sentences consisting of 5-10 words.

The correlation coefficients (see Table 3.5.3 and 3.5.4, bottom row) show that the scores of some groups of phonemes have a fair relation with the corresponding sentence intelligibility, but that the scores of other phoneme types (initial consonants, vowels) have a poor to bad relation. It can therefore be expected that intelligibility measures based on consonants or vowels only (rhyme test, digits, alphabet) will also have a poor relation with sentence intelligibility.

Phoneme and word tests

With the results of the CVC-word test, as used for the evaluation of the objective intelligibility measuring method (chapters 2, 4 and 5), various analyses were made:

(1) We analyzed the effect of several parameters concerning the experimental design of a subjective intelligibility test. The goal of this analysis is to optimize this experimental design.

(2) The relation between the scores of the 43 different phonemes was analyzed and could be reduced to four groups of phonemes, each group with a similar response at various transmission conditions.

re 1. Experimental design

Tests with an open response design and based on nonsense words require extensive training of the listeners. Consequently, as argued in section 3.3.5, a balanced experimental design is required in order to minimize systematic effects of learning by the listeners.

Speaker-listener scores differ significantly, see Table 3.4.1. In general, with 16 speaker-listener pairs a standard error of the mean score of 1-3% is obtained. This may be acceptable for a comparative study.

The amount of variance introduced by the speakers and by the listeners is about equal, therefore the number of speakers and listeners used for an experiment should also be equal and high.

As the results obtained for male and female speakers differ significantly, separate experiments for both groups may be considered. In the past many experiments, mainly concerning military communication, were performed with male speech only.

re 2. Phoneme grouping

It was found that the degradation of the 43 different phonemes, obtained at many different transmission conditions, can be grouped into four sub-sets. These sub-sets are fricatives, plosives, vowel-like consonants, and vowels. A similar grouping, focused on production features, was obtained earlier by Miller and Nicely (1950). The relation between the scores of these four group responses might be useful to characterize an (unknown) communication channel. For some results see Steeneken (1986), Bos and Steeneken (1991).

6.2 Objective intelligibility prediction

Prediction by the AI and STI of the intelligibility of speech signals, degraded by band-pass limiting and noise, is based on the sum of weighted contributions of individual frequency bands. We have found that this simple additive model leads to erroneous results for conditions with certain types of band-pass limiting (reduction at the high and low frequency range) and for conditions with gaps in the frequency transfer. A redundancy correction was proposed in order to account for the mutual dependency between adjacent octave bands (see section 2.3.2). This extension of the model gave a major improvement of the accuracy of the prediction of intelligibility. We verified the revised model with results of earlier studies by Kryter (1960) and Steeneken and Houtgast (1980). For both studies a major improvement of the relation between the objective prediction and the corresponding observed intelligibility scores was obtained.

Frequency-weighting factors were found, representing the relative contributions and relative redundancy corrections for the frequency bands considered. These frequency-weighting factors were found to be similar in independent studies for male and female speech and for speech signals combined with masking noise at different signal-to-noise ratios. We also found a robust relation between signal-to-noise ratio and the contribution to the objective measure.

All verifications in this thesis are based on the relation between the CVC-word scores and the objective intelligibility prediction. The definition of the signal-to-noise ratios for the various transfer conditions used in the initial part of this study was based on the long-term speech spectrum. This means that these signal-to-noise ratios are not valid for specific speech items such as different types of phonemes, since the corresponding speech spectra may differ significantly. Therefore the obvious extension of the model to predict the scores for different groups of phonemes (phoneme-group-specific STI) could not be

made directly. Either the predefined signal-to-noise ratios had to be corrected according to the mean spectrum of the phoneme group considered, or the signal-to-noise ratios had to be measured by making use of test signals with a representative frequency spectrum. We preferred the latter method and we measured the STI for all the conditions used in chapter 2 with four different test-signal spectra matched to the phoneme groups as found in chapter 3. The optimal frequency contribution and redundancy corrections found are different for the four phoneme groups but similar for male and female speech. The shape of these frequency-weighting functions might be helpful for an optimal design of a specific communication system with a limited frequency transfer (e.g. public address systems with horn loudspeakers) or for hearing impaired with a reduced frequency range of the hearing organ. Also, the different frequency-weighting functions described in the literature can be unified by this speech-type-dependent frequency weighting (see section 4.3.1).

Individual prediction of the phoneme-group scores allows prediction of nonsense-word scores by a weighted combination of the phoneme-group probabilities. This was demonstrated in section 4.3.3. and extends the prediction of the word score for the equally balanced CVC-type words, as used in this study, to other types of nonsense words. The draft ANSI standard (ANSI 1992) proposes five different intelligibility tests and gives the corresponding frequency-weighting factors. The approach described here is more general and allows for any combination of phonemes as long as the word scores can be predicted by the individual phoneme group probabilities. As proposed by Bronkhorst et al. (1992) this procedure may even be extended to meaningful words.

The experiments described in this study were based on spectral filtering in octave bands. Between adjacent octave bands at both the low and the high frequency ends of the speech spectrum, relatively high redundancy corrections were required for optimal prediction by the extended model. This reflects a high correlation between these bands and, consequently, the frequency resolution may be higher than required. On the other hand, for the octave band with a centre frequency of 2 kHz a high frequency-weighting factor and a low redundancy correction were found. This means that an improvement might be obtained by dividing this frequency band into two or more frequency bands. The design of the present experiments, in which the conditions are defined in terms of octave bands, did not allow the evaluation of such an extension. The optimal solution would probably be a fairly flat frequency-weighting function and a small redundancy correction for all frequency bands. The frequency-weighting and redundancy correction as obtained for the Kryter data (see Fig. 2.4.2) give an indication in this direction. Unfortunately, the frequency range of his

experiments was limited to between 250 Hz and 3000 Hz, and the data were based on one male speaker only.

6.3 Validation of the STI_r

We applied the revised model, as discussed above, to various types of transmission channels with distortions which are representative of communication channels. For application to channels with band-pass limiting and noise, a verification was performed with the revised model and the redefined parameters values. We found the same relation between the measured STI_r values and the observed CVC-word scores as was found before with the optimization of the STI_r model. The vertical spread of the data points around the optimal relation found before is s = 4.6% for male speech and s = 4.5% for female speech, whereas for the present independent set of data s = 4.4% and s = 6.6% respectively.

Two separate experiments were performed to test channels with nonlinear distortion (peak clipping, centre clipping, and wave-form coders), and channels with distortion in the time domain (automatic gain control and echoes). According to the STI measuring procedure (see appendix A1) an independent parameter can be used to fit the results of the conditions with nonlinear distortion to the existing relation between STI and CVC-word score. This was performed in an iterative procedure, by adjusting the relative level of the test signal in the octave bands containing the random fluctuations. The vertical spread around the optimal relation found before is s = 6.46% for male speech and s = 7.83% for female speech.

Distortion in the time domain was adjusted by a proper choice of the range of the modulation frequencies included in the modulation transfer function (see appendix A1). We measured the complete matrix of 7 octave bands times 14 modulation frequencies (according to Fig. A.1.2) and optimized the weighting function of the modulation transfer for an optimal fit with the existing relation. The vertical spread around the optimal relation found before is s = 6.93% for male speech and s = 8.21% for female speech. In this relation the data points related to centre clipping are omitted. We are presently unable to fit these in the existing model.

6.4 Application of the STI_r

The STI_r measurements of this study were performed with the existing measuring device, based on fixed frequency bands as designed in 1980. Recent developments with digital signal processors making use of a floating-point hardware, which are available as add-on boards for a PC, allows integration of the STI_r measuring procedure within such a system. In this way a flexible measuring set-up can be obtained and the method can be made available by software algorithms. Two identical systems are currently under development, one real-time version making use of a digital signal processor and one off-line version working on any C-language oriented computer.

The accuracy of the prediction of the CVC-word score or the phoneme-group score was expressed by the vertical spread of the data points around the best fitting third-order polynomial between the STI_r and the observed intelligibility scores. This accuracy can also be expressed by a corresponding change of the signal-to-noise ratio. For the relation between STI_r and CVC-word scores the standard deviation for male and female speech is approximately 4.5%. This is equivalent to a variation of the speech level of approximately 1.5 dB. This value is very small, also in comparison with the various uncontrolled level variations as caused by speakers, background noise, microphone positions, and other system parameters.

It should be noted that the STI method as described and extended in this study is applicable to most types of transmission systems which are based on the transmission or coding of the wave-form of the speech signal. Certain types of distortion are not covered by the STI concept, such as centre clipping, frequency shifts, pitch variation, and coding of speech spectra. Hence vocoders and speech synthesis systems cannot be evaluated with the present STI_r method.

It should also be noted that the evaluation was restricted to nonsense words and phonemes pronounced according to the Dutch language. It may be expected, however, that the results are valid for all West European languages, as was found before with the evaluation of the RASTI method (Houtgast and Steeneken, 1984).

6.5 Relations with other topics

Assessment methods for automatic speech-recognition systems range from methods based on a representative test corpus to methods based on artificial

(speech-like) signals. A range of methods was described by Steeneken and van Velden (1991). In general, however, most research projects focus on a representative data-base with highly redundant words of variable length. This requires extensive test runs to discriminate between systems with a fairly equal performance. These tests are in general not very diagnostic, and do not indicate systematic response errors. Therefore the use of test items which discriminate at phoneme level are more appropriate. A CVC-word data-base has been proposed consisting of words which only differ in the initial consonant, resulting in a data-base with 17 test words corresponding with the 17 different initial consonants for the Dutch language (Steeneken and van Velden, 1989). The same can be done for the final consonant or the vowel. The evaluation of the test results, consisting of phoneme scores and confusions, is similar to the method described in section 3.4. A two-dimensional representation is obtained which indicates the discriminative properties between phonemes by the evaluated condition or system.

The speech communication channels as described in appendices A2 and A3 were carefully calibrated with respect to the CVC-word scores and phoneme (group) scores. This was performed for male and female speech. These transmission conditions could be used to assess speech-recognition systems in adverse conditions. As the communication channels were realized by means of digital signal processing techniques, a conversion could be made to a software simulation which performs off-line processing of speech data-bases. This will improve the international standardization of the assessment of speech systems (speech recognizers, speech synthesizers, and speech communication systems). A first step in this direction was made by the Esprit SAM group.

Humes et al. (1986) demonstrated the ability of the AI and STI to describe the speech-recognition performance of hearing-impaired listeners. An accurate prediction in the frequency domain by the AI, and an accurate prediction in the time domain by the STI were obtained. He proposed a combination of the two methods for application to the hearing impaired. The improvement of the STI concept with respect to the frequency weighting and redundancy correction can also be applied to the recognition of speech by hearing-impaired listeners. As some types of hearing disorder show similar gaps in the frequency domain as we used for the study described in chapter 2, it may be feasible for the revised model to offer a more accurate prediction than that reported by Humes.

6.6 Proposal for future research

Several research questions, raised by the interpretation of the results of the experiments described in this thesis, are still unsolved and require more attention.

The relatively high contribution to intelligibility of the octave band with a centre frequency of 2 kHz and the corresponding small redundancy correction with the adjacent frequency bands indicate the high and unique information content of this octave band. Separation of this frequency band into two or more bands would provide a better resolution in this frequency range and a more uniform frequency-weighting and redundancy function.

Another point of interest is the huge effect on intelligibility of a speech signal degraded by centre clipping, which is not reflected in the STI_r value (see section 5.4.2). It is of interest to study the spectral variation introduced by this type of distortion, in relation to the type of speech signal (different groups of phonemes).

The STI model was further extended with respect to speaker variability. Independent, significantly different, frequency-weighting functions and redundancy-correction factors were found for four male and four female speakers. These functions are useful to predict speaker variability for a specific communication channel. It is noticed that these results have been obtained for a limited number of speakers and therefore only give a global indication. It is more general than the speaker-proficiency factor as proposed by Fletcher and Galt (1950).

6.7 Conclusions

Subjective intelligibility measures

- A good relation was found between sentence intelligibility and a weighted combination of phoneme-group scores ($R \approx 0.85$).

- Significantly different intelligibility scores were found for male and female speech at various transmission conditions.

- Four groups of phonemes (fricatives, plosives, vowel-like consonants, and vowels) can be identified, each group consisting of phonemes with a similar response to various types of transmission conditions and degradation.

- Given the systematic differences between the scores of these different groups of phonemes for various transmission conditions, a robust intelligibility measuring method should include all these groups. Therefore tests based on CVC words or CV words are more appropriate than tests predominantly based on either consonants (DRT) or vowels (digits, alphabet).

Objective intelligibility measures

- The extension of the original STI model, which was based on the independent contribution of a number of frequency bands, to a model which includes the interaction between adjacent frequency bands (redundancy correction), leads to a more accurate prediction of the CVC-word score and phoneme-group scores.

- The frequency-weighting factors and corresponding redundancy-correction factors obtained for male and female speech at various signal-to-noise ratios are also very robust.

- Different frequency-weighting factors and corresponding redundancy-correction factors were found for different types of speech. This may explain the variation in frequency-weighting functions found in some other studies.

- A reproducible relation between signal-to-noise ratio and the contribution to the STI was found for male and female speech.

- A validation of the revised STI_r concept with a set of independent transfer conditions indicated a high reproducibility of the original relation between STI_r and CVC-word score.

7 REFERENCES

Anderson, B.W., and Kalb, J.T. (1987). "English verification of the STI method for estimating speech intelligibility of a communications channel," J. Acoust. Soc. Am. 81, 1982-1985.

ANSI (1969). Ansi S3.5-1969, American national standard methods for the calculation of the articulation index, American National Standards Institute, New York.

ANSI (1992). Ansi draft WG S3.79, American national standard methods for the calculation of the articulation index, American National Standards Institute, New York.

ASA (1960). ASA S3.2-1960, American Standard Method of Measurement of Monosyllabic Word Intelligibility, American Standards Association, New York.

Benoit, C. (1990). "An intelligibility test using semantically unpredictable sentences: towards the quantification of linguistic complexity," Speech Communication 9, 293-304.

Beranek, L.L. (1947). "The design of speech communication systems," Proc. of the Institute of Radio Engineers 35, 880-890.

Beranek, L.L. (1954). *Acoustics* (McGraw-Hill, New York).

Berry, R.W. (1971). "Speech volume measurements on telephone circuits," Proc. IEE 118(2), 335-338.

Bos, C.S.G.M., and Steeneken, H.J.M. (1991). "Phoneme confusions in distorted speech: a diagnostic study," Report IZF 1991 I-4, TNO Institute for Perception, Soesterberg, The Netherlands.

Bosman, A.J. (1989). "Speech perception by the hearing impaired," Doctoral dissertation, Rijks-Universiteit, Utrecht.

Bolt, R.H., and MacDonald, A.D. (1949). "Theory of speech masking by reverberation," J. Acoust. Soc. Am. 21, 577-580.

Brady, P.T. (1965). "A statistical basis for objective measurement of speech levels," Bell System Techn. J. 44, 1453-1486.

Brady, P.T. (**1968**). "Equivalent Peak Level: A threshold-independent speech-level measure," J. Acoust. Soc. Am. **44**, 695-699.

Broecke, M.P.R. van den, Aerts, A., Reizevoort, J., Lammens, J., and Elstrodt, M. (**1987**). "Type and token frequencies of word classes, phonemes, and phoneme pairs in Dutch," Progress report Institute of Phonetics, Utrecht, **12** (1), 1-15.

Bronkhorst, A.W., Bosman, A.J., and Smoorenburg, G.F., (**1992**). "A model for context effects in speech recognition," submitted to J. Acoust. Soc. Am.

Caroll, J.D., and Chang, J.J. (**1970**). "Analysis of individual differences in multidimensional scaling via an n-way generalization of the 'Eckhart-Young'-decomposition," Psychometrika **35**, 283-319.

Cronbach, L.J. (**1951**). "Coefficient Alpha and the internal structure of tests," Psychometrika **16**, 297-334.

Duggirala, V., Studebaker, G.A., Pavlovic, C.V., and Sherbecoe, R.L. (**1988**). "Frequency importance functions for a feature recognition test material," J. Acoust. Soc. Am. **83**, 2372-2382.

Dijkhuizen, J.C. van, Anema, P.C., and Plomp, R. (**1987**). "The effect of varying slope of the amplitude-frequency response on the masked speech-reception threshold of sentences." J. Acoust. Soc. Am. **81**, 465-469.

Dunn, H.K., and White, S.D. (**1940**). "Statistical measurements on conversational speech", J. Acoust. Soc. Am. **11**, 278-288.

Egan, J.P. (**1944**). "Articulation testing methods," OSRD report No. 3802.

Fairbanks, G. (**1958**). "Test of phonetic differentiation: The Rhyme Test," J. Acoust. Soc. Am. **30**, 596-600.

Feldmann, H. (**1960**). "Die geschichtliche Entwicklung der Hörprüfungs-methoden. Kurze Darstellung und Bibliographie von den Anfängen bis zur Gegenwart." In *Zwanglose Abhandlung aus dem Gebiet der Hals-Nasen-Ohren-Heilkunde*, edited by H. Leicher, R. Mittermaier, and G. Theissing (Georg Thieme Verlag, Stuttgart).

Fletcher, H., and Steinberg, J.C. (**1929**). Bell Sys Tech. J. **8**, 806.

Fletcher, H., and Galt, R.H. (**1950**). "The perception of speech and its relation to telephony," J. Acoust. Soc. Am. **22**, 89-151.

Fletcher, H. (**1953**). *Speech and Hearing in Communication* (D. van Nostrand, New York).

French, N.R., and Steinberg, J.C. (**1947**). "Factors governing the intelligibility of speech sounds," J. Acoust. Soc. Am. **19**, 90-119.

Goodman, D.J., and Nash, R.D. (**1984**). "Subjective quality of the same speech transmission conditions in seven different countries," IEEE Trans Comm. **30**, 642-654.

Grant, K.W., and Braida, L.D. (**1991**). "Evaluating the articulation index for auditory-visual input," J. Acoust. Soc. Am. **89**, 2952-2960.

Greenspan, S.L., Bennett, R.W., and Syrdal, A.K. (**1989**). "A study of two standard speech intelligibility measures," J. Acoust. Soc. Am. **85**, S43(A).

Hecker, M.H.L., Bismarck G. von, and Williams, C.E. (**1986**). "Automatic evaluation of time-varying communications systems," IEEE Trans. on Audio and Electroacoustics **AU-16**, 100-106.

House, A.S., Williams, C.E., Hecker, M.H.L., and Kryter, K.D. (**1965**). "Articulation testing methods: Consonantal differentiation with a closed-response set," J. Acoust Soc. Am. **37**, 158-166.

Houtgast, T., and Steeneken, H.J.M. (**1971**). "Evaluation of speech transmission channels by using artificial signals," Acustica **25**, 355-367.

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., 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.

Houtgast, T., and Steeneken, H.J.M. (**1984**). "A multi-lingual evaluation of the Rasti-method for estimating speech intelligibility in auditoria," Acustica **54**, 185-199.

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

Houtgast, T., and Verhave, J. (**1991**). "A physical approach to speech quality assessment: correlation patterns in the speech spectrogram," Proc. Eurospeech '91, Genova, 285-288.

Humes, L.E., Dirks, D.D., Bell, T., Ahlstrom, C., and Kincaid, G.E. (**1986**). "Application of the articulation index and the speech transmission index to the recognition of speech by normal-hearing and hearing-impaired listeners," J. Speech Hear. Res., 29, 447-462.

IEC-report (**1988**). "The objective rating of speech intelligibility in auditoria by the 'RASTI' method," Publication IEC 268-16.

IEEE (**1969**). "Speech quality measurements," IEEE Transactions on Audio and Electroacoustics, September, 227-246.

Keurs, M. ter, and Houtgast, T. (**1988**). "Spectral synchrony in running speech," Report IZF 1987 I-6, TNO Institute for Perception, Soesterberg, The Netherlands.

Klein, W., Plomp, R., and Pols, L.C.W. (**1970**). "Vowel spectra vowel spaces and vowel identification," J. Acoust. Soc. Am. **34**, 1217-1223.

Kryter, K.D. (**1970**). *The effects of noise on man* (Academic Press).

Kryter, K.D. (**1960**). "Speech bandwidth compression through spectrum selection," J. Acoust. Soc. Am. **32**, 547-556.

Kryter, K.D. (**1962a**). "Methods for the calculation and use of the articulation index," J. Acoust. Soc. Am. **34**, 1689-1697.

Kryter, K.D. (**1962b**). "Validation of the articulation index," J. Acoust. Soc. Am. **34**, 1698-1702.

Kryter, K.D., and Ball, J.H. (**1964**). "SCIM -- A meter for measuring the performance of speech communication systems," Techn. Doc. report No. ESD-TDR-64-674.

Licklider, J.C.R. (**1959**). "Three auditory theories," in *Psychology: A Study of Science, Vol. I,* edited by S. Koch (McGraw-Hill, New York), pp 41-144.

Licklider, J.C.R., Bisberg, A., and Schwartzlander, H. (**1959**). "An electronic device to measure the intelligibility of speech," Proc. Natl. Electronic Conf. **15**, 329-334.

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

Payne, J.A., and McManamon, P.M. (**1973**). "An objective speech quality measurement of a communication channel," OT report 73-14, Department of Commerce, Office of Telecommunications.

Pavlovic, C.V., and Studebaker, G.A. (**1984**). "An evaluation of some assumptions underlying the articulation index," J. Acoust. Soc. Am. **75**, 1606-1612.

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

Peckels, J.P., and Rossi, M. (**1973**). "Le test diagnostique par paires minimales," Revue d'Acoustique **27**, 245-262.

Peutz, V.M.A. (**1971**). "Articulation loss of consonants as a criterion for speech transmission in a room," J. Aud. Eng. Soc. **19**, 915-919.

Plomp, R., and Mimpen, A.M. (**1979**). "Improving the reliability of testing the speech reception threshold for sentences," Audiology **8**, 43-52.

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.

Pollack, I. (**1948**). "Effect of high pass and low pass filtering on the intelligibility of speech in noise," J. Acoust. Soc. Am. **20**, 259-266.

Pols, L.C.W. (**1983**). "Three-mode principal-component analysis of confusion matrices, based on the identification of Dutch consonants, under various conditions of noise and reverberation," Speech Comm. **2**, 275-293.

Pols, L.C.W. (**1991**). "Quality assessment of text-to-speech synthesis by rule." In *Advances in Speech Signal Processing*, edited by S. Furui and M. Mohan Sondhi, (Marcel Dekker, Inc.), pp. 387-415.

Quackenbush, S.R., Barnwell, T.P., and Clements, M.A. (1988). *Objective Measures of Speech Quality* (Prentice Hall, New Jersey).

Raaij, J.L. van, and Steeneken, H.J.M. (1991). "Digital simulation of speech transmission channels," Report IZF 1991-A7, TNO Institute for Perception, Soesterberg, The Netherlands.

Riemersma, J.B.J. (1974). "Development of an individual multidimensional scaling program," unpublished.

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.

SAMPA, ESPRIT-SAM Project Nr. 1541 (1987). Phonetic alphabet (1989). See Wells, J., "Computer coded phonetic transcription." J. of the International Phonetic Association **17**(2), 94-114.

Schwartzlander, H. (1959). "Intelligibility evaluation of voice communications," Electronics **29**, 88-91.

Smoorenburg, G.F. (1989). "Development of a statistical software package, STATPAC," unpublished.

Sotscheck, J. (1982). "Ein Reimtest fur Verständlichkeitsmessungen mit Deutscher Sprache als ein verbessertes Verfahren zur Bestimmung der Sprachguteübertragung," Fernmeldeingenieur **36**, 1-84.

Spiegel, M.F., Altom, M.J., Macchi, K., and Wallace, K.L. (1989). "A monosyllabic test corpus to evaluate the intelligibility of synthesized and natural speech," Proceedings ESCA Workshop, Noordwijkerhout, The Netherlands, 1.2.1-1.2.5.

Spiegel, M.F., Altom, M.J., Macchi, M.J., and Wallace, K.L. (1990). "Comprehensive assessment of the telephone intelligibility of synthesized and natural speech," Speech Communication **9**, 279-291.

Steeneken, H.J.M., and Houtgast, T. (1973). "Intelligibility in telecommunication derived from physical measurements," Proc. Symp. Intelligibilité de la Parole, Liège, 73-80.

Steeneken, H.J.M., and Houtgast, T. (**1978**) "Comparison of some methods for measuring speech levels," Report IZF 1978-22, TNO Institute for Perception, Soesterberg, The Netherlands.

Steeneken, H.J.M., and Houtgast, T. (**1979**). "Measuring ISO-intelligibility contours in auditoria," Proc. 3rd Symp of FASE on building Acoustics, Dubrovnik, 85-88.

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

Steeneken, H.J.M., and Agterhuis, E. (**1982**). "Description of STIDAS II D, Part 1, General system and program description," Report IZF 1982-29, TNO Institute for Perception, Soesterberg, The Netherlands (in Dutch).

Steeneken, H.J.M., and Houtgast, T. (**1982**). "Some applications of the Speech Transmission Index (STI) in auditoria," Acustica **51**, 229-234.

Steeneken, H.J.M. (**1982**). "Ontwikkeling en toetsing van een Nederlandstalige diagnostische rijmtest voor het testen van spraakkommunikatiekanalen," Report IZF 1982-13, TNO Institute for Perception, Soesterberg, The Netherlands (in Dutch).

Steeneken, H.J.M., and Houtgast, T. (**1983**). "The temporal envelope spectrum of speech and its significance in room acoustics," Proc. 11th International Congress on Acoustics, Paris, **Vol. 7**, 85-88.

Steeneken, H.J.M., and Houtgast, T. (**1986**). "Comparison of some methods for measuring speech levels," Report IZF 1986-20, TNO Institute for Perception, Soesterberg, The Netherlands.

Steeneken, H.J.M. (**1987a**). "Diagnostic information of subjective intelligibility tests," Proc. IEEE ICASSP, Dallas, 131-134.

Steeneken, H.J.M. (**1987b**). "Comparison among three subjective and one objective intelligibility test," Report IZF 1987-8, TNO Institute for Perception, Soesterberg, The Netherlands.

Steeneken, H.J.M., and Geurtsen, F.W.M. (**1988**). "Description of the RSG-10 noise data-base," Report IZF 1988-3, TNO Institute for Perception, Soesterberg, The Netherlands.

Steeneken, H.J.M., and Velden, J.G. van (**1989**). "Objective and diagnostic assessment of (isolated) word recognizers," Proc. IEEE ICASSP, Glasgow, 540-543.

Steeneken, H.J.M., Geurtsen, F.W.M., and Agterhuis, E. (**1990**). "Speech database for intelligibility and speech quality measurements," Report IZF 1990 A-13, TNO Institute for Perception, Soesterberg, The Netherlands.

Steeneken, H.J.M., and Houtgast, T. (**1991**). "On the mutual dependency of octave-band specific contributions to speech intelligibility," Proc Eurospeech '91, Genova, 1133-1136.

Steeneken, H.J.M., and Velden, J.G. van (**1991**). "RAMOS-recognizer assessment by means of manipulation of speech applied to connected speech recognition," Proc. Eurospeech '91, Genova, 529-532.

Steeneken, H.J.M. (**1992**). "Quality evaluation if speech processing systems," Chapter 5 in *Digital Speech Coding: Speech coding, Synthesis and Recognition*, edited by Nejat Ince, (Kluwer Norwell USA), 127-160.

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

Studebaker, G.A., and Sherbecoe, R.L. (**1991**). "Frequency-importance and transfer functions for recorded CID W-22 word lists," J. Speech Hear. Res., 34, 427-438.

Terken, J.M.B., and Collier, R. (**1989**). "Automatic synthesis of natural-sounding intonation for text-to-speech conversion in Dutch," Proc. Eurospeech '89, Paris, 26-28.

Velden, J.G. van (**1991**). "Speech level meter, version 2, SAM_SLM". Esprit-SAM, document SAM-TNO-043, TNO Institute for Perception, Soesterberg, The Netherlands.

Voiers, W.D. (**1977a**). "Diagnostic evaluation of speech intelligibility." In *Speech Intelligibility and Speaker Recognition*, Vol. 2. Benchmark papers in Acoustics, edited by M.E. Hawley (Dowden, Hutchinson, and Ross, Stroudburg), 374-384.

Voiers, W.D. (**1977b**). "Diagnostic acceptability measure for speech communication systems," Proc. IEEE ICASSP, Hartford CT, 204-207.

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.

Zwicker, E., and Feldtkeller, R., (**1967**). *Das Ohr als Nachrichtenempfanger*, (Hirzel Verlag, Stuttgart), 187-200.

8 SUMMARY

The intelligibility of speech, degraded by a speech-communication system, has been the topic of many studies in the past 70 years. Already between 1920 and 1930, Fletcher and Steinberg developed several methods to determine intelligibility. They also found a relation between the transmission quality and several physical aspects of the transmission channel. Mainly bandwidth and signal-to-noise ratio were considered. The second world war had a great impact on the evaluation of speech communication. Many papers appeared just after the war on the subjective and objective assessment of speech-communication systems (Egan, 1944; French en Steinberg, 1947; Beranek, 1947; Fletcher en Galt, 1950).

Subjective intelligibility measures

The intelligibility of sentences is an obvious measure for quantifying the quality of speech communication. However, a sentence-intelligibility score already reaches 100% at a poor-to-fair transmission quality and is therefore limited in its use. A more generally applicable measuring method is based on nonsense syllables of the CVC-word type (consonant-vowel-consonant). This type of test discriminates between transmission conditions over the full range from bad to excellent and was also used in the present study (chapter 3).

The significance of the test results for several parameters of the test conditions such as speakers, listeners, and speaker sex was analyzed. The relation between the CVC-word scores and the individual phoneme-group scores (initial consonants, vowels, and final consonants), and phoneme types (fricatives, plosives, vowel-like consonants, and vowels) was also analyzed. To support the grouping of certain phonemes, a principal-component analysis on the phoneme scores for a wide variety of conditions, as well as multi-dimensional scaling on the phoneme-confusion matrices, were used. It was found that at the phoneme level, four groups can be identified, each with a fairly similar distribution of responses for various transmission conditions. The differentiation between the phoneme-group responses can be used as a diagnostic tool and to improve predictive intelligibility measurements.

We obtained, for a number of reference conditions, the optimal relation between sentence intelligibility and phoneme-group scores. A reliable prediction of sentence intelligibility was obtained by a weighted combination of phoneme-group scores.

Objective intelligibility measures

Rather than measuring intelligibility of degraded speech with subjective intelligibility measures, the effect on intelligibility of a transmission channel can also be predicted by considering the physical properties of such a channel. The accuracy of this prediction as obtained by two existing models (AI and STI) could be improved.

The Articulation Index (AI; French and Steinberg, 1947) and Speech Transmission Index (STI; Steeneken and Houtgast, 1980) are based on a linear summation of the contributions of individual frequency bands to intelligibility. There is evidence that this assumption that frequency bands are independent from each other is not correct for conditions with gaps or selective masking in the frequency domain.

We designed an experiment in which the contribution of individual frequency bands, and the question of mutual dependency, could be studied. Evaluation of the observed scores and the corresponding physical specifications resulted in a *revised model*, which accounts for the mutual dependency between adjacent octave bands by the introduction of a so-called redundancy correction factor. The weighting factors proved to be identical for male and female speech and for various signal-to-noise ratios. This robust model for prediction of intelligibility gives a significant improvement of the prediction accuracy in comparison to the original model (chapter 2).

However, the optimal frequency weighting and redundancy correction also depends on the type of speech considered. Therefore, the four groups of phonemes with a similar response at various transmission conditions (fricatives, plosives, vowel-like consonants, and vowels) were used to specify the frequency weighting for various types of speech. For each group the optimal set of frequency-weighting factors and the optimal redundancy-correction factor were determined separately (chapter 4).

The predicted phoneme-group scores can be used to predict the word score of several types of nonsense words. This is performed in two steps. For instance, to calculate the CVC-word score, the scores of the initial consonant, the vowel, and the final consonant are calculated first by a summation of the phoneme-group scores (weighted according to the frequency of occurrence of the phonemes). The word score is obtained by the product of the (normalized) consonant and vowel scores.

The revised STI_r model for the prediction of intelligibility was validated for a set of independent transmission channels. It was concluded that the present STI_r model provides a reliable measure, applicable to a wide range of transmission conditions (chapter 5).

HET METEN EN VOORSPELLEN VAN DE SPRAAKVERSTAANBAARHEID

9 NEDERLANDSE SAMENVATTING

Het meten en voorspellen van de verstaanbaarheid van spraaksignalen die zijn overgedragen via een communicatiekanaal houdt onderzoekers al meer dan 70 jaar bezig. In de twintiger jaren ontwikkelden Fletcher en Steinberg verscheidene methoden om de transmissiekwaliteit van een telefoonsysteem te bepalen. Ook probeerden zij de transmissiekwaliteit te relateren aan bepaalde fysische eigenschappen van het transmissiekanaal zoals bandbreedte en signaal-ruisverhouding. De tweede wereldoorlog heeft veel invloed gehad op de evaluatie van communicatiesystemen. In de tweede helft van de veertiger jaren verschenen dan ook veel publikaties die zowel subjectieve (met sprekers en luisteraars) als objectieve (op fysische metingen gebaseerde) meetmethoden voor de spraaktransmissiekwaliteit beschrijven (Egan, 1944; French en Steinberg, 1947; Beranek, 1947; Fletcher en Galt, 1950).

Subjectieve verstaanbaarheidsmaten

Het meten van de spraakverstaanbaarheid met zinnen is een voor de hand liggende methode om communicatiekanalen te evalueren. Een zinsverstaanbaarheid van 100% wordt echter reeds bereikt bij een slechte tot matige transmissiekwaliteit. Mede om deze reden zijn dan ook gevoeliger meetmethoden, meestal op basis van nonsenswoorden, ontwikkeld. De éénlettergrepige CVC-woorden (Consonant-Vocaal-Consonant) zijn hiervan een voorbeeld. Ook voor de in dit proefschrift beschreven studie is van CVC-tests gebruik gemaakt en werd de betrouwbaarheid van de test, bijvoorbeeld met betrekking tot de verschillen tussen mannelijke en vrouwelijk sprekers, het benodigd aantal sprekers en luisteraars, en scoringsmethode, getoetst. Tevens werd de individuele foneemscore bepaald en een groepsindeling van fonemen gemaakt (fricatieven, plosieven, klinkerachtige medeklinkers en klinkers) die eenzelfde mate van degradatie in de score ondervinden bij verschillende soorten vervorming. Deze indeling werd verkregen op basis van groepering van de individuele foneemscores door een principale componentenanalyse en op basis van verwisselingen tussen fonemen met multidimensionele analyses.

Voor een aantal referentiecondities werd de optimale relatie tussen zinsverstaanbaarheid en foneemgroepscores bepaald. Tevens blijkt op grond van de foneemgroepen een betrouwbare voorspelling van de zinsverstaanbaarheid mogelijk te zijn voor eenvoudige (Nederlandse) zinnen (hoofdstuk 3).

Objectieve verstaanbaarheid

In hoofdstuk 1 van dit proefschrift wordt vastgesteld dat het principe waarop de twee belangrijkste bestaande objectieve meetmethoden voor spraakverstaanbaarheid zijn gebaseerd nader onderzocht dient te worden. Deze methoden, de Articulatie Index (AI; French en Steinberg, 1947; Kryter, 1963) en de Spraak Transmissie Index (STI; Houtgast en Steeneken, 1973; Steeneken en Houtgast, 1980), zijn gebaseerd op de individuele bijdrage aan de verstaanbaarheid van bepaalde frequentiegebieden. De individuele bijdragen worden, na normalisatie, gesommeerd en leveren respectievelijk de AI of de STI. Dit concept, waarbij de bijdragen van de frequentiegebieden als onderling onafhankelijk worden beschouwd, blijkt niet juist te zijn voor condities waarbij een discontinuïteit in de frequentie-overdracht optreedt of selectieve maskering aanwezig is.

Het in hoofdstuk 2 beschreven experiment werd ontworpen om de relatieve bijdrage en de wederzijdse afhankelijkheid van frequentiegebieden in het spraaksignaal, ten aanzien van de spraakverstaanbaarheid, nader te onderzoeken. Er werd vastgesteld dat er enige mate van redundantie tussen de toegepaste frequentiegebieden (octaafbanden) aanwezig is. De gevonden weegfactoren en de redundantiecorrectie zijn vrijwel identiek voor mannelijke en vrouwelijke spraaksignalen en voor spraaksignalen gemaskeerd door stoorsignalen bij verschillende stoorniveaus. Dit levert een robuust model voor het voorspellen van de spraakverstaanbaarheid en een aanzienlijke verbetering van de voorspelkracht ten opzichte van het oorspronkelijke model. De weegfactoren blijken echter verschillend te zijn voor verschillende typen spraak. Er is daarom een analyse uitgevoerd om een optimale indeling van spraakklanken (fonemen) te verkrijgen. Fonemen die eenzelfde afname van de herkenning opleveren bij verschillende typen storing en vervorming worden als gelijkwaardig beschouwd. Dit leverde een indeling in vier foneemgroepen: fricatieven, plosieven, klinkerachtige medeklinkers en klinkers. Voor elke groep zijn verschillende frequentie-weegfactoren en redundantiecorrecties gevonden (hoofdstuk 4).

Het voorspellen van de individuele foneemgroepverstaanbaarheid maakt het mogelijk ook de woordscore voor nonsenswoorden te voorspellen. Dit geschiedt in twee stappen. Bijvoorbeeld voor CVC-woorden worden eerst de beginmedeklinker-, klinker- en eindmedeklinker-scores bepaald door een gewogen sommatie van de foneemgroepscores. De woordscore wordt verkregen uit het produkt van deze (genormaliseerde) medeklinker- en klinkerscores.

Als verificatie van de herziene methode voor het objectief voorspellen van de spraakverstaanbaarheid werd de methode gevalideerd met onafhankelijke referentiecondities. Er werd vastgesteld dat het huidige STI-model een betrouwbare voorspelling geeft van de spraakverstaanbaarheid, en toegepast kan worden bij de meeste typen transmissiekanalen (hoofdstuk 5).

Blank 131/132

APPENDICES

A1 Measurement and calculation of the STI

The STI is an objective measure, based on the weighted contribution of the *effective* signal-to-noise ratio of a number of frequency bands within the frequency range of speech signals (see also section 2.1). This signal-to-noise ratio is called effective because it may be determined by several factors. The most obvious one is background noise, which has a direct contribution to the signal-to-noise ratio. However, distortions in the time domain and nonlinearities are also considered as noise. This is due to the specific design of the test signal. In Fig. A1.1 an illustration is given of the estimation of the signal-to-noise ratio within each frequency band. The test signal consists of a noise signal with a frequency spectrum equal to the long-term frequency spectrum of the speech signal. Each octave-band is modulated with a periodic signal in such a way that the *intensity*[1] is modulated sinusoidally. This is indicated in Fig. A1.1 for the octave band with centre frequency 250 Hz. The modulation index (m) in this example is m = 1.

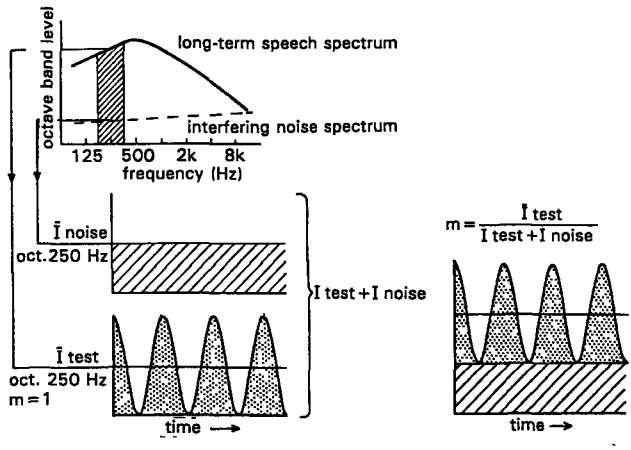

Fig. A1.1 Illustration of the effect of interfering noise on the modulation index m of a test signal.

[1] The summation of uncorrelated signals (echoes, reverberation, and masking noises) is based on the energy content, therefore the intensity envelope is considered. For instance, the summation of two sinusoidally modulated signals (same modulation frequency) with uncorrelated carriers will consist of a signal with a sinusoidal envelope modulation being the vector summation of the sinusoidal envelope of the two primary signals. This statement is only valid for intensity modulations, and not for amplitude modulations.

Noise may be added to the test signal and the resulting envelope is obtained by addition of the intensity of both signal envelopes. Hence, the resulting envelope of this example is defined by a steady noise envelope (of a stationary noise signal) and the test-signal envelope. The resulting modulation index, being the test-signal intensity divided by the total intensity (test signal and noise), is directly related to the signal-to-noise ratio (SNR) according to:

$$SNR = 10 \log \frac{m}{1 - m} \text{ dB} \qquad \text{(A1.1)}$$

As described in section 5.2 the envelope function of a speech signal contains a range of fluctuation frequencies, representing the succession of speech events from the shortest speech items up to words and sentences. Due to distortion in the time domain (reverberation, echoes, and automatic gain control) part of this fluctuation pattern may be affected, in this way reducing intelligibility. This is modelled in the STI procedure by determining the modulation transfer function for the range of relevant fluctuation frequencies present in natural speech signals. As described before (Steeneken and Houtgast, 1980) a relevant choice for these modulation frequencies is the range from 0.63 Hz up to 12.5 Hz with 1/3-octave steps, hence 14 modulation frequencies are considered. This results in a measuring procedure according to Fig. A1.2 where the modulation transfer index, m, for each octave band (125 Hz - 8 kHz) and each modulation frequency (0.63 - 12.5 Hz) is determined separately. The figure gives the measuring set-up for one octave band. A noise signal with the required frequency spectrum (normally the long-term speech spectrum) is amplitude modulated by a signal $\sqrt{\{1 + \cos(2\pi.f_m.t)\}}$ which results in a sinusoidal intensity modulation $I\{1 + \cos(2\pi.f_m.t)\}$. This modulation function can be obtained digitally and can be generated by a computer. At the receiving side, octave-band filtering and (intensity) envelope detection is applied. From this resulting envelope function the modulation index, due to the original envelope modulation, is determined by a Fourier analysis. This procedure is repeated for each cell of the matrix given in Fig. A1.2. It should be noted that the block diagram of Fig. A1.2 represents only one channel corresponding with one octave band. The original set-up consists of a set of separate channels for all octave bands considered.

With the test signal as described above, distortions such as band-pass limiting, and noise masking, as well as distortion in the time domain can be accounted for. Nonlinear distortions, however, have to be modelled additionally. If a speech signal is passed through a system with a nonlinear transfer (such as obtained with peak clipping or quantization), harmonic distortion components and inter-modulation components will be produced.

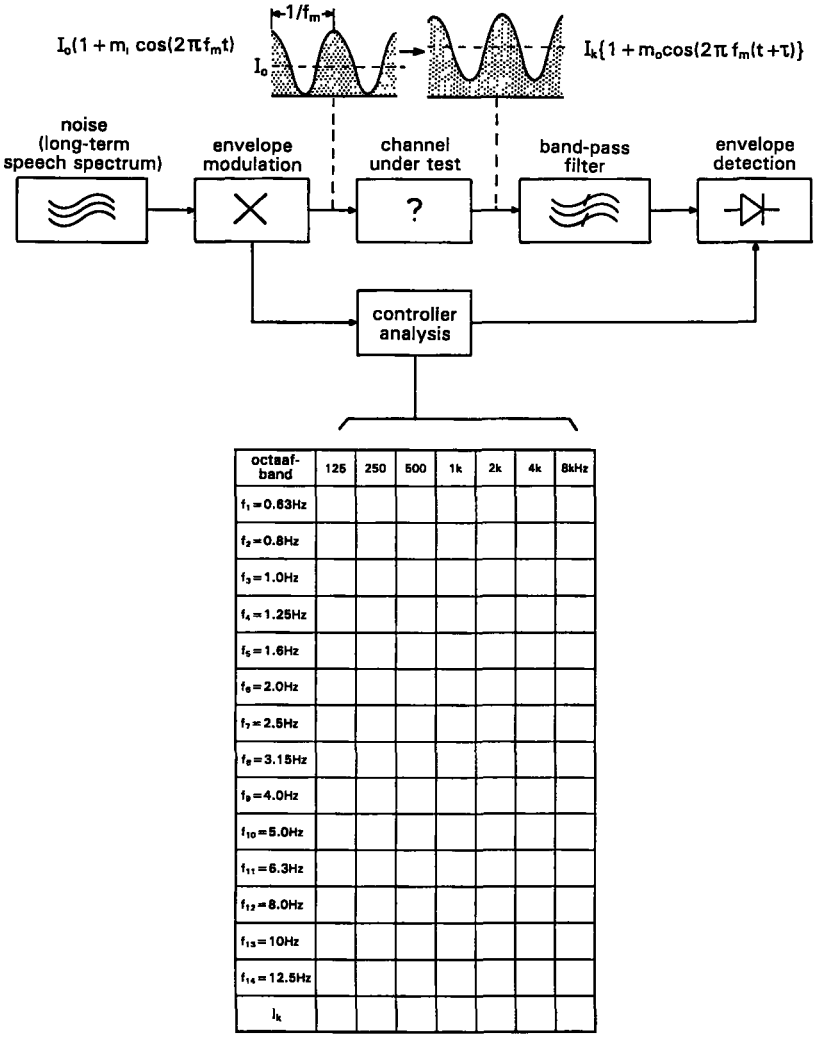

Fig. A1.2 General block diagram of the measuring set-up. The modulation index reduction at the output (m_i/m_o) is determined for all cells of the matrix (7 octave bands and 14 modulation frequencies). Also the octave levels are obtained, for calculation of the auditory spread of masking.

For this reason the test signal should not be modulated with one and the same modulation frequency for all octave bands simultaneously. Otherwise, nonlinear distortion components cannot be discriminated from the modulated test signal in the frequency band considered. Therefore, in the case of nonlinear distortion, all frequency bands, except the one under test, are modulated with uncorrelated signals in order to introduce distortion components of which the fluctuations are not

correlated with the test-signal envelope in the octave band under test. Such distortion components are then considered as noise (they add to the noise in the octave band under test) and reduce the effective signal-to-noise ratio in a similar way as would occur with speech signals. The relative levels of the test signal in the octave bands with the uncorrelated (speech-like) envelope were adjusted for optimal prediction of intelligibility in the nonlinear transfer condition. The consequence of this procedure is a successive measurement of the seven octave bands rather than a simultaneous measurement as can be applied with channels with a linear transfer.

Just as the masking introduced by the noise in the transmission channel, an additional auditory masking phenomenon[2] has to be taken into account. This effect is modelled as an imaginary masking noise which leads to a decrease of the effective signal-to-noise ratio and a corresponding reduction of the modulation transfer index m. For this purpose not only the modulation transfer has to be determined but also the signal levels in the frequency bands have to be considered. In Fig. A1.3 the effect of the masking for frequency band (k-1) upon frequency band k is indicated. The masking effect, as modelled in the STI approach, does not depend on the frequency band considered nor on the level. The slope of masking decreases with 35 dB/oct, corresponding to an auditory masking factor (amf) of the intensity of the primary masking signal of amf = 0.000316. As the masking effect of only the lower adjacent frequency band is considered, the intensity of the masking signal becomes:

$$I_{am,k} = I_{k-1} * amf \qquad \textbf{(A1.2)}$$

where $I_{am,k}$ represents the intensity level of the auditory masking signal for octave band k, and I_{k-1} represents the signal intensity of octave band (k-1).

The auditory spread of masking is accounted for by a reduction in the modulation index given by:

$$m'_{k,f} = m_{k,f} \frac{I_k}{I_k + I_{am,k}} \qquad \textbf{(A1.3)}$$

where $m_{k,f}$ represents the modulation index for octave band k and modulation frequency f, m' the corrected modulation index.

---

[2] Auditory spread of masking is the effect, introduced by the hearing organ, that the perception of a tone or narrow-band signal may be reduced by a strong masker in a lower frequency range. The amount of masking depends on the level difference between masker and masked signal, on the absolute level of the masker, and on their frequency distance. A detailed description is given by Zwicker and Feldtkeller (1967).

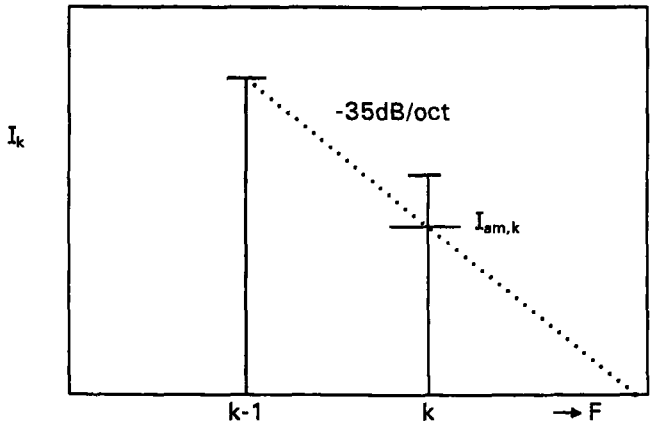

Fig. A1.3 Auditory masking of octave band k-1 upon the next higher octave band k. The slope of the masking effect versus frequency band corresponds to -35 dB/oct. This is similar to an auditory masking factor of amf = 0.000316.

The effective signal-to-noise ratio for octave band k and modulation frequency f then becomes:

$$SNR_{k,f} = 10 \log \frac{m'_{k,f}}{1 - m'_{k,f}} \text{ dB} \qquad \text{(A1.4)}$$

According to the STI concept a signal-to-noise ratio between -15 dB and 15 dB is related to a contribution to intelligibility of between 0 and 1. Therefore, the effective signal-to-noise ratio is converted to transmission index ($TI_{k,f}$), specific for octave band (k) and modulation frequency (f), by the equation:

$$TI_{k,f} = \frac{SNR_{k,f} + \text{shift}}{\text{range}}, \quad \text{where } 0 \le TI_{k,f} \le 1.0 . \qquad \text{(A1.5)}$$

The shift equals 15 dB and the range equals 30 dB. In this way a relation between the effective signal-to-noise ratio and the TI is obtained as shown in Fig. A1.4.

All 14 transmission indices, related to modulation frequencies between 0.63 and 12.5 Hz, are obtained for each octave band. The mean of these linear weighted indices results in the modulation transfer index (MTI_k) and is specific for the contribution of octave band k. The MTI_k is given by:

$$MTI_k = \frac{1}{14} \sum_{f=1}^{14} TI_{k,f} . \qquad \text{(A1.6)}$$

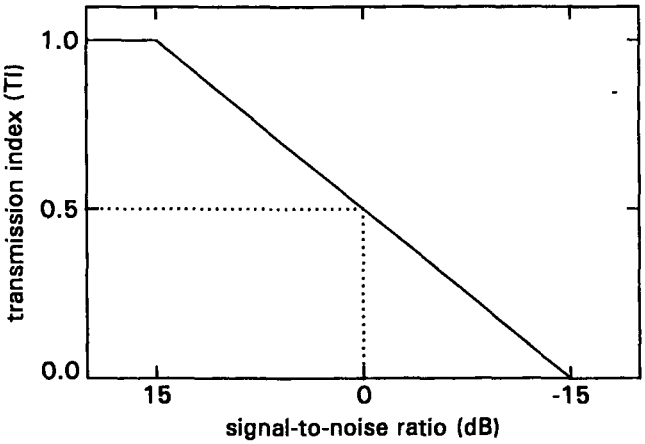

Fig. A1.4 Relation between the effective signal-to-noise ratio and the transmission index for a shift of 15 dB and a range of 30 dB.

Finally, according to the original concept, the STI is obtained by a weighted summation of the modulation transfer indices for all seven octave bands. This is given by:

$$STI = \sum_{k=1}^{7} (\alpha_k * MTI_k) , \quad \text{with} \quad \sum_{k=1}^{7} \alpha_k = 1.0 . \qquad \text{(A1.7)}$$

Where α_k represents the octave-weighting factor. According to the results of the study described in chapter 2, a better prediction is obtained by the introduction of a so-called redundancy factor β_k, which is related to the contribution of adjacent frequency bands. The optimal weighting factors and redundancy factors for male and female speech and different groups of phonemes are described in section 4.3.

Some simplifications of the procedure described above were made in order to decrease the measuring time, but these simplifications restrict the range of applicability. The measurement of a complete matrix of 96 m-values according to Fig. A1.2 and a measuring time for each m-value of 5s results in a total measuring time of 8 minutes.

A reduction of the 14 modulation frequencies to only three modulation frequencies results in a total measuring time of less than 2 minutes, but as a consequence no complete modulation transfer is obtained. This means that distortions in the time domain are not accounted for correctly. Therefore, this method is

normally used only for communication channels with no degradation due to echoes or reverberation.

Also, the number of octave bands considered may be reduced. This is the case with the RASTI method, where only the contributions of the modulation transfer for the octave bands with centre frequencies 500 Hz and 2 kHz are considered. This can be used as a screening approach for most acoustical applications.

Another simplification can be applied to the test signal when the uncorrelated modulations, required for the correct interpretation of nonlinear distortions, are omitted. This opens the possibility of applying a simultaneous modulation and parallel processing of all frequency bands, thus decreasing the measuring time by a factor of seven. This procedure is used by the STITEL measuring method which requires a measuring time of 15s.

All these simplifications have led to different measuring schemes which can be applied for specific applications. These corresponding measuring programs all make use of specific hardware and are therefore not generally available. The results of the study described in this thesis, and the conversion of the measuring procedure to a standard signal processor controlled by a personal computer, will increase applicability.

A2 Description of the measuring conditions of the experiment on band-pass limiting and noise masking

Description of the measuring conditions of the experiment on band-pass limiting and noise for MALE speech. The table gives the condition number, which octave bands were included (0 vs. 1) in the frequency transfer, and the CVC-word score based on 16 (male) speaker-listener pairs for three signal-to-noise ratios.

Rippled envelope

No	Octave-band centre frequency 125	250	500	1k	2k	4k	8kHz	CVC-word score at signal-to-noise ratio 15	7.5	0 dB
1	1	1	1	1	0	0	0	32.5	22.9	10.7
2	0	0	0	0	1	1	1	63.6	50.7	33.7
3	1	1	0	0	0	1	1	36.2	25.2	14.8
4	0	0	1	1	1	0	0	69.8	61.6	26.7
5	1	1	0	0	1	1	0	60.0	49.6	26.6
6	0	0	1	1	0	0	1	61.6	52.5	25.1
7	1	0	1	0	1	0	1	79.5	66.4	40.6
8	0	1	0	1	0	1	0	65.1	53.9	27.9

Adjacent triplets

No	Octave-band centre frequency 125	250	500	1k	2k	4k	8kHz	CVC-word score signal-to-noise ratio 15	7.5	0 dB
9	1	1	1	0	0	0	0	10.8	8.2	2.9
10	0	1	1	1	0	0	0	36.8	24.9	12.1
as 4	0	0	1	1	1	0	0			
11	0	0	0	1	1	1	0	78.6	56.1	31.9
as 2	0	0	0	0	1	1	1			

Isolated triplets (selected from a total of 35 possible triplets with the restriction of no adjacent bands and each band being included not more than three times in the list)

No	Octave-band centre frequency 125	250	500	1k	2k	4k	8kHz	CVC-word score signal-to-noise ratio 15	7.5	0 dB
12	1	0	1	0	1	0	0	65.3	44.9	17.8
13	1	0	1	0	0	1	0	58.6	39.1	14.6
14	1	0	0	1	0	1	0	53.6	43.0	13.1
as 8	0	1	0	1	0	1	0			
15	0	1	0	1	0	0	1	50.2	41.1	25.0
16	0	1	0	0	1	0	1	56.7	45.5	31.1
17	0	0	1	0	1	0	1	72.9	63.8	44.1

Contiguous bands from 4 octaves to 7 octaves

No	Octave-band centre frequency							CVC-word score signal-to-noise ratio		
	125	250	500	1k	2k	4k	8kHz	15	7.5	0 dB
18	0	1	1	1	1	0	0	79.8	66.7	35.0
19	0	0	1	1	1	1	0	83.6	72.1	38.5
20	0	0	0	1	1	1	1	77.9	70.0	44.4
21	1	1	1	1	1	0	0	78.3	65.3	34.3
22	0	1	1	1	1	1	0	85.0	75.2	48.9
23	0	0	1	1	1	1	1	83.1	76.6	54.7
24	1	1	1	1	1	1	0	88.5	79.7	54.0
25	0	1	1	1	1	1	1	91.5	82.8	61.9
26	1	1	1	1	1	1	1	89.8	82.5	66.5

Description of the measuring conditions of the experiment on band-pass limiting and noise for FEMALE speech. The table gives the condition number, the octave-bands included (1) in the frequency transfer, and the CVC-word score based on 16 (female) speaker-listener pairs for three signal-to-noise ratios.

Rippled envelope

No	Octave-band centre frequency							CVC-word score signal-to-noise ratio		
	125	250	500	1k	2k	4k	8kHz	15	7.5	0 dB
1	-	1	1	1	0	0	0	22.4	19.9	5.9
2	-	0	0	0	1	1	1	69.6	47.2	25.6
3	-	1	0	0	0	1	1	40.7	21.7	8.7
4	-	0	1	1	1	0	0	65.3	47.1	16.7
5	-	1	0	0	1	1	0	63.2	50.6	17.2
6	-	0	1	1	0	0	1	50.5	43.3	18.8
7	-	0	1	0	1	0	1	69.2	58.0	29.4
8	-	1	0	1	0	1	0	66.1	39.7	20.2

Adjacent triplets

No	Octave-band centre frequency							CVC-word score signal-to-noise ratio		
	125	250	500	1k	2k	4k	8kHz	15	7.5	0 dB
as 1	-	1	1	1	0	0	0			
as 4	-	0	1	1	1	0	0			
9	-	0	0	1	1	1	0	75.2	61.6	24.4
as 2	-	0	0	0	1	1	1			

Isolated triplets (selected from a total of 35 possible triplets with the restriction of no adjacent bands and each band being included no more than three times in the list)

No	Octave-band centre frequency							CVC-word score signal-to-noise ratio		
	125	250	500	1k	2k	4k	8kHz	15	7.5	0 dB
as 8	-	1	0	1	0	1	0	40.9	29.9	14.1
10	-	1	0	1	0	0	1			
11	-	1	0	0	1	0	1	56.5	40.1	17.2
as 7	-	0	1	0	1	0	1			

Contiguous band from 4 octaves to 7 octaves

No	Octave-band centre frequency							CVC-word score signal-to-noise ratio		
	125	250	500	1k	2k	4k	8kHz	15	7.5	0 dB
12	-	1	1	1	1	0	0	67.6	59.9	24.0
13	-	0	1	1	1	1	0	81.9	68.3	31.9
14	-	0	0	1	1	1	1	81.1	68.9	36.0
15	-	1	1	1	1	1	0	83.7	72.8	47.7
16	-	0	1	1	1	1	1	86.9	73.3	43.4
17	-	1	1	1	1	1	1	86.4	80.1	50.2

A3 Description of the measuring conditions of the experiment on communication channels

This appendix describes the physical composition of the 68 transmission channels used for the experiments on communication channels (chapter 4). The transmission channels are divided into five groups; each group is representative of a typical distortion: band-pass limiting, nonlinear distortion, automatic gain-control, echoes, and digital wave-form coding. For each group, combinations with a masking noise at various signal-to-noise ratios were made.
Specification and coding of the individual distortion conditions:

- band-pass limiting:
  - BP1 frequency transfer 50 Hz - 10.5 kHz (-3 dB),
  - BP2 frequency transfer 300 Hz - 3400 Hz (-3 dB).

- noises:
  - N1 white noise, equal energy per unit of frequency (Hz),
  - N2 pink noise, equal energy per relative unit of frequency (octave),
  - N3 very low-frequency noise, white noise with -12 dB/oct shaping from 250 Hz,
  - N4 speech noise, noise frequency spectrum according to the long-term speech spectrum,
  - SNR RMS-A signal-to-noise ratio in dB, based on spectral weighting according to the A-curve, speech threshold -14 dB *re* RMS (see appendix A7).

- nonlinear distortion:
  - PC1 peak clipping, clip level -24 dB below the 1% speech peak level.
  - CC1 centre clipping, clip level -24 dB below the 1% speech peak level.
  - CC2 centre clipping, clip level -21 dB below the 1% speech peak level.

- automatic gain-control:
  - AGC1 attack time 50 ms, decay time 1 s, threshold -30 dB,
  - AGC2 attack time 10 ms, decay time 30 ms, threshold -30 dB.

- echoes:
  - E1 echo delay 50 ms, *re* level -3 dB,
  - E2 echo delay 100 ms, *re* level -3 dB,
  - E3 echo delay 200 ms, *re* level -3 dB.

- digital waveform coding:
  - PCM1 pulse-code modulation word length 8 bit, bit rate 64 kBit/s,
  - DM1 fixed-step delta modulation, bit rate 32 kBit/s,
  - DM2 variable-step delta modulation, bit rate 32 kBit/s,
  - BER bit error rate % of random errors between coder/decoder.

No	cond.	filter	noise	SNR dB
Band-pass limiting and noise				
1		BP1	--	∞
2		BP1	N1	0
3		BP1	N1	-8
4		BP1	N2	0
5		BP1	N2	-8
6		BP1	N3	3
7		BP1	N3	-3
8		BP1	N4	3
9		BP1	N4	-3
10		BP2	--	∞
11		BP2	N1	0
12		BP2	N1	-8
13		BP2	N2	0
14		BP2	N2	-8
15		BP2	N3	6
16		BP2	N3	0
17		BP2	N4	6
18		BP2	N4	0
Nonlinear distortion				
19	PC1	BP1	--	∞
20	PC1	BP1	N1	12
21	PC1	BP1	N1	6
22	PC1	BP1	N4	9
23	PC1	BP1	N4	3
24	PC1	BP2	--	∞
25	PC1	BP2	N1	12
26	PC1	BP2	N1	6
27	PC1	BP2	N4	12
28	PC1	BP2	N4	6
29	CC1	BP1	N4	∞
30	CC1	BP1	N4	6
31	CC1	BP1	N4	0
32	CC2	BP1	N4	∞

No	cond.	filter	noise	SNR dB
Automatic gain control				
33	AGC1	BP1	--	∞
34	AGC1	BP1	N4	<15
35	AGC1	BP1	N4	<9
36	AGC1	BP2	--	∞
37	AGC1	BP2	N4	<18
38	AGC1	BP2	N4	<12
39	AGC2	BP1	--	∞
40	AGC2	BP1	N4	<18
41	AGC2	BP1	N4	<12
42	AGC2	BP2	--	∞
43	AGC2	BP2	N4	<18
44	AGC2	BP2	N4	<12
Echoes				
45	E1	BP1	--	∞
46	E1	BP1	N4	12
47	E1	BP1	N4	6
48	E2	BP1	--	∞
49	E2	BP1	N4	12
50	E2	BP1	N4	6
51	E3	BP1	--	∞
52	E3	BP1	N4	12
53	E3	BP1	N4	6
54	E3	BP2	--	∞
55	E3	BP2	N4	15
56	E3	BP2	N4	9

No	cond.	filter	noise	SNR dB	BER
Digital waveform coding					
57	PCM1	BP2	--	∞	0
58	PCM1	BP2	N4	12	0
59	PCM1	BP2	--	∞	0.005
60	PCM1	BP2	--	∞	0.02
61	DM1	BP2	--	∞	0
62	DM1	BP2	N4	9	0
63	DM1	BP2	--	∞	0.06
64	DM1	BP2	--	∞	0.12
65	DM2	BP2	--	∞	0
66	DM2	BP2	N4	9	0
67	DM2	BP2	--	∞	0.06
68	DM2	BP2	--	∞	0.12

A4 Intelligibility scores for the conditions of the experiments on band-pass limiting and on communication channels

Table A4.1 Mean and median values of word scores and phoneme scores of two groups of two MALE speakers each and four listeners (16 speaker-listener pairs) for each of the conditions of the experiment on band-pass limiting and noise. The condition numbers are according the description given in appendix A2.

No	CVC	CVC_{med}	C_i	V	C_f	$C_{i,fr}$	$C_{i,pl}$	$C_{i,vl}$	$C_{f,fr}$	$C_{f,pl}$	$C_{f,vl}$
1	32.5	31.4	64.7	73.5	58.8	37.5	70.0	79.2	41.7	52.8	72.7
2	22.9	23.5	51.0	74.5	53.9	22.9	40.8	76.2	33.0	42.4	68.5
3	10.7	10.8	34.3	60.8	40.2	15.4	31.3	54.2	18.1	22.2	58.1
4	63.6	65.7	89.2	77.5	90.2	86.7	82.9	93.5	96.5	100.0	80.6
5	50.7	47.1	80.4	69.6	86.3	81.3	78.7	78.9	99.3	92.7	75.2
6	33.7	34.3	62.7	61.8	70.6	82.9	56.7	51.8	96.5	76.0	52.9
7	36.2	35.3	88.2	42.2	85.3	86.2	96.3	83.9	97.6	96.2	71.3
8	25.2	24.5	80.4	39.2	78.4	83.3	88.9	70.2	95.1	91.7	58.3
9	14.8	14.7	59.8	33.3	62.7	78.3	62.1	40.5	92.4	73.3	39.2
10	69.7	67.6	82.4	94.1	91.2	72.1	84.2	89.6	94.8	91.7	86.3
11	61.6	60.8	71.6	96.1	86.3	66.3	75.8	74.1	90.3	76.4	89.0
12	26.7	22.5	40.2	85.3	52.9	37.9	52.1	46.4	57.3	36.8	58.8
13	60.0	59.8	89.2	75.5	90.2	80.0	93.3	90.2	95.8	97.2	81.5
14	49.6	48.0	80.4	69.6	83.3	80.4	82.9	77.4	94.4	88.9	71.3
15	26.6	26.5	54.9	53.9	64.7	64.2	63.3	46.4	78.1	65.6	56.0
16	61.6	60.8	83.3	82.4	87.3	80.8	77.9	87.2	91.7	93.8	77.7
17	52.5	52.0	76.5	80.4	80.4	85.3	75.0	70.2	93.1	92.0	68.3
18	25.1	23.5	48.0	64.7	64.7	65.0	54.6	37.5	89.6	64.9	50.1
19	79.5	78.4	91.2	93.1	93.1	84.2	93.8	93.2	97.6	98.6	89.9
20	66.4	63.7	86.3	86.3	84.3	83.3	90.0	84.8	95.8	89.9	72.7
21	40.6	37.3	62.7	81.4	72.5	73.3	61.7	57.4	88.5	73.6	59.4
22	65.1	63.7	87.3	82.4	88.2	82.9	92.1	87.5	93.8	95.8	77.3
23	53.9	56.9	78.4	77.5	85.3	74.6	83.8	77.7	92.4	86.1	74.2
24	27.9	29.4	54.9	65.7	67.6	67.9	47.5	45.5	80.9	66.7	54.8
25	10.8	11.8	44.1	47.1	50.0	22.1	34.6	70.5	22.9	32.6	63.7
26	8.2	7.8	36.3	46.1	46.1	15.8	22.9	60.4	21.5	23.6	58.5
27	2.9	3.9	24.5	35.3	35.3	10.2	16.7	38.7	16.7	24.0	44.4
28	36.8	34.3	64.7	76.5	62.7	43.7	65.8	80.7	53.1	62.8	76.0
29	24.9	23.5	52.9	72.5	56.9	37.1	45.4	67.6	36.5	50.0	72.7
30	12.1	11.8	31.4	61.8	43.1	16.3	30.0	46.4	24.3	27.1	57.5
31	78.6	77.5	89.2	92.2	95.1	84.6	86.7	93.8	96.9	98.6	91.0
32	56.1	57.8	74.5	86.3	86.3	78.3	67.9	76.8	93.8	87.5	79.0
33	31.9	31.4	54.9	76.5	68.6	65.8	50.8	50.3	71.2	62.5	58.5
34	65.3	66.7	83.3	92.2	86.3	70.2	80.0	90.8	84.7	81.6	86.3
35	44.9	44.1	62.7	90.2	72.5	60.8	60.8	69.6	57.6	68.8	78.1
36	17.8	16.7	40.2	76.5	48.0	34.6	43.3	44.9	50.3	26.7	51.9
37	58.6	58.8	87.3	75.5	82.4	84.6	87.1	89.3	91.0	77.1	77.0
38	39.1	38.2	66.7	68.6	77.5	72.9	65.8	66.1	84.7	72.2	72.1
39	14.6	13.7	42.2	59.8	54.9	55.4	35.0	33.6	68.8	50.0	40.6

No	CVC	CVC_{med}	C_i	V	C_f	$C_{i,fr}$	$C_{i,pl}$	$C_{i,vl}$	$C_{f,fr}$	$C_{f,pl}$	$C_{f,vl}$
40	53.6	55.9	82.4	71.6	87.3	79.6	85.0	78.9	95.5	88.5	72.9
41	43.0	40.2	70.6	73.5	79.4	78.3	68.8	63.7	88.9	86.8	65.2
42	13.1	10.8	41.2	52.0	53.9	54.2	40.4	26.5	65.6	44.4	43.5
43	50.2	51.0	88.2	65.7	84.3	84.2	85.0	85.1	91.3	89.9	76.3
44	41.1	41.2	72.5	65.7	83.3	77.1	78.8	64.6	94.8	84.7	72.1
45	25.0	25.5	63.7	51.0	63.7	77.9	62.5	47.9	89.2	69.1	50.2
46	56.7	55.9	89.2	70.6	89.2	83.8	93.8	88.4	95.8	94.8	79.6
47	45.5	47.1	80.4	61.8	82.4	80.0	79.2	79.8	94.4	92.4	69.4
48	31.1	32.4	59.8	58.8	72.5	76.3	61.3	45.2	90.3	69.1	60.0
49	72.9	71.6	89.2	95.1	86.3	86.7	83.7	90.8	88.9	95.5	84.2
50	63.8	62.7	79.4	91.2	82.4	82.1	81.3	80.1	88.2	94.1	79.2
51	44.1	40.2	60.8	83.3	74.5	77.1	63.8	61.0	86.8	73.6	60.9
52	79.8	77.5	90.2	96.1	90.2	78.3	92.9	95.5	89.2	99.7	85.4
53	66.7	68.6	82.4	94.1	86.3	73.3	78.3	90.8	83.0	80.9	85.2
54	35.0	32.4	52.0	89.2	72.5	40.0	52.9	61.6	66.7	60.8	74.0
55	83.6	83.3	90.2	96.1	96.1	87.5	88.3	95.5	99.3	96.5	94.0
56	72.1	71.6	82.4	98.0	90.2	83.8	82.5	84.2	95.5	93.8	85.4
57	38.5	39.2	59.8	89.2	70.6	60.8	61.3	53.0	74.3	56.6	64.4
58	77.9	76.5	90.2	88.2	94.1	86.7	87.5	97.3	99.3	98.6	88.3
59	70.0	72.5	85.3	91.2	90.2	85.0	78.8	89.0	99.3	94.4	82.3
60	44.4	43.1	67.6	80.4	76.5	77.5	55.8	65.2	96.5	75.7	64.0
61	78.3	75.5	89.2	95.1	91.2	81.7	94.6	94.0	92.0	92.7	87.1
62	65.3	66.7	80.4	94.1	84.3	75.0	77.8	89.5	83.7	78.5	86.5
63	34.3	32.4	50.0	88.2	69.6	45.0	48.8	56.0	64.2	53.8	76.7
64	85.0	85.3	93.1	94.1	96.1	87.5	97.5	94.0	97.6	100.0	93.1
65	75.2	74.5	87.3	95.1	94.1	81.3	81.3	93.2	97.2	98.6	87.7
66	48.9	49.0	63.7	90.2	79.4	67.9	62.1	69.3	78.5	81.9	73.3
67	83.1	84.3	90.2	94.1	98.0	83.8	87.9	98.5	100.0	100.0	94.8
68	76.7	76.5	90.2	94.1	92.2	91.3	82.5	91.1	98.6	95.1	86.7
69	54.7	52.9	74.5	92.2	77.5	81.3	75.0	68.2	91.0	66.0	72.5
70	88.5	88.2	96.1	96.1	96.1	89.6	99.2	98.2	99.0	100.0	89.2
71	79.7	78.4	89.2	96.1	94.1	85.0	86.7	92.6	98.6	100.0	89.2
72	54.0	54.9	70.6	89.2	84.3	71.3	66.7	72.6	84.0	75.3	81.2
73	91.5	91.2	98.0	96.1	98.0	91.1	98.3	100.0	100.0	100.0	96.4
74	82.8	82.4	92.2	98.0	91.2	90.4	91.7	94.9	98.3	96.5	85.0
75	61.9	61.8	78.4	94.1	83.3	85.4	76.3	71.4	97.2	83.3	71.3
76	89.8	91.2	96.1	96.1	98.0	89.2	96.7	99.4	100.0	99.3	95.8
77	82.5	81.4	91.2	95.1	94.1	88.3	96.3	89.9	98.7	96.9	89.4
78	66.5	65.7	82.4	83.1	84.3	88.3	85.8	74.1	95.1	86.8	75.8

Table A4.2 Mean and median values of word scores and phoneme scores of two groups of two MALE speakers each and four listeners (16 speaker-listener pairs) for each of the conditions of the experiment on communication channels. The condition numbers are according the description given in appendix A3.

No	CVC	CVC_{med}	C_i	V	C_f	$C_{i,fr}$	$C_{i,pl}$	$C_{i,vl}$	$C_{f,fr}$	$C_{f,pl}$	$C_{f,vl}$
1	90.3	90.2	96.1	98.0	97.1	88.8	97.9	99.1	100.0	99.7	94.2
2	58.0	56.9	83.3	86.3	78.4	75.8	84.6	80.4	84.7	79.2	75.0
3	35.0	31.4	65.7	72.5	66.7	49.6	57.1	75.9	53.8	56.3	75.8
4	40.1	40.2	59.8	84.3	68.6	49.6	58.8	74.4	64.6	59.4	72.7
5	12.5	12.7	35.3	60.8	41.2	25.8	35.4	41.7	28.5	26.7	51.9
6	83.8	87.3	93.1	94.1	96.1	87.5	92.1	94.6	100.0	99.3	91.7
7	68.4	70.6	83.3	91.2	88.2	88.3	83.3	81.3	94.1	97.9	76.9
8	71.3	70.6	89.2	92.2	90.2	83.7	91.7	84.8	94.4	91.7	83.8
9	43.0	40.2	64.7	86.3	73.5	80.0	58.3	56.5	81.3	67.7	66.0
10	89.5	91.2	98.0	96.1	98.0	88.8	99.6	98.8	97.6	100.0	96.9
11	43.9	42.2	62.7	87.3	67.6	43.3	65.8	81.5	56.3	53.5	77.7
12	18.3	18.6	42.2	74.5	52.0	16.7	40.0	60.4	23.6	37.2	73.8
13	37.1	33.3	56.9	85.3	66.7	47.5	54.6	71.7	62.5	54.5	72.5
14	7.6	8.8	25.5	59.8	37.3	14.6	26.7	33.6	23.3	22.9	46.5
15	79.4	81.4	88.2	95.1	94.1	85.4	83.3	91.4	98.3	95.8	88.8
16	69.6	70.6	81.4	94.1	89.2	77.1	80.0	88.4	88.5	87.5	86.7
17	65.6	65.7	81.4	95.1	84.3	75.4	82.5	83.6	78.5	86.8	85.2
18	47.8	47.1	72.5	92.2	66.7	63.8	71.7	78.9	69.8	55.9	70.1
19	70.1	71.6	91.2	84.3	90.2	87.1	97.5	89.6	100.0	100.0	80.2
20	51.1	54.9	86.3	71.6	81.4	81.7	90.8	76.8	98.6	96.9	65.6
21	47.8	45.1	72.5	70.6	80.4	78.3	83.3	68.5	95.8	87.5	66.7
22	44.7	43.1	75.5	71.6	78.4	83.3	80.4	63.7	99.3	89.9	60.2
23	23.8	23.5	56.9	56.9	66.7	82.5	62.1	35.4	97.9	68.8	44.0
24	65.4	64.7	89.2	81.4	86.3	81.3	93.8	87.2	100.0	95.1	72.3
25	44.4	46.1	73.5	68.6	75.5	63.3	80.4	67.0	84.7	87.2	67.3
26	27.6	24.5	49.0	58.8	62.7	53.3	48.3	54.8	67.7	64.6	56.9
27	41.4	35.3	67.6	66.7	68.6	74.6	77.9	63.1	82.6	84.4	58.3
28	27.9	23.5	52.0	60.8	66.7	64.5	59.6	42.3	71.9	70.8	52.3
29	43.1	41.2	67.6	91.2	65.7	31.3	65.8	95.8	26.7	61.8	86.7
30	35.3	34.3	61.8	92.2	58.8	35.0	60.0	83.0	21.5	54.5	82.3
31	24.5	24.5	53.9	82.4	50.0	30.8	50.8	63.1	21.5	47.9	66.5
32	39.2	39.2	66.7	94.1	60.8	19.6	64.6	91.4	12.8	60.1	84.0
33	92.2	92.2	98.0	96.1	100.0	89.2	100.0	98.8	99.3	100.0	97.1
34	72.7	75.5	90.2	89.2	88.2	88.3	94.6	89.3	99.3	96.9	80.0
35	48.9	47.1	69.6	80.4	72.5	82.1	83.3	53.6	89.9	86.8	58.3
36	89.4	89.2	99.0	95.1	97.1	90.6	100.0	100.0	98.6	98.1	94.2
37	77.8	78.4	94.1	97.1	90.2	85.8	96.7	92.3	88.9	93.1	84.4
38	49.6	47.1	69.6	86.3	74.5	75.4	80.8	62.2	66.3	89.6	67.5
39	91.3	93.1	97.1	95.1	99.0	89.6	98.8	99.1	100.0	99.3	98.1
40	72.4	74.5	89.2	91.2	87.3	89.2	98.8	83.6	99.3	100.0	75.6
41	69.5	67.6	85.3	85.3	85.3	91.7	95.0	74.4	97.9	97.9	72.7
42	90.8	93.1	96.1	97.1	98.0	88.3	99.6	98.5	99.3	99.7	95.4
43	70.3	67.6	87.3	91.2	86.3	86.7	94.2	82.4	95.1	98.3	70.4
44	47.3	43.1	70.6	81.4	73.5	77.5	82.1	60.7	93.8	83.7	57.5

No	CVC	CVC_{med}	C_i	V	C_f	$C_{i,fr}$	$C_{i,pl}$	$C_{i,vl}$	$C_{f,fr}$	$C_{f,pl}$	$C_{f,vl}$
45	87.4	88.2	97.1	94.1	96.1	90.8	97.5	95.8	100.0	100.0	93.5
46	72.2	72.5	87.3	94.1	89.2	86.3	86.3	84.8	96.9	89.9	85.4
47	51.7	48.0	69.6	86.3	82.4	85.0	63.3	64.3	93.8	78.8	76.9
48	88.0	88.2	94.1	96.1	98.0	84.2	99.2	97.9	96.7	100.0	94.4
49	76.0	78.4	89.2	94.1	90.2	84.6	90.4	87.2	98.6	94.8	81.3
50	49.0	47.1	67.6	88.2	80.4	76.3	67.9	55.7	85.4	78.5	74.6
51	80.5	80.4	92.2	94.1	96.1	86.3	95.0	91.4	98.6	99.3	90.6
52	69.5	68.6	81.4	92.2	90.2	87.9	82.9	78.6	93.8	89.6	86.3
53	42.3	45.1	62.7	86.3	74.5	70.8	54.6	60.9	79.2	69.1	66.9
54	78.2	77.5	85.3	96.1	92.2	80.4	88.8	89.9	98.3	93.1	91.3
55	60.8	63.7	76.5	96.1	84.3	71.3	75.8	79.8	75.7	85.1	79.4
56	39.6	40.2	58.8	86.3	70.6	56.7	59.6	65.5	66.0	61.8	74.1
57	82.1	81.4	95.1	94.1	93.1	86.7	92.1	97.6	98.6	97.6	86.9
58	73.9	73.5	84.3	96.1	90.2	80.0	81.3	90.2	92.7	95.5	85.8
59	76.0	76.5	87.3	96.1	90.2	81.7	81.7	94.0	95.8	91.3	89.0
60	40.8	40.2	64.7	86.3	72.5	63.0	47.9	77.1	75.3	57.6	70.6
61	77.1	76.5	90.2	91.2	92.2	83.3	96.3	92.9	97.6	98.3	88.8
62	66.3	65.7	82.4	92.2	85.3	77.1	90.4	82.1	93.1	90.3	78.5
63	73.4	72.5	90.2	92.2	89.2	80.4	93.8	94.3	92.4	92.7	82.3
64	63.4	63.7	85.3	86.3	83.3	74.6	84.2	87.2	80.9	85.8	80.4
65	81.3	81.4	95.1	93.1	92.2	84.2	97.9	98.2	99.0	96.5	88.1
66	71.7	72.5	88.2	94.1	85.3	71.7	95.0	91.4	88.9	87.5	84.8
67	71.6	69.6	86.3	93.1	89.2	82.5	90.4	84.5	94.8	97.9	80.4
68	59.3	58.8	78.4	89.2	81.4	77.5	77.5	83.3	77.8	81.6	76.3

Table A4.3 Mean and median values of word scores and phoneme scores of two groups of two FEMALE speakers each and four listeners (16 speaker-listener pairs) for each of the conditions of the experiment on band-pass limiting and noise. The conditions are numbered according the description given in appendix A2.

No	CVC	CVC_{med}	C_i	V	C_f	$C_{i,fr}$	$C_{i,pl}$	$C_{i,vl}$	$C_{f,fr}$	$C_{f,pl}$	$C_{f,vl}$
1	22.4	21.6	59.8	67.6	54.9	40.0	52.5	76.2	42.7	45.8	64.0
2	19.9	18.6	51.0	66.7	53.9	37.5	39.6	68.2	38.5	45.5	64.0
3	5.9	5.9	33.3	49.0	33.3	13.8	29.6	47.6	19.4	17.7	46.5
4	69.6	69.6	89.2	83.3	91.2	84.6	87.5	94.0	99.3	89.9	84.6
5	47.2	47.1	74.5	72.5	82.4	80.4	69.6	74.7	96.5	87.2	66.7
6	25.6	24.5	52.9	60.8	62.7	67.1	54.2	45.8	81.3	59.7	43.1
7	40.7	43.1	88.2	50.0	85.3	86.6	97.1	84.8	98.6	93.1	67.7
8	21.7	22.5	74.5	44.1	67.6	76.3	82.5	68.8	86.5	79.9	47.9
9	8.7	8.8	50.0	28.4	52.0	70.8	48.8	34.5	83.7	50.7	26.7
10	65.3	66.7	84.3	90.2	84.3	75.8	82.1	90.2	80.6	90.3	82.3
11	47.1	47.1	71.6	90.2	72.5	59.2	75.0	78.3	79.2	66.7	69.8
12	16.7	16.7	41.2	76.5	50.0	34.0	50.4	40.2	41.0	37.8	54.4
13	63.2	64.7	91.2	82.4	87.3	82.1	92.1	90.8	91.7	95.8	75.8
14	50.6	52.9	77.5	72.5	78.4	74.1	80.8	75.3	94.4	83.7	64.0
15	17.2	16.7	49.0	57.8	52.0	53.8	49.6	42.9	62.2	43.1	39.2
16	50.5	49.0	82.4	70.6	82.4	78.8	86.7	81.5	89.2	84.4	75.8
17	43.3	45.1	73.5	74.5	73.5	74.7	80.8	66.6	89.6	68.1	65.6
18	18.8	16.7	52.0	52.0	53.9	59.6	55.0	39.3	65.6	54.5	40.4
19	69.2	69.6	88.2	85.3	89.2	84.2	92.5	89.0	89.6	95.1	82.5
20	58.0	58.8	80.4	82.4	81.4	74.2	91.3	78.3	94.1	84.7	68.5
21	29.4	28.4	57.8	74.5	60.8	67.5	65.8	45.8	71.9	51.7	53.1
22	66.1	65.7	87.3	83.3	89.2	80.4	94.6	87.2	96.5	88.2	80.8
23	39.7	41.2	72.5	71.6	72.5	75.6	75.5	67.6	80.6	74.0	57.7
24	20.2	19.6	46.1	53.9	56.9	55.4	53.8	40.2	64.2	45.8	45.8
25	75.2	74.5	89.2	88.2	92.2	87.5	87.5	92.0	99.3	96.5	86.9
26	61.6	63.7	82.4	87.3	86.3	78.0	77.1	82.7	89.9	88.9	82.7
27	24.4	22.5	48.0	69.6	53.9	52.7	47.5	42.9	63.5	56.9	46.0
28	40.9	43.1	79.4	64.7	76.5	83.3	83.3	74.1	83.0	84.7	66.3
29	29.9	28.4	71.6	54.9	70.6	72.9	85.0	50.0	83.7	80.2	54.4
30	14.1	13.7	47.1	44.1	52.0	65.4	48.8	36.3	67.0	49.0	38.3
31	56.5	55.9	84.3	77.5	85.3	77.5	91.7	83.0	93.8	89.6	74.6
32	40.1	46.1	71.6	67.6	71.6	68.3	83.3	62.5	89.9	80.9	51.0
33	17.2	16.7	52.0	51.0	54.9	65.4	55.0	39.3	67.4	50.0	45.2
34	67.6	67.6	82.4	94.1	87.3	77.9	81.7	89.0	83.7	89.9	86.0
35	59.9	58.8	79.4	90.2	78.4	67.5	80.4	87.5	80.2	80.6	80.8
36	24.0	23.5	51.0	80.4	53.9	36.7	50.8	58.6	49.7	36.8	64.4
37	81.9	83.3	92.2	94.1	96.1	84.6	90.8	96.7	100.0	97.2	90.0
38	68.3	68.6	87.3	92.2	86.3	83.8	86.7	87.5	93.1	86.1	79.4
39	31.9	31.4	58.8	84.3	65.7	64.2	66.3	47.3	66.3	58.3	55.6
40	81.1	80.4	90.2	94.1	95.1	85.0	87.9	94.3	98.6	98.3	89.2
41	68.9	70.6	87.3	89.2	90.2	83.7	82.5	89.9	97.2	99.0	79.0
42	36.0	35.3	67.6	75.5	62.7	79.6	71.7	52.7	88.2	64.2	48.5
43	83.7	84.3	92.2	95.1	96.1	85.6	95.8	96.4	99.3	96.5	90.0
44	72.8	74.5	90.2	92.2	88.2	83.8	92.9	89.0	93.1	84.7	82.7

No	CVC	CVC_{med}	C_i	V	C_f	$C_{i,fr}$	$C_{i,pl}$	$C_{i,vl}$	$C_{f,fr}$	$C_{f,pl}$	$C_{f,vl}$
45	47.7	47.1	75.5	86.3	71.6	68.7	76.2	74.4	77.4	62.5	66.3
46	86.9	86.3	94.1	96.1	96.1	87.5	95.4	98.2	100.0	97.2	92.1
47	73.3	72.5	88.2	92.2	90.2	84.6	91.3	86.0	100.0	90.3	82.5
48	43.4	43.1	67.6	84.3	70.6	72.9	77.5	57.4	91.7	70.1	58.5
49	86.4	87.3	96.1	94.1	97.1	87.9	97.5	98.8	99.0	96.9	95.0
50	80.1	80.4	91.2	94.1	94.1	80.8	96.7	95.5	100.0	95.5	87.5
51	50.2	48.0	79.4	85.3	72.5	81.3	79.6	73.8	77.8	72.2	64.2

Table A4.4 Mean and median values of word scores and phoneme scores of two groups of two FEMALE speakers each and four listeners (16 speaker-listener pairs) for each of the conditions of the experiment on communication channels. The conditions are numbered according the description given in appendix A3.

No	CVC	CVC_{med}	C_i	V	C_f	$C_{i,fr}$	$C_{i,pl}$	$C_{i,vl}$	$C_{f,fr}$	$C_{f,pl}$	$C_{f,vl}$
1	89.3	90.2	96.1	94.1	100.0	85.0	97.1	99.4	99.0	99.0	98.1
2	44.1	44.1	73.5	78.4	74.5	70.0	85.0	70.8	77.4	70.8	71.7
3	27.3	26.5	69.6	58.8	57.8	57.1	64.2	66.7	46.5	58.3	61.7
4	29.3	29.4	58.8	69.6	61.8	53.8	51.3	66.4	54.9	43.1	62.7
5	5.0	5.9	33.3	39.2	37.3	26.9	33.8	40.8	35.1	21.5	39.2
6	83.8	83.3	92.2	95.1	94.1	86.7	95.0	96.1	99.3	99.0	87.9
7	73.2	74.5	85.3	94.1	90.2	81.7	83.8	88.7	97.9	92.0	83.1
8	60.7	60.8	84.3	87.3	81.4	80.8	93.8	84.8	92.4	87.2	66.9
9	40.6	39.2	65.7	77.5	66.7	73.8	72.9	56.3	78.1	69.8	55.4
10	85.3	86.3	94.1	94.1	98.0	85.8	96.3	97.0	100.0	96.2	93.5
11	22.1	21.6	52.9	64.7	50.0	36.3	50.4	68.8	27.4	36.8	64.6
12	9.8	9.8	38.2	52.9	35.3	22.9	30.4	56.8	23.3	29.2	42.9
13	18.0	18.6	47.1	65.7	49.0	30.1	45.8	61.6	35.4	43.1	60.2
14	3.4	3.9	28.4	35.3	24.5	16.7	23.8	40.2	19.4	16.0	34.8
15	72.3	72.5	86.3	92.2	89.2	81.3	86.7	90.5	97.9	91.3	83.8
16	58.6	60.8	81.4	89.2	78.4	75.8	81.3	86.3	77.8	68.4	80.2
17	51.8	52.9	74.5	90.2	70.6	68.8	75.4	82.1	72.6	69.8	72.3
18	27.5	25.5	52.9	78.4	55.9	42.9	53.8	58.6	54.5	46.5	57.3
19	64.6	62.7	92.2	77.5	91.2	86.7	94.6	91.1	99.3	98.3	83.8
20	36.0	32.4	77.5	55.9	77.5	79.8	90.8	65.5	98.3	94.1	60.2
21	25.0	18.6	61.8	39.2	60.8	71.7	78.3	52.1	82.3	70.8	42.1
22	30.8	29.4	70.6	48.0	70.6	87.4	86.3	53.6	95.1	84.7	52.7
23	18.5	18.6	62.7	36.3	64.7	78.8	68.8	39.0	94.1	70.8	44.8
24	46.7	43.1	84.3	58.8	84.3	79.6	88.8	80.7	89.2	97.9	71.5
25	21.7	19.6	68.6	44.1	66.7	52.9	76.3	67.0	66.3	65.3	59.0
26	13.1	13.7	52.9	47.1	47.1	43.7	50.4	54.8	50.3	38.9	45.4
27	30.6	28.4	68.6	55.9	65.7	72.9	75.8	62.2	79.5	75.3	54.2
28	15.2	13.7	51.0	49.0	52.0	59.6	59.2	34.5	60.8	52.8	44.6
29	35.7	37.3	62.7	84.3	59.8	32.9	60.4	89.3	16.7	50.3	77.1
30	30.5	32.4	62.7	80.4	52.0	34.2	60.4	83.9	20.8	55.2	72.1
31	23.2	21.6	55.9	72.5	44.1	35.8	61.3	64.9	18.4	43.1	61.9
32	28.3	30.4	58.8	78.4	52.9	24.2	45.0	90.8	11.1	41.0	82.3
33	88.0	88.2	95.1	94.1	100.0	85.4	99.6	99.1	100.0	100.0	97.9
34	71.1	70.6	88.2	89.2	88.2	82.9	97.5	82.7	100.0	93.8	76.5
35	49.1	49.0	79.4	74.5	74.5	80.8	95.8	67.9	94.8	89.2	54.8
36	86.9	86.3	93.1	94.1	99.0	83.4	95.0	97.3	99.0	99.3	96.9
37	62.3	60.8	85.3	86.3	79.4	75.8	91.7	84.2	83.7	91.7	73.5
38	40.3	41.2	75.5	77.5	65.7	78.3	82.1	62.2	70.8	72.9	54.6
39	88.6	90.2	93.1	98.0	99.0	80.8	99.6	99.7	100.0	100.0	96.0
40	71.7	71.6	89.2	87.3	86.3	83.0	99.2	86.5	100.0	99.0	70.0
41	48.8	46.1	75.5	74.5	71.6	79.2	92.9	61.9	97.9	90.3	50.2
42	82.4	83.3	94.1	90.2	98.0	83.7	97.5	99.7	96.5	99.3	94.4
43	53.2	50.0	85.3	76.5	79.4	78.3	88.3	81.8	89.9	86.5	68.1
44	30.4	32.4	62.7	69.6	65.7	72.1	75.4	44.9	91.0	74.0	44.2
45	84.7	86.3	95.1	92.2	96.1	85.4	98.3	99.1	98.6	99.7	92.3

No	CVC	CVC_{med}	C_i	V	C_f	$C_{i,fr}$	$C_{i,pl}$	$C_{i,vl}$	$C_{f,fr}$	$C_{f,pl}$	$C_{f,vl}$
46	70.1	68.6	86.3	90.2	88.2	79.6	96.2	86.9	100.0	92.7	77.3
47	44.4	50.0	74.5	82.4	78.4	74.6	72.9	61.0	88.5	68.8	66.7
48	84.4	84.3	94.1	94.1	98.0	84.2	99.6	94.3	99.7	100.0	94.0
49	64.7	68.6	84.3	88.2	86.3	82.9	91.7	78.0	99.3	91.3	75.0
50	43.9	46.1	67.6	82.4	77.5	72.5	75.4	55.1	87.2	70.5	66.3
51	77.6	82.4	90.2	93.1	96.1	82.1	95.0	92.3	97.2	99.3	88.5
52	61.6	63.7	80.4	91.2	82.4	79.6	90.8	76.2	87.2	85.1	69.4
53	46.7	47.1	70.6	81.4	77.5	75.4	75.8	58.0	84.7	83.3	64.6
54	71.4	70.6	84.3	93.1	90.2	72.5	90.0	89.6	88.2	90.6	87.7
55	45.6	48.0	72.5	88.2	71.6	66.3	68.3	76.8	72.6	59.0	77.9
56	26.7	24.5	51.0	76.5	53.9	50.0	53.8	53.9	54.2	51.0	51.5
57	76.2	76.5	88.2	92.2	94.1	77.5	89.6	95.2	98.6	94.4	87.9
58	59.9	58.8	81.4	90.2	78.4	72.9	80.8	89.9	86.8	70.1	82.1
59	67.2	66.7	86.3	92.2	84.3	80.4	78.3	92.6	98.3	80.2	79.2
60	32.0	30.4	56.9	83.3	60.8	45.5	60.4	65.2	59.4	41.3	66.3
61	73.7	72.5	89.2	92.2	92.2	77.9	87.5	97.9	93.8	97.6	84.8
62	55.8	56.9	80.4	88.2	73.5	71.7	87.3	84.2	77.1	76.4	73.5
63	56.0	57.8	82.4	83.3	80.4	80.0	85.8	82.7	91.3	78.5	76.0
64	52.1	50.0	77.5	80.4	77.5	76.3	81.7	80.4	74.7	75.7	77.5
65	79.3	79.4	92.2	94.1	94.1	84.2	92.1	95.8	97.6	95.1	87.9
66	56.9	58.8	80.4	92.2	78.4	65.8	81.3	87.2	78.1	75.3	78.1
67	56.6	55.9	79.4	86.3	82.4	75.0	80.0	82.1	81.3	86.5	76.3
68	45.6	46.1	70.6	80.4	75.5	67.5	77.9	64.6	80.2	68.4	67.5

Table A4.5 Mean INITIAL CONSONANT scores (m%) and standard deviation (s%), for male and female speech and the two experiments on band-pass limiting and communication channels.

Phoneme	Band-pass limiting				Communication channels			
	male		female		male		female	
	m	s	m	s	m	s	m	s
initial consonants								
p	72.2	22.9	74.6	20.1	80.4	19.6	77.5	20.0
t	81.7	22.5	86.4	20.0	81.7	21.9	80.2	23.6
k	79.4	22.0	82.9	18.3	86.1	18.6	86.5	20.4
b	55.8	26.5	56.4	23.1	69.0	22.6	68.4	20.8
d	69.6	20.7	76.9	20.5	79.3	19.1	77.9	20.1
f	64.7	23.2	68.6	18.4	62.6	27.0	60.1	25.7
s	81.6	27.3	82.1	23.4	85.9	19.6	80.0	20.4
v	50.6	20.6	43.3	18.7	53.6	21.7	41.2	20.0
z	71.5	22.7	76.8	18.4	76.9	16.5	76.2	16.8
x	84.7	19.3	85.5	14.4	84.7	24.7	85.1	21.1
m	67.9	24.1	62.3	23.1	74.7	20.1	67.7	21.0
n	74.8	19.0	75.5	19.8	74.3	18.8	70.0	21.4
l	71.9	25.5	67.7	25.7	80.0	19.1	76.0	21.3
R	76.0	21.6	76.6	24.7	78.5	22.0	79.1	22.4
w	65.0	24.1	60.6	25.1	71.4	21.3	70.1	21.1
j	85.0	16.5	82.2	18.5	91.1	11.9	86.5	15.7
h	76.7	17.8	75.4	19.7	83.7	15.8	80.6	16.9

Table A4.6 Mean VOWEL scores (m%) and standard deviation (s%), for male and female speech and the two experiments on band-pass limiting and communication channels.

Phoneme	Band-pass limiting				Communication channels			
	male		female		male		female	
	m	s	m	s	m	s	m	s
vowels								
aa	90.3	16.9	90.9	13.1	94.6	7.7	90.8	11.5
au	66.6	28.0	68.7	22.4	83.2	9.4	71.8	17.3
a	83.8	22.9	80.0	18.3	95.2	6.2	85.3	14.8
ee	77.0	22.6	79.7	19.2	85.7	11.7	80.8	18.0
ei	78.7	25.3	84.9	18.8	91.6	9.8	84.8	19.8
eu	77.5	17.9	71.2	23.4	81.8	15.7	77.7	19.5
e	80.6	21.5	78.3	18.5	87.6	12.8	78.2	22.4
ie	80.3	19.9	74.8	19.1	88.0	14.2	78.7	17.7
i	70.3	23.9	68.1	23.2	80.6	14.9	71.2	18.9
oo	72.5	20.3	68.3	23.0	79.4	11.8	76.8	13.5
oe	85.2	15.4	83.0	12.8	86.6	16.7	75.9	20.9
o	72.8	22.3	63.1	22.9	84.3	16.1	71.8	20.7
uu	69.3	23.5	65.6	23.1	74.7	21.5	64.7	20.7
ui	85.9	17.4	79.8	19.4	91.2	10.9	81.5	20.4
u	75.7	20.4	62.4	24.7	77.0	19.4	67.0	21.2

Table A4.7 Mean FINAL CONSONANT scores (m%) and standard deviation (s%), for male and female speech and the two experiments on band-pass limiting and communication channels.

Phoneme	Band-pass limiting				Communication channels			
	male		female		male		female	
	m	s	m	s	m	s	m	s
final consonants								
p	72.0	25.0	59.5	30.0	79.4	20.1	68.2	25.3
t	86.2	23.9	87.7	20.6	84.7	22.0	76.3	26.4
k	74.7	24.4	75.6	20.8	82.8	21.4	80.7	21.9
f	79.5	25.3	77.6	23.5	75.5	29.0	69.0	31.3
s	86.1	26.6	87.4	21.7	90.7	21.6	85.4	22.7
x	81.5	20.8	80.3	18.6	80.0	23.8	78.0	25.7
m	63.9	16.9	55.9	19.6	67.8	17.3	59.5	19.7
ng	58.9	21.1	52.7	23.6	68.9	18.1	60.9	21.5
n	62.6	17.7	59.6	17.6	65.9	16.6	63.8	17.3
l	88.2	12.5	79.0	18.4	90.1	11.6	82.7	15.5
R	92.8	10.1	86.5	14.7	94.4	9.0	85.6	15.1

A5 Example of the output of the scoring program

The scoring program gives, for each listener response file, the individual speaker-listener results and the mean speaker results. The overall results are based on all listener response files and include in this example four speakers and four listeners, hence 16 speakers-listener combinations. The overall results include the word score, the C_i, C_f, and V scores, the confusion matrices for C_i, C_f, and V, and some weighted means of the individual phoneme scores.

Program output:

- Session information - - - - - - - - - - - - -
From result file : AM001017.RC1
Date Time : 08-JUN-90 10:17:34
Length of the list : 51
Comparison file : cvclist.017
Parameter file : am001017.par
Translation file : \CVCEQB.TRA

Percentage Correct Score (orig. distrib.)

Subjects		C	V	C	Word
1 H v G	:	70.6	72.5	64.7	31.4
2 D S	:	74.5	72.5	64.7	37.3
3 J C	:	66.7	72.5	51.0	29.4
4 T C	:	56.9	64.7	56.9	23.5
Subjects' mean	:	67.2	70.6	59.3	30.4
standard deviation	:	7.6	3.9	6.7	5.7

- Session information - - - - - - - - - - - - -
From result file : AM001074.RC1
Date Time : 12-JUN-90 11:04:58
Length of the list : 51
Comparison file : cvclist.074
Parameter file : am001074.par
Translation file : \CVCEQB.TRA

Percentage Correct Score (orig. distrib.)

Subjects		C	V	C	Word
1 H v G	:	66.7	70.6	58.8	29.4
2 D S	:	60.8	76.5	70.6	41.2
3 J C	:	60.8	78.4	62.7	33.3
4 T C	:	68.6	82.4	62.7	39.2
Subjects' mean	:	64.2	77.0	63.7	35.8
standard deviation	:	4.0	4.9	4.9	5.4

- Session information - - - - - - - - - - - - -
From result file : AM001331.RC2
Date Time : 27-JUN-90 13:48:18
Length of the list : 51
Comparison file : cvclist.331
Parameter file : am001331.par
Translation file : \CVCEQB.TRA

Percentage Correct Score (orig. distrib.)

Subjects		C	V	C	Word
1 B W	:	60.8	68.6	58.8	29.4
2 W V	:	68.6	70.6	51.0	29.4
3 M v W	:	58.8	80.4	58.8	29.4
4 I S	:	66.7	82.4	58.8	33.3
Subjects' mean	:	63.7	75.5	56.9	30.4
standard deviation	:	4.7	6.9	3.9	2.0

- Session information - - - - - - - - - - - - -
From result file : AM001394.RC2
Date Time : 27-JUN-90 16:50:55
Length of the list : 51
Comparison file : cvclist.394
Parameter file : am001394.par
Translation file : \CVCEQB.TRA

Percentage Correct Score (orig. distrib.)

Subjects		C	V	C	Word
1 B W	:	58.8	58.8	58.8	27.5
2 W V	:	64.7	76.5	56.9	31.4
3 M v W	:	64.7	74.5	64.7	41.2
4 I S	:	64.7	74.5	62.7	33.3
Subjects' mean	:	63.2	71.1	60.8	33.3
standard deviation	:	2.9	8.2	3.6	5.8

- Overall results -
Results based on 16 talker-listener pairs

Percentage correct score (original distribution)

	C	V	C	Word	Phoneme product
Mean score	64.6	73.5	60.2	32.5	28.6
Standard deviation	4.8	6.2	5.1	5.0	
Standard error	1.2	1.6	1.3	1.3	
Median score	64.7	73.5	58.8	31.4	28.0
25% Quartile	60.8	70.6	57.8	29.4	
75% Quartile	67.6	77.5	65.7	35.3	

Percentage correct score

	C	V	C	Phoneme product
Mean eq. bal.	64.2	71.9	58.8	27.1
Mean phon. bal.	66.9	72.6	60.5	29.4

	fric.	plos.	vow.
Mean initial cons.	37.5	70.0	79.2
Mean final cons.	41.7	52.8	72.7

Response / Stimulus	P	T	K	F	S	G	M	NG	N	L	R	W	J	H	B	D	V	Z	??	Tot	%
1 P	29	-	10	4	-	-	-	-	-	-	-	-	-	1	2	1	1	-	-	48	60.4
2 T	3	29	11	-	2	-	-	-	-	-	-	-	-	1	-	-	2	-	-	48	60.4
3 K	-	1	46	-	-	-	-	-	-	-	-	-	-	-	-	-	1	-	-	48	95.8
4 F	5	1	-	21	1	12	-	-	-	-	-	-	-	-	-	-	7	-	1	48	43.8
5 S	2	12	1	12	6	1	-	-	-	-	-	-	1	1	-	1	4	3	-	44	13.6
6 G	-	-	2	4	-	45	-	-	-	-	-	-	-	-	-	-	1	-	-	52	86.5
7 M	-	-	-	-	-	-	25	-	12	5	-	-	-	4	2	-	-	-	-	48	52.1
8 NG	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	—.-
9 N	-	-	-	-	-	-	8	-	39	1	-	-	-	-	-	-	-	-	-	48	81.3
10 L	-	1	-	-	-	-	-	-	10	36	1	-	-	-	-	-	-	-	-	48	75.0
11 R	-	-	-	-	-	1	-	-	-	-	44	-	-	3	-	-	-	-	-	48	91.7
12 W	-	-	-	-	-	-	-	-	1	1	-	37	-	2	2	4	1	-	-	48	77.1
13 J	-	-	-	-	-	-	-	-	-	-	-	3	45	-	-	-	-	-	-	48	93.8
14 H	-	-	-	2	-	-	-	-	1	-	-	4	-	40	-	-	1	-	-	48	83.3
15 B	-	-	-	-	-	-	-	-	-	-	-	10	2	-	26	10	-	-	-	48	54.2
16 D	-	-	1	-	-	1	-	-	-	4	-	-	-	-	4	38	-	-	-	48	79.2
17 V	1	3	-	8	-	2	-	-	-	-	2	4	3	4	1	-	14	6	-	48	29.2
18 Z	1	2	-	1	3	-	-	-	-	-	3	5	18	2	1	1	4	7	-	48	14.6
Confusion	12	20	25	31	6	17	8	0	24	11	6	26	24	18	12	17	22	9	1	289	

Response / Stimulus	AA	AU	A	EE	IJ	EU	E	IE	I	OO	OE	O	UU	UI	U	??	Tot	%
1 AA	47	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	48	97.9
2 AU	-	42	5	-	-	-	1	-	-	-	-	-	-	-	-	-	48	87.5
3 A	-	-	95	-	-	-	-	-	-	-	-	-	-	-	-	1	96	99.0
4 EE	-	-	-	23	-	15	1	-	3	4	-	2	-	-	-	-	48	47.9
5 IJ	-	-	-	-	17	-	1	-	-	3	-	-	-	27	-	-	48	35.4
6 EU	-	-	-	2	-	36	-	1	3	-	-	1	1	-	4	-	48	75.0
7 E	-	-	4	-	2	-	70	-	1	-	-	5	-	1	13	-	96	72.9
8 IE	-	-	-	1	-	-	-	21	-	-	9	-	17	-	-	-	48	43.8
9 I	1	-	-	-	-	-	1	-	19	-	6	2	3	-	16	-	48	39.6
10 OO	-	-	-	-	-	1	1	-	-	35	1	10	-	-	-	-	48	72.9
11 OE	-	-	-	-	-	-	-	1	-	2	41	1	1	-	2	-	48	85.4
12 O	-	-	-	-	-	-	2	-	-	5	1	40	-	-	-	-	48	83.3
13 UU	-	-	-	1	-	-	-	16	-	-	1	-	30	-	-	-	48	62.5
14 UI	-	-	-	-	-	1	-	-	-	-	-	-	-	47	-	-	48	97.9
15 U	-	-	-	-	-	3	1	-	2	-	2	-	3	-	37	-	48	77.1
Confusion	1	0	10	4	2	20	8	18	9	14	20	21	25	28	35	1	216	

Response / Stimulus	P	T	K	F	S	G	M	NG	N	L	R	W	J	H	B	D	V	Z	??	Tot	%
1 P	21	4	21	1	-	-	-	-	-	1	-	-	-	-	-	-	-	-	-	48	43.8
2 T	22	44	16	10	1	1	-	-	1	-	1	-	-	-	-	-	-	-	-	96	45.8
3 K	5	6	33	3	-	1	-	-	-	-	-	-	-	-	-	-	-	-	-	48	68.8
4 F	2	6	6	13	-	20	-	-	-	-	1	-	-	-	-	-	-	-	-	48	27.1
5 S	6	17	10	23	14	21	-	-	-	1	3	-	-	-	-	-	-	-	1	96	14.6
6 G	1	1	-	5	1	40	-	-	-	-	-	-	-	-	-	-	-	-	-	48	83.3
7 M	-	-	-	-	-	-	65	2	28	-	-	-	-	-	-	-	-	-	1	96	67.7
8 NG	-	-	-	-	-	-	4	23	21	-	-	-	-	-	-	-	-	-	-	48	47.9
9 N	-	-	-	-	-	-	36	5	55	-	-	-	-	-	-	-	-	-	-	96	57.3
10 L	-	-	-	-	-	1	1	-	-	87	1	6	-	-	-	-	-	-	-	96	90.6
11 R	-	-	-	-	-	-	-	-	-	-	96	-	-	-	-	-	-	-	-	96	100.0
12 W	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	—.-
13 J	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	—.-
14 H	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	—.-
15 B	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	—.-
16 D	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	—.-
17 V	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	—.-
18 Z	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0	—.-
Confusion	36	34	53	42	2	44	41	7	50	2	6	6	0	0	0	0	0	0	2	325	

A6 Mean frequency spectra of the speech signals of the four phoneme groups used in this study for male and female speakers

The frequency spectra given below are based on a 5-second sentence from ten male and ten female speakers. The phonemes were annotated and separated into four groups (fricatives, plosives, vowel-like consonants, and vowels). For each group the 1/3-octave spectrum of the required (annotated) phoneme groups was determined. All silent periods before and after the phoneme signal burst were removed, either by the annotation or by applying a level threshold. The filtering and level estimation was performed digitally. Due to the limited sampling frequency (12.5 kHz), the highest 1/3-octave band was restricted to 5 kHz. From the 1/3-octave spectra the 1/1 octave spectra were calculated. These octave-band spectra, together with the A-weighted level, are given in Figs A6.1 and A6.2 for male and female speech respectively. Based on the weighted contributions of the four phoneme-group spectra the spectrum of phonetically balanced speech was also calculated.

The speech material and analysis procedure is part of a study on the assessment of speech recognition systems by manipulation of speech, and is described by Steeneken and Van Velden (1989, 1991).

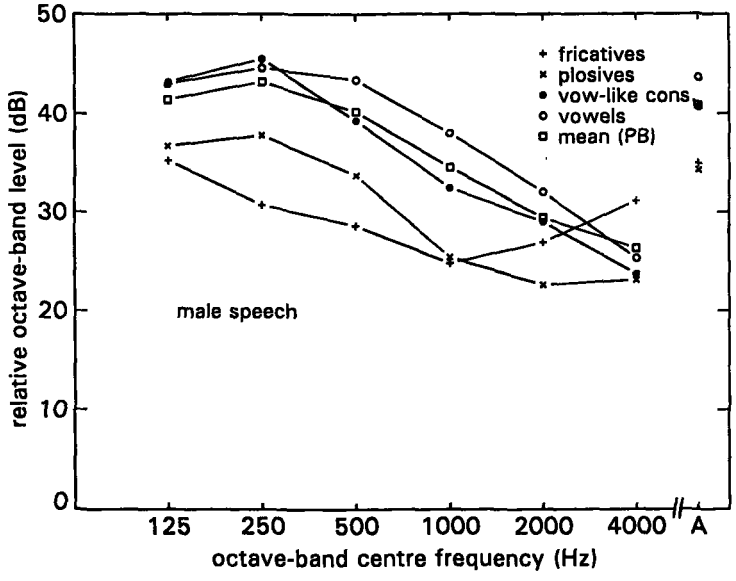

Fig. A6.1 Octave-band spectra obtained from speech signals of fricatives, plosives, vowel-like consonants, and vowels and for MALE speech. The spectrum labelled phonetically balanced (PB) is based on the weighted mean of the spectra of the four phoneme groups.

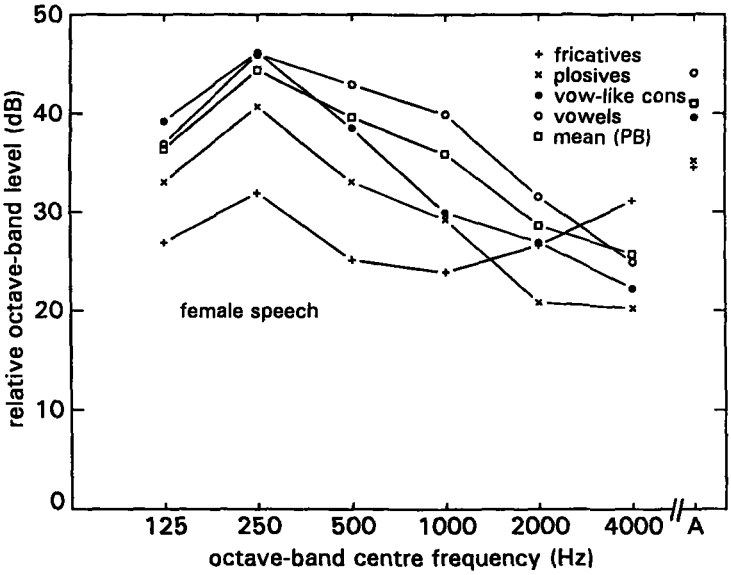

Fig. A6.2 Octave-band spectra obtained from speech signals of fricatives, plosives, vowel-like consonants, and vowels and for FEMALE speech. The spectrum labelled phonetically balanced (PB) is based on the weighted mean of the spectra of the four phoneme groups.

A7 Speech level measurement

For reproducible experiments concerning the effect of noise on speech transmission quality, it is important to specify the speech levels, the noise levels and the corresponding signal-to-noise ratios.

Various studies (Brady, 1965; Kryter, 1970; Berry 1971; Steeneken and Houtgast, 1978, 1986) concerning speech intelligibility have shown that a signal-to-noise ratio variation of only 1-2 dB may have the same effect on the results as typical speaker and inter-listener variations. We therefore specified a method for measuring speech levels and noise levels which offers such a resolution. The measure should be robust for the various speech types (male/female, connected discourse/isolated words), recording conditions (background noise, frequency transfer), and should also be applicable to noise signals. We have developed such a measure (Steeneken and Houtgast, 1978, 1986) mainly for adjusting the test signal level of the STI method to the speech level for similar conditions. We incorporated the measuring method into the existing specific hardware of the STI measuring device. Recently the measuring method was made generally available by converting the hardware solution into a digital signal-processing algorithm.

Description of the method

A high correlation was found between the speech level and the speech intelligibility for level measures based on frequency-weighted speech signals with a reduced contribution of frequency components below approx. 250 Hz (Kryter, 1970; Steeneken and Houtgast, 1978, 1986). The standardized frequency-weighting function according to the A-filter was used for this purpose (standardized for acoustical measurements).

After filtering, the running (intensity) envelope is determined by squaring and low-pass filtering the waveform. From this envelope function the envelope distribution histogram is obtained: the RMS value can be computed from this histogram. The advantage is that the RMS value can also be obtained for values above a certain level after sampling. In order to compare levels of short speech tokens with long silent periods in between (single words) and of connected discourse, a level threshold for suppression of the silent periods is required. Hence, this threshold is applied to the envelope function of the speech signal rather than to the waveform and therefore does not affect each zero-crossing of the speech signal. The threshold level is defined to be 14 dB below the resulting RMS level and therefore is signal-related and does not depend on other effects such as background noise, shape of the envelope distribution, etc. The same

principle can be applied to stationary noises but in that case the threshold function is not effected.

The relation between various level measures obtained from two speech samples (connected discourse, and CVC words in a short carrier phrase) is given in Fig. A7.1. The level measures are: the 1% peak level (1% overflow criterion), the mean of the peak deflections of a sound level meter set to "fast" (dB(A) fast), the RMS values obtained with a linear detector from the envelope function (RMS_{lin}, no true rms), the RMS values obtained with a squaring detector from the envelope function (RMS_{sq}, true rms), the RMS values obtained with direct sampling and by squaring the wave-form samples (RMS_{dir}, true rms, but application of a threshold is incorrect), and the equivalent peak-level (EPL) according to Brady (1968). The last method is not applicable to noise. The RMS-$A_{thr,sq}$ is representative of the level measure as applied in the Esprit-SAM group. For all these measures the addition of A-filtering is indicated by "A" and the application of a threshold by (thr).

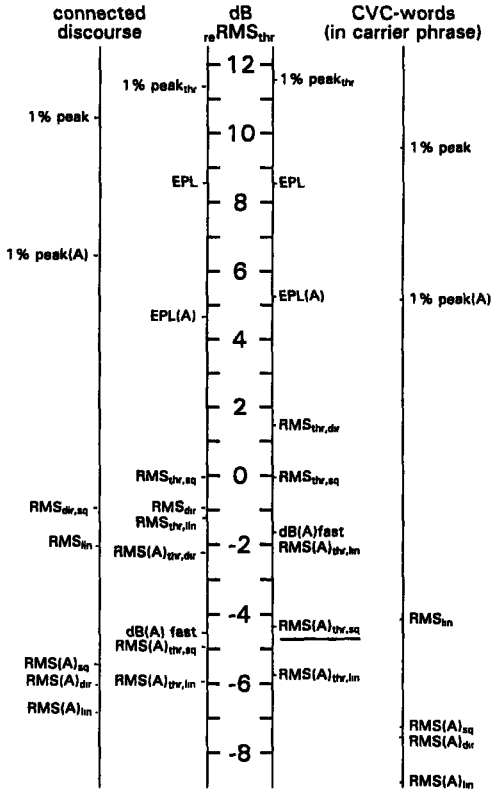

Fig. A7.1 Relative speech levels for different speech level measures applied to connected discourse and embedded CVC words. The values are relative to the $RMS_{thr,sq}$ value.

Blank 163/164

CURRICULUM VITAE

Herman J.M. Steeneken werd op 20 oktober 1939 te Delft geboren. In 1962 haalde hij het einddiploma HTS voor Elektronica te Hilversum. Daarna vervulde hij gedurende twee jaar de militaire dienstplicht als reserve-officier bij de Koninklijke Luchtmacht.

In 1965 ving het dienstverband aan met TNO, bij de Afdeling Audiologie van het Instituut voor Zintuigfysiologie TNO te Soesterberg. Oorspronkelijk was zijn primaire verantwoordelijkheid het beheren van de meetapparatuur en het ontwikkelen van nieuwe (analoge) apparatuur. In de loop der jaren is de methode van signaalanalyse geleidelijk gedigitaliseerd en is de rekenmachine een integraal deel van de meetopstelling geworden.

In 1968 legde hij het examen voor inschrijving in het toenmalige Ing-register met lof af.

Naast de hierboven beschreven werkzaamheden nam het aandeel in het onderzoek geleidelijk toe. Dat resulteerde in een aanstelling als wetenschappelijk medewerker. De voornaamste projecten betreffen spraaktransmissiekwaliteit, evaluatie van automatische spraakherkenners en actieve geluidreductiesystemen.

Buiten TNO geeft hij reeds meer dan 25 jaar les aan de Hogere Elektronica Opleiding (HEO) Rens & Rens te Hilversum. Tevens presenteerde hij in 1976 een Teleac cursus over moderne elektronica. In 1983 stelde hij een Teleac cursus samen over het werken met audio-apparatuur.

Hij onderhoudt internationale contacten o.a. via een Europees Esprit project betreffende de evaluatie van spraaksystemen, en in NAVO verband als Nederlands afgevaardigde in een werkgroep die smalband-vercijferde spraaksystemen bestudeert, alsmede in een "Research Study Group" op het gebied van de spraaktechnologie. Van deze laatste groep is hij tevens voorzitter.