EINDHOVEN UNIVERSITY OF TECHNOLOGY

Psychophysical and signal-processing aspects of speech representation

Citation for published version (APA):
Ma, C. (1992). *Psychophysical and signal-processing aspects of speech representation*. [Phd Thesis 1 (Research TU/e / Graduation TU/e), Industrial Engineering and Innovation Sciences]. Technische Universiteit Eindhoven. https://doi.org/10.6100/IR381327

DOI:
10.6100/IR381327

Document status and date:
Published: 01/01/1992

Document Version:
Publisher's PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the "Taverne" license above, please follow below link for the End User Agreement:

www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:

openaccess@tue.nl

providing details and we will investigate your claim.

Download date: 16. Nov. 2023

Psychophysical and Signal-processing Aspects of Speech Representation

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Eindhoven,
op gezag van de Rector Magnificus, prof. dr. J.H. van Lint,
voor een commissie aangewezen door het College van Dekanen
in het openbaar te verdedigen
op woensdag 23 september 1992 om 16.00 uur

door

Changxue Ma

geboren te Hubei (P.R. China)

druk: wibro dissertatiedrukkerij, helmond

Dit proefschrift is goedgekeurd door de promotoren
prof. dr. A.J.M. Houtsma
prof. dr. Y. Kamp

en de copromotor
dr. A. Kohlrausch

To: Xiaohung

Contents

§Paper with Y. Kamp and L.F. Willems submitted for publication to *SPEECH COMMUNICATION*

¶Paper with L.F. Willems published in *Signal Processing V: Theories and Applications*, edited by Torres, T., Masgrau, E., and lagunas, M.A. (Elsevier Science Publisher), pp.1171-1174, 1990.

‖Paper with Y. Kamp and L.F. Willems submitted for publication to J. Acoust. Soc. Am.

**Paper published in the IEEE Trans. on Signal Processing, Vol.39, pp.989-992, 1991.

††Parts of this chapter were published in the Proc. Eurospeech-91, Genova, Italy, 1991, pp.1125-1128 and as a poster at the Royal Society discussion meeting *Auditory Processing of Complex Sounds*, London, Dec. 4-5, 1991.

Chapter 1

Introduction

SPEECH communication research in general is the study of the production, transmission, manipulation and perception of speech. This research has been very dynamic and active, especially, since digital technology offers efficient and readily available tools for a variety of applications. Speech communication comprises many research areas, ranging from linguistics, phonetics, speech signal analysis and synthesis, speech coding and speech recognition to speech perception. In this thesis, we are mainly concerned with some signal-processing and psychophysical aspects of speech representation.

Speech sound in the form of the vibrations of the air is produced by the vibrations of vocal folds, which are driven by air from the lungs, and the movement of the vocal tract. The vocal tract can be described as an acoustical tube, whose nonuniform cross-sectional areas are manipulated by the movements of the lips, jaw, tongue and velum. The movements of these speech organs are controlled by the central nervous system such that these movements can convey distinctive information.

The acoustic waves of speech can be received by communication systems such as telephones or transformed into digital and analogue electric signal for further processing. Most importantly, human beings as ultimate receivers perceive the vibrations of speech waves. These speech waves are spectrally shaped by the outer and the middle ears and transformed into the vibration of the inner ear–the cochlea. The cochlea acts as a spectral analyzer (a bank of bandpass filter) with limited spectral resolution. The responses of bandpass filters are then transformed into a flow of spikes in the nerve fibers. The central nervous system decodes the trains of neural spikes into meaningful concepts.

Speech analysis is a very important way of achieving suitable signal representations for speech communication. In the signal-processing approach, speech production is described by a source-filter model. This model is considered to be time invariant on the short-time basis because the speech organs move slowly due to physical constraints. For

voiced sounds, the source is mainly situated in the vibrating vocal folds which modulate the air flow from the lungs. The vibration frequency of the vocal folds is called the fundamental frequency. The unvoiced sound source consists of the turbulent flow formed somewhere in the constricted vocal tract. The filter is of the all-pole type, whose coefficients represent an optimum linear prediction coding (LPC) of the signal (Flanagan, 1972). Therefore, the vowels typically have line spectra and the envelopes of the line spectra are modelled by the LPC model. The envelopes of vowel spectra show peaks and valleys. These peaks, called formants, correspond to the resonance frequencies of the vocal tract.

Modelling speech by the LPC analysis is very popular now. The LPC analysis and synthesis of a female voice, however, have not been successful (Klatt and Klatt, 1990). The reason is that the higher fundamental frequencies of women and children make it difficult to estimate formant parameters. In other words, for high-pitch sounds, the number of speech samples which are the output of the vocal tract filter without excitations, are relatively small in the closed glottis regions. Another reason for the lack of naturalness of LPC speech is that the excitation source is often simulated by a series of impulses or by white noise. The use of an elaborated model of excitation waveform has been a successful approach to improve the naturalness of synthetic vowels.

Improving the signal processing aspects of speech is just one way of improving the quality of speech coding and synthesis. The psychophysical study of speech provides a research direction to deal with the perceptually important aspects of speech. When we listen to speech sounds which are digitally processed and reproduced, both the intelligibility and the perceptual quality of the speech sounds are important for conveying information such as the identity of the speaker. Signal processing of speech always introduces distortion into the speech sounds. These distortions can be produced by quantization, or by parameterization of speech signals and produce different auditory sensations. On the other hand, speech generated by speech synthesizers also lacks naturalness. The evaluation of the processed and synthesized speech is closely related to the auditory perception of complex sounds. The understanding of the perception of complex sounds is therefore helpful to improve the quality of the processed sounds. Most of the time, this perception study is of course related to many aspects of working mechanisms of the central nervous system. For example, the perceptual evaluation of the quality of a text-to-speech system can involve the intelligibility, the naturalness of prosody and sound quality of the computer generated speech. The

psychophysical study of speech is relatively easy and provides discrimination thresholds because it makes a comparison between the original speech and the speech that results from the signal processing techniques such as speech coding.

In this dissertation we study ways to improve the LPC analysis/synthesis techniques. This includes an improved way to estimate the LPC parameters, especially for high-pitched voices and to estimate the time instants of glottal closure. The determination of instants of glottal closure has become a very important step for segmenting voiced sounds into successive pitch periods. In Chapter 2 a novel weighted LPC analysis of speech is investigated. In this approach, a weighting function is derived from the short-time-energy function of the speech signal. Speech samples are selectively weighted based on how well they match the speech production model. The estimates of the LPC coefficients by this novel LPC analysis are therefore more accurate than those obtained from the conventional LPC analysis.

In Chapter 3 the relation between the covariance linear prediction (CLP) analysis of a frame of a speech signal and the CLP analysis of its subframes is established. The results of CLP analysis derived from a set of subframes of speech samples are equivalent to those of a residual-weighted CLP analysis of the complete frame and the solutions of the residual-weighted CLP are the same as those of the generalized weighted average of subframe CLP. Those subframes which best reflect the filter model of the speech production can be chosen to improve the accuracy of the estimate of the LPC parameters.

The detection of glottal closure instants has been a necessary step in several applications of speech processing, such as speech coding, speech prosody manipulation and speech synthesis. Speech processing needs efficient and robust glottal closure detection methods. In Chapter 4 a singular value decomposition (SVD) approach is developed to detect the glottal closure instant in the speech signal. The proposed SVD method is equivalent to the calculation of the Frobenius norm of signal matrices and is therefore computationally efficient.

The spectral modelling of speech sound can be realized by linear prediction analysis, by which formant frequencies of the vocal tract are estimated from the peaks of the spectral envelope. The direct use of the phase spectrum of speech signal to estimate the formant frequencies has some advantages (Yegnanarayana et al., 1978). This method of estimation of formant frequencies can be applied if the speech signal can be uniquely determined (to a factor) from the phase of its Fourier

transform. In Chapter 5, we discuss a new approach for ascertaining whether a signal is uniquely determined by the phase of its Fourier transform. It will be shown that uniqueness corresponds to the non-singularity of a matrix which can be formed from the finite length real sequence.

The perceptual study of speech sounds in Chapter 6 is mainly concerned with auditory masking. The experiments are intended to make a contribution to the understanding of the perceptual aspects of speech processing, such as speech coding, speech synthesis, and speech manipulation. In contrast to most psychoacoustical masking studies, the targets in the measurement are narrow- or wide-band noise signals. The results can show the limitations of the auditory system in perceiving the distortions introduced by speech processing such as quantization noise in bit-compressed coding of audio signals or speech and by phase manipulation of speech.

References

J.L. Flanagan, (**1972**), *Speech Analysis, Synthesis and Perception* (Springer-Verlag, New York), 2nd edition.

D.H. Klatt and L.C. Klatt, (**1990**), "Analysis, synthesis, and perception of voice quality variations among female and male talkers", J. Acoust. Soc. Am. **87**, 820-857.

B. Yegnanarayana, G. Duncan and H.A. Murthy, (**1978**), "Improving formant extraction from speech using minimum phase group delay spectra", in *Signal Processing IV: Theories and Application*, ed. by J.L. Lacoume et al. (Elsevier, Amsterdam), pp.447-450.

Chapter 2

Robust signal selection for linear prediction analysis of speech *

Abstract

This paper investigates a weighted LPC analysis of speech. In view of the speech production model, the weighting function is either chosen to be the short-time energy function of the preemphasized speech sample sequence with certain delays, or is obtained by thresholding the short-time energy function. In this method, speech samples are selectively weighted on the basis of how well they match the speech production model. Therefore the estimates of the LPC coefficients obtained by this novel LPC analysis are more accurate than those obtained from the conventional LPC analysis. They are also less sensitive to the values of the fundamental frequency than is the case in the conventional LPC.

2.1 Introduction

THE source-filter model of speech production can be characterized by linear prediction equations (Markel and Gray, 1970; Makhoul, 1975) and two types of sources. The source for voiced sounds is a quasi-periodical glottal pulse train over a short interval of time and is produced by the vibration of vocal folds. The source for unvoiced sounds consists of the turbulent flow formed somewhere in the constricted vocal tract. The estimates of the predictor coefficients can be obtained by either an autocorrelation linear prediction method or a covariance linear prediction (CLP) method (Flanagan, 1972). The autocorrelation approach is a general probabilistic approach to the spectral analysis of a stationary Gaussian process (Itakura and Saito, 1970; Markel, 1972). When this method is applied to the speech signal, the nonstationary and the quasi-periodic characteristics of the speech signal are neglected. As a frequency-domain approach, the

*Paper with Y. Kamp and L.F. Willems submitted for publication to *SPEECH COMMUNICATION*

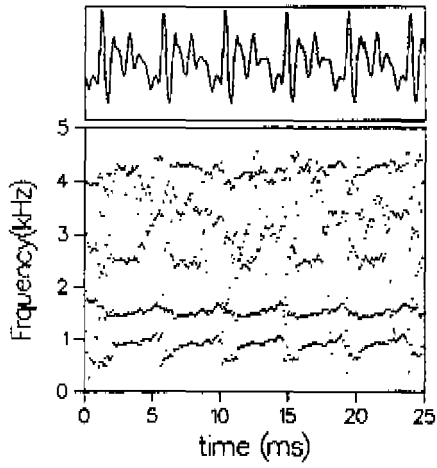


Figure 2.1: Formant tracks obtained by a covariance LPC of order 10. The top trace shows the speech waveform. Formant traces are shown as dotted lines. For further details see text.

autocorrelation method requires relatively long speech segments to provide adequate spectral resolution. Due to the harmonic structure of the voiced speech, this spectral match method does not perform well when the number of the harmonics is small as is the case for high-pitched voices. On the other hand, the covariance analysis is a nonstationary formulation of the estimation problem. In this method, speech is directly considered to be the output of the vocal tract filter with excitation sources, and thus specified by the time-varying transfer function of the filter and the characteristics of the source function (Atal and Hanauer, 1971). Therefore the covariance method directly models the speech wave rather than its spectrum. This time-domain approach to the estimation problem can be flexibly applied to short speech segments where the vocal-tract model is best fitted, avoiding the influence of the source.

In view of the speech-production model, one expects that speech segments containing source excitations will not be good candidate data for the estimation of LPC parameters. This can be seen from Fig. 2.1, which shows the estimated parameters from a natural vowel. The top trace shows the speech waveform and the dotted lines show the formant frequencies, which are linearly associated with the angular values of the zeros of prediction polynomials. These prediction polynomials are obtained by a covariance

analysis of order 10 and a sliding 3-ms rectangular window. The speech signal was preemphasized by a filter having a transfer function $(1-0.9z^{-1})$. Indeed, it is easy to observe from Fig. 2.1 that the estimated formant frequencies vary significantly in the region containing excitations. Therefore covariance analysis in these regions does not yield reasonable pole estimates. A more interesting observation from Fig. 2.1 is that the variations in the estimates of the high formant frequencies (F3, F4 and F5) are significantly greater than those of the low formant frequencies, even those estimated from the excitation-free regions. From careful examination it is found that this problem is due to the fact that the portion of the waveform associated with high formants decays rapidly because of the large formant bandwidth, so that background noise can become dominant. It is therefore expected that the estimates of the LPC parameters can be improved by choosing data segments which have a high signal-to-noise ratio and are not affected by the glottal pulse.

One approach to improvement of the parameter estimation is either to select or overweight those speech samples which are excitation-free and are thus expected to fit the LPC model better. There are several ways in which selection or over-weighting can be performed (Miyoshi et al., 1987; Lee, 1988; Ma and Willems, 1990). Pitch-synchronous LPC analysis is a particular method of getting rid of the influence of glottal pulses by using a short window to select excitation-free portions, such as the closed-glottis portion, and to improve the estimates of formant parameters (Steiglitz and Dickinson, 1977; Pinson, 1978; Kuwabara, 1984). It is not always easy, however, to choose those excitation-free portions in voiced sounds uttered by females or children because the pitch period is short. The results are dependent on the data available in the pitch period and are also sensitive to the window position (Larar et al., 1985). In addition, the signal of a voiced sound is quasi-periodic. Differences between the successive pitch periods are due either to noise or to other factors from the glottal source. The pitch-synchronous LPC analysis is limited to a single period and does not benefit from the time averaging of the speech data over several periods. Another selection technique examined earlier (Miyoshi et al., 1987; Ma and Willems, 1990) is the generalized sample-selective LPC. There, a preliminary conventional LPC analysis provides an approximation for the residual excitation signal. In a second LPC analysis only those speech samples are kept for which the residual in the first LPC analysis lies below a certain threshold. This method conceivably increases the computational burden. Moreover, due to the inaccuracy of the first LPC analysis, the LPC inverse filtering can give rise to significant pulses in the residual, other than those on the ex-

citation moments (Ananthapadmanabha and Yegnanarayana, 1979). The peaks of the residual error, which are assumed to correspond to the instants of glottal closure, are not always very prominent and, as a result, selection windows might be misplaced. A more complex procedure (Lee, 1988) is to minimize a more elaborate loss function of the LPC residual signal which discriminates between the residual samples exceeding a threshold and those below this threshold. In general, these weight functions are imposed on the residuals to reduce the effects of the glottal pulses and to improve the estimates of the LPC parameters. They need, however, complex procedures to locate the pitch pulses and to synchronize selection windows (Miyoshi et al., 1987; Lee, 1988).

The robust selection and weighting techniques described in this paper are based on the observation that pre-emphasized vowel signals show clear peaks just after, and clear valleys just before, the moments of excitations, which also correspond to the peaks in the LPC residual. This is an indication that the short-time energy function (STE) of the signal could be taken either as a selection criterion or as a weighting function. These two possibilities are further developed in this paper and compared with the sample-selective method of Miyoshi et al. (1987). The short-time energy of the signal is computed over a short window which has a constant lag with respect to the speech samples considered for the computation of the LPC parameters. In this way, speech samples that fit the LPC model well and produce small LPC residuals are over-weighted, and speech samples that do not fit the model are down-weighted. Therefore the estimates of the LPC coefficients obtained by this method are more accurate than those obtained from the conventional LPC analysis and they are also less sensitive to the values of the fundamental frequency. Comparison of experimental results show that the proposed LPC analysis is attractive from the point of view of computational efficiency, estimation accuracy and selection of speech samples. Finally, a stability analysis of the linear predictor computed with the short-time energy as weighting function is presented, based on the theory of the numerical range of a linear operator.

2.2 Weighted LPC analysis

The speech-production model can be generally described by the following equations:

$$s_n = \sum_{i=1}^{p} s_{n-i} a_i + e_n \quad n = n_1, n_1 + 1, n_1 + 2, \ldots, n_2 \tag{2.1}$$

where s_n denotes the nth sample of a speech wave, e_n is the nth sample of an excitation wave, a_i the ith predictor coefficient and p is the order of the prediction equations. In the autocorrelation case, $n_1 = 1$; $n_2 = N + p$ and the speech signal is assumed to be zero outside the interval $[1, N]$. In the covariance case, $n_1 = p + 1$ and $n_2 = N$. Here a weight function W_n is introduced to select or weight speech samples for the LPC analysis. The estimates of the LPC parameters can then be obtained by minimizing the weighted residual energy

$$E = \sum_{n=n_1}^{n_2} (s_n - \sum_{i=1}^{p} s_{n-i} a_i)^2 W_n \tag{2.2}$$

The parameters a_i can be obtained by setting the derivative of E with respect to a_j to zero. Then we obtain the following p equations:

$$\begin{array}{l} \sum_{n=n_1}^{n_2} s_{n-1} W_n \sum_{i=1}^{p} s_{n-i} \hat{a}_i = \sum_{n=n_1}^{N} s_n s_{n-1} W_n \\ \sum_{n=n_1}^{n_2} s_{n-2} W_n \sum_{i=1}^{p} s_{n-i} \hat{a}_i = \sum_{n=n_1}^{n_2} s_n s_{n-2} W_n \\ \qquad \cdots \\ \sum_{n=n_1}^{n_2} s_{n-p} W_n \sum_{i=1}^{p} s_{n-i} \hat{a}_i = \sum_{n=n_1}^{n_2} s_n s_{n-p} W_n \end{array} \tag{2.3}$$

For the sake of simplicity we will use a vector notation to represent these linear equations, namely

$$\sum_{n=n_1}^{n_2} W_n \vec{S}_n \vec{S}_n^T \vec{\hat{a}} = \sum_{n=n_1}^{n_2} W_n \vec{S}_n s_n \tag{2.4}$$

where T represents matrix transposition; $\vec{\hat{a}} = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p)^T$, and $\vec{S}_n = (s_{n-1}, s_{n-2}, \ldots, s_{n-p})^T$. Therefore, the estimated value $\vec{\hat{a}}$ and the "true value" $\vec{a} = (a_1, a_2, \ldots, a_p)^T$ are related by the following relation, which is easily obtained by substituting equation (2.1) into equation (2.4).

$$\vec{\hat{a}} - \vec{a} = (\sum_{n=n_1}^{n_2} W_n \vec{S}_n \vec{S}_n^T)^{-1} (\sum_{n=n_1}^{n_2} W_n \vec{S}_n e_n) \tag{2.5}$$

It is easily seen from equation (2.5) that the estimates of the LPC parameters can indeed be improved by choosing a proper weight function to make the item on the right-hand side of equation (2.5) small. As long as the matrix $C = \sum_{n=n_1}^{n_2} W_n \vec{S}_n \vec{S}_n^T$ is nonsingular, it would then be desirable to make the sum $\sum_{n=n_1}^{n_2} W_n \vec{S}_n e_n$ as small as possible. For natural speech this cannot easily be fulfilled because speech samples in $\vec{S}_n$ are delayed outputs of the production equation (2.1) under the input e_n which is not an

idealized pulse train. The weighting function W_n, however, can take on a low value or be zero when e_n is large, and take on a high value when e_n is small. Consequently, the difference between the estimated predictive coefficient vector $\hat{\vec{a}}$ and the "true value" $\vec{a}$ decreases. In other words, the speech samples that fit the LPC model are over-weighted and those samples that do not fit the model are down-weighted.

2.3 Choosing the weighting functions

In the Introduction it has been mentioned that the sample-selective method (Miyoshi et al., 1987) suffers from several shortcomings, in particular, that it is computationally expensive and that the selection of speech samples is still unreliable. To avoid these difficulties we now propose a method which only requires a single LPC analysis but achieves essentially the same objective by using either an appropriate sample-selection window (i.e. W_n=1 or 0) or an appropriate weight function W_n. In both cases the weight W_n is based on the STE (the short-time energy), $\sum_{i=0}^{M-1} s_{n-i-k}^2$, computed over a window of M samples and with a certain lag k with respect to the prediction residual e_n which is multiplied by W_n in equation (2.2). This choice of W_n is based on the following observations. In Figure 2.2 it can be seen that the pre-emphasized speech signals for vowel /a/, spoken by a female and a male, show clear peaks and valleys. The peaks are due to the strong excitations that are produced by rapid closing of the vocal folds and the valleys result from the decay of the ringing of the vocal tract filter. These strong excitations generally also correspond to the peaks in the LPC residuals indicating that the LPC model does not fit the speech samples in these regions. Due to these peaks and valleys the short-time energy function of these signals, calculated with a window of a size less than a half period, will over-weight the speech samples which follow the main excitations while down-weighting those containing the excitations. In other words, the speech samples that fit the LPC model well are over-weighted and the samples that do not fit the model are down-weighted (Lee, 1988; Ma and Willems, 1990). This will be further clarified in the tests described in the next section.

In the spirit of these consideration, we will consider two alternatives. In the first case, we select the speech samples s_n for which the STE function exceeds a certain threshold T_d. In the following, this case is referred to as "STE-thresholded" and the function W_n in (2.4) is accordingly defined as

$$W_n = \begin{cases} 1 & \text{if } \sum_{i=0}^{M-1} s_{n-i-k}^2 > T_d \\ 0 & \text{otherwise} \end{cases} \tag{2.6}$$

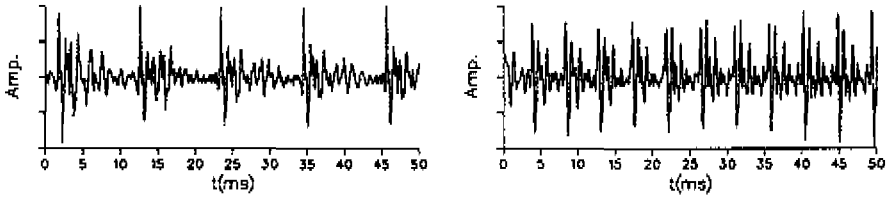


Fig.2.2(a) Fig.2.2(b)

Figure 2.2: Preemphasized speech waveforms. (a) vowel /a/ uttered by a male. (b) vowel /a/ uttered by a female.

In the second case, called "STE-weighted", all speech samples are considered for the LPC analysis and the weight function W_n in (2.4) is the STE itself, i.e.

$$W_n = \sum_{i=0}^{M-1} s_{n-i-k}^2 \tag{2.7}$$

This short-time energy function will be used as a weighting function to over-weight the speech samples which follow the main excitations and to down-weight those containing excitations.

2.4 Performance evaluation

The STE-thresholded and the STE-weighted LPC analyses defined in the preceding section are applied to synthetic vowels and natural vowels. They will be compared with the conventional autocorrelation LPC and against the sample-selective method (Miyoshi et al., 1987). In all cases the speech was sampled at 10 kHz and was then pre-emphasized by a filter $(1-0.9z^{-1})$. The prediction order was 10 and the speech data are refreshed every 10 ms with an analysis frame of 25-ms duration. Formant frequencies were obtained by solving for the zeros of the estimated LPC polynomials.

In the autocorrelation case a Hamming window of 25-ms duration was used. The sample-selective, the STE-thresholded and the STE-weighted solutions were obtained from the covariance equations. The sample-selective method implemented here uses rectangular windows of 12 samples width; their left edges are set one sample ahead of the time instants at which the residual of the conventional LPC analysis just falls below 50% of the peak

value in the current-analysis frame. For our solutions, the short-time energy STE is calculated over a window of size $M = 12$ and then delayed by one sample, i.e. $k = 1$ in (2.6) and (2.7). For the STE-weighted solution, the weighting function W_n is the STE itself; for the STE-thresholded solution, W_n is defined by (2.6) with threshold T_d equal to 50% of the peak value of the short-time energy function in the current analysis frame.

These four LPC solutions, being the conventional LPC, the sample-selective LPC, the STE-thresholded LPC and the STE-weighted LPC, were first applied to four synthetic vowels of about 1.2 seconds duration, which were produced by using two different excitations: single-pulse excitation for vowels V1 and V2 and LF-modelled excitation for vowels V3 and V4 (Fujisaki and Ljungqvist, 1986). The influence of high fundamental frequency on the estimation accuracy was investigated. The formant frequencies and bandwidths for the four vowel sounds are listed in Table I. The fundamental frequency of the vowel sounds increases linearly in the specified range listed in Table I on a logarithmic scale.

The means and standard deviations of the estimated five formant frequencies are listed in Tables II- V for the different LPC solutions. It can be seen from Tables II and III for vowels V1 and V2 which have the

Vowel	$F_0 - range$	(F_1, B_1)	(F_2, B_2)	(F_3, B_3)	(F_4, B_4)	(F_5, B_5)
V1	100 – 250	(500, 50)	(1500,150)	(2500,250)	(3500,350)	(4500, 450)
V2	250 – 400	(500, 50)	(1500,150)	(2500,250)	(3500,350)	(4500, 450)
V3	100 – 250	(790, 50)	(1300,150)	(2565,250)	(3500,350)	(4500, 450)
V4	250 – 400	(790, 50)	(1300,150)	(2565,250)	(3500,350)	(4500, 450)

Table I: Formant frequencies and bandwidths in Hz for two synthetic vowels in two fundamental frequency ranges.

LPC	F_1	F_2	F_3	F_4	F_5
Conventional	498(16)	1497(9)	2498(8)	3500(5)	4500(5)
Sample – selective	500(1)	1498(9)	2499(5)	3499(1)	4500(2)
STE – thresholded	500(0)	1499(6)	2499(3)	3500(3)	4500(3)
STE – weighted	499(3)	1499(3)	2499(3)	3500(3)	4500(4)

Table II: Estimated formant frequencies and their standard deviations (in parentheses) for vowel V1.

same formant structure, but different fundamental frequency range, that the formant frequencies obtained from the autocorrelation LPC are more scattered than those obtained by the other three LPC solutions. Comparing Tables II and Table III, it can be seen that the formant frequencies estimated by the conventional LPC analysis are even more scattered for vowel V2 than for vowel V1, due to the higher fundamental frequencies of vowel V2. The difference between the estimated and the original F1 reaches as much as 8 percent for vowel V2, which is significantly greater than the just-noticeable difference of formant frequency (3-4% for formant one) reported by Flanagan (1972), who used the same vowel stimuli for the listening test. The formant frequencies obtained by the sample-selective and the STE-thresholded LPCs are very close to the correct values. The formant frequencies obtained by the STE-weighted LPC is also clustered around the correct values. The formant frequencies for vowels V3 and V4 are shown in Tables IV and V, respectively. One observes from Table IV that the three low-frequency formants obtained by the sample-selective and the two STE-based LPC's are narrowly distributed around the correct values, while those obtained by the conventional LPC are relatively scattered. Comparing the results for vowels V3 and V4, one can see that the for-

LPC	F_1	F_2	F_3	F_4	F_5
Conventional	499(39)	1496(31)	2500(22)	3498(15)	4500(9)
Sample − selective	500(0)	1499(1)	2499(1)	3499(1)	4499(0)
STE − thresholded	500(0)	1499(0)	2499(0)	3499(0)	4499(0)
STE − weighted	501(9)	1501(7)	2500(4)	3498(3)	4499(2)

Table III: Estimated formant frequencies and their standard deviations for vowel V2.

LPC	F_1	F_2	F_3	F_4	F_5
Conventional	777(15)	1267(7)	2581(29)	3626(59)	4868(219)
Sample − selective	774(5)	1278(7)	2580(4)	3535(22)	4500(29)
STE − thresholded	789(1)	1303(2)	2559(5)	3503(7)	4548(23)
STE − weighted	781(5)	1284(3)	2535(3)	3505(24)	4705(171)

Table IV: Estimated formant frequencies and their standard deviations for vowel V3.

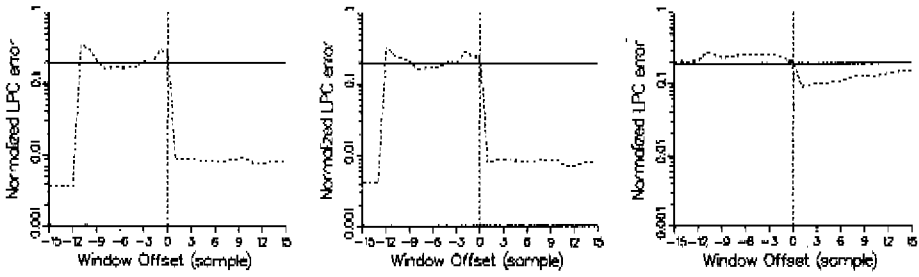


Fig.2.3(a) Fig.2.3(b) Fig.2.3(c)

Figure 2.3: Normalized LPC error as a function of the position of signal selection window for vowel V2. Solid line shows the error for conventional LPC analysis. (a) Sample-selective LPC analysis. (b) STE-thresholded LPC analysis. (c) STE-weighted LPC analysis. Analysis condition: LPC order=10, Preemphasis=0.9, Window width=12 samples.

mant frequencies F_1 and F_2 obtained by the conventional LPC are more scattered for the vowel V4 with higher fundamental frequency. This again indicates that the conventional LPC is much more sensitive to the excitation pulses. Generally, the results obtained by the sample-selective method and the STE-thresholded LPC are the best ones; those obtained from the STE-weighted LPC are slightly poorer but still show much improvement compared with the conventional LPC analysis. From the viewpoint of computation complexity, both STE-based LPCs are much more economical than the sample-selective LPC method.

In the second test the normalized total squared LPC residual error is

LPC	F_1	F_2	F_3	F_4	F_5
Conventional	781(31)	1290(27)	2524(22)	3457(36)	4391(25)
Sample − selective	787(3)	1297(6)	2571(6)	3523(20)	4516(42)
STE − thresholded	784(11)	1297(11)	2564(6)	3502(13)	4539(33)
STE − weighted	784(15)	1291(11)	2532(9)	3466(17)	4480(47)

Table V: Estimated formant frequencies and their standard deviations for vowel V4.

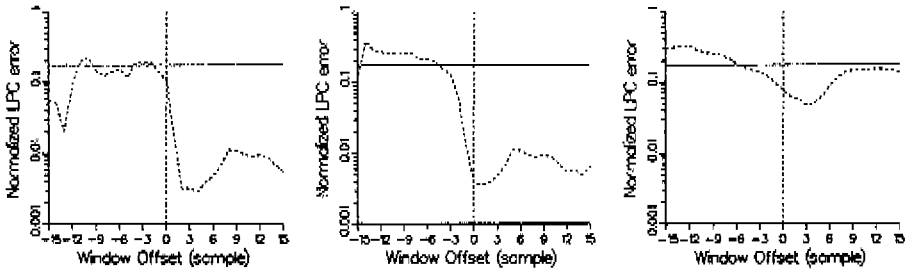


Fig.2.4(a) Fig.2.4(b) Fig.2.4(c)

Figure 2.4: Same as Figure 2.3, but for vowel V4.

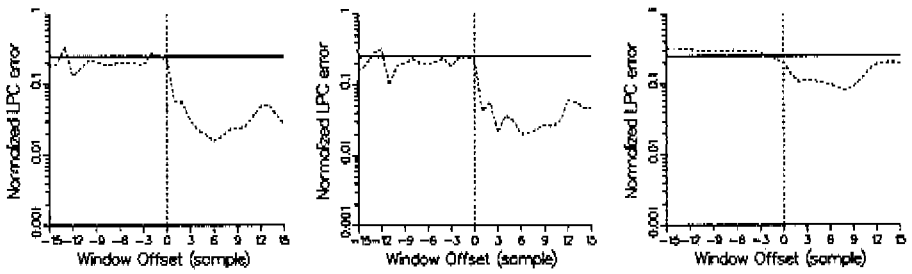


Fig.2.5(a) Fig.2.5(b) Fig.2.5(c)

Figure 2.5: Same as Figure 2.3, but for natural vowel /a/.

calculated as a function of the amount of offset of the windows or the weighting functions. For the sample-selective LPC the shift of the windows is the number of samples between the left edge of the window and the point where the residual just falls below 50% of the peak value. For both STE-based LPC's the shift is simply the value of k in the expressions (2.6) and (2.7). The experimental conditions are the same as in the first experiment. Synthetic vowel segments and natural vowel segments of 250 samples were used in the test. Synthetic vowel segments were taken from vowels V2 and V4 (fundamental frequency 250 Hz) and two natural vowel segments /a/ and /e/ were spoken by a female. The results obtained from synthetic vowels V2 and V4 are plotted in Figs. 2.3 and 2.4, respectively

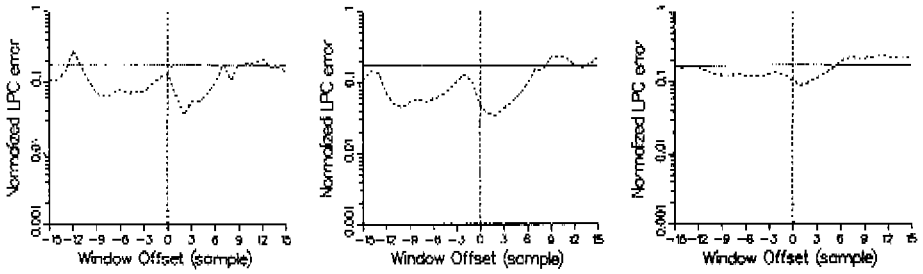


Fig.2.6(a) Fig.2.6(b) Fig.2.6(c)

Figure 2.6: Same as Figure 2.3, but for natural vowel /e/.

and those from natural vowels /a/ and /e/ are shown in Figs. 2.5 and 2.6, respectively. In Figs. 2.3- 2.6, solid lines represent the normalized total squared LPC error from the conventional LPC analysis and dotted curves represent the error obtained from the weighted LPC analysis. In each figure, panel (a), (b) and (c) show the results from the sample-selective, the STE-thresholded and the STE-weighted LPC analysis, respectively. It can be seen from panels (a) in Figs. 2.3- 2.6 that the normalized error from the sample-selective LPC analysis sharply decreases as the window is advanced by one sample and that the amount of the decrement is dependent on the vowel sounds. From panels (b) in Figs. 2.3- 2.6 we see that normalized error curves are similar to those shown in panels (a), except that the sharp decrease may be offset to the right by one or two samples. Panels (c) in Figs. 2.3- 2.6 also show that normalized errors decrease as the window advances. The decrement is, however, less than shown in panels (a) and (b) due to the fact that the STE-weighted LPC uses a continuous weight function and does not make hard decisions in choosing speech samples. The normalized error functions obtained by the sample-selective LPC and the STE-thresholded LPC are quite similar. The smallest normalized errors obtained by the STE-weighted LPC are higher than those obtained by the sample-selective LPC or the STE-weighted LPC, but they are still below the normalized error obtained by the conventional LPC.

In the third test, five natural vowels, /a/, /e/, /u/, /i/ and /o/, spoken by a female were used for the four types of LPC analyses. The average fundamental frequency of the vowels is about 200 Hz. For comparison, the spectral envelopes of the LPC filters obtained by the conventional LPC

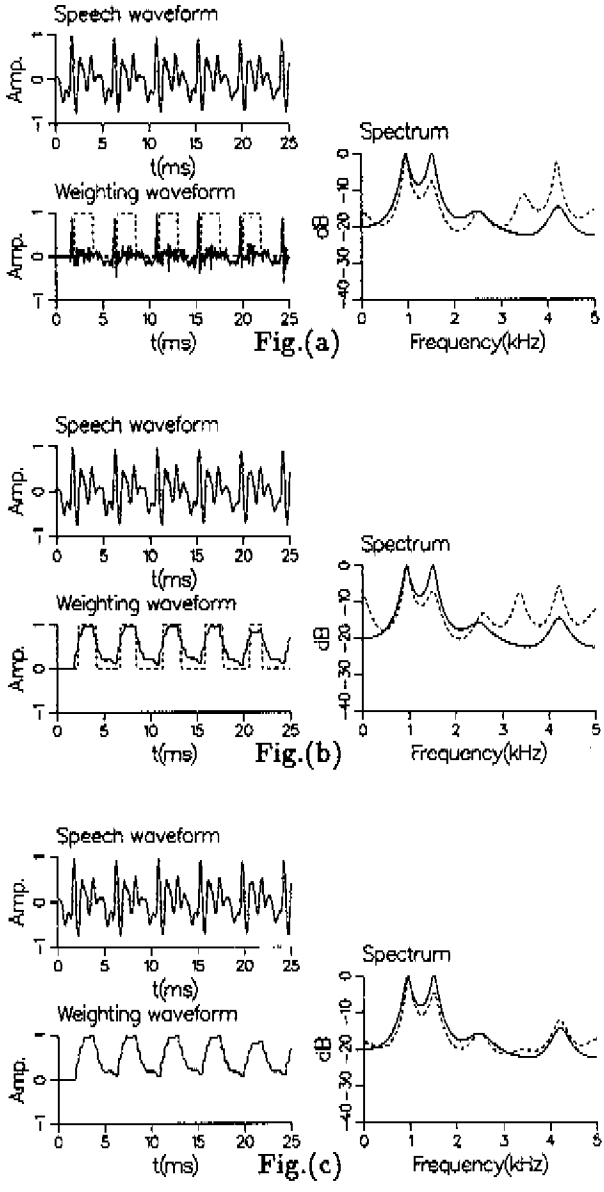


Figure 2.7: Results for vowel /a/. (a) Sample-selective LPC. (b) STE-thresholded LPC. (c) STE-weighted LPC. Spectral envelopes for conventional LPC are shown as solid lines.

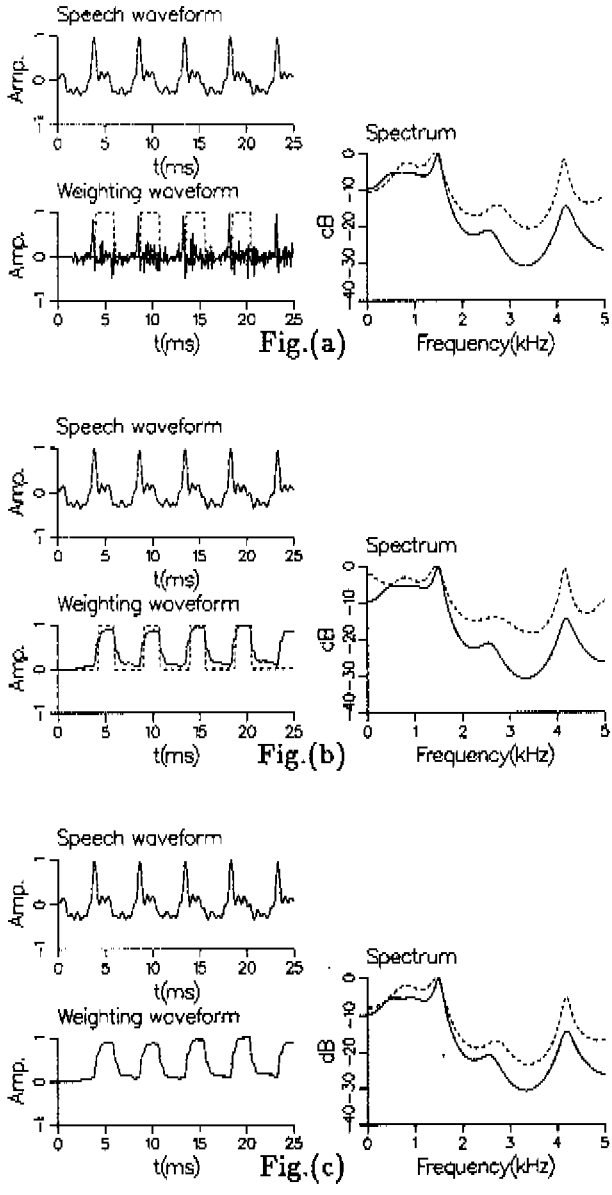


Figure 2.8: Same as Figure 2.7, but for vowel /e/.

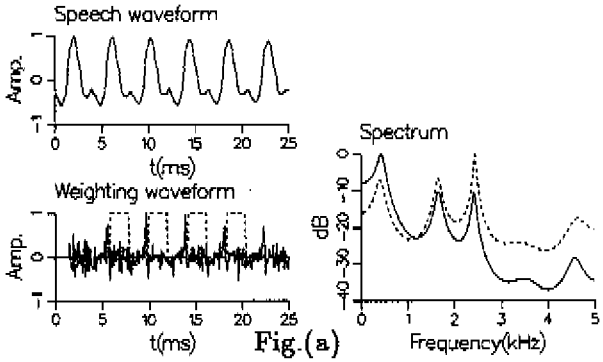


Fig.(a)

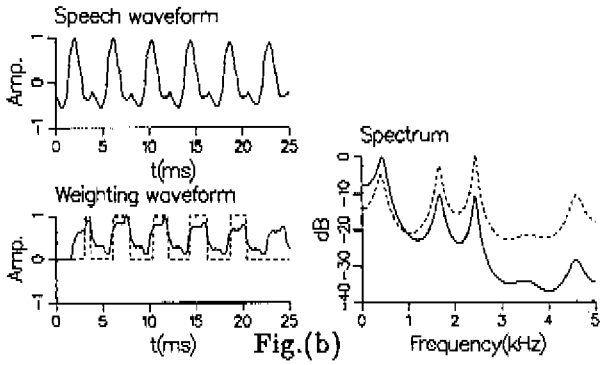


Fig.(b)

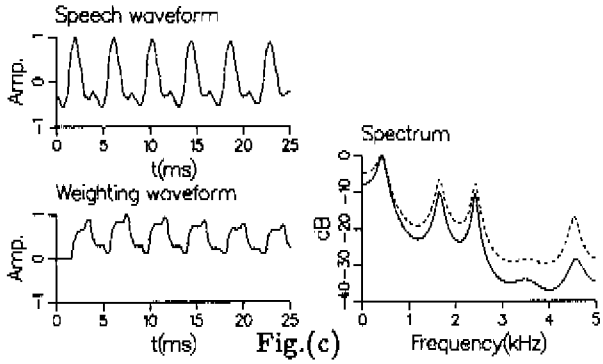


Fig.(c)

Figure 2.9: Same as Figure 2.7, but for vowel /u/.

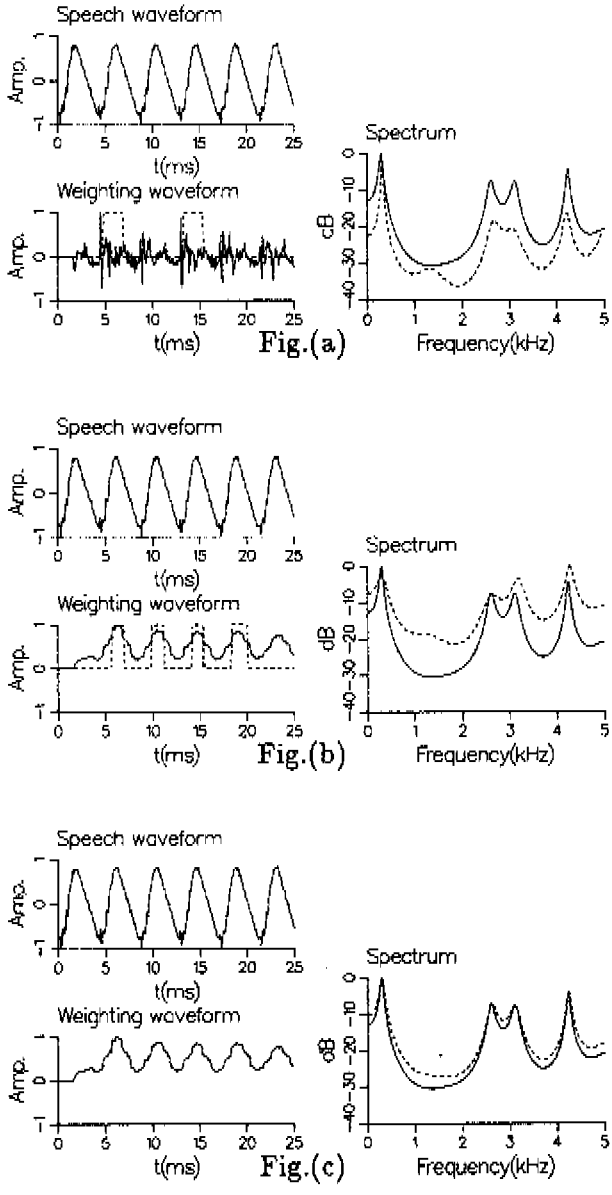


Figure 2.10: Same as Figure 2.7, but for vowel /i/.

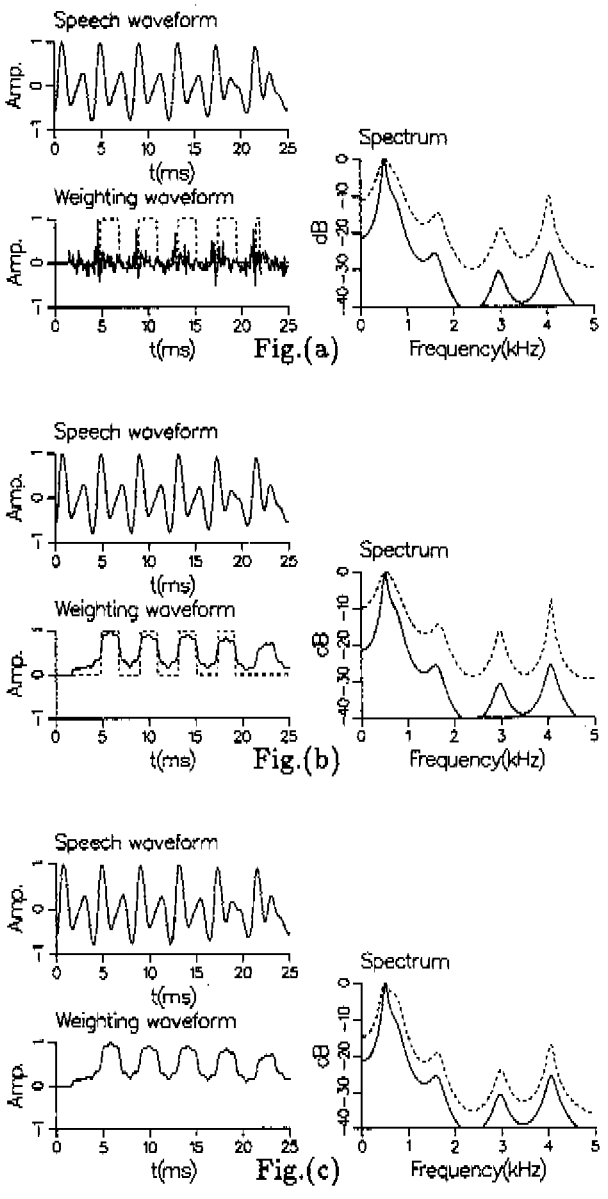


Figure 2.11: Same as Figure 2.7, but for vowel /o/.

are shown in each panel by solid lines in Figs. 2.7- 2.11. The spectral envelopes obtained by the sample-selective, the STE-thresholded and the STE-weighted LPC analyses of order 12 are plotted as dotted lines in panels (a), (b) and (c), respectively. On the left-hand side of each panel the speech waveform is plotted at the top. At the bottom either the LPC residual or the short-time energy is plotted as a solid line and the selection window as a dotted line. In this test a window size of 20 samples was used for the sample-selective method and for the calculation of the short-time energy function, since the average period of the vowel signals is about 50 samples. The threshold value was set at 70% of the peak value of the residual error or the short-time energy function and the sample-selective function was set to be zero in the end portion (30 samples) of the analysis frame. It can be seen from Figs. 2.7- 2.11 that the spectral envelopes of the LPC filters obtained by the sample-selective LPC and both STE-based LPC are quite similar and that the peaks in the spectral envelopes are more prominent than those obtained from the conventional LPC. Also, their estimated formant bandwidths are generally narrower than those from conventional LPC. This is due to the fact that the excitation, which contributes to the widening of the formant bandwidth, is down-weighted in these LPC analyses. One sees that for vowel /e/ in Fig. 2.8 the first formant does not appear from the spectral envelope of the conventional LPC, while the three other LPC solutions give rise to a clear first formant. One also observes from Fig. 2.10 (a) by using the sample-selective function based on this residual that the pulse-like excitation in the LPC residual is not always prominent and that the estimated bandwidth for the first formant is unrealistically narrow. This could be due to the fact that the LPC residual is not well estimated in the LPC analysis of the first step. A low threshold value can be used to avoid missing the selection of speech samples, but the rectangular window will be generally located in a somewhat irregular manner due to the irregularity of the pulse excitation and to erroneous pulses in the LPC residual resulting from the inaccuracy of the first LPC analysis. However, it can be seen that the short-time energy function shows good periodicity just like the speech waveform. Therefore the positions of windows or weighting functions based on the short-time energy appear to be more regular, with the consequence that speech samples with similar positions in each period will be over-weighted. This could be an advantage in the analysis of natural vowels where the LPC model is only approximately valid and where in a period of the LPC residual, there is often more than one pulse or no prominent pulse at all. Obviously these techniques, based on the short-time energy of the signal, are robust in the selection of speech samples and computationally

less expensive. The estimate obtained by the STE-thresholded LPC is as accurate as that derived from the sample-selective method based on the two-step LPC analysis. The estimate of the STE-weighted LPC is somewhat less accurate, but it is quite attractive, taking into consideration the saving in computation time.

2.5 Stability analysis of STE-weighted autocorrelation-based LPC

In this section, a stability analysis for the STE-weighted autocorrelation-based LPC will be presented, although this cannot be performed for the convariance LPC analysis. In the autocorrelation case, the STE-weighted LPC equations (2.4) of section II can be rewritten as

$$Y^T Y(1, \hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p)^T = (E, 0, 0, \ldots, 0)^T \tag{2.8}$$

where

$$Y = \begin{pmatrix} w_1 s_1 & 0 & 0 & \cdots & 0 \\ w_2 s_2 & w_2 s_1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{p-1} s_{p-1} & w_{p-1} s_{p-2} & w_{p-1} s_{p-3} & \cdots & 0 \\ w_p s_p & w_p s_{p-1} & w_p s_{p-2} & \cdots & w_p s_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ w_{N-1} s_{N-1} & w_{N-1} s_{N-2} & w_{N-1} s_{N-3} & \cdots & w_{N-1} s_{N-p} \\ w_N s_N & w_N s_{N-1} & w_N s_{N-2} & \cdots & w_N s_{N-p+1} \\ 0 & w_{N+1} s_N & w_{N+1} s_{N-1} & \cdots & w_{N+1} s_{N-p+2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & w_{N+p} s_N \end{pmatrix} \tag{2.9}$$

and $w_n = \sqrt{W_n}$ (see equation (2.4)). If the columns of Y are denoted by $y_0, y_1, \ldots, y_p$ then one observes that these columns can be generated via the formula

$$y_k = A^{p-k} y_p \quad k = 0, 1, \ldots, p-1 \tag{2.10}$$

with A a constant matrix of order $N + p$ defined as

$$A = \begin{pmatrix} 0 & w_1/w_2 & 0 & \dots & 0 \\ 0 & 0 & w_2/w_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & w_{N+p-1}/w_{N+p} \\ \omega & 0 & \dots & 0 & 0 \end{pmatrix} \tag{2.11}$$

where ω is arbitrary. In fact, it turns out that the value of ω plays no useful role in the following and we will thus put $\omega = 0$. The important consequence of (2.10) is that the zeros of the prediction polynomial $P(z) = \sum_0^p a_{p-k} z^k$ belong to the numerical range of the matrix A (Delsarte et al., 1987). By definition, the numerical range or the field of values $F(A)$ of a square matrix A is the set of complex numbers $\tilde{\eta} A \eta$ for $||\eta|| = 1$, where the tilde denotes the conjugate transpose (Delsarte et al., 1987).

Let us now turn to the actual computation of the numerical range $F(A)$. Following (Delsarte et al., 1987), we observe that $F(A)$ has circular symmetry around the origin. Indeed, in view of the particular form of (2.11) of matrix A, one has

$$\tilde{\eta} A \eta = \sum_{k=1}^{n+p-1} \tilde{\eta}_k \frac{w_k}{w_{k+1}} \eta_{k+1} \tag{2.12}$$

and therefore the substitution $\eta_k \to e^{ik\phi}\eta_k$ transforms $F(A)$ into $e^{i\phi}F(A)$. Consequently, it is sufficient to find the intersection of $F(A)$ with the real axis and this, in turn, is given by the numerical range of the symmetric matrix $D = (A + A^T)/2$. Since D is symmetric it can be diagonalized by some unitary transformation and therefore its numerical range coincides with the interval $[-\lambda_{max}, \lambda_{max}]$ where λ_{max} is the largest eigenvalue of D.

According to (2.11), matrix D is a tridiagonal nonnegative matrix of order N+p with the following expression

$$D = \frac{1}{2} \begin{pmatrix} 0 & w_1/w_2 & 0 & \dots & 0 \\ w_1/w_2 & 0 & w_2/w_3 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & w_{N+p-1}/w_{N+p} \\ 0 & 0 & \dots & w_{N+p-1}/w_{N+p} & 0 \end{pmatrix} \tag{2.13}$$

For the maximal eigenvalue of a nonnegative matrix several upper and lower bounds are known (Minc, 1988), which are all based on the minimal and maximal row (or column) sum of the matrix. Let r_i denote the ith row

sum of matrix $D = (d_{ij})$, that is, $r_i = \sum_{j=1}^{N+P} d_{ij}$. The most straightforward of these bounds is given by

$$\lambda_{max} \leq \max_i\{r_i\} \quad r = 1, 2, \ldots, N + p \tag{2.14}$$

In view of the detailed expression (2.13) of matrix D, we obtain the following result.

Theorem: The zeros of the STE-weighted linear predictor defined by equation (2.1) are all located inside a circle with centre at the origin and with radius $\frac{1}{2}\max_n(w_n/w_{n+1} + w_{n+1}/w_{n+2})$ for $n = 1, 2, \ldots, N + p - 2$.

Using tighter bounds on the maximal eigenvalue of a nonnegative matrix such as provided in (Minc, 1988) gives correspondingly tighter estimates on the location of the zeros of the STE-weighted predictor.

Ideally, we would have liked to show that the predictor polynomial for the STE-weighted LPC is stable, i.e. that all its zeros lie in the open unit disk $|z| < 1$. Although we have not succeeded in doing so, we have derived an upper bound for the modulus of the largest zero. It appears that this upper bound is directly related to the weight ratios w_n/w_{n+1} $(n = 1, 2, \ldots, N + p - 1)$. In particular, if the largest weight ratio is less than or equal to unity then the predictor is proved to be stable. This is the case for linear prediction with exponential forgetting factor (Lee et al., 1981), since the weight ratio is then a constant less than unity. In our case, where the short-time energy of the speech signal over a lagged window is taken as a weighting function, two successive weight factors w_n and w_{n+1} are not significantly different since they represent the STE over windows which are shifted by one sample only. The difference becomes small as the window size increases. In our experiments, it turned out that the weight ratio w_n/w_{n+1} was at most 1.2. According to the Theorem, this gives 1.2 as an upper bound for the predictor zero with largest modulus which is, of course, insufficient to guarantee stability. It may however provide an explanation for the experimental observation that, in practice, the STE-weighted predictors computed according to section II turn out to be almost always stable.

2.6 Conclusion

We have derived a generalized STE-based LPC analysis under the linear least square criterion. The sample selection window or the weighting function in this algorithm are based on the short-time energy of the speech sig-

nal. Their effect is to over-weight the speech samples that fit the LPC model well and to down-weight the others. This novel LPC approach produces less deviating estimates of the formant frequencies than those obtained from the conventional LPC and is less sensitive to the values of the fundamental frequency. From the experimental observations, the STE-thresholded LPC solution is preferable to the sample-selective method based on two-step LPC analyses in terms of computation efficiency and robustness in the selection of speech samples and preferable to the STE-weighted LPC from the viewpoint of estimation accuracy.

Acknowledgement

The authors gratefully acknowledge the constructive comments and valuable suggestions of A.J.M. Houtsma and A. Kohlrausch of IPO.

References

T. V. Ananthapadmanabha and B. Yegnanarayana(1979), "Epoch extraction from linear prediction residual for identification and closed glottis interval", IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-27, pp. 309-319.

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

B.S. Atal and M.R. Schroeder(1978), "Linear prediction analysis of speech based on a pole-zero representation", J. Acoust. Soc. Am., Vol. 64, pp. 1310-1318.

P. Delsarte, Y. Genin and Y. Kamp(1987), " Stability of linear predictors and numerical range of a linear operator", IEEE on Information Theory, Vol.IT-33, pp.412-415.

J.L. Flanagan(1972), *Speech Analysis, Synthesis and Perception* (Springer-Verlag, New York), 2nd edition.

H. Fujisaki and M. Ljungqvist(1986), "Proposal and evaluation of models for the glottal source waveform", Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing, pp.1605-1608.

F. Itakura and S. Saito(1970), "A statistical method for estimation of speech spectral density and formant frequencies", Electron. and Commun., Vol. 53-A, pp.36-43.

H. Kuwabara(1984), "A pitch-synchronous analysis/synthesizer system to independently modify formant frequencies and bandwidth for voiced speech", Speech comm., Vol. 3, pp. 211-220.

J.N. Larar, Y.A. Alsaka and D.G. Childers(1985), "Variability in closed phase analysis of speech", Proc. ICASSP, pp.29.2.1-29.2.4.

Chin-Hui Lee(1988), "On robust linear prediction of speech", IEEE Trans. Acoust., Speech, Signal Processing, Vol.ASSP-36, pp. 642-650.

D.T.L. Lee, M. Morf and B. Friedlander(1981), "Recursive least squares ladder estimation algorithms", IEEE Trans. Acoust., Speech, Signal Processing, Vol.ASSP-29, pp. 627-641.

C. Ma and L.F. Willems(1990), "A generalized sample-selective linear prediction analysis", in *Signal Processing V: Theories and Applications, edited by Torres, T., Masgrau, E., and lagunas, M.A.* (Elsevier Science Publisher), pp.1171-1174.

J. Makhoul(1975), Linear Prediction: A Tutorial Review. Proc. IEEE, Vol. **63**, No. 4, 561-580.

J.D. Markel and A.H. Gray(1970), *Linear Prediction* (The Mouton, The Hague), 2nd edition.

J.D. Markel(1972) "Digital inverse filtering-A new tool for formant trajectory estimation", IEEE Trans. Audio Electroacoust., Vol. AU-20, pp. 129-137.

H. Minc(1988), *Nonnegative matrix*, (John Wiley& Sons, New York).

Y. Miyoshi, K. Yamato, R. Mizoguchi, M. Yanagida and O. Kakusho(1987), "Analysis of speech signals of short pitch period by a sample-selective linear prediction", Trans. IEEE ASSP-35, No.9, pp.1233-1239.

E.N. Pinson(1978), "Pitch-synchronous time-domain estimation of formant frequencies and bandwidths", J. Acoust. Soc. Am., Vol. 35, pp. 1264-1273.

K. Steiglitz and B. Dickinson(1977), "The use of time-domain selection from improved linear prediction", IEEE Trans. Acoust., Speech, Signal Processing, Vol.ASSP-25, pp. 34-39.

Chapter 3

A generalized sample-selective linear prediction analysis *

Abstract

In this paper, we consider the relation between the covariance linear prediction (CLP) analysis of a frame of a speech signal and the CLP analysis of its subframes. The results of CLP analysis derived from a set of subframes are equivalent to those of a residual-weighted CLP analysis of the complete frame and the solutions of the residual-weighted CLP are the same as those of the generalized weighted average of subframe CLP. A generalized sample-selective CLP analysis is proposed. Those subframes which best reflect the filter model of the speech production can be chosen to improve the accuracy of the estimate of the LPC parameters.

3.1 Introduction

THE process of speech production can be simplified as a source-filter model [1]. The filter can be characterized by an all-pole model represented by the linear prediction equations [2][3]. For voiced sounds, the source is situated in the vibrating vocal folds which modulate the air flow from the lungs. We refer to the vibration frequency of the vocal folds as the fundamental frequency. The unvoiced sound source consists of the turbulent flow formed somewhere in the constricted vocal tract .

In estimating the predictor coefficients, the methods of autocorrelation linear prediction (ALP) and covariance linear prediction (CLP) analysis have become very important. The CLP, in particular, is often used for very short segments of sampled data, for instance in pitch synchronous analysis and closed-glottis-period analysis. When the analysis

*Paper with L.F. Willems published in *Signal Processing V: Theories and Applications*, edited by Torres, T., Masgrau, E., and lagunas, M.A. (Elsevier Science Publisher), pp.1171-1174, 1990.

window is quite wide, for example, covering more than two pitch periods, the performance of CLP is close to that of ALP, but that is not the case for very short segments of sampled data. In order to give a better description of the process of speech production, researchers have paid much attention to the fine structure of formants by means of very short window CLP analysis, or an analysis of only the excitation-free portions, such as the closed glottis portion, to estimate the parameters of the linear prediction model of speech production. But it is not always easy to choose those excitation-free portions, for example, in voiced sounds uttered by females or children, because the pitch period is short. The results are dependent on the data available in the pitch period and are sensitive to the window position [5]. The estimation accuracy of the parameters can be improved by sample selective linear prediction (SSLP) [4], proposed by Yoshiaki Miyoshi et al. In the following sections we shall show that SSLP is a special form of the generalized weighted average CLP analysis.

In practice it is always preferable to obtain an accurate estimate of the LPC parameters so that the source and the filter can be well separated. One example is the glottal inverse filtering technique, which derives the glottal pulse from the speech signal. Improving the estimate of the LPC parameters is one of the main goals in speech processing. The signal of a voiced sound is quasi-periodic. The differences between the successive pitch periods are due to noise or other factors from the glottal source. The pitch synchronous LPC analysis does not benefit from the correlation of the successive pitch periods. However, there always exists some more or less excitation -free portion which best reflects the parameters of the filter model. We cannot use these portions to do pitch synchronous analysis separately, but we can use them in combination to obtain a good estimate of LPC parameters.

In section 2 we shall present the relation between the results of the residual-weighted CLP of a frame and that of its subframe CLP. In section 3 the relation between the frame CLP and the subframe CLP is given. In section 4 a generalized sample-selective CLP method for speech analysis is discussed.

The conclusions are that 1) the results of CLP analysis derived from a set of subframes are equivalent to those of the residual-weighted CLP analysis of the whole frame and 2) the solutions of the residual-weighted CLP are the same as those of the generalized weighted average CLP of the individual subframes. From this we obtain a generalized sample-selective CLP analysis to improve the estimate of LPC parameters.

3.2 The Residual-weighted CLP of a frame and its subframe CLP

The speech production model can be generally described by the following equations:

$$s_n = \sum_{i=1}^{p} s_{n-i} a_i + e_n \tag{3.1}$$

where s_n denotes the nth sample of a speech wave, e_n is the nth sample of an excitation wave, and a_i the ith predictor coefficient. CLP analysis is based on the minimization of the following sum of squared prediction residuals,

$$E = \sum_{n=n1}^{n2} (s_n - \sum_{i=1}^{p} s_{n-i} a_i)^2 \tag{3.2}$$

For the sake of simplicity we use a matrix form to represent it. The prediction equations and the error for the CLP are therefore as follows

$$\begin{pmatrix} s_{n1-1} & s_{n1-2} & \cdots & s_{n1-p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{n2-1} & s_{n2-2} & \cdots & s_{n2-p} \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} s_{n1} \\ s_{n1+1} \\ \vdots \\ s_{n2} \end{pmatrix} \tag{3.3}$$

and

$$E = (S\mathbf{a} - \mathbf{s})^T (S\mathbf{a} - \mathbf{s}) \tag{3.4}$$

S and s stand for the left-hand matrix and the column matrix of the s_n, respectively, and T represents matrix transpose; $\mathbf{a}^T = (a_1, a_2, \ldots, a_p)$. The Least Square Solution of (3.3) is

$$\mathbf{a} = (S^T S)^{-1} S^T \mathbf{s} \tag{3.5}$$

The equation above is the normal CLP analysis for a frame of signal samples running from $n1 - p$ to n2.

We now choose a window W with a frame length $n2 - n1 + p + 1$ and some subwindows W_k running from $bk - p$ to ek ($bk \geq n1$ and $ek \leq n2$). This is illustrated in Fig. 3.1. For each subframe W_k, we obtain a set of prediction equations. Putting all subframe equations together, we obtain what we shall call the residual-weighted CLP equations. In this case the total energy of the residual error can be represented by

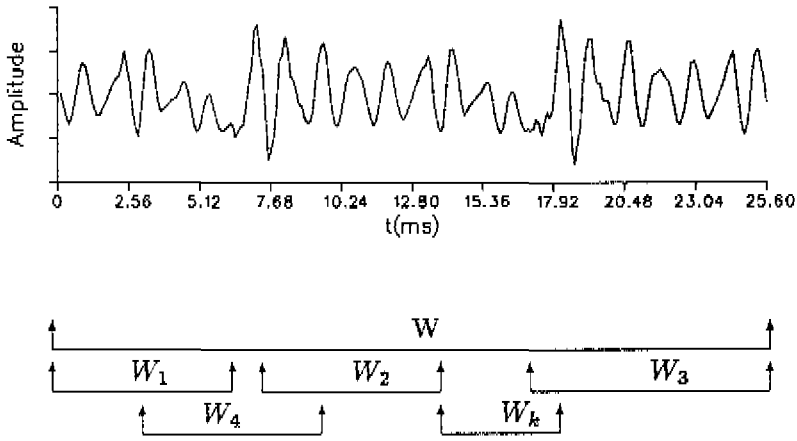


Figure 3.1: Top: a speech signal. Bottom: an illustration of how the subframes are chosen.

$$\hat{E} = \sum_{n=n1}^{n2} A_n(s_n - \sum_{i=1}^{p} s_{n-i}\hat{a}_i)^2 \tag{3.6}$$

where A_n is the number of times that the prediction equation $s_n = \sum_{i=1}^{p} s_{n-i}a_i$ appears in the set of CLP equations.

In order to relate the predictor $\hat{\mathbf{a}}$ which minimizes (3.6) to the solution (3.5), we construct the augmented matrix $(\hat{S}, \hat{s})$ from the augmented matrix (S,s). This matrix is constructed such that the k-th prediction equation is explicitly represented A_n times. It is easy to prove that the Least Squares solution of (3.3) will be

$$\hat{\mathbf{a}} = (\hat{S}^T\hat{S})^{-1}\hat{S}^T\hat{\mathbf{s}} \tag{3.7}$$

or

$$\hat{\mathbf{a}} = (S^TQ^TQS)^{-1}S^TQ^TQ\mathbf{s} \tag{3.8}$$

where $\hat{\mathbf{a}}^T = (\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_p)$, and Q^TQ is a $(n2 - n1 + 1) \times (n2 - n1 + 1)$ diagonal matrix containing the number of times that the predictor appears, which depends on how we choose the subwindows.

From equation (3.6)-(3.8), we can see that the results of CLP analysis derived from a set of subframes are equivalent to those obtained by weighting the residual error function e_n of the whole frame with a window in which the amplitude is the element $Q(n,n)$ of matrix Q.

3.3 The frame CLP and the subframe CLP

The following equations are derived from the augmented matrix which is partitioned according to row.

$$\begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_M \end{pmatrix} \begin{pmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \vdots \\ \hat{a}_p \end{pmatrix} = \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \vdots \\ \mathbf{s}_M \end{pmatrix} \tag{3.9}$$

In the above equation

$$S_k = \begin{pmatrix} s_{bk-1} & s_{bk-2} & \cdots & s_{bk-p} \\ \vdots & \vdots & \vdots & \vdots \\ s_{ek-1} & s_{ek-2} & \cdots & s_{ek-p} \end{pmatrix} \tag{3.10}$$

$$\mathbf{s}_k = \begin{pmatrix} s_{bk} \\ s_{bk+1} \\ \vdots \\ s_{ek} \end{pmatrix} \tag{3.11}$$

Because every submatrix S_k and $\mathbf{s}$ represent a subframe CLP analysis in which the speech samples are from $bk - p$ to ek ($bk \geq n1$ and $ek \leq n2$), we can rewrite equation (3.7) as follows:

$$\begin{pmatrix} S_1^T & S_2^T & \ldots & S_M^T \end{pmatrix} \begin{pmatrix} S_1 \\ S_2 \\ \vdots \\ S_M \end{pmatrix} \hat{\mathbf{a}}$$

$$= \begin{pmatrix} S_1^T & S_2^T & \ldots & S_M^T \end{pmatrix} \begin{pmatrix} \mathbf{s}_1 \\ \mathbf{s}_2 \\ \vdots \\ \mathbf{s}_M \end{pmatrix} \tag{3.12}$$

From the product of two partitioned matrices we have

$$\sum S_k^T S_k \hat{\mathbf{a}} = \sum S_k^T \mathbf{s}_k \tag{3.13}$$

Each individual subframe analysis, on the other hand, has a solution $\mathbf{a}_k$ given by

$$S_k^T S_k \hat{\mathbf{a}}_k = S_k^T \mathbf{s}_k \tag{3.14}$$

where $\mathbf{a}_k$ are the prediction coefficients obtained from the analysis of the k-th subframe. Of course, for each different subframe CLP analysis we find different $\mathbf{a}_k$ coefficients. Comparing the above two formulas, we have

$$\sum S_k^T S_k \hat{\mathbf{a}} = \sum S_k^T S_k \mathbf{a}_k \tag{3.15}$$

That is

$$\hat{\mathbf{a}} = \left(\sum S_k^T S_k\right)^{-1} \left(\sum S_k^T S_k \mathbf{a}_k\right) \tag{3.16}$$

It is obvious that the solutions of the residual-weighted CLP are a generalized weighted average of the solutions to the individual subframe CLP. We call this the generalized weighted average CLP. The covariance matrix $\sum S_k^T S_k$ is the weighting factor. The error can be calculated by

$$\hat{E} = \sum \mathbf{s}_k^T \mathbf{s}_k - \sum \mathbf{s}_k^T S_k \hat{\mathbf{a}} \tag{3.17}$$

Note that the $\hat{\mathbf{a}}$ coefficients are from either the residual-weighted CLP analysis (3.8) or the windowed signal (3.16).

3.4 The generalized sample-selective CLP

So far we have arrived at the relation (3.16) between the frame CLP and its subframe CLP. The generalized sample-selective CLP will be discussed in this section.

It is useful to analyze a frame of the speech signal with a group of subframes CLP analysis in which only the $\mathbf{a}_k$ coefficients with a small excitation influence are kept. That amounts to an analysis of the speech signal by a generalized sample-selective linear prediction. We can see that SSLP is just a special form of this generalized weighted average CLP. The influences of the excitation are included in the analysis frame in SSLP[4]. The generalized weighted average CLP gives us more freedom to choose several subframes to compensate for the scarcity of data and to reduce the noise influence. We can, for example, choose those subframes which do not include any excitation influence. This subframe scheme was also used by P. D. Welch to estimate power spectra in the nonstationary case [7]. The choice of subframes is related to the model of speech production. We will discuss this in the following part.

It is noted that there is a strong correlation from pitch period to pitch period in voiced sounds (exceptions are the voice onset and offset portions). To take advantage of this correlation we can choose subwindows

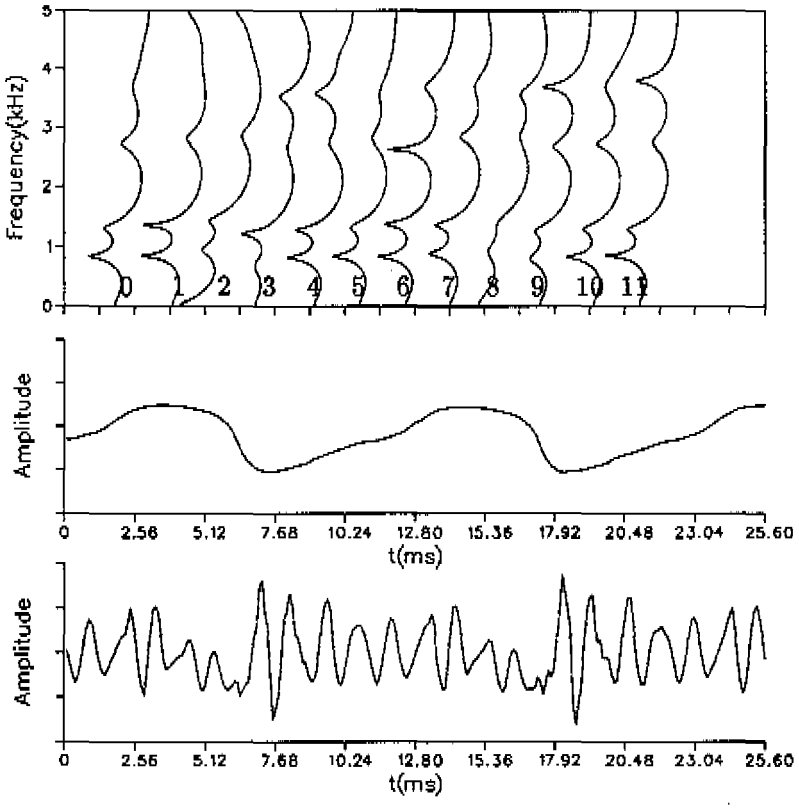


Figure 3.2: Top: spectrum of the CLP of a 40-sample subframe. Middle: electroglottogram. Bottom: speech signal. The electroglottogram and speech signal are plotted on the same time scale. In the top panel curve 1 corresponds to the result of the CLP of data in the window from 0 ms to 4 ms and curve 2 to the result in the window from 2 ms to 6 ms, and so on. Curve 0 is the generalized average of curves 6 and 11.

so that they just cover the region where the excitations are relatively small. The information for determining the formants has to be taken from these subwindows, and the result can be optimally obtained from the generalized average of these subframe CLPs.

From Fig. 3.2, we can see how formants change according to subframe position. The electroglottogram indicates the status of the vocal folds; the spectral curves are numbered from 0 to 11. When the window contains the main excitation the results are unacceptable, for instance, those

indicated by curves 3 and 8. When the window is located in the closed-glottis portion, good formant estimates are obtained, as illustrated by curves 6 and 11. The analysis conditions for this experiment were as follows. The speech signal was sampled at 10 kHz. The pre-emphasis parameter was -0.9 and the window length was 40 samples. The window was moved forward 20 samples every time. Formant frequencies in the open-glottis portions deviate from those in the closed-glottis regions. Dividing every pitch period approximately into an open-glottis portion and a closed-glottis portion, we just analyze the data in the closed-glottis portion and calculate the average according to equation (3.16). Due to the correlation between pitch periods the averaging process can also reduce some noise influence. To take an example, curve 0 in Fig. 3.2 shows the results from the average for the windows corresponding to spectral curves 3 and 10. As can be seen, an estimation of the formant parameters which fits the speech production model better is obtained.

From the above discussion, we know that we have more freedom in the choice of subframe data than with the SSLP, in addition, the relationship between long-frame CLP , short-frame CLP and the generalized weighted-average CLP is now established.

The conclusions are 1) the results of CLP analysis derived from a set of subframes are equivalent to those of the residual-weighted CLP analysis of the whole frame and 2) the solutions of the residual-weighted CLP are the same as those of the generalized weighted-average CLP of the individual subframe. From this we obtained a generalized sample-selective CLP analysis to improve the estimate of LPC parameters.

Acknowlegement

The constructive criticism and the helpful comments of W. Verhelst and R. Veldhuis are gratefully acknowledged. The authors thank Berry Eggen for providing the speech material and for helpful discussions.

References

[1] J.L. Flanagan, Speech Analysis, Synthesis and Perception. 2nd ed. New York: Springer-Verlag, 1972.

[2] J.D. Markel and A.H. Gray, Linear Prediction. 2nd ed. The Hague, The Netherlands:Mouton, 1970.

[3] J. Makhoul, Linear Prediction: A Tutorial Review. Proc. IEEE, Vol. 63, No.4, pp.561-580, 1975.

[4] Yoshiaki Miyoshi, et al., Analysis of Speech Signals of Short Pitch Period by a Sample-Selective Linear Prediction. *Trans. IEEE ASSP-35, No.9*, pp.1233-1239, 1987.

[5] J.N. Larnar, Y.A. Alsaka, and D.G. Childers, Variability in Closed Phase Analysis of Speech. *Proc. ICASSP*, pp.29.2.1-29.2.4, 1985.

[6] J.N. Holmes, Requirements for Speech Synthesis in the Frequency Range 3-4 kHz. F.A.S.E. Symposium on Acoustics and Speech, Venice, Vol.1, pp.169-172, 1981.

[7] P.D. Welch, The Use of Fast Transform for the Estimation of Power Spectra: A Method Based on Time Averaging Over Short, Modified Periodograms. In: Modern Spectrum Analysis Ed. by D.G. Childers, IEEE Press, pp.17-20, 1978.

Chapter 4

A singular value decomposition approach to glottal closure determination from the speech signal *

Abstract

The detection of glottal closure instants has been a necessary step in several applications of speech processing, such as speech coding, speech prosody manipulation and speech synthesis. The proposed methods to date, in particular, the methods proposed by Strube and by Wong et al. are deficient in some aspects. Speech processing needs efficient and robust glottal closure detection methods. In this paper, we propose to use the singular value decomposition (SVD) approach to detect the glottal closure instant from the speech signal. The proposed SVD method amounts to calculate the Frobenius norms of signal matrices and therefore is computationally efficient. Moreover, it produces well-defined and reliable peaks that indicate the glottal closure instants. Finally, with the introduction of the total least squares technique, the two methods proposed by Strube and Wong are reinvestigated and unified into the SVD framework.

Introduction

THE process of speech production can be simply described by a source-filter model (Flanagan, 1972). The filter can be characterized as linear (Markel and Gray, 1970; Makhoul, 1975). For voiced sounds, the source is situated in the vibrating vocal folds which modulate the air flow from the lungs and produce glottal pulses. The unvoiced

*Paper with Y. Kamp and L.F. Willems submitted for publication to J. Acoust. Soc. Am.

sound source consists of the turbulent flow formed somewhere in the constricted vocal tract.

Present speech research shows a great interest in analyzing the voice sound period by period over an interval which is delimited by two successive instants of glottal closure. For the sake of simplicity, we call the instants of glottal closure in the speech signal the epochs. Determination of the epochs plays an important part in applications, such as in inverse glottal pulse analysis to extract speaker characteristics (Hedelin, 1984; Kuwabara, 1984; Eggen, 1989), prosody manipulations of speech sounds by means of the PSOLA technique (Moulines and Di Francesco, 1990), and speech synthesis and speech coding (Hedelin, 1984; Eggen, 1989).

During the past decades, several epoch detection methods have been proposed for the speech signal. One such method is to detect the discontinuities of the differentiated speech signal (Ananthapadmanabha and Yegnanarayana, 1975). It is a simple and effective technique for very clean vowels with sharp glottal closures, but as it is a high-pass filter operation, it is thereby understandably sensitive to the noise excitations in sounds like voiced frictave and contaminating noise. Epoch detection based on the residual signal of the LPC analysis, as is described by Ananthapadmanabha and Yegnanarayana (Ananthapadmanabha and Yegnanarayana, 1979), also cannot produce reliable results, because the LPC inverse filtering can give rise to significant pulses or predictive errors in the residual other than those on the excitation moments. Moreover, the separation of the source and the system by using the popular LPC method is strongly influenced by the shape of the glottal pulse and its repetition rate, and therefore does not work well in some cases such as in female and children's voice sounds. Due to errors in the LPC analysis there often is more than one impulse at a closure instant in the LPC residuals. Then it often fails to produce accurate epochs by detecting those impulses in the residual signals.

At present, the two following methods are better known because they can produce a reliable glottal closure detection. The first one is proposed by Strube to calculate determinants of the autocovariance matrices, which can produce satisfactory detection of the epochs. However, it cannot easily be normalized (Strube, 1974). The second approach, proposed by Wong, Markel and Gray (1979), directly makes use of the speech production model with a clearly defined glottal pulse. Here the epoch is defined as the minimum of the total LPC residual energy calculated from rather short analysis frames. Unfortunately, total LPC

residual energy tends to be noisy and therefore needs to be smoothed which implies some loss in location resolution.

Another group of related techniques is based on the analysis of a long speech segment with the aim of determining the length of the pitch period but not determining epochs. Among these techniques are the pitch detection algorithms AMDF (Ross, et al. 1974), SIFT (Markel, 1972), DWS (Duifhuis, et al. 1982) and SHS (Hermes, 1988), Discussion of these algorithms, however, is out of the scope of this paper.

Many epoch detection methods, among which are the two important methods of Strube and Wong et al., are, in essence, based on the idea that the linear prediction model fits better and, consequently, its prediction error is smaller within a short segment (less than one pitch period) of the speech signal which contains no excitations (Strube, 1974; Wong, Markel, and Gray, 1979; Cheng and O'Shaughnessy, 1989; Moulines and Di Francesco, 1990). When the instant of glottal closure or main excitation is included in the data segment, the linear prediction model does not fit the data well and the prediction error will be large. These large prediction errors are indications of the glottal closure instants.

The main contribution of this investigation is to establish a framework of the epoch detection, to compare the results from different approaches and, finally, to propose a new singular value decomposition (SVD) approach to the epoch detection problem. This approach leads to a better formulation and has clear advantages over the two abovementioned methods, as it is computationally very efficient and robust against noise. The resulting measure has a dimension of energy and can be easily normalized and thresholded. We are also able to show the relationship between Strube's method, Wong's method and our SVD approach and the advantages of the latter.

In the next section, we introduce our approach more explicitly with a brief description of the notions of the singular value decomposition (SVD) technique and the linear least squares, and present the Total Linear Least Squares (TLLS) approach. In the third section, we propose our new SVD-based approach for epoch detection. Finally, our method is compared with two others, and examples are given.

4.1 The SVD as unifying framework for epoch detection

Epoch detection has often been based on a source-filter model of the speech production. In either parametric or statistical approaches, the all-pole system assumption is usually made (Cheng and O'Shaughnessy,

1989; Moulines and Di Francesco, 1990). The source of the system is assumed to have an open glottis portion and a closed glottis portion in each pitch period of a voiced sound. The rate of transition from the closed to the open glottis portion is much slower than that from the open to the closed glottis portion and thus the main excitation occurs at the instant of the glottal closure. The differentiation of the main excitation results in a very sharp pulse at the instant of glottal closure. Epoch determination from the speech signal is based on the fact that there is strong and abrupt change of the glottal flow at the instant of glottal closure. A vocal tract is approximately a time invariant linear system over a short duration of time. When the system parameters are well estimated, its excitation should be small in the closed glottis regions and large at glottal closure instants.

Therefore, the amount of deviation from the linear prediction is a primary criterion used in different epoch detection approaches. The largest deviation is expected to happen at the glottal closure instant. The question , however, how to extract the linear predictability or how to identify the linear relation from the speech signal has a significant influence on the quality of the detection schemes. Moreover, speech sounds are dynamic in nature and the source-filter model of the speech production is inevitably accompanied by the presence of unknown disturbances, parameter variations and other uncertainties. Therefore, the linear model will only hold approximately and the solution will depend on the error criterion used. In practice, a particular solution is obtained by imposing additional constraints on the problem , such as least squares, maximum likelihood or $l1$-norm and accordingly, a variety of estimation schemes are utilized. Among the most popular estimation schemes for linear relation from noisy data, are the Linear Least Squares (LLS) and the Total Linear Least Squares (TLLS) schemes. As we shall see, the SVD method can provide a unifying framework in identifying linear relations from data and it makes the formulation of the problem explicit and guarantees the robustness of the numerical solutions. In these estimation schemes, the data matrix or measurement data are in fact modified to meet linear relations imposed on the data. In other words, the data matrix is decomposed into the sum of two matrices, one of which consists of the linear dependent column vectors and another consists of error elements. The constraints mentioned above are used because they approximately meet the physical requirements of the problem and produce a tractable mathematical solution.

4.1.1 Singular Value Decomposition(SVD)

The SVD method has been used in several applications of digital signal processing (Vandewalle and De Moor, 1988). The SVD of a certain data matrix, allows a particularly robust separation of signal and noise and is very effective in dealing with noisy data (Vandewalle and De Moor, 1988).

Consider a sequence of measurements or observation vectors, consisting of segments of a speech signal, obtained by advancing a rectangular window of length p+1 samples one sample further successively. The following data matrix can then be formed :

$$S = \begin{pmatrix} s_{p+1} & s_p & s_{p-1} & \cdots & s_1 \\ s_{p+2} & s_{p+1} & s_p & \cdots & s_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{p+m} & s_{p+m-1} & s_{p+m-2} & \cdots & s_m \end{pmatrix} \tag{4.1}$$

We shall assume that $m \geq p+1$ and that the data matrix has full column rank, i.e. p+1. Under these assumptions, it is well known (Golub and Van Loan, 1983) that there exist orthogonal matrices

$$U = (\bar{u}_1, \bar{u}_2, ..., \bar{u}_m)$$

$$V = (\bar{v}_1, \bar{v}_2, ..., \bar{v}_{p+1}),$$

such that

$$S = \Sigma_{i=1}^{p+1} \sigma_i \bar{u}_i \bar{v}_i{}^t \tag{4.2}$$

where

$$U^t U = V^t V = V V^t = I_{p+1},$$

and

$$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_{p+1} > 0,$$

where σ_i are called the singular values, the superscript t denotes matrix transposition and I_{p+1} is an identity matrix of order $p+1$. The column vector $\bar{u}_i$ of the matrix U is a normalized eigenvector associated with the eigenvalue σ_i^2 of matrix SS^t. In the same manner the column vector $\bar{v}_i$ is a normalized eigenvector associated with the eigenvalue σ_i^2 of matrix S^tS. Equation (4.2) is called the singular value decomposition (SVD).

It is clear that the SVD method decomposes a data matrix into the sum of (p+1) rank one matrices. The matrix S^tS is the autocovariance matrix of the speech signal and its determinant can be rewritten as

$$det(S^tS) = \Pi_{i=1}^{p+1}\sigma_i^2. \tag{4.3}$$

Moreover, the Frobenius norm of a matrix $S = \{s_{ij} : 1 \leq i \leq m, 1 \leq j \leq p+1\}$ is defined as

$$\|S\|_F = (\Sigma_{i=1}^{m}\Sigma_{j=1}^{p+1}s_{ij}^2)^{1/2} \tag{4.4}$$

and it is known (Golub and Van Loan, 1983) that it can be expressed in terms of the singular values as

$$\|S\|_F = (\sum_{i=1}^{p+1}\sigma_i^2)^{1/2}. \tag{4.5}$$

Hence, the Frobenius norm of S is the square-root of the sum of its squared singular values.

4.1.2 Strube's method for epoch detection

Let $\bar{s}_i$ denote the *ith* column vector of matrix S. In the absence of excitation, the linear filter model of order p imposes a linear dependence between the vectors $\bar{s}_1$, $\bar{s}_2$, ..., $\bar{s}_{p+1}$. Consequently, the determinant of the matrix S^tS as a function of time will increase sharply when the data matrix contains an impulse excitation and it will decrease (ideally it should become zero) when the data matrix does not contain any impulse excitation. Therefore, the determinant value can be used as a way to detect the location of epochs in the signal. This is, in essence, Strube's method for the detection of epochs (Strube, 1974), which in view of (4.3) is equivalent to computing the product of all squared singular values of matrix S. The Cholesky factorization of S^tS provides however an efficient recursive scheme to actually perform this calculation (Strube, 1974).

4.1.3 Wong's approach to epoch detection and LLS

The source-filter model used for linear predictive coding (LPC) is based on the assumption that the vocal tract can be approximated by an all-pole filter of order p. Accordingly, the first column of the data matrix S in (4.1) is assumed to be a linear combination of the other columns and any deviation from this particular linear dependence is attributed to the excitation produced by the source. This viewpoint is expressed by the following set of equations:

$$S \begin{pmatrix} 1 \\ -a_1 \\ -a_2 \\ \vdots \\ -a_p \end{pmatrix} = \begin{pmatrix} e_{p+1} \\ e_{p+2} \\ e_{p+3} \\ \vdots \\ e_{p+m} \end{pmatrix} \tag{4.6}$$

where e_n is the nth sample of the glottal excitation wave, and a_i the ith predictor coefficient. The least squares solution of the equations above can be obtained from

$$S^t S \begin{pmatrix} 1 \\ -a_1 \\ -a_2 \\ \vdots \\ -a_p \end{pmatrix} = \begin{pmatrix} E_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{4.7}$$

where $E_1 = \sum_{i=p+1}^{m+p} e_i^2$ is the LPC residual energy which can be computed by

$$E_1 = \frac{det(S^t S)}{det(S^t S)_{11}} \tag{4.8}$$

where $(S^t S)_{11}$ denotes the principal submatrix obtained by removing the first row and column in matrix $S^t S$. The epoch detection proposed by Wong, Markel and Gray (Wong, Markel, and Gray, 1979) is essentially based on the minimum of this normalized residual energy E_1. In practice, the LPC residual energy E_1 is sequentially calculated from the speech samples covered by a short analysis window. When the analysis window advances through a glottal pulse, the residual energy will increase first and then sharply decrease (in principle to zero) when the window just leaves the glottal pulse. However, these minima may not be well defined in real speech due to the fact that the LPC model does not perfectly fit the speech samples and the speech samples are corrupted by noise. Owing to the poor prediction of the vocal tract resonances, the residual does not become zero after the glottal pulse and the minima may not correspond to the instants of the glottal closures. This has been demonstrated by Kuwabara (1984) and Larar et al. (1985), and further discussion about this method will be developed in the following sections.

Finally, let us also observe that the residual energy can also be expressed in terms of the SVD of matrix S. Indeed equation (4.7) yields

$$\begin{pmatrix} 1 \\ -a_1 \\ -a_2 \\ \vdots \\ -a_p \end{pmatrix} = (S^t S)^{-1} \begin{pmatrix} E_1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{4.9}$$

and from the SVD of S one obtains

$$(S^t S)^{-1} = \Sigma_{i=1}^{p+1} \sigma_i^{-2} \bar{v}_i \bar{v}_i^t$$

which finally gives

$$E_1 = \frac{1}{\sum_{k=1}^{p+1} \frac{v_{1k}^2}{\sigma_k^2}}, \tag{4.10}$$

where v_{1k} are the elements of the first row of the matrix V. If the smallest singular value σ_{p+1} is significantly smaller than the others, the equation above is approximately

$$E_1 \approx \frac{\sigma_{p+1}^2}{v_{1,p+1}^2} \tag{4.11}$$

The LPC solution is in fact a particular case in a whole family of estimation schemes for linear relation between noisy data. Indeed, p other estimations can be conceived, similar to LPC, but where each column of S in turn is considered to be a linear combination of the p remaining columns. This set of estimations is known as the Linear Least Squares (LLS) family (Vandewalle and De Moor, 1988). Let E_i denote the residual energy of the *ith* LLS solution where the column i of the data matrix S is considered to be a linear combination of the other columns. In view of the singular value decomposition (4.2) of S one has

$$E_i = \frac{1}{\sum_{k=1}^{p+1} \frac{v_{ik}^2}{\sigma_k^2}}, \tag{4.12}$$

and expression (4.10) corresponds thus to the case $i = 1$ in the LLS family.

4.1.4 Total Linear Least Squares (TLLS)

Each of the LLS solutions considered above can be looked upon as a modification of the original data matrix S such that a rank reduction

from p+1 to p results. For instance, the LPC (or 1st LLS) equations (4.6) can be rewritten as

$$\begin{pmatrix} s_{p+1} - e_{p+1} & s_p & s_{p-1} & \cdots & s_1 \\ s_{p+2} - e_{p+2} & s_{p+1} & s_p & \cdots & s_2 \\ s_{p+3} - e_{p+3} & s_{p+2} & s_{p+1} & \cdots & s_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ s_{p+m} - e_{p+m} & s_{p+m-1} & s_{p+m-2} & \cdots & s_m \end{pmatrix} \begin{pmatrix} 1 \\ -a_1 \\ -a_2 \\ \vdots \\ -a_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{4.13}$$

With the equation under this form, one can thus consider that the LPC solution achieves the rank reduction by modifying only the first column of S while all other columns remain unchanged (Vandewalle and De Moor, 1988). Similarly, the *ith* LLS solution can be interpreted as modifying only the *ith* column. In speech signals, however, all data in the matrix S could be contaminated by noise or deviations from the model. In addition, the same elements in the first column of the matrix S occur in other columns as well and should therefore also be changed. Even in pitch synchronous speech analysis (Kuwabara, 1984; Eggen, 1989), the closed-glottis portions of speech, which are often considered excitation free, deviate from the linear model because of noise and nonlinearity. Therefore it may be unrealistic to modify one column only in order to fit the linear production model and it would be more reasonable to modify all elements of the matrix S. This is the point of view adopted by the Total Linear Least Squares (TLLS) which modifies all data columns. In other words, every element of the data matrix can be changed or perturbed in order to fit the linear relation model.

Let $\hat{S}$ be a perturbed matrix and the $\|S - \hat{S}\|_F$ be the perturbation energy. The TLLS solution to the linear relation model is then obtained by modifying matrix S into $\hat{S}$ such that the following set of equations

$$\sum_{i=0}^{p} \hat{s}_{n-i}\alpha_i = 0 \tag{4.14}$$

where index n runs from $p+1$ to $m+p$, or equivalently

$$\begin{pmatrix} \hat{s}_{p+1} & \hat{s}_p & \hat{s}_{p-1} & \cdots & \hat{s}_1 \\ \hat{s}_{p+2} & \hat{s}_{p+1} & \hat{s}_p & \cdots & \hat{s}_2 \\ \hat{s}_{p+3} & \hat{s}_{p+2} & \hat{s}_{p+1} & \cdots & \hat{s}_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \hat{s}_{p+m} & \hat{s}_{p+m-1} & \hat{s}_{p+m-2} & \cdots & \hat{s}_m \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \tag{4.15}$$

is exactly solvable and the perturbation energy is minimized. We thus want to find $\hat{S}$ such that

$$rank(\hat{S}) \leq p$$

and

$$||S - \hat{S}||_F$$

is minimized. The solution to this problem is well known (Golub and Van Loan, 1983) and obtained by deleting in (4.2) the contribution of the smallest singular value, i.e. σ_{p+1}, assuming for simplicity that $\sigma_p > \sigma_{p+1}$. Thus

$$\hat{S} = \Sigma_{i=1}^{p} \sigma_i \bar{u}_i \bar{v}_i^t. \tag{4.16}$$

The perturbation error $||S - \hat{S}||_F = \sigma_{p+1}^2$, and the linear model is $\bar{v}_{p+1}$ since, in view of the orthogonality of the vectors $\bar{v}_i$, the following holds:

$$\hat{S}\bar{v}_{p+1} = 0. \tag{4.17}$$

Finally, the residual error signal is given by

$$S\bar{v}_{p+1} = \sigma_{p+1}\bar{u}_{p+1}. \tag{4.18}$$

If, instead of the approach described above, we delete from (4.2) the contribution of a different singular value $\sigma_i \neq \sigma_{p+1}$, then, we obtain a different perturbed matrix $\hat{S}_i$, namely,

$$\hat{S}_i = \Sigma_{j \neq i} \sigma_j \bar{u}_j \bar{v}_j^t$$

for which a different linear model holds, i.e.

$$\hat{S}_i \bar{v}_i = 0,$$

but with a relatively larger perturbation error $||S - \hat{S}_i|| = \sigma_i^2$. Each singular value thus measures the deviation from some corresponding linear model and the sum $\frac{1}{(p+1)} \sum_{i=1}^{p+1} \sigma_i^2$ can be considered as an "average" of the deviation from any linear model.

4.2 The new epoch detection based on SVD

The new criterion for epoch detection proposed here is the arithmetic mean of the squared singular values, namely, $C = \frac{1}{(p+1)} \sum_{i=1}^{p+1} \sigma_i^2$. Although it does not seem possible to establish a rigorous and direct connection between the numerical value of C and the instant of glottal

closure, the theoretical argument given below shows why this criterion makes sense. Additional support is provided by a series of experiments presented at the end of this section and further developed in section IV. These experimental results show that the maxima of C indeed nicely correlate with the instants of glottal closure.

We start from expression (4.12) for the residual energy of the *ith* LLS solution which is rewritten as

$$\frac{1}{E_i} = \frac{1}{\sigma_i^2}\sum_{k=1}^{p+1} v_{ik}^2 \frac{\sigma_i^2}{\sigma_k^2}, \tag{4.19}$$

Since $\frac{\sigma_i^2}{\sigma_k^2} \leq \frac{\sigma_1^2}{\sigma_{p+1}^2}$ one has

$$\frac{1}{E_i} \leq \frac{1}{\sigma_i^2}\frac{\sigma_1^2}{\sigma_{p+1}^2}\sum_{k=1}^{p+1} v_{ik}^2, \tag{4.20}$$

and, in view of the orthogonality of the matrix V which implies $\sum_{k=1}^{p+1} v_{ik}^2 = 1$, the latter inequality reduces to

$$\sigma_{p+1}^2 \leq \sigma_i^2 \leq \frac{\sigma_1^2}{\sigma_{p+1}^2} E_i. \tag{4.21}$$

On the other hand, the inequality between geometric and arithmetic means yields

$$(\Pi_{i=1}^{p+1}\sigma_i^2)^{\frac{1}{p+1}} \leq \frac{1}{p+1}\sum_{i=1}^{p+1}\sigma_i^2. \tag{4.22}$$

Considering (4.21) and (4.22) we finally get

$$(\Pi_{i=1}^{p+1}\sigma_i^2)^{\frac{1}{p+1}} \leq C = \frac{1}{p+1}\sum_{i=1}^{p+1}\sigma_i^2 \leq \frac{\sigma_1^2}{\sigma_{p+1}^2}\frac{1}{p+1}\sum_{i=1}^{p+1} E_i. \tag{4.23}$$

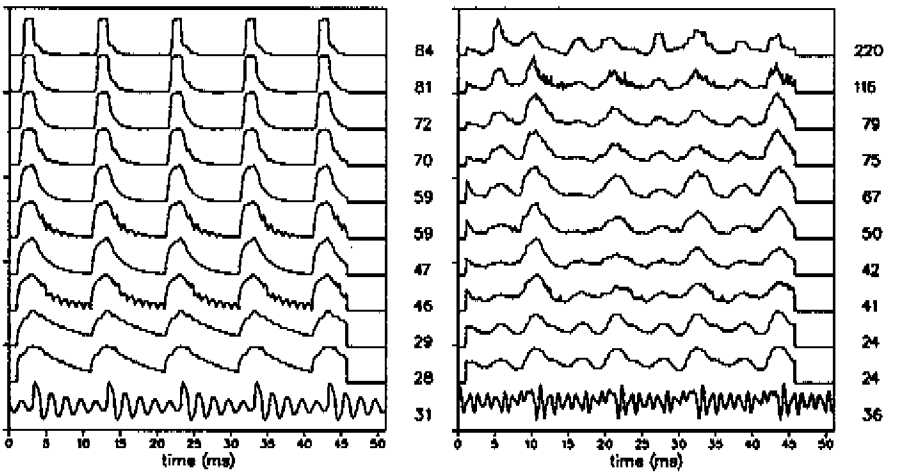
This double inequality provides the rationale for the new criterion. Indeed, C lies between an upper and a lower bound, both of which can be considered as measuring the deviation of the speech data from the linear dependence model. The lower bound, $(\Pi_{i=1}^{p+1}\sigma_i^2)^{\frac{1}{p+1}}$, is in fact Strube's criterion in view of (4.3). On the other hand, except for the scaling factor $\frac{\sigma_1^2}{\sigma_{p+1}^2}$, the upper bound is the arithmetic mean of the residual energies

E_i associated with each of the LLS solutions and it can thus be considered as an "average" deviation of the data from linear dependence. By definition, these lower and upper bounds will both increase in the open glottis region of the speech signal and will both decrease in the closed glottis portion when the linear dependence between the columns of the data matrix is better realized. Consequently, one can reasonably expect that the new criterion C will follow a similar behaviour in view of the fact that it lies between these bounds.

One observes incidentally that a similar argument also holds for the individual singular values in view of the double inequality (4.21). Indeed, the lower bound σ_{p+1}^2 measures the deviation from the linear model for the TLLS solution and the upper bound is, within a scaling factor, the residual energy E_i of the *ith* LLS solution. Both can be expected to increase when the excitation is present, i.e. when the data do not comply with the linear model. It is then not surprising that the experimental observations at the end of this section support the fact that, indeed, each singular value increases in the open glottis portion of the signal. Finally, it should be noted that (4.23) provides a tighter lower bound than would result from straightforward addition of the inequalities (4.21) over index i, since σ_{p+1}^2 is of course smaller than the geometric mean of the squared singular values.

It should be stressed that the new criterion is very efficient from a computational point of view. First, relation (4.4) and (4.5) show that the numerical value of C can be obtained simply by calculating the Frobenius norm of the data matrix S without the need of actually performing a singular value decomposition. Moreover, our criterion can easily be updated when a new sample comes in the observation window. The sequential computation of the Frobenius norm of the matrix reduces to adding the sum of the squared entries of the last row of the matrix and to subtracting the sum of the squared entries of the first row of the preceeding matrix.

The arguments presented hereabove are corroborated by experimental evidences as can be seen from Fig. 4.1 (a) for a synthetic vowel with impulse excitation and in Fig. 4.1 (b) for a natural vowel. The speech signal is displayed on the bottom of the figure and all of the singular values are scaled in the display and ordered such that the smallest singular value is displayed on the top row of the figure. All singular values exhibit local maxima when the analysis interval just comes across the excitation. Of course, the new criterion being the arithmetic mean of the squared values, will also show maxima which coincide with the occurrences of



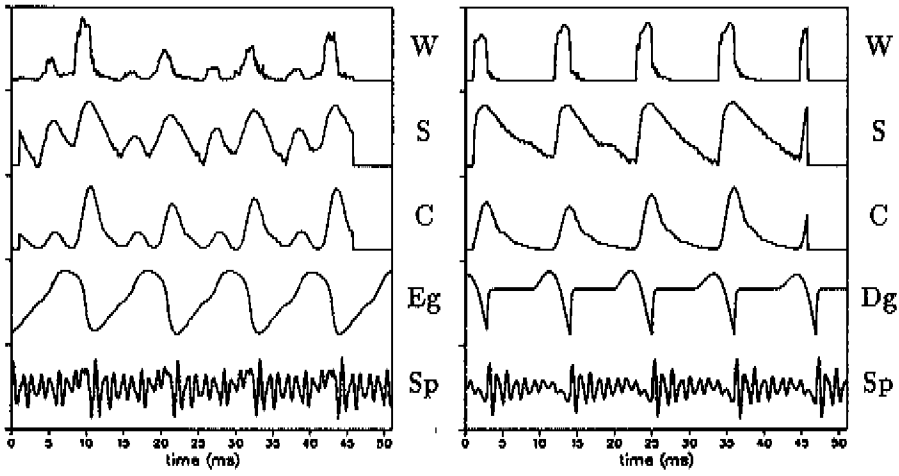
(a) (b)

Figure 4.1: (a) and (b). From top to bottom, 10 singular value curves obtained from a synthetic vowel with impulse excitation and a natural vowel, respectively. The singular values are ordered and scaled (indicated by the numbers in the figures), the smallest one on the top, the speech waveform on the bottom.

the glottal pulses.

We decide to locate the time position of the maxima in the Frobenius norm of the signal segment at time $t = p + 1$ given that the signal segment extends from $t = 1$ to $t = m + p$, where m is the number of the equations and p the order of the linear model. The reason for this is that a maximum in the Frobenius norm appears when the excitation point just enters the first row of the data matrix S. This can be seen from equation (4.1): when s_{p+1} is the excitation point, and when the analysis interval shifts further, there will be fewer rows of the data matrix that contain the excitation point. Thus, beyond $t = p + 1$, the perturbation energy starts to decrease. This is clearly illustrated in Fig. 4.2 (b) for a synthetic vowel. As a consequence, the maxima have been delayed with respect to the speech signal. The amount of the delay is equal to the number m of prediction equations and this delay has been compensated for in all the figures.

The instant of glottal closure can be determined a priori via the



(a) (b)

Figure 4.2: (a). From top to bottom, the results from Wong's method (W), Strube's method (S), and the new method (C), the electroglottal waveform (Eg) and the speech waveform (Sp) for a natural vowel /a/. (b) From top to bottom, the results from Wong's method (W), Strube's method (S), and the new method (C), the differentiated glottal pulses (Dg) and the speech waveform (Sp) for a synthetic vowel /a/.

electroglottal waveform for natural vowels or the excitation waveform for synthetic vowels. Fig. 4.2 (a) shows, from top to bottom, the results of Wong's method, Strube's method, the new method, the electroglottal waveform and speech signal. In the same manner, Fig. 4.2 (b) shows, from top to bottom, the results of Wong's method, Strube's method, the new method, the differentiated glottal pulse waveform and the speech signal. The speech signals were sampled at a sampling frequency of 10 kHz, and preemphasized with a filter $(1 - 0.9z^{-1})$ that differentiates glottal pulses and produces sharp impulses. The analysis interval was 30 samples long in total and the prediction order p was 10. As a rule, the analysis interval should be shorter than the pitch period and we found that the order of the predictor should be about 10. From the electroglottal waveform (in Fig. 4.2(a)) or from the differentiated glottal pulses (in Fig. 4.2(b)) we can determine exactly the instants of glottal

closures. It can be seen that the new criterion produces clear peaks at the these instants.

4.3 Comparison and examples

Let us first observe that Strube uses in fact the logarithm of the determinant of the autocovariance matrix to determine the epochs. In view of (3), the actual criterion he uses is thus $\Sigma_{i=1}^{p+1}\log\sigma_i^2$ whereas our criterion is $\Sigma_{i=1}^{p+1}\sigma_i^2$. Consequently, the dynamics of the singular values is nonlinearly compressed in Strube's method with the consequence that the peaks will be less prominent.

As seen from the simplified expression (4.11), Wong's criterion relies essentially on the smallest singular value. On the other hand, we observe from Figs. 4.1 and 4.2 that, as a function of time, the smallest singular value tends to exhibit flat tops and bottoms, and therefore both the maxima and the minima of the curve may not be well defined. Consequently the exact location of the epoch derived from the minima may not be possible with Wong's method in this situation. It can also be seen from the top trace in Fig. 4.1 (b) that the smallest singular value as a function of time is noisy and therefore in view of (4.11), Wong's method will be sensitive to noise. A FIR lowpass filtering, with linear phase, can reduce this noise effect. At the same time, this filtering of Wong's criterion produces distinct maxima in the smoothed curve. It is observed from Fig. 4.1 and 4.2 that these maxima correspond to the sharp down-going edges of the time function of the smallest singular value and which indicate the instants of the glottal closures (Eggen, 1989). But the positions of these maxima may depend on the filtering, as will be illustrated in the following.

It can be seen from Fig. 4.3 (a) and (b) that a FIR lowpass filtering of Wong's criterion produces nice maxima, corresponding to the instants of glottal closures, but the minima on which Wong's method is based may not be well defined. The FIR filter was designed according to the window method with a Kaiser window. The cutoff frequency of the filter was 1kHz and the length was 30 samples for Fig. 4.3 (a) and 20 samples for Fig. 4.3(b). Fig. 4.3 (a) and (b) were obtained from vowels uttered by a male and a female respectively and we have chosen the length of the analysis window to be 30 for the male voice (Fig. 4.2 (a)) and 21 for the female voice (Fig. 4.3(b)) because the pitch period of the male voices is in general longer than that of the female voices. However, Fig. 4.3(c), and (d) demonstrate that the positions of the maxima obtained from

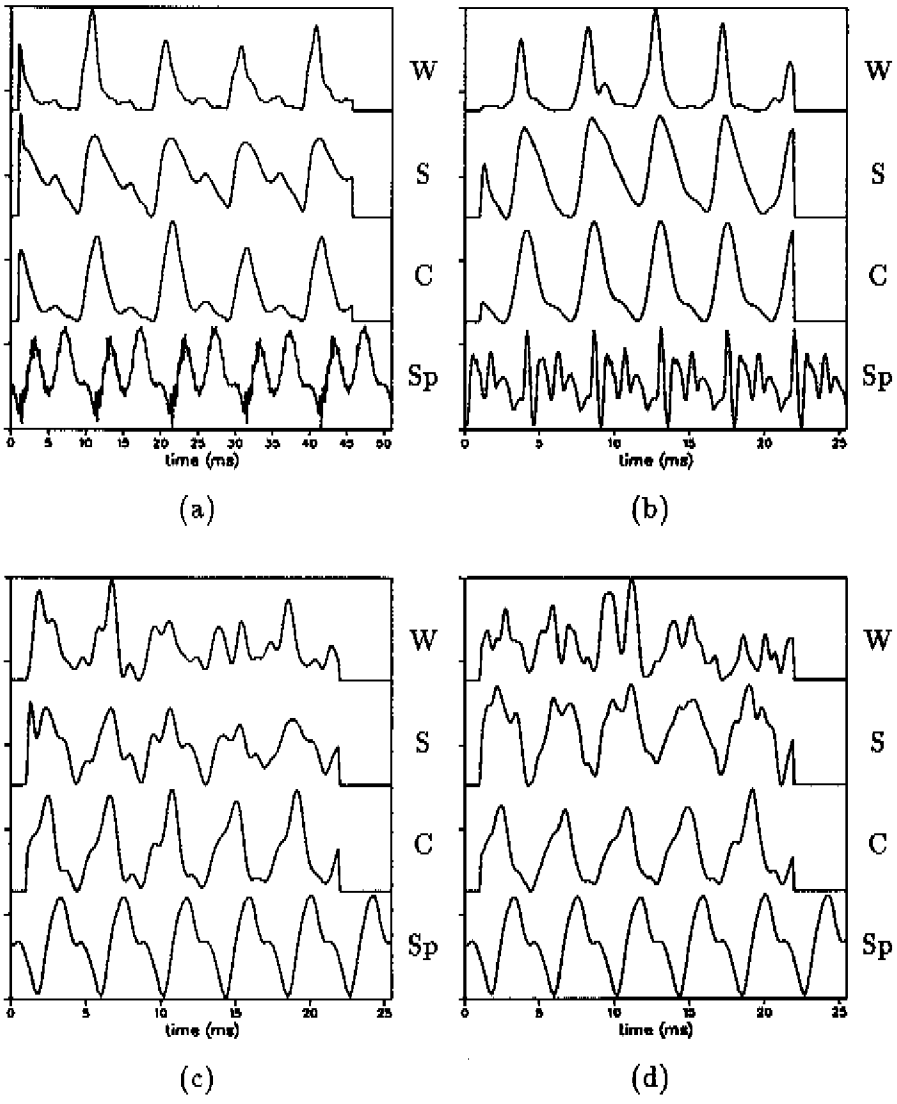
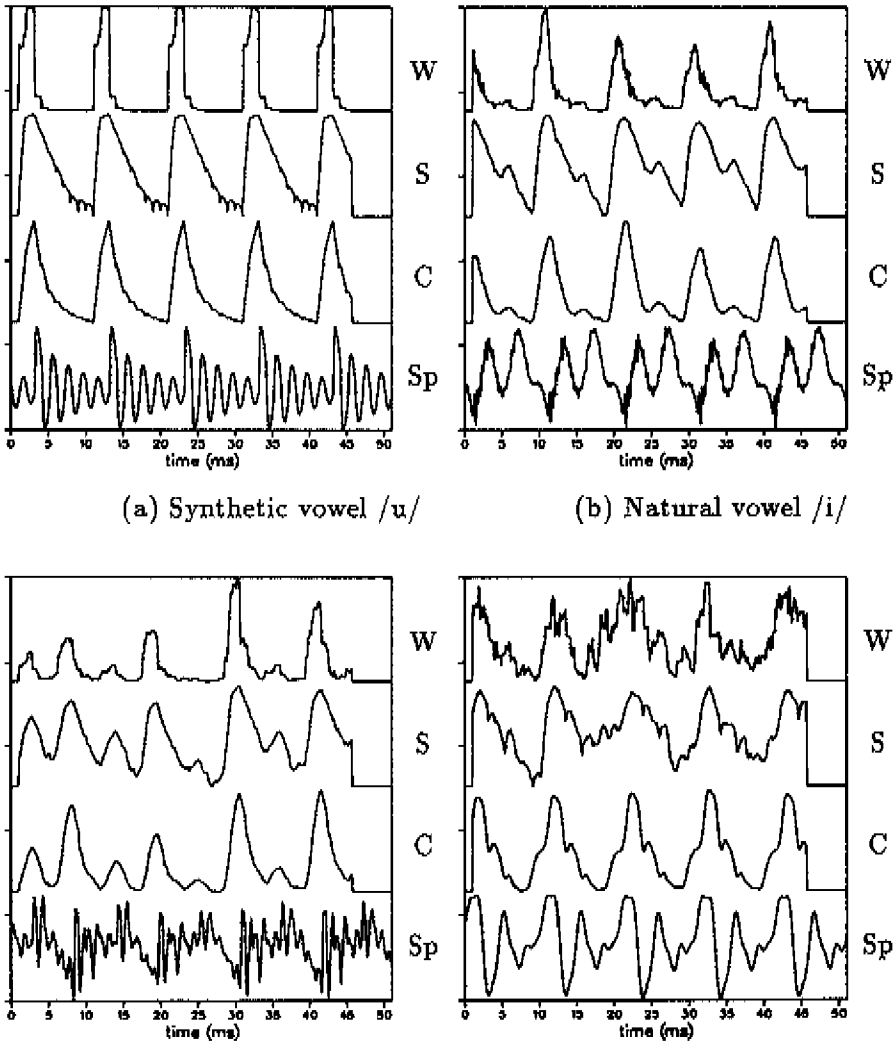


Figure 4.3: (a) and (b). FIR lowpass filtered results from the three methods (labeled as W, S, and C) for clean speech uttered by a male and a female, respectively. The cutoff frequency of the filter was 1 kHz and the filter length of the filter was 30 samples in (a) and 20 in (b). Figures (c) and (d) show the lowpass filtered results from the three methods for noisy speech with filter length of 40, and 10 samples, respectively. The SNR of the noisy speech was 20 dB.

noisy speech by the three methods are affected by the filtering in quite different ways. The noisy speech was produced by adding white noise to the speech signal with a signal-to-noise ratio of 20 dB. Fig. 4.3(c) and (d), show the results of the lowpass filtering of the three criteria with filter length of 40, and 10, respectively. It can be seen that the lowpass filtering changes the position of the maxima obtained by Wong's and Strube's methods. But the lowpass filtering has less or almost no influence on the positions of the maxima obtained by the new proposed method. From (4.4) one can see that the Frobenius norm is actually equivalent to multiplying the square of preemphasized speech samples s_n, $(s_n, n = 1, 2, \ldots, p+m)$ with a trapezoidal window $w(n)$, $(w(n) = n, for \quad n = 1, 2, \ldots, p+1; w(n) = p+1, for \quad p+1 < n < m; w(n) = p+m+1-n, for \quad n = m, m+1, \ldots, p+m)$. Since the frequency response of the trapezoidal window is similar to a lowpass filter, this explains why our method is inherently robust to noise and thus produces stable epoch detections.

In brief, the new SVD based method is to add all squared singular values on a linear scale which produces a very smooth curve with well-defined maxima. The result is a very clear picture of the glottal closure instants. Lower secondary peaks can easily be distinguished from the main peaks and the epoch detection is stable even under noisy conditions.

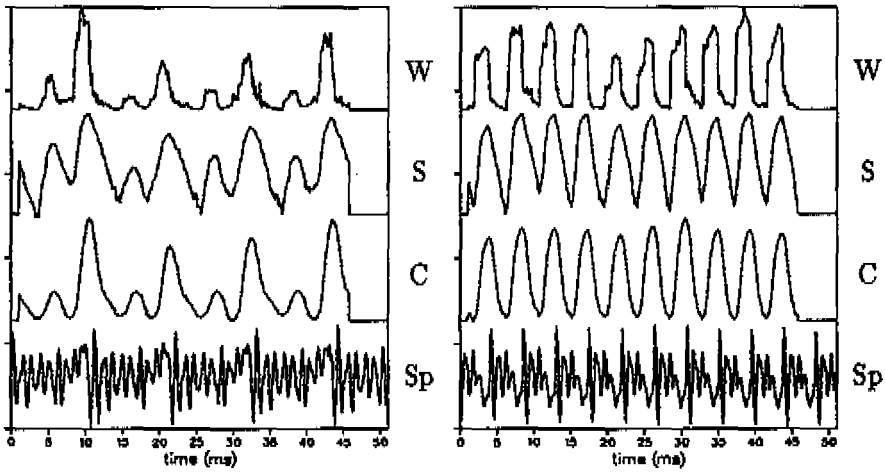
The advantage of the proposed method can be seen from more examples shown in Fig. 4.4 and comparisons can be made between the results obtained by the three methods discussed in this paper. Fig. 4.4 (a), (b), (c), (d) and (e), give the results of these three methods for a synthetic vowel /u/, and the natural vowels /i/, /a/, /u/ and /a:/, uttered by a male. Fig. 4.4 (f), (g) and (h) show the results from natural vowels /a/, /i/ and /u/, uttered by a female. The order of the predictor was 10 and the analysis window was 30 samples for both the male voices and the female voices. In each panel, from top to bottom the curves 1, 2 and 3, respectively, represent the results of Wong's and Strube's methods and the new approach. Curve 3 for the new criterion shows very clearly-defined peaks at the impulse excitation instants for the synthetic vowel (Fig. 4.2 (a)). However, curves 1 and 2, which respectively represent the results of Wong's and Strube's methods, show no peaks, but rather clear transitions at the excitation instants. The results for vowel /a:/ in panel (e) also show that curve 3 has distinct maxima, but curve 1 is noisy and curve 2 has relatively high secondary peaks at the instant of glottal opening. These relatively high secondary peaks are due to



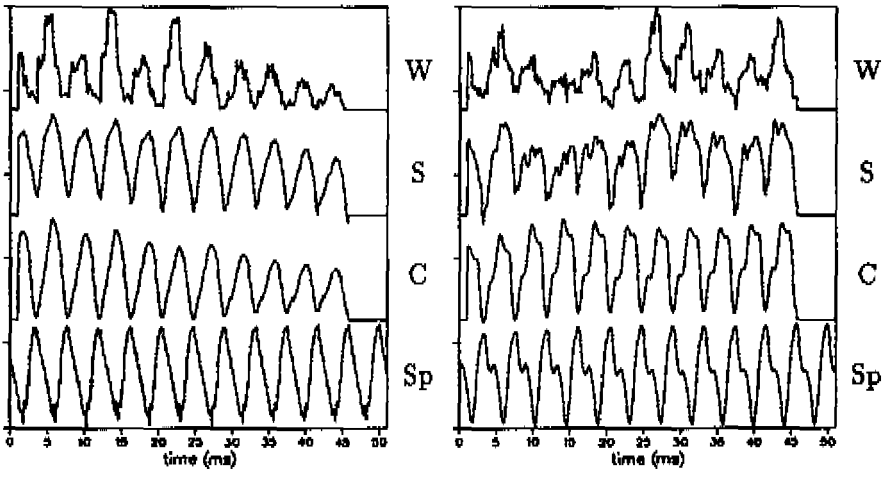
(a) Synthetic vowel /u/ (b) Natural vowel /i/

(c) Natural vowel /a/ (d) Natural vowel /u/

Figure 4.4: Comparison of the results obtained by Wong's method (W), Strube's method (S), the new method (C). Figures (a), (b), (c), (d), and (e) show results obtained from a synthetic vowel /u/ and natural vowels /i,a,u,a:/, uttered by a male, respectively. Figures (f), (g), and (h) show results obtained from vowels /a, i, u/ respectively uttered by a female. Sampling frequency is 10 kHz, analysis length 30 samples, the order of predictor 10. All results obtained from the preemphasized speech signal with a filter $1 - 0.9z^{-1}$.



(e) Natural vowel /a:/ (f) Natural vowel /a/



(g) Natural vowel /i/ (h) Natural vowel /u/

Figure 4.4: See caption to panel (a).

the logarithmic compression in Strube's method. For the same situation, the new approach produces very distinct peaks at glottal closure instants and significant lower secondary peaks that can be easily distinguished from the main excitation points for the other vowels. It can be seen from panel (h) for the vowel /u/, however, that the peaks produced by the new method do not agree with those from Strube's method and that the distance between the successive local maxima in Strube's criterion is not so regular for this vowel although the repetition rate of the excitation is quite regular. This is due to the fact that, for this vowel with a short pitch period, the relatively high secondary peaks strongly interfere with the main peaks. It can also be seen from the results for vowel /u/ uttered by a male in panel (d) and by a female in panel (h) that Wong's criterion shows quite noisy pictures. For all other vowels, one can observe that Wong's method produces sharp down-going edges which correspond to the local maxima in both Strube's criterion and the new criterion.

4.4 Application to sentences

In the preceding section, examples have been given to emphasize the advantageous aspects of the new method for detecting glottal closure instants. In those simple examples, short stationary speech segments were used and the peaks produced by the new SVD-based method in Fig. 4.4 were easy to pick out by a peak-picking algorithm. In order to facilitate the peak selection in more realistic situations, we first use a threshold to separate these peaks into isolated regions and then pick the local maximum in each of the regions. Since the short time energy of speech changes drastically, adaptive schemes should be implemented. In order to isolate the peaks we can consider two schemes. One is to use an adaptive threshold, and the other is to normalize the curve and then to select peaks from it. The latter is similar to the solution adopted by Wong who used the normalized residual energy in his glottal closure detection method (Wong, Markel, and Gray, 1979).

We shall resort to the adaptive threshold solution, and the short time energy of speech will provide the basis of thresholding technique. In the application to sentences, a relatively long interval of L samples is used to calculate the short-term energy ε_l, and the Frobenius norm ε_F is calculated from a relatively short interval of M samples. The short interval is located in the middle of the long interval, which can avoid using a lower threshold in voice onset region and a higher threshold in

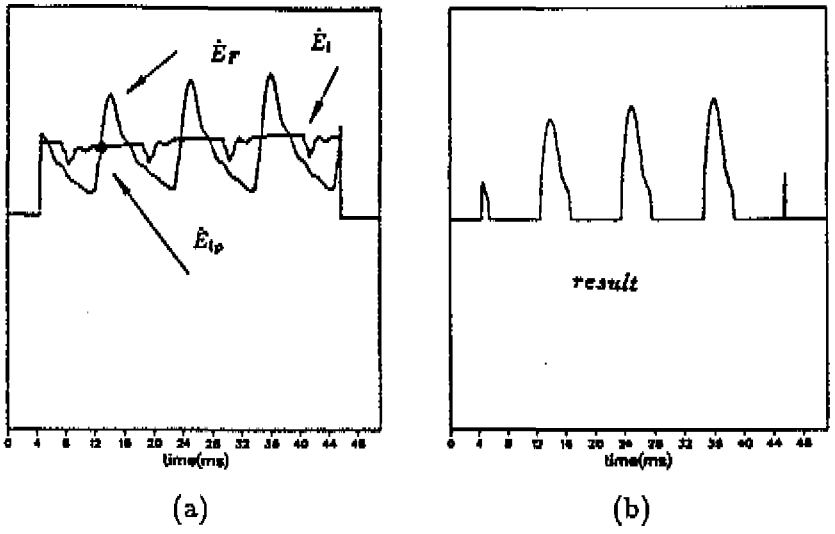


Figure 4.5: Adaptive threshold procedure. Curve $\hat{E}_F$ thresholded by curve $\hat{E}_l$. A constant threshold $\hat{E}_{lp}$ is used to avoid the distortion of the peaks. (a) The threshold method. (b) Isolated regions containing the instants of glottal closure.

voice offset region. Therefore, the average energy per sample can be obtained:

$$\hat{E}_l = \epsilon_l / L$$

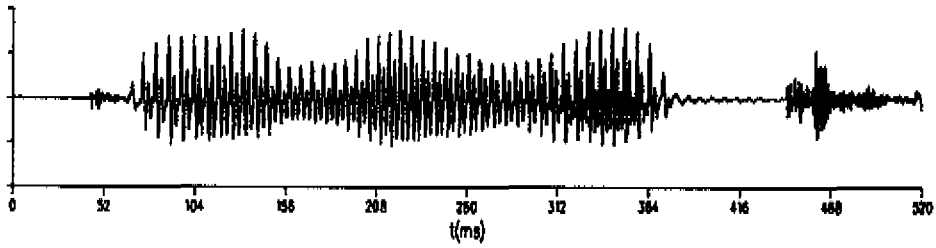
and

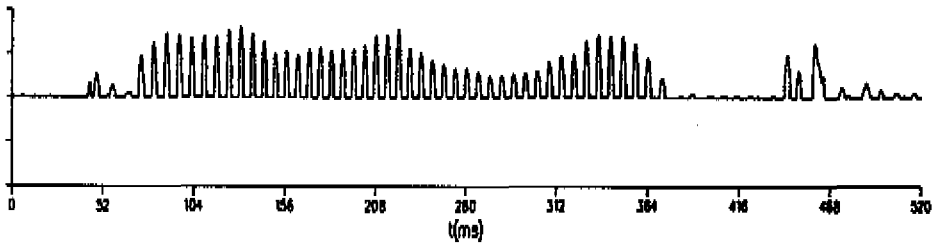
$$\hat{E}_F = \epsilon_F / N,$$

where N is the number of elements in the data matrix (4.1) . Therefore, the following algorithm can be implemented to select peaks. i. e.

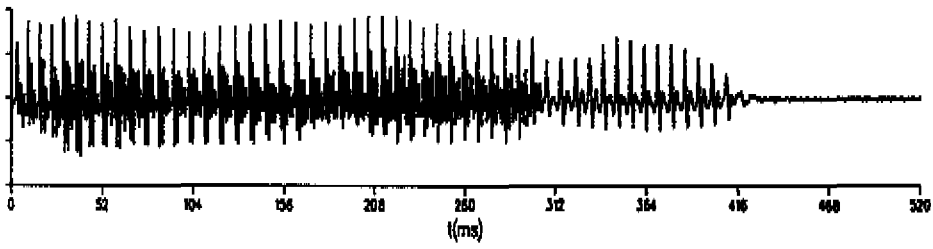
if $(\hat{E}_F - \beta\hat{E}_l > 0)$ then $result := sqrt(\hat{E}_F - \beta\hat{E}_l)$ else $result := 0$.
where sqrt represents square root operation and β is a scale factor to change the threshold. Note that the long interval average $\hat{E}_l$ is less than or equal to the maxima of the short-interval average $\hat{E}_F$ in the above algorithm. Because $\hat{E}_l$ changes much more smoothly than $\hat{E}_F$, the algorithm does not displace the peaks. Of course, the following algorithm can also be used:

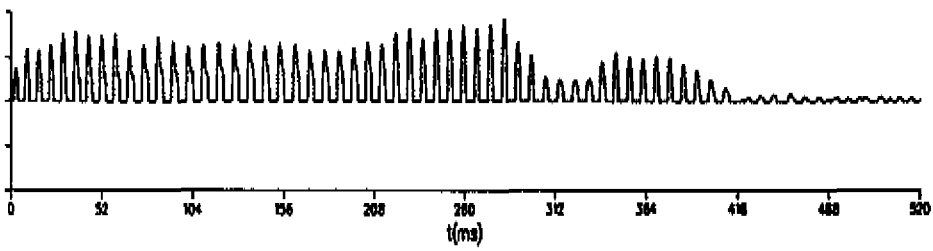
if $(\hat{E}_F - \beta\hat{E}_{lp} > 0)$ then $result := sqrt(\hat{E}_F - \beta\hat{E}_{lp})$ else $result := 0$.
where $\hat{E}_{lp}$ is the value of the short time energy when $\hat{E}_F$ starts to become larger than $\beta\hat{E}_l$. This threshold value is then kept fixed until $\hat{E}_F$ is falling





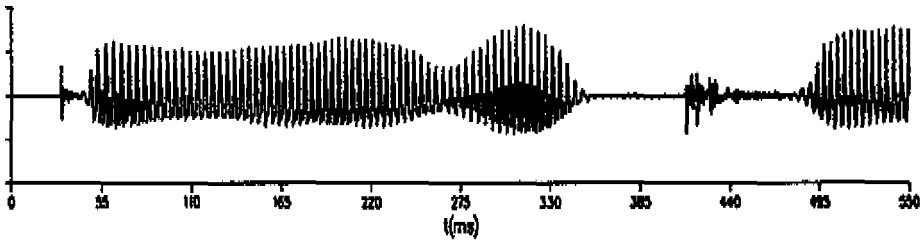
(a)

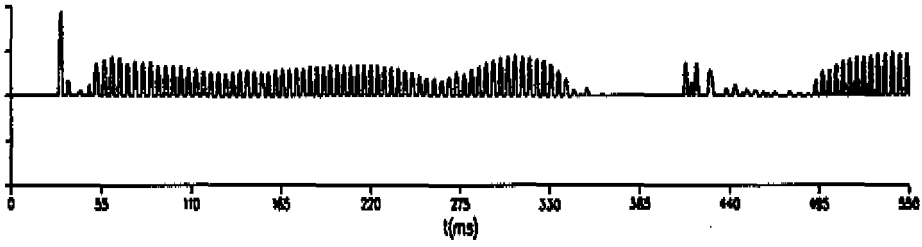




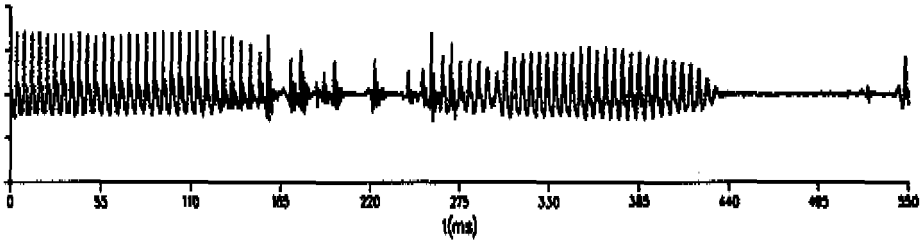
(b)

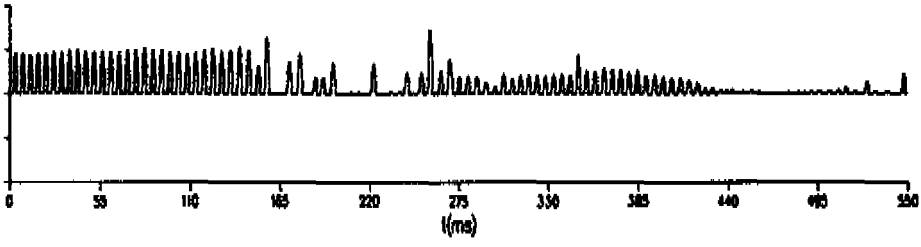
Figure 4.6: (a) and (b). Results from a male speech, "Do you require any further transaction?". The upper trace is the speech signal and the lower trace the new criterion with the peaks indicating the glottal closure instants. Sampling frequency is 10kHz, analysis length 30 samples, the order of predictor 12. All results obtained from the preemphasized speech signal with a filter $1 - 0.9z^{-1}$.





(c)





(d)

Figure 4.6: Same as panel (a) and (b), but with a female voice.

below the threshold. Fig. 4.5 illustrates the latter algorithm. Thus, the threshold is constant in the selected peak regions and the peaks are not distorted. This is why this second method of thresholding is preferable.

Choosing a proper threshold strategy is a delicate matter. In general, the value $\beta = 1$ produces satisfactory results for most of the voiced sounds. However, in some parts, it also selects secondary excitations, which correspond to the glottal opening instants. These secondary excitation instants can be deleted by considering that they have smaller amplitude than the peaks nearest to them (or by choosing $\beta > 1$) and they seriously deviate from the global pitch or long-term pitch period measured by, for instance, the DWS method (Duifhuis, et al. 1982).

Fig. 4.6 (a) and (b) show the results obtained from a sentence, "Do you require any further transaction?", which is uttered by a male and Fig. 4.6 (c) and (d) show the result from the same sentence uttered by a female. For the male voices the length of the long interval L is 100 samples, the length of the short interval M is 30 samples, and the order of the predictor p is 12, and $\beta = 1$. For the female voices only the length of the long interval is changed to 70 samples because of its shorter pitch period. It can be seen that well-defined peaks clearly indicate the glottal closure instants in the voiced regions, however, epochs are not defined in the unvoiced segments.

4.5 Conclusion

In summary, a new epoch detection technique is proposed in which only the Frobenius norm of the linear predictive matrix has to be computed. The sequential computation of the Frobenius norm of the matrix is reduced to just the addition of the sum of the squared entries of the last row of the matrix and the subtraction of the sum of the squared entries of the first row of the preceeding matrix. Therefore, the new method is computationally very attractive and more efficient than those of Strube and Wong. As an additional benefit, the new method is less sensitive to noise. Finally all three methods are interpreted in the unifying framework of singular value decomposition.

Acknowledgement

The constructive criticism and the helpful comments of Prof. A.J.M. Houtsma and Dr. A. Kohlrausch are gratefully acknowledged.

References

Ananthapadmanabha, T.V., and Yegnanarayana, B. (**1975**). "Epoch Extraction of Voiced Speech," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-**23**, 562-570.

Ananthapadmanabha, T.V., and Yegnanarayana, B. (**1979**). "Epoch Extraction from Linear Prediction Residual for Identification of Closed Glottis Interval," IEEE Trans. Acoust. , Speech, Signal Processing, Vol. ASSP- **27**, 309-319.

Charpentier, F., and Moulines, E. (**1989**). "Pitch-synchronous Waveform Processing Techniques for Text-to-speech Synthesis Using Diphones," Proceedings EUROSPEECH-89, Vol. **2**, 13-19.

Cheng, Y.M., and O'Shaughnessy, D. (**1989**). "Automatic and Reliable Estimation of Glottal Closure Instant and Period, " IEEE Trans. Acoust., Speech, Signal Processing, Vol.ASSP-**37**, 1805-1815.

Duifhuis, H., Willems, L., and Sluyter, R.J. (**1982**). "Measurement of Pitch in Speech: An Implementation of Goldstein's Theory of Pitch Perception," J. Acoust. Soc. Am. **72**, 1568-1580.

Eggen, J.H. (**1989**). "A Glottal-excited Speech Synthesizer," IPO Annual Progress Report **24**.

Flanagan, J.L. (**1972**). *Speech Analysis, Synthesis and Perception.* (Springer-Verlag, New York), 2nd edition.

Golub, G.H., and Van Loan, C.F. (**1983**). *Matrix Computation*, (Johns Hopkins University Press, Baltimore).

Hedelin, P. (**1984**). "High Quality Glottal LPC-vocoding," Proc. Int. Conf. ICASSP-86, Tokyo, 465-468.

Hermes, D.J. (**1988**). "Measurement of Pitch by Subharmonic Summation," J. Acoust. Soc. Am. **83**, 527-264.

Kuwabara, H. (**1984**). "A Pitch-synchronous Analysis/synthesizer System to Independently Modify Formant Frequencies and Bandwidth for Voiced Speech," Speech Communication, Vol. **3**, 211-220.

Larar, J.N., Alsaka, Y.A., and Childers, D.G. (**1985**). "Variability in Closed Phase Analysis of Speech," Proc. Int. Conf. ICASSP-85, 29.2.1-29.2.4.

Makhoul, J. (**1975**). Linear Prediction: A Tutorial Review. Proc. IEEE, Vol. **63**, No. 4, 561-580.

Markel, J.D., and Gray, A.H. (**1970**). *Linear Prediction*. (Mouton, The Hague, The Netherlands).

Markel, J.D. (**1972**). "The SIFT Algorithm for Fundamental Frequency Estimation," IEEE Trans. Audio ElectroAcoust., Vol. AU-**20**, 367-377.

Moulines, E., and Di Francesco, R. (**1990**). "Detection of the Glottal Closure by Jumps in the Statistical Properties of the Speech Signal," Speech Communication, Vol. **9**, 401-418.

Ross, M.J., Shaffer, H.L., Cohen, A., Freudberg, R., and Manley, H.J. (**1974**). "Average Magnitude Difference Function Pitch Extractor," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-**22**, 353-362.

Strube, H.W. (**1974**). "Determination of the Instant of Glottal Closures from the Speech Wave," J. Acoust. Soc. Am. Vol. **56**, 1625-1629.

Vandewalle, J., and De Moor, B. (**1988**). "A Variety of Applications of Singular Value Decomposition in Identification and Signal Process," in *SVD and signal processing*, edited by F. Deprettere, (Elsevier, Amsterdam, North-Holland), 43-91.

Wong, D.Y., Markel, J.D., and Gray, jr. A.H. (**1979**). "Least Squares Glottal Inverse Filtering from the Acoustic Speech Waveform," IEEE Trans. Acoust., Speech, Signal Processing, Vol. ASSP-**27**, No. 4, 353-362.

Chapter 5

Novel criteria of uniqueness for signal reconstruction from phase *

Abstract

In this paper we propose a new approach for ascertaining whether a signal is uniquely determined by its Fourier transform phase. It is shown that uniqueness corresponds to the non-singularity of a matrix which can be formed from the finite length real sequence.

5.1 Introduction

RECONSTRUCTION of a signal from its Fourier transform phase or magnitude are of special interest in the area of speech processing, geophysical signal processing and image processing. Generally speaking, we cannot determine a signal only from its Fourier transform magnitude or phase. It is obvious that the magnitude of a signal sequence with a determined magnitude , when it passes through an all-pass filter, is unchanged, and that if a signal sequence with a determined phase is convolved with a zero phase sequence, the resultant sequence has the same phase as the original. In 1980 Hayes, Lim and Oppenheim gave some conditions under which it is possible to reconstruct the signal sequence from its phase or magnitude uniquely [1] [2]. These conditions were given in the Z-transform domain. Other studies were focused on the zero distribution of the Z-transform of the reconstructed signal, and algorithms for signal reconstruction under certain constraints were proposed [7] [8]. In order to guarantee the uniqueness (to within a scale factor) of signal reconstruction, constraints must be imposed on the Z-transform of the signal or its zero distribution, or on the algorithms.

Generally speaking, we do not know the zero distribution of the Z-transform of the finite length real sequence. To find the zeros of poly-

*Paper published in the IEEE Trans. on Signal Processing, Vol.39, pp.989-992, 1991.

nomials of higher degree than five is a genuine nonlinear problem and one must apply numerical methods, which are time-consuming and complicated [3]. The problem of accurately finding the zeros of high-order polynomials (of degree greater that 100) is an extremely difficult one and the accuracy can not be guaranteed.

In this paper, novel criteria are proposed for determining the uniqueness of the reconstruction of a signal from its Fourier transform phase. We can decide whether the reconstructed signal sequence is unique (to within a scale factor) by determining whether a matrix formed by the reconstructed finite length real sequence is singular. Thus, only elementary transformations such as Doolittle factorization, are needed to determine the singularity [4]. This method shows clear advantages over that of Hayes in numerical stability and computation time. In section 2, we discuss the criterion of uniqueness for reconstructing a one-dimensional finite length sequence from its phase. In section 3, we present the criterion of uniqueness for reconstructing a multi-dimensional finite length sequence.

5.2 Uniqueness of a one-dimensional finite length sequence

5.2.1 Reconstruction from a continuous phase function

Let $\{x_n, n = 0, 1, \ldots, N-1\}$ be a finite length real signal sequence. Its discrete Fourier transform is

$$X(j\omega) = \sum_{n=0}^{N-1} x_n exp(-j\omega n)$$

and

$$X(j\omega) = |X(j\omega)| exp(j\psi_x(\omega)).$$

From the above definitions, we have

$$\tan\psi_x(\omega) = -\frac{\sum_{n=0}^{N-1} x_n \sin\omega n}{\sum_{n=0}^{N-1} x_n \cos\omega n}. \tag{5.1}$$

To reconstruct a sequence from a known phase $\tan\phi_x(\omega)$, we also express $\tan\phi_x(\omega)$ as

$$\tan\phi_x(\omega) = -\frac{\sum_{n=0}^{N-1} a_n \sin\omega n}{\sum_{n=0}^{N-1} a_n \cos\omega n}, \tag{5.2}$$

where $\{a_n, n = 0, 1, \ldots, N-1\}$ is the reconstructed finite length real sequence in some sense. For reconstruction, the available relation is

$$\tan \phi_x(\omega) = \tan \psi_x(\omega).$$

Equivalently,

$$\frac{\sum_{n=0}^{N-1} x_n \sin \omega n}{\sum_{n=0}^{N-1} x_n \cos \omega n} = \frac{\sum_{n=0}^{N-1} a_n \sin \omega n}{\sum_{n=0}^{N-1} a_n \cos \omega n}. \tag{5.3}$$

After reduction, then

$$\sum_{n=1}^{N-1} x_n \sum_{i=0}^{N-1} a_i \sin \omega(n-i) = x_0 \sum_{i=0}^{N-1} a_i \sin \omega i. \tag{5.4}$$

For simplicity, we define the following vector S and matrix B:

$$S = (\sin \omega, \sin 2\omega, \sin 3\omega, \ldots, \sin(N-1)\omega)^T$$

and

$$B = \begin{pmatrix} a_0 & 0 & \ldots & 0 & 0 \\ a_1 & a_0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{N-3} & a_{N-4} & \ldots & a_0 & 0 \\ a_{N-2} & a_{N-3} & \ldots & a_1 & a_0 \end{pmatrix} - \begin{pmatrix} a_2 & a_3 & \ldots & a_{N-1} & 0 \\ a_3 & a_4 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{N-1} & 0 & \ldots & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 0 \end{pmatrix}.$$

In the same way, sequences $\{x_n, n = 1, 2, \ldots, N-1\}$ and $\{a_n, n = 1, 2, \ldots, N-1\}$ can be written as follows:

$$X^T = (x_1, x_2, \ldots, x_{N-1})$$

and

$$A^T = (a_1, a_2, \ldots, a_{N-1}),$$

where T is a transpose operation. Note that x_0 and a_0 are not included in the above vectors. Therefore, equation (5.4) can be changed into a matrix form as

$$(X^T B - x_0 A^T) S = 0. \tag{5.5}$$

We know that function series $\{\sin \omega i, i = 1, 2, \ldots, N-1\}$ are linearly independent in the interval $0 < \omega < \pi$. That is, the necessary and sufficient condition, under which the linear combination $\sum_{i=1}^{N-1} c_i \sin i\omega$ is equal to zero, is $\{c_i = 0, i = 1, 2, \ldots, N-1\}$. Because the left-hand side

of equation (5.5) is a linear combination of $\{\sin \omega i, i = 1, 2, \ldots, N-1\}$, we can conclude that

$$(X^T B - x_0 A^T) = (0, 0, \ldots, 0).$$

Thus,

$$B^T X = x_0 A. \tag{5.6}$$

It is apparent that equation (5.6) is linear with respect to X, and that $X = \frac{x_0}{a_0} A$ is one of the solutions, which can be easily proved. Here we have made an assumption that x_0 is not equal to zero. If matrix B is nonsingular, then $X = \frac{x_0}{a_0} A$ is the only solution vector. Thus the reconstructed sequence is identical to the original sequence within a scale factor. If matrix B is singular, the solution is not unique. That is, the sequence reconstructed from phase is not unique.

Conversely, suppose that two finite length real sequences have the relation, $\{a_i^{(1)} \neq \beta a_i^{(2)}\}$, β is a real number, and $\tan \psi_1(\omega) = \tan \psi_2(\omega)$. Matrices B_1 and B_2 are nonsingular. Then we have

$$B_1^T A_2 = a_0^{(2)} A_1$$

and

$$B_2^T A_1 = a_0^{(1)} A_2.$$

Consequently, $A_1 = \beta A_2$.

Therefore, the nonsingularity of matrix B is a necessary and sufficient condition for unique reconstruction of a finite length sequence from it Fourier transform phase.

5.2.2 Reconstruction from discrete phase values

Now suppose we are given the value of the phase function $\tan \psi_x(\omega)$ only at (N-1) different frequency point ω_k in the interval $0 < \omega_k < \pi$. Substituting the value ω_k into equation (5.4), we have (N-1) equations.

$$S_\omega^T B^T X = x_0 S_\omega^T A, \tag{5.7}$$

where

$$S_\omega = \begin{pmatrix} \sin \omega_1 & \sin \omega_2 & \ldots & \sin \omega_{N-1} \\ \sin 2\omega_1 & \ldots & \ldots & \sin 2\omega_{N-1} \\ \vdots & \vdots & \vdots & \vdots \\ \sin(N-1)\omega_1 & \ldots & \ldots & \sin(N-1)\omega_{N-1} \end{pmatrix}.$$

We have assumed that $\omega_i \neq \omega_j$, if $i \neq j$, and $\omega_k \subset (0, \pi)$. In addition, functions $\{\sin(i\omega), i = 1, 2, \ldots, N-1\}$ are linearly-independent in the interval $\omega \subset (0, \pi)$. The matrix S_ω is nonsingular, therefore equation (5.7) and equation (5.6) are equivalent.

From the above reasoning, we conclude that: (1), if and only if the matrix B, which is formed from a finite length real sequence, is nonsingular, then the sequence is uniquely determined by its phase function, and (2), if the phase function or its samples $\tan\psi_x(\omega_k)$ is given, the sequence can be reconstructed in several ways, such as an iterative method with constraints [6]. Even if we do not know whether the reconstructed sequence is unique, we can check its matrix B to determine the uniqueness.

5.2.3 Determination of singularity of matrix B and examples

To determine the singularity of the matrix B, we can use the well known Doolittle factorization method [4], which is numerically stable. Considering the structure of the matrix B, We can speed up the computation. Let B be partitioned in the form

$$B = \begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix},$$

where B_{11} and B_{22} are square submatrices, so that

$$B_{22} = \begin{pmatrix} a_0 & 0 & \ldots & 0 \\ a_1 & a_0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ a_m & a_{m-1} & \ldots & a_0 \end{pmatrix}, B_{12} = \begin{pmatrix} a_{N-m} & \ldots & a_{N-1} & 0 \\ a_{N-m+1} & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ a_{N-1} & 0 & \ldots & 0 \end{pmatrix},$$

are triangular matrices, where $m = (N-1)/2 - 1$, If $(N-1)$ is even; otherwise $m = (N-2)/2$. Because $a_0 \neq 0$, B_{22} is nonsingular and the following identity

$$\begin{pmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{pmatrix} = \begin{pmatrix} C & B_{12} \\ 0 & B_{22} \end{pmatrix} \begin{pmatrix} I & 0 \\ P & I \end{pmatrix},$$

$$C = B_{11} - B_{12} B_{22}^{-1} B_{21}, P = B_{22}^{-1} B_{21},$$

can be obtained, where I is a unit matrix and B_{22}^{-1} is the inverse of B_{22}. It evidently follows that the singularity of B is determined by that of matrix C, which is about half the size of the matrix B. Both B_{12} and

B_{22}^{-1} are triangular matrices, therefore we can save more computation time.

We give two special cases in this section. When $N = 1$, the determinant of matrix B, $\|B\|$ equals a_0. The Z-transform of the sequence a_n is

$$X(z) = a_0 + a_1 z^{-1}.$$

If $a_0 \neq 0$, then the sequence is uniquely determined by its Fourier transform phase.

When $N = 3$, $\|B\| = a_0(a_0 - a_2)$. If $a_0 \neq 0$ and $a_0 \neq a_2$, the sequence can be uniquely reconstructed. Its Z-transform is

$$X(z) = a_0 + a_1 z^{-1} + a_2 z^{-2}.$$

If $a_0 = a_2$, the two roots, z_1 and z_2, of polynomial $X(z)$ have a relation such that $z_1 z_2 = a_0/a_2 = 1$. That means they are the zeros in reciprocal pairs or on the unit circle. Theoretically, any Z-polynomial of the finite length real sequence can be factorized into the product of a number of second order polynomials. So if a polynomial $X(z)$ has no zeros in reciprocal pairs or on the unit circle, the sequence is uniquely specified by its phase, and vice versa. Suppose, for example, that one of the second-order polynomials has a pair of reciprocal zeros, then the polynomial $X(z)$ can be written as

$$X(z) = (1 + \alpha z^{-1} + z^{-2}) \sum_{i=0}^{N-3} c_i z^{-i}.$$

It is easy to show that the matrix B becomes

$$B = \left\{ \begin{pmatrix} c_0 & 0 & \dots & 0 & 0 \\ c_1 & c_0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{N-3} & c_{N-4} & \dots & c_0 & 0 \\ 0 & c_{N-3} & \dots & c_1 & c_0 \end{pmatrix} - \begin{pmatrix} c_0 & c_1 & \dots & c_{N-3} & 0 \\ c_1 & c_2 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{N-3} & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix} \right\}$$

$$* \begin{pmatrix} 1 & 0 & 0 & 0 & \dots & 0 & 0 & 0 \\ \alpha & 1 & 0 & 0 & \dots & 0 & 0 & 0 \\ 1 & \alpha & 1 & 0 & \vdots & 0 & 0 & 0 \\ 0 & 1 & \alpha & 1 & \vdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & \alpha & 1 \end{pmatrix}.$$

Note that the first columns of the two first matrice are equal. Therefore, the determinant value of the matrix B is equal to zero. This is in agreement with Hayes' conclusions [1].

5.3 Uniqueness of a multidimensional finite length real sequence

5.3.1 Reconstruction from a continuous phase function

Let $\{x(n_1, n_2, \ldots, n_m), n_k = 0, 1, \ldots, N-1, k = 1, 2, \ldots, m\}$ be an m-dimensional finite length real sequence (m-D) with Fourier transform

$$X(\vec{\omega}) = \sum_{\vec{n}} x(\vec{n}) \exp(-j\vec{\omega} \bullet \vec{n}),$$

where vector $\vec{\omega} = (\omega_1, \omega_2, \ldots, \omega_m)$, $\vec{n} = (n_1, n_2, \ldots, n_m)$, and $\vec{\omega} \bullet \vec{n}$ denotes the inner product of $\vec{\omega}$ and $\vec{n}$.

In a reasoning similar to that in the above section, we have

$$\sum_{\vec{n} \neq 0} x(\vec{n}) \sum_{\vec{i}} a(\vec{i}) \sin\{\vec{\omega} \bullet (\vec{n} - \vec{i})\} = x(\vec{0}) \sum_{\vec{i}} a(\vec{i}) \sin(\vec{\omega} \bullet \vec{i}), \qquad (5.8)$$

where $\vec{0} = (0, 0, \ldots, 0)$.

In order to write Eq. (5.8) in matrix form, we define a mapping relation between natural numbers and vectors. For any vector $\vec{n} = (n_1, n_2, \ldots, n_m)$, there exists a number M, so that

$$M = n_1 + n_2 N + \ldots + n_m N^{m-1},$$

where $0 \leq M \leq L$, and $L = N^m - 1$. Conversely, for any $M \in (0, L)$, there exists an $\vec{n}$. For example, if $m = 2, N = 3$ and $\vec{n} = (1, 2)$, then $M = 1 + 2 * 3 = 7$. When $M = 3$, $\vec{n} = (0, 1)$.

According to the mapping relation, we can rewrite the m-D sequences $x(\vec{n})$ and $a(\vec{n})$ as

$$X_m^T = \{x(\vec{1}), x(\vec{2}), \ldots, x(\vec{L})\}$$

and

$$A_m^T = \{a(\vec{1}), a(\vec{2}), \ldots, a(\vec{L})\},$$

Note that $x(\vec{0})$ and $a(\vec{0})$ are not included in the above vectors.

We can now define a matrix B_m and a vector S_m as:

$$B_m = \begin{pmatrix} a(\vec{0}) & 0 & \dots & 0 \\ a(\vec{1}) & a(\vec{0}) & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ a(\vec{L-1}) & \dots & \dots & a(\vec{0}) \end{pmatrix} - \begin{pmatrix} a(\vec{2}) & \dots & a(\vec{L}) & 0 \\ \vdots & \vdots & \vdots & \vdots \\ a(\vec{L}) & \dots & \dots & 0 \\ 0 & \dots & 0 & 0 \end{pmatrix}$$

and

$$S_m = \{\sin(\vec{\omega} \bullet \vec{1}), \sin(\vec{2} \bullet \vec{\omega}), \sin(\vec{3} \bullet \vec{\omega}), \dots, \sin(\vec{L} \bullet \vec{\omega})\}^T,$$

where $\vec{i}$ denotes a vector which is mapped from i. Since the functions $\{\sin\vec{\omega} \bullet \vec{n}\}$ are linearly independent, we have

$$B_m^T X_m = x(\vec{0}) A_m. \tag{5.9}$$

5.3.2 Reconstruction from discrete phase values

If the values of the phase function $\tan\psi_x(\vec{\omega})$ are given at L different frequency points ω_k $(0 < \omega_k < \pi)$ in vector space $\vec{\omega}$, equation (5.8) yields L equations. That is,

$$S_\omega^T B_m^T X_m = x(\vec{0}) S_\omega^T A_m, \tag{5.10}$$

where

$$S_\omega = \begin{pmatrix} \sin\vec{1} \bullet \vec{\omega_1} & \sin\vec{1} \bullet \vec{\omega_2} & \dots & \sin\vec{1} \bullet \vec{\omega_L} \\ \sin\vec{2} \bullet \vec{\omega_1} & \dots & \dots & \sin\vec{2} \bullet \vec{\omega_L} \\ \vdots & \vdots & \vdots & \vdots \\ \sin\vec{L} \bullet \vec{\omega_1} & \dots & \dots & \sin\vec{L} \bullet \vec{\omega_L} \end{pmatrix}.$$

The matrix S_ω is nonsingular, so equation (5.10) reduces to (5.9).

Therefore, an m-D finite length real sequence can be uniquely specified by the Fourier transform phase, if matrix B_m is nonsingular.

5.4 Conclusion

In this paper, the uniqueness of reconstructing a finite length real sequence from the Fourier transform phase is determined by the singularity of matrix B or matrix B_m. This criterion is the same for the reconstruction from the phase function or from discrete phase values. To determine the singularity of the matrix B or B_m, only elementary transformations

such as Doolittle factorization are needed. The numerical treatment of determining the rank of matrix B is much easier than as is used in another metheod. The properties of the matrix B or B_m also make the calculation easier. When an iterative method is used to reconstruct a sequence under certain constraints, we can determine the uniqueness or the effectiveness of the constraints in this way. For a given sequence, we can also find out whether there is a unique mapping between the phase and the sequence (to within a scale factor).

References

[1] M.H. Hayes, J.S. Lim and A.V. Oppenheim, "Signal reconstruction from phase or magnitude", IEEE trans. Acoust., speech and signal processing, Vol. ASSP-28, No. 26, pp. 672-680, Dec. 1980.

[2] M.H. Hayes, "The reconstruction of a multidimentional sequence from the phase or magnitude of its Fourier transform", IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-30, pp. 140-154, April 1982.

[3] A.S. Householder, The numerical treatment of a single nonlinear equation. New York: McGraw-Hill, 1970.

[4] J.H. Wilkinson and C. Reinsch, Linear Algebra, Berlin: Springer-Verlag, pp. 93-110, 1971.

[5] P.L. Van Hove, M.H. Hayes, J.S. Lim and A. V. Oppenheim, "Signal reconstruction from signed Fourier transform magnitude", IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-31, No. 5, pp. 1286-1289, Oct. 1983.

[6] T. F. Quatieri, JR. and A.V. Oppenheim, " Iterative techniques for minimum phase signal reconstruction from phase or magnitude", IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-29, No.6, pp. 1187-1193, Dec. 1981.

[7] J. L. C. Sanz and T.S. Huang, "Polynomial system of equations and its applications to the study of the effect of noise on multidimennsional Fourier transform phase", IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-33, No. 4, pp. 997-1004, Aug. 1985.

[8] S.R. Curtis, A.V. Oppenheim and J.S. Lim, "Signal reconstruction from Fourier transform sign information", IEEE Trans. Acoust., Speech and Signal Processing, Vol. ASSP-33, No. 3, pp. 643-657, June 1985.

Chapter 6

Masking of noise by broadband harmonic complex sounds: implications for the processing of complex sounds *

6.1 Introduction

6.1.1 Auditory masking and the auditory system

WHEN we listen to two competing simultaneous or successive sounds, one sound can reduce the ability of the ear to perceive the other. The threshold at which a target sound is just audible is raised due to the presence of the masker sound. Depending on the temporal relationship between the two sounds, if the two sounds are presented simultaneously we speak about simultaneous or spectral masking, and about nonsimultaneous or temporal masking, if the two sounds are presented in succession. Nonsimultaneous masking is further classified as forward masking and backward masking, depending on whether the masker is presented before or after the target.

In response to the incoming sounds, the auditory system first transforms the sound pressure waves into traveling waves on the basilar membrane. The vibration of the basilar membrane is then transformed into neural activity which is sent to higher levels of the auditory system through nerve fibers. The masking behaviour reflects both the mechanical properties of the basilar membrane and the electrical firing characteristics of nerve fibers in the auditory system. The masking of one sound by another is therefore often used as a tool to explore the auditory system and it can reveal the spectral and the temporal resolution of the system.

In view of the relationships between masker and target mentioned

*Parts of this chapter were published in the Proc. Eurospeech-91, Genova, Italy, 1991, pp.1125-1128 and as a poster at the Royal Society discussion meeting *Auditory Processing of Complex Sounds*, London, Dec. 4-5, 1991.

above, two extreme choices for masker and target are often made to simplify the experiment. The signals chosen are either concentrated in the time domain such as clicks or concentrated in the frequency domain such as sinusoids. As in a general system analysis, signals with very compact distribution in the frequency domain (sinusoids) are often used as maskers or targets so that an appropriate frequency resolution can be measured. On the other hand, sounds with compact time distribution are utilized in the masking experiments in order to measure temporal resolution.

Masking experiments of pure tones by pure tones and by narrow-band noise (target signal concentrated in the frequency domain) have revealed that the auditory periphery resembles a bank of bandpass filters (Wegel and Lane, 1924; Fletcher, 1940; Egan and Hake, 1950; Schafer et al., 1950; Ehmer, 1959; Small, 1959). This filter bank forms the basis for a frequency analysis of incoming sounds and the resolution of this frequency analysis is related to the bandwidth of the filters. These bandwidths, called critical bands (Fletcher, 1929), have been measured psychophysically in several ways (Bos and de Boer, 1966; Patterson and Moore, 1986) and were found to be systematically ordered, with the widest one corresponding locally to the base of the cochlea and the smallest one to the apex (Greenwood, 1961). The filter-bank structure and the values of the critical-band width are approximately in line with direct measurements of mechanical frequency separation of the cochlea along the cochlear partitions (Bekesy, 1960; Yates, 1986). They are further supported by the measurement of neural activity of auditory nerve fibers that is described by tuning curves (Pickles, 1986).

This spectral masking behaviour can also be described by considering that the excitation pattern of a pure tone on the basilar membrane is quite spread. Therefore, the excitation patterns produced by the masker and the target overlap and interact with each other (Zwicker, 1970). Moreover, the pure-tone masking behaviour tells us that the system is nonlinear and the spread of masking towards higher frequencies is strongly dependent on masker level. This is manifested by the fact that the threshold of the target at the high-frequency side of the masker increases faster than the sound pressure level of the masker (Wegel and Lane, 1924; Ehmer, 1959; Schöne, 1977).

Forward masking and backward masking patterns, on the other hand, are obtained by using short-duration target signals (for review see e.g. Duifhuis, 1973; Fastl, 1976, 1976/77, 1979). In the experimental configuration, the target signal starts after the offset of the masker (forward

masking) or ends before the onset of the masker (backward masking). The target threshold is typically determined as a function of the frequency of the target signal, the masker level and the time delay between the masker and the target. Broadband noise, pure tones and narrowband noise are often used as maskers and pulsed pure tones as targets in the forward-masking paradigm (Duifhuis, 1973; Fastl, 1976). Even pulsed pure tones with a short duration of 2 or 5 ms, producing wide spectral splatter, can be used to evaluate temporal fine structures in high-frequency regions (Fastl, 1976; Zwicker, 1976b). The temporal masking patterns reflect the temporal resolution of the auditory system. Two important processes of the auditory pathway are revealed in these patterns. One is the ringing of the auditory filter in response to the masker. This ringing response persists after the offset of the masker and can overlap temporally with the target signal. The other process is due to neural adaptation because the inner hair cells and the auditory nerve fibers deplete their energy in response to the masker and only slowly recover from it (Duifhuis, 1973; Jesteadt et al., 1982; Moore and Glasberg, 1983a). In general, target thresholds in forward masking increase monotonically with masker level and decrease with masker-signal delay. In contrast to simultaneous masking, the target threshold in forward masking increases less than proportional with an increase of the masker level (Jesteadt et al., 1982; Moore and Glasberg, 1983a).

In this paper, we are concerned with the masking properties of harmonic complex sounds. It is therefore worthwhile to consider dynamic stimuli in terms of time-frequency representation (or spectro-temporal representation) in the masking experiments. The auditory system can be considered as a multi-resolution device which has a decreasing frequency resolution and an increasing temporal resolution with the increase of the center frequency of the channel.

6.1.2 Auditory behavior and speech processing

Auditory masking can work against our ability to perceive a useful or meaningful sound in the presence of competing sounds. The same masking properties, on the other hand, can work favorably by making quantization noise inaudible in bit-compressed coding of audio signals or speech. One purpose of speech coding is to reduce the bit rate of information flow in communication channels and storage media. Simplifying the representation of the speech signals almost always results in quantization noise. Ideally, this quantization noise should be inaudible in

the reproduced signal. This can be achieved by taking care that the quantization noise is masked by the speech sounds. As we know, speech signals are spectrally complex and dynamic. The global spectra of the signals over a short interval of time show peaks and valleys, but the quantization noise generally has a flat spectrum. Therefore, spectral weighting has been used to shape the spectrum of the noise such that its power spectrum is similar to that of the speech and the noise can be masked effectively (Atal, 1988). This spectral weighting of the noise is called perceptual weighting. In the coding of wideband signals, the use of perceptual weights has been remarkably successful, allowing wideband signal representation with an average of four bits per sample (Johnston, 1988).

Speech signals are often manipulated or transformed for different purposes without damaging their subjective sound quality, such as in pitch manipulation, phase equalization and dispersion. The long-term spectrum of the speech is usually preserved after these manipulations, but the phase spectrum has often been totally changed for each pitch period of the speech (Strube, 1982; Moriya and Honda, 1986; Charpentier and Moulines, 1989; Quatieri et al., 1990). Such manipulations and transformations take advantage of the fact that the human auditory system seems rather insensitive to phase. The phase spectrum does play a role, however, in judging the sound quality (Goldstein, 1967; Plomp and Steeneken, 1969) and the pitch (Houtsma and Smurzynski, 1990), but within limits. Therefore, it is a very interesting and practical question to what degree and how phase plays a role in determining speech sound quality.

Finally, there also has been the question of how to choose the excitation signal in the source-filter model to produce a natural sound. It is a well-known fact that the LPC synthesizer with pulse excitation produces mechanically sounding speech (Markel and Gray, 1976). If an excitation function more similar to the glottal pulse shape is chosen, more natural sounds can be obtained (Rosenberg, 1971; Holmes, 1973). Since the shape of the glottal pulse is determined by its Fourier phase, given a flat amplitude spectrum, these results show that the phase spectrum plays an important role in judging the quality of speech sounds.

We have seen that auditory masking phenomena have a great impact on the processing of speech such as in speech coding. However, the perceptual weights, used in speech coding so far, have been based on the masking patterns of pure tones by noise bands or masking of noise bands by pure tones only. The response of the auditory pathway to

complex signals like speech, however, cannot easily be predicted from the response to simple sinusoids or the results from simple masker-target setups.

It is the intention of the present masking study to make a contribution to the understanding of the perceptual correlates of speech processing, such as speech coding, speech synthesis, and speech manipulation. The masking technique is used not only to provide threshold values of target signals in complex maskers, but also to shed light on how the auditory system processes complex sounds. In speech coding, for instance, quantization noise is the target, speech sounds are the maskers, and the noise should be inaudible in the coded speech. In order to achieve our goal, a series of experiments with periodic pulses and synthetic vowels as maskers and noise bands as the targets are performed. The maskers are still simple compared to speech signals. This simplification, however, makes it possible to systematically study some important aspects of a signal and to facilitate the experiment. The spectrally flat signals are, for instance, also used as the excitations in speech synthesizers for voiced sounds. Based on the experiments, some auditory correlates of speech, related to limitations of the auditory system such as phase sensitivity and masking, are discussed.

6.2 Experimental method

6.2.1 General procedure

A two interval, two alternative, forced choice (2I2AFC), adaptive procedure was used to determine thresholds in all experiments (Levitt, 1971). Each interval contained either 200 ms of masker alone or 200 ms of masker plus target, both intervals including 25-ms sinusoidal onset and offset ramps. The pause between the two sound intervals was 500 ms and the order of two intervals was varied randomly. The level of the target was initially well above the expected threshold. In a two-down one-up procedure, shown in Fig. 6.1, the target level was decreased after two consecutive correct responses at the same signal level, and was increased after each incorrect response. The amount of level change was determined by a varying step size. For each block of trials, a step size of 8 dB was used until the first reversal, which was defined as a transition from down to up or vice versa, as shown in Fig. 6.1. A step size of 4 dB was then used until the second reversal, in order to quickly reach the threshold value. After the third reversal the step size was 2 dB.

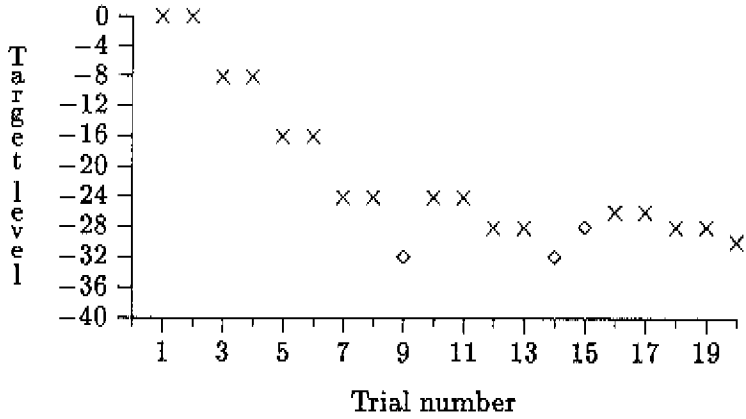


Figure 6.1: Two-down one-up 2I2AFC procedure. × shows a correct response and ○ shows an incorrect response.

The average of the midpoints between consecutive reversals, excluding the first three points, was taken as the threshold level. This procedure theoretically estimates the 70.7% correct response point of a psychometric function. Fourteen reversals were taken for each data point and this procedure was repeated three times for each parameter and subject. The response time was controlled by the subjects.

6.2.2 Stimuli

All stimuli were generated by a computer and had a dynamic range of more than 90 dB (16 bits). In order to reduce spectral shaping by the sample-hold device in the D/A converters, they were operated at a sampling frequency of 20 kHz.

Maskers were synthesized by adding up harmonics of a certain fundamental frequency, according to the following formula:

$$m(t) = \sum_{i=1}^{M} A_i \cos(2\pi i F_0 t + \psi_i) \tag{6.1}$$

where F_0 is the fundamental frequency and M was chosen such that the spectrum of $m(t)$ covered the frequency range up to 10 kHz. When the signal was delivered to the headphones, its upper spectral edge was limited by a lowpass filter with a cutoff frequency of 7.8 kHz (see section Apparatus and subjects).

In the experiments, the global spectral shape and the phase relationship between the harmonics of the masker were manipulated. The

spectral slopes used were 0 dB/oct (flat spectrum), -3 dB/oct, and -6 dB/oct, which corresponded respectively to choosing $A_i = 1$, $A_i = 1/\sqrt{i}$, and $A_i = 1/i$. The phase relationship between the harmonics of the masker was chosen as follows. For zero-phase maskers, ψ_i was equal to zero for all i. For cosine-sine alternating-phase stimuli, ψ_i was equal to $\pi/2$ for odd harmonic numbers and zero for even harmonic numbers. Maskers with two Schroeder-phase conditions were also used where $\psi_i = -i(i+1)\pi/M$ and $\psi_i = +i(i+1)\pi/M$, which will be respectively called masker m_- and masker m_+ for convenience. In Fig. 6.2, examples of the masker waveforms are plotted which have been normalized to the same RMS value. It can be seen that the zero-phase masker has a much larger peak factor than the Schroeder-phase maskers. In addition, the peak factor also becomes smaller for complexes with a tilted spectral slope.

The target signal was either a narrowband or a broadband noise of 200-ms duration. These target signals were calculated by adding equal-amplitude sinusoids with a spectral spacing of 4 Hz:

$$s(t) = C \sum_i cos(2\pi i F_n t + \phi_i) \tag{6.2}$$

where ϕ_i is the phase angle randomly distributed over the range $(-\pi, \pi)$, F_n was equal to 4 Hz, and i was chosen such that the above formula could produce a particular narrow- or wideband noise.

The threshold of the noise band in a specified frequency region was calculated as the ratio of the average energy of the noise to that of the masker in a 1-Hz band. In other words, the threshold, TD, of the noise target was defined as the ratio of the spectral densities between the target and the masker, expressed in decibels:

$$TD = 10\log(\frac{C^2 F_0}{A^2 F_n}) \tag{6.3}$$

where $A^2/2F_0$ and $C^2/2F_n$ are the spectral power densities of the masker and the target at the center frequency of the noise band.

6.2.3 Apparatus and subjects

Stimuli were generated through two equal D/A converters and filtered by two lowpass filters. The cutoff frequency of the lowpass filters was 7.8 kHz, with an attenuation of 90 dB/octave. Programmable analog attenuators were used to control the levels of the maskers and targets.

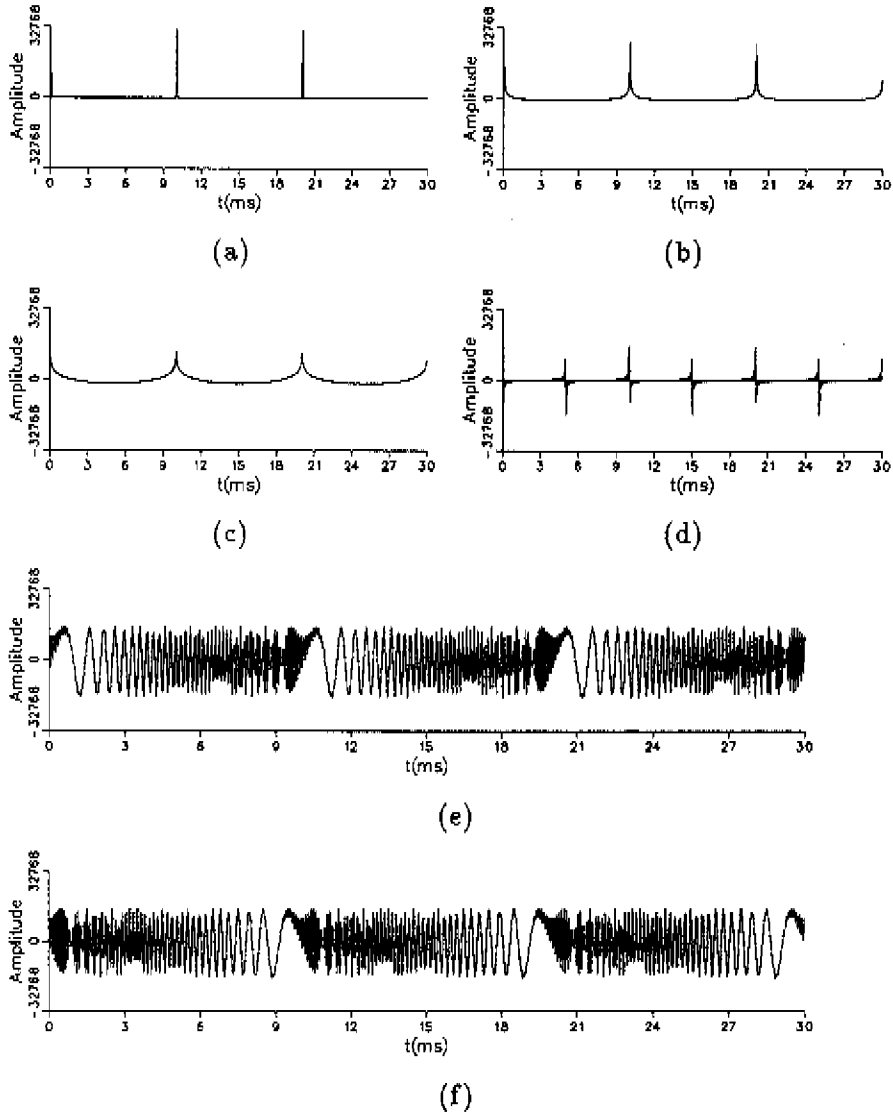


Figure 6.2: (a) Waveform of the masker with a flat spectrum and zero phase. (b) masker with a spectral slope -3 dB/oct and zero phase. (c) masker with a spectral slope -6 dB/oct and zero phase. (d) masker with a flat spectrum and alternating phase. (e) the m_- masker. (f) the m_+ masker. Waveforms in (a)-(d) are normalized to the same RMS value, while the RMS value for the Schroeder-phase maskers is a factor 5 larger. For the latter two, the time scale is also increased by a factor 2.

The stimuli were presented diotically through ETYMŌTIC RESEARCH ER-2 insert earphones, which have a flat spectral response up to 10 kHz. Colleagues from the laboratory as well as paid subjects participated in the experiments. They all were about at age 30 and had normal hearing.

6.3 Experiment 1

Detection of narrowband noise of critical-band width in spectrally-flat and zero-phase harmonic complexes

This experiment is concerned with the threshold of narrowband noise as a function of fundamental frequency of the masker and of the center frequency of the target. The broadband maskers (0-10 kHz) were spectrally-flat harmonic complex sounds with the initial phases of the harmonics set to zero. The fundamental frequencies of the complexes were 100, 150, 200, 250 and 400 Hz.

Noise bands with critical-band widths served as targets. The values of the critical-band width were computed according to the formula proposed by Zwicker and Terhardt (1980). In the low-frequency region, the spectral spacing of the masker components was larger than the bandwidth of the target signal. In this region, the spectrum of the target was either centered on a specific harmonic or placed between two successive harmonics. When the bandwidths of the noise targets were greater than the spacing of two successive harmonics, the noise targets were added without consideration of the harmonic structures of the maskers. The maskers were presented at a sound pressure level of 80 dB.

6.3.1 Results

Three subjects participated in this experiment. Since their results were similar, only the averages of the measurements are presented. Figs. 6.3 (a-e) show the results for masker fundamental frequencies of 100, 150, 200, 250 and 400 Hz, respectively.

One sees from panels (a-c) that in the high-frequency region, the threshold of the target decreases with an increase of the center frequency of the noise band. On the other hand, in the low-frequency region in panels (b) and (c), and to a certain extent in panel (a), the threshold of the noise target increases globally with an increase of the center frequency of the noise band until the center frequency reaches a critical point. This critical point corresponds to the maximum of the threshold and is dependent on the fundamental frequency of the masker.

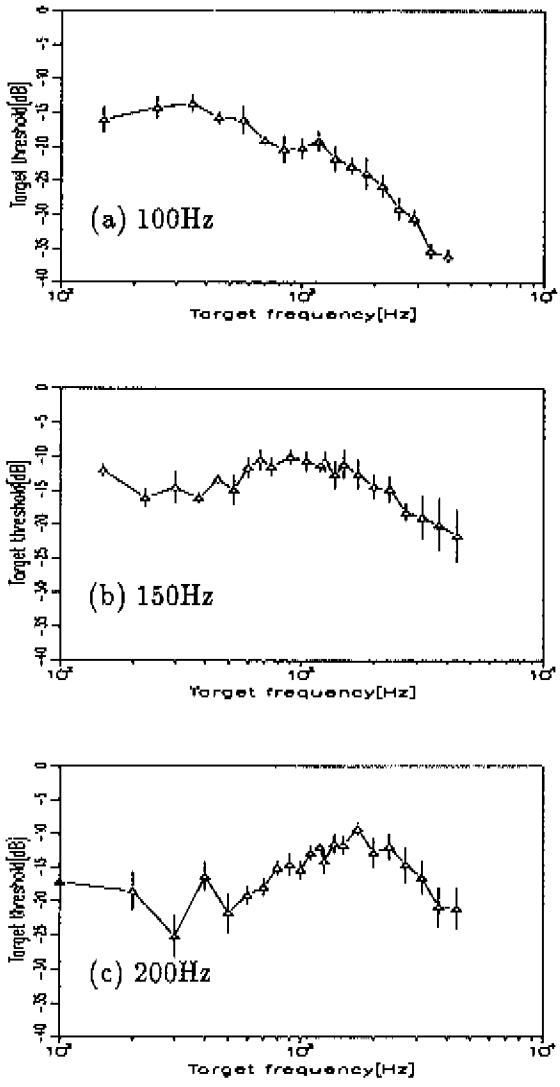


Figure 6.3: See next page.

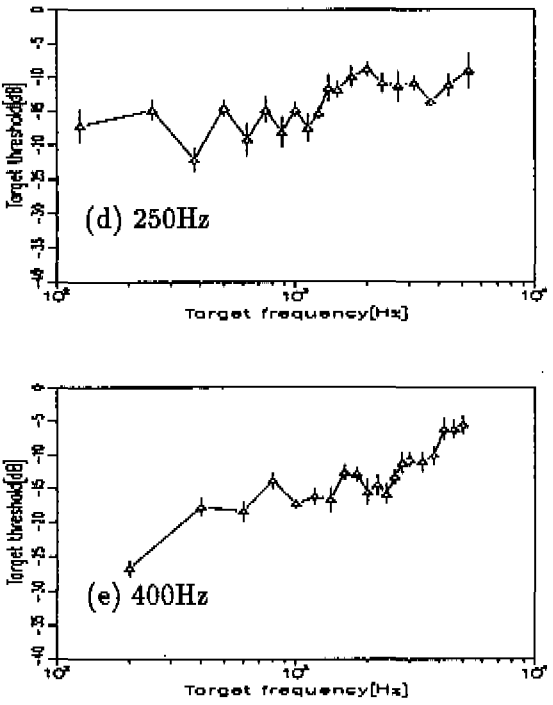


Figure 6.3: Thresholds of noise bands of critical-band width in flat-spectrum, zero-phase maskers are plotted as a function of the center frequency of the noise band. Parameter in the panels is the fundamental frequency of the masker. (a) the 100-Hz masker. (b) the 150-Hz masker. (c) the 200-Hz masker. (d) the 250-Hz masker. (e) the 400-Hz masker. Target threshold represents the ratio of the spectral densities of target and masker expressed in decibels. Masker level is 80 dB SPL.

Besides the global increase of the thresholds of the noise targets towards high frequencies, the thresholds of the noise bands in the low-frequency region show also peaks and dips (most clearly seen in panels c and d). Peaks reflect the thresholds of the noise targets which have a center frequency equal to the frequency of a harmonic of the masker and dips reflect the thresholds of the noise targets which are situated between two successive masker harmonics.

6.3.2 Discussion

The masking patterns in this experiment are strongly dependent on the relationships between the fundamental frequency of the masker and the bandwidth of the auditory filter. In the low-frequency region, if the bandwidth of the filter is smaller than or close to the fundamental frequency of the masker, the threshold for noise targets is predominantly determined by the sharpness of spectral resolution. In the high-frequency region, auditory filters with a wide bandwidth pass through more than three harmonics and the interaction of these harmonics produces a temporally modulated waveform. The detection of noise targets can then easily be realized by temporally analyzing the masker.

Thresholds which are mainly determined by spectral resolution reflect the harmonic structure of the maskers. In the low-frequency region, the masking patterns show clear peaks and dips for the pulse trains with high fundamental frequencies because the critical-band widths at low frequencies are smaller than the space between harmonics. Individual masker harmonics therefore become visible in the masking patterns. For maskers with high fundamental frequencies, the detection is dominated by the spectral resolution up to very high frequencies (see Figs. 6.3 d and 6.3 e).

For maskers with a low fundamental frequency, the threshold of the targets at high frequencies decreases as a consequence of the increasingly better temporal resolution of the auditory channels and as a result of the energy increase of the critical-band noises towards high frequencies. In these situations, the response of the auditory filter to the pulsed maskers decays fast and the filtered waveform are therefore more deeply modulated. At a fixed channel, this modulation becomes shallower for maskers with higher fundamental frequencies. Consequently, the thresholds of the noise bands in maskers with higher fundamental frequencies are higher.

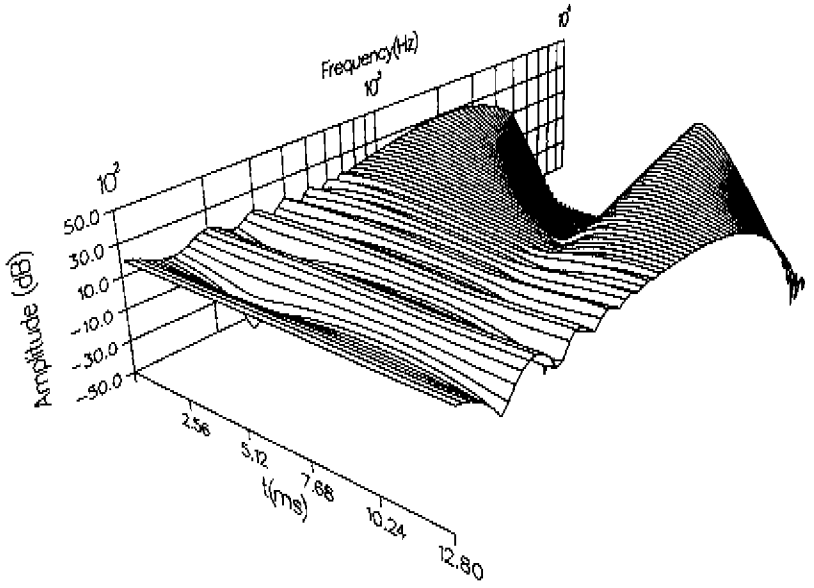
As the critical-band width increases monotonically with the increase

of frequency, the spectral resolution degrades towards high center frequencies and the temporal resolution improves. Masking patterns therefore show global maxima in the middle frequency region in Figs. 6.3 (b-c), but the global maxima are less pronounced in the Figs. 6.3 (a) and (d), corresponding to maskers with fundamental frequencies of 100 and 250 Hz.

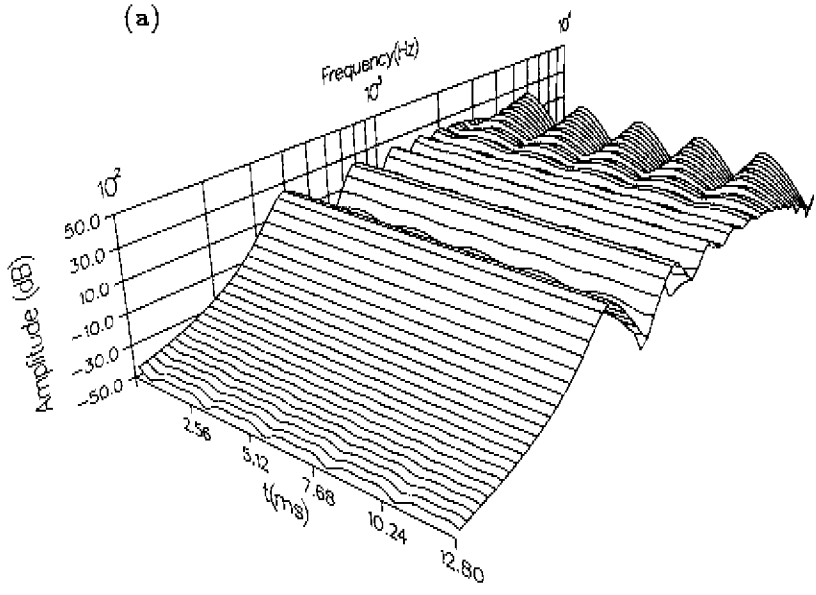
The results suggest that the detection of targets in harmonic complex tones is optimally realized by listening to either the spectral valleys of the masker or the valleys in the temporal envelopes of the masker. Since the envelopes of the auditory filter responses represent a distribution of the energy of the masker in the time-frequency plane, the masking patterns could be qualitatively explained by examining the valleys in the time-frequency distribution of the masker energy. Figs. 6.4 (a) and (b) show the envelopes of the responses of a cochlear-filter bank to the 100 and 400-Hz maskers. Each filter has an impulse response of a gammatone filter (Patterson, 1987). The center frequencies of the in total 128 filters were linearly distributed in the range from 0 to 5 kHz. The envelopes were obtained by Hilbert transform and are represented on a decibel scale. By looking at the low frequency region of both panels in Fig. 6.4, the *spectral* composition of the two maskers becomes apparent. Due to the wider spacing of the harmonics of the 400-Hz masker, the spectral valleys are much wider and deeper for the 400-Hz masker than for the 100-Hz masker. At high frequencies, on the other hand, the *temporal* modulation is the obvious feature. Due to the longer period of the 100-Hz masker, the temporal valleys are much wider and deeper for the 100-Hz masker than for the 400-Hz masker. Finally, one can see that the transition from spectral valleys to temporal valleys occurs at a much higher frequency for the 400-Hz masker than for the 100-Hz masker.

Figs. 6.5 (a) and (b) show two samples of temporal envelopes at 4 kHz for the two maskers. The solid line represents the envelope of the response to the masker alone and the dotted line the response envelope for the masker plus noise target at threshold. This analysis suggests that subjects are indeed able to "listen into the deep temporal valleys" of the 100-Hz masker and that therefore less target energy is necessary to reach the threshold.

These results are, in principle, in agreement with the masking patterns of pure-tone targets masked by harmonic complex sounds (Duifhuis, 1970; Schroeder and Mehrgardt, 1982; Kohlrausch, 1992). Duifhuis (1970) examined the ability of the auditory system to perform spectral and temporal analysis by using zero-phase harmonic complexes

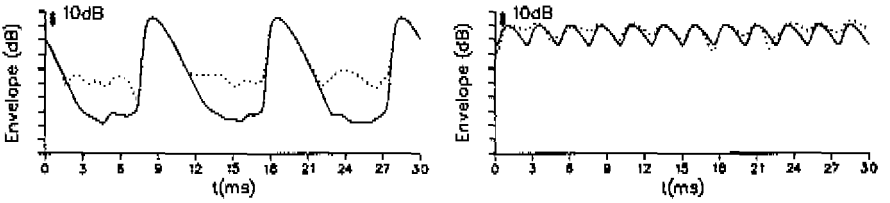


(a)



(b)

Figure 6.4: (a) Envelopes of gamma-tone filter responses to the 100-Hz masker with a flat spectrum and zero phase. (b) Envelopes for the 400-Hz masker with a flat spectrum and zero phase.



(a) (b)

Figure 6.5: Envelopes of gamma-tone filter responses at 4 kHz for (a) 100-Hz and (b) 400-Hz maskers with a flat spectrum and zero phase respectively. Dotted lines show the envelopes for masker plus noise signal at threshold.

as maskers. In one of his experiments he found that a pure tone, which was added in phase with the harmonic component of the masker, was detected by spectral analysis of the maskers for low harmonic numbers, and by temporal analysis of the maskers for high harmonic numbers. In addition, the masking period pattern of the harmonic complexes obtained with pulsed tones as targets (Duifhuis, 1971) clearly showed that the detection of the target was realized by listening during the valleys of the cochlear-filter responses to the maskers.

The masking of a pure tone by a harmonic complex with harmonic amplitude proportional to $1/f$ and zero phase also revealed the ability of the auditory system to perform a temporal analysis of the masker (Schroeder and Mehrgardt, 1982). In this experiment, the threshold of a target tone at 1200 Hz monotonically increased with the increase of the fundamental frequency of the masker, until the fundamental frequency reached 150 Hz, and remained constant for higher fundamental frequencies. Thus, for fundamental frequencies below 150 Hz, the detection of the 1200-Hz target is dominated by temporal analysis. This corresponds to our findings that the threshold of the noise band at 1200 Hz is located at the threshold plateau in the 150-Hz masker, i.e. at the transition between temporal and spectral analysis of the 150-Hz complex.

Finally, a quantitative comparison can be made between the thresholds of noise bands and thresholds of pure-tone signals masked by complex tones of 20 harmonics with a fundamental of 100 Hz (Kohlrausch, 1992). For a comparison, the thresholds from the different experiments have to be converted into ratios of spectral densities of targets and

maskers as defined in our experiment. We can then easily compare the thresholds on the basis of the target power in a critical band due to the fact that the energy of the target in a critical band is integrated for detection. The thresholds from the present measurements expressed as the total power of the narrowband noise targets relative to the spectral level of the masker are 3 dB and 2 dB at frequencies of 500 Hz and 1 kHz respectively. This compares well with the corresponding level of 1 dB and -2 dB found for the pure tone targets (Kohlrausch, 1992).

So far, the masking patterns in this experiment were assumed to be given by the envelope of the output of the auditory filters. No attempt was made to account for nonlinear characteristics and phase dispersion of the auditory system, although they are very important factors to quantitatively describe the masking pattern of temporal masking (Jesteadt et al., 1982; Schroeder, 1975). These two factors will be studied in the experiments that follow. The envelopes of waveforms in the outputs of the filters represent their short-time energies and were already used as a decision criterion to explain masking data (Martens, 1982; Kohlrausch, 1988). It is interesting to see from panels (a) and (b) in Fig.6.5 that the valleys in the envelope of the 100-Hz masker are more than 40 dB below the valleys in the envelope of the 400-Hz masker. The threshold difference, however, is only about 25 dB. It is most likely to assume that further lowpass filtering and neural adaptation processes would reduce the difference in the depth of the valleys. Alternatively one could think of a much narrower bandwidth of the auditory filter than that given by the gamma-tone filter bank. However, in order to explain the experimental difference by narrowing the filter bandwidth only, one would need a reduction factor for the bandwidth of more than three. Thus we conclude that the thresholds are indeed influenced by neural adaptation. The data obtained from the present experiments can therefore only be qualitatively explained from the spectro-temporal representation shown in Figs. 6.4 and 6.5.

6.4 Experiment 2

Masking of broadband noise by spectrally-flat and zero-phase harmonic complex sounds

Masking of critical-band noise by broadband maskers has been studied in the previous experiment. Since the energy of the noise targets was concentrated in one critical band, the detection of the noise targets is

predominantly determined by the auditory response of a single channel. In practice, noise as produced by quantization is a wideband signal. For a broadband noise target, the energy in different critical bands will influence its threshold in different ways, depending on the spectral and temporal structures of maskers. Based on the previous experiment, one can expect that the thresholds of broadband noise in a masker with a low fundamental frequency will be mainly determined by the temporal analysis. For a masker with a high fundamental frequency, on the other hand, thresholds will be determined by auditory frequency analysis. To investigate this experimentally, the energy of a broadband noise target can be limited to contain only high frequencies or low frequencies by highpass or lowpass filtering of the noise.

Broadband noise signals were generated in this experiment by adding sinusoids from 10 Hz to 5 kHz with a frequency spacing of 4 Hz. The targets were produced from these broadband noise samples by lowpass or highpass filtering through a programmable DIFA filter that had a cutoff attenuation of 180 dB/oct. The thresholds of the targets were determined as a function of the cutoff frequency of the filter. Maskers were spectrally-flat and zero-phase harmonic complexes with fundamental frequencies of 100, 200, and 400 Hz. Maskers were presented at a sound pressure level of 80 dB.

6.4.1 Results

The results, being the average from two subjects, are shown in Fig. 6.6. The panels (a-f) show the thresholds of the filtered noise target in the presence of maskers with fundamental frequencies of 100, 200 and 400 Hz, respectively. The three panels to the left show results for lowpass-filtered and the three panels to the right show results for highpass-filtered noise (Δ in the figures). For a comparison, the diamonds ($\diamond$) in the figure show threshold values which are estimated from the threshold patterns of the critical-band noise targets for the three maskers in Fig.6.3. The estimation is based on the minimum of the critical-band-noise thresholds within the passband of the broadband noise.

One sees from panel (a) in Fig. 6.6 that the threshold of the lowpass-filtered noise in the 100-Hz masker increases sharply with decreasing cutoff frequency. For the maskers with fundamental frequencies of 200 and 400 Hz, however, thresholds remains rather constant in the cutoff-frequency range from 500 Hz to 5 kHz, as can easily be seen in panels (b) and (c).

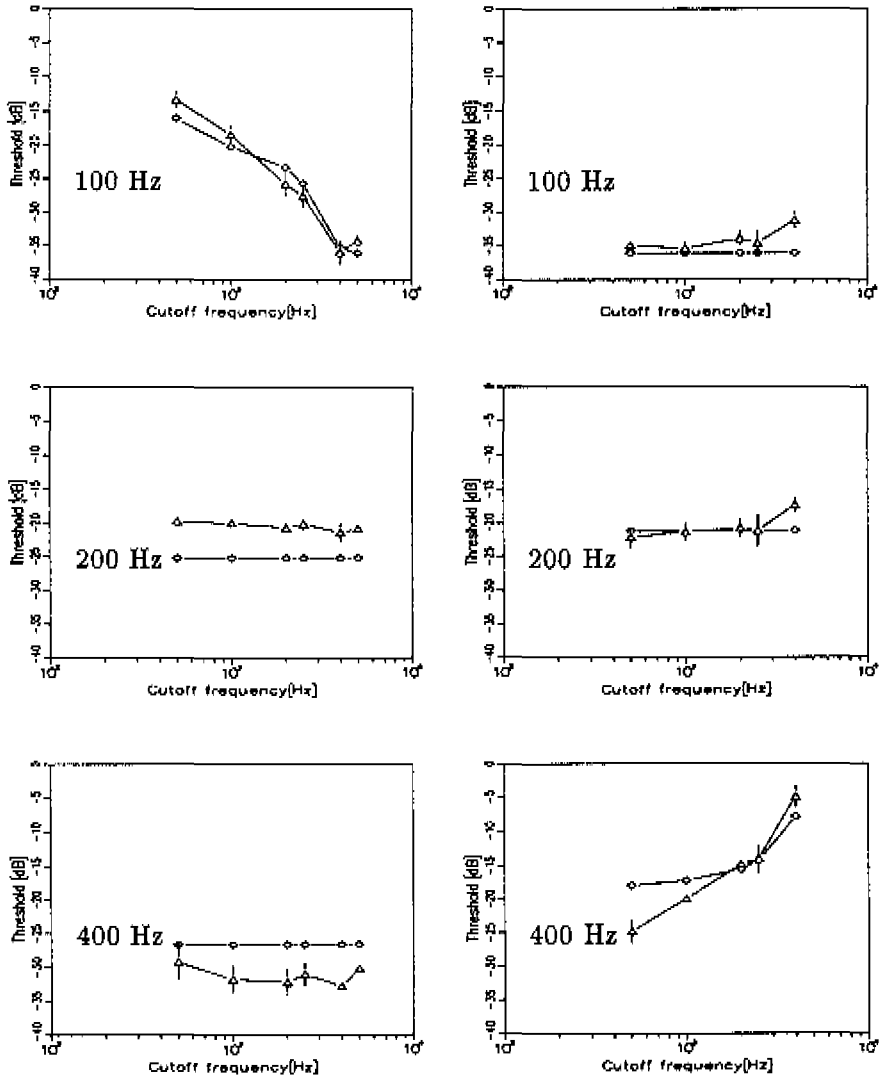


Figure 6.6: Left panels: Thresholds of lowpass-filtered noise targets (△) are plotted as a function of the cutoff frequency. Right panels: Thresholds for highpass-filtered noise targets (△). Threshold values (◇) are estimated from the minima in the pass-band regions of the masking patterns in Fig.6.3. Parameter is the fundamental frequency of the maskers. Masker level is 80 dB SPL.

The threshold of the highpass-filtered noise target in maskers with a fundamental frequency of 400 Hz decreases significantly with a decrease of the cutoff frequency (panel f). For maskers with fundamental frequencies of 100 and 200 Hz, thresholds show little change in the cutoff frequency range from 500 Hz to 4 kHz.

6.4.2 Discussion

As predicted from Experiment 1, the masking of broadband noise targets by harmonic complex tones is mainly determined by the detection of these noise targets in specific frequency regions. In these regions, the masker is best resolved either in the frequency domain or in the time domain. For lowpass-filtered noise targets, the threshold in the masker with a fundamental frequency of 100 Hz increases with a decrease of the cutoff frequency of the lowpass filter (see △, Fig. 6.6 a). This is in line with the experimental finding in Experiment 1, as can be seen from the nearly identical course of the prediction (see ◇). Because the threshold minimum for the narrowband noise targets in a masker with a fundamental of 200 Hz was situated in the low-frequency region (see Fig. 6.3c), lowpass filtering of a wideband noise target does not reduce the noise energy in this region and therefore does not significantly influence its threshold. This argument also applies to the threshold for lowpass-filtered noise targets in the 400-Hz masker, where the detection of noise target is also mainly determined by the presence of noise energy in the low frequency region.

For highpass-filtered noise targets, where the detection of targets in the 100-Hz masker is predominantly determined by its high-frequency energy, the decrease of the cutoff frequency does not influence detection and the threshold remains constant (see △ in Fig. 6.6d). On the other hand, when the threshold of the noise target is mainly determined by its lowest spectral component, as is true for the 400-Hz masker, detection of noise targets is further improved by a decrease of the cutoff frequency and the threshold decreases (see △ in Fig. 6.6 f). The thresholds of highpass-filtered noise targets agree well with the threshold minima (◇) of the noise bands, which are estimated from the region above the cutoff frequency in Fig.6.3.

6.5 Experiment 3

Masking of narrowband noise by broadband harmonic complex sounds as a function of spectral tilt and level

The detection of noise targets in maskers composed of equal-amplitude harmonics has been studied in the two previous experiments. Since natural sounds such as speech have in general some spectral tilt, it is more realistic to investigate the detection of noise targets in such maskers. Tilting the spectrum of the masker while keeping its overall level constant will redistribute the energy of the masker in the frequency domain. The local spectral level of the masker is therefore a function of component frequency. In addition, the waveform of the maskers is more dispersed in time than for the flat-spectrum signal (see Fig. 6.2). Therefore, for the purpose of separating the influence of masker level and of spectral tilt on threshold, the maskers with spectral tilts are presented at different sound pressure levels.

In a pilot experiment that was concerned with spectral slope effects only, the narrowband noise targets were centered at 1000, 1414, 2000, 2828, and 4000 Hz. Their bandwidth was equal to 10% of their center frequency, which is somewhat less than their corresponding critical-band width. The maskers were spectrally tilted by 0, -3, and -6 dB/oct and consisted of zero-phase harmonics with a fundamental frequency of 100 Hz. The maskers were presented at an overall sound pressure level of 64 dB.

In the experiment that dealt with masker level effects, narrowband noise targets were centered at frequencies of 500, 1000, 2000 and 4000 Hz, with bandwidths of 100, 100, 200, and 400 Hz, respectively. The sound pressure level of the masker was changed from 44 dB to 64 dB in steps of 5 dB for maskers with a spectral slope of 0 dB/oct, and in steps of 10 dB for maskers with spectral slopes of -3 and -6 dB/oct.

An important remark is necessary with respect to the expected level effects. First of all, all thresholds are expressed relative to the spectral density of the masker. Thus, if target thresholds (expressed in dB SPL) vary in the same way as the masker level, the (relative) target thresholds remain constant. If target thresholds increase less than the masker level, the (relative) thresholds decrease, and we say that the target becomes better audible at higher masker levels, This behaviour is expected for temporal resolution. In the case of spectral resolution, on the other hand, target thresholds increase faster than the masker level. Thus the

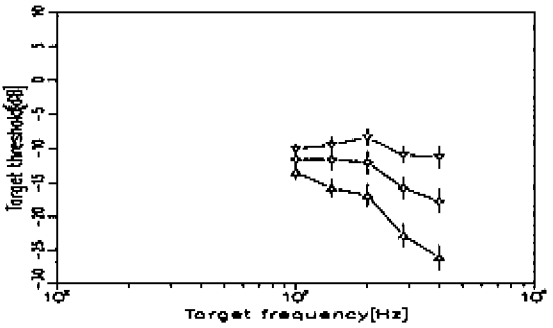


Figure 6.7: Thresholds of noise bands in the 100-Hz maskers with spectral slopes of 0 dB/oct (△), -3 dB/oct (◇) and -6 dB/oct (▽). Masker level is 64 dB SPL.

(relative) thresholds increase and we say that the target becomes less audible at higher masker levels.

For spectrally tilted maskers, the spectral level is of course a function of frequency. The threshold of the narrowband noise targets was therefore defined as the spectral density relation between the target and the masker *at the center frequency of the target noise*, expressed in decibels.

6.5.1 Results

Results for the pilot experiment were obtained from three subjects. The average thresholds are plotted in Fig. 6.7. One sees that the threshold for the noise targets in the spectrally-flat masker is the lowest and that it decreases strongest towards high frequencies. The threshold for the masker with a spectral slope -6 dB/oct is the highest and remains rather constant in the whole frequency region.

The measurements for variable masker level were performed with two subjects. The average thresholds are plotted in Figs. 6.8 (a-c) against the center frequency of the narrowband noise target, with the sound pressure level as a parameter. One sees from Figs. 6.8(a-c) that the threshold of the narrowband noise signal increases when the sound pressure level decreases from 64 to 44 dB. The threshold increase is the largest for the spectrally-flat masker. The increase of thresholds is also the largest at high frequencies. In general, the increase of thresholds due to the decrease of masker level is reduced in the maskers with a spectral slopes of -3 and -6 dB/oct. Towards low frequencies, the level effect becomes

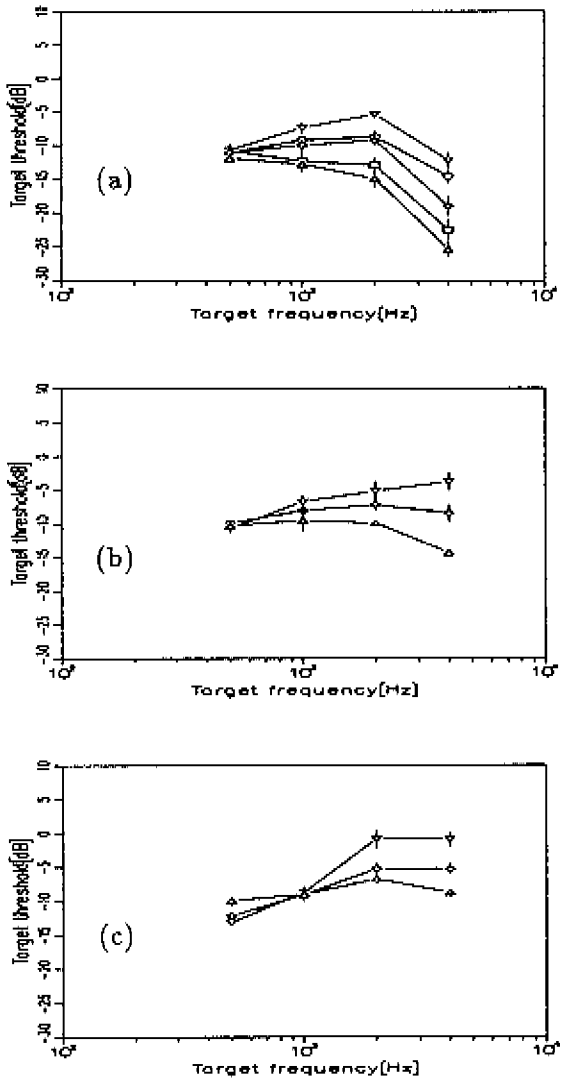


Figure 6.8: Thresholds of noise bands in the 100-Hz zero-phase maskers with spectral slopes of (a) 0 dB/oct, (b) -3 dB/oct, and (c) -6 dB/oct. The masker levels are 64 dB (△), 59 dB (□), 54 dB (◇), 49 dB (○) and 44 dB (▽).

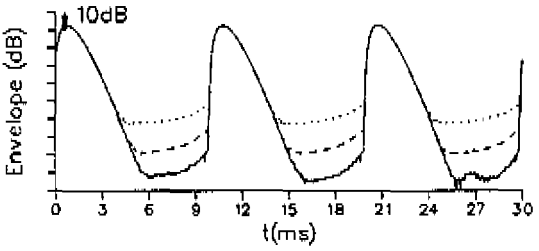


Figure 6.9: Envelopes of gamma-tone-filter responses at 4 kHz for the 100-Hz maskers with three different spectral slopes: 0 dB/oct (solid line), -3 dB/oct (dashed line) and -6 dB/oct (dotted line). The maskers have the same harmonic amplitude at 4 kHz.

small for all three maskers. In addition, for the masker with a spectral slope of -6 dB/oct, one sees that the threshold of the target at 500 Hz *increases* slightly with the increase of the masker level.

6.5.2 Discussion

The results shown in Fig. 6.7 indicate that the spectral slope of the masker has significant influence on the thresholds of narrowband noise targets. The differences between the thresholds in the three spectrally-tilted maskers increase with an increase of the center frequency of the noise signal. There are two factors contributing to these threshold differences. On the one hand, the modulation depth in the temporal waveform at the output of auditory filters is reduced as a result of spectral tilt. This can be seen from Fig. 6.9, where the envelopes of the auditory filter responses to the three maskers at 4 kHz are plotted on a decibel scale. The auditory filter is here simulated by a gamma-tone filter and the spectral levels of the three maskers at 4 kHz are identical. The envelopes for the three maskers with spectral slopes of 0, -3 and -6 dB/oct are shown by solid, dashed and dotted lines, respectively.

On the other hand, due to the spectral tilt, the local spectral levels are different if the three maskers are presented at the same overall level. As the detection of narrowband noise targets in the measurements is predominantly determined by temporal resolution of the auditory system, similar level effects as seen in temporal masking will influence the noise thresholds in our measurements. It has been shown that in forward masking the ratio between changes in threshold of the target and

changes in sound pressure level of the masker is less than one (Jesteadt et al., 1982). In other words, the target threshold does not correspond to a constant target-to-masker ratio; it decreases with an increase of masker level. Since, in our measurements, a steeper slope of the masker level is associated with a "lower masker level", these level effects are a second contribution to the high thresholds for sloped vs. flat-spectrum maskers.

In contrast to thresholds at high frequencies, the threshold of a noise target at 500 Hz in a masker with a spectral slope of -6 dB/oct *increases* somewhat with an increase of masker level (see Fig.6.8 c). This behaviour is expected, if spectral resolution and, especially, upward spread of masking plays a dominant role (Wegel and Lane, 1924; Ehmer, 1959; Schöne, 1977). For a masker with a spectral slope of -6 dB/oct, the spectral level at low frequencies is quite high. For example, the first harmonic of the masker is 18 dB higher than in the spectrally-flat masker. Due to the upward spread of masking, the threshold should therefore increase with an increase of masker level.

We have suggested two possible factors which contribute to the threshold differences at high frequencies in Fig. 6.7 for the spectrally-tilted maskers. It is not clear, however, how much each of the two factors contributes to the differences. This leads us to the discussion of the results obtained by varying the overall masker level. The results shown in Figs. 6.8(a-c) indicate that the threshold of the noise target (expressed relative to the spectrum level of the masker) changes systematically as the sound pressure level changes. The rate of the threshold change is strongly dependent on the center frequency of the noise target. This is clearly shown in Fig. 6.10 for the spectrally-flat masker. The thresholds are replotted as a function of sound pressure level of the maskers with the center frequency of the noise target as a parameter. The data in Fig. 6.10 are well fitted by straight lines. One sees from Fig. 6.10 that the slope of the straight lines becomes steeper with an increase of the center frequency of the noise target. This implies that the influence of level effects on the target thresholds decreases as the auditory responses to the masker become less modulated.

We have shown that the change of threshold can be caused by the change of the temporal waveforms and the change of local sound pressure levels. Since the thresholds of the narrowband noise targets in the spectrally-flat masker as a function of masker level are fitted quite well by straight lines, we can interpolate these data to an arbitrary sound pressure level in the region we have used. In this way, we can compensate

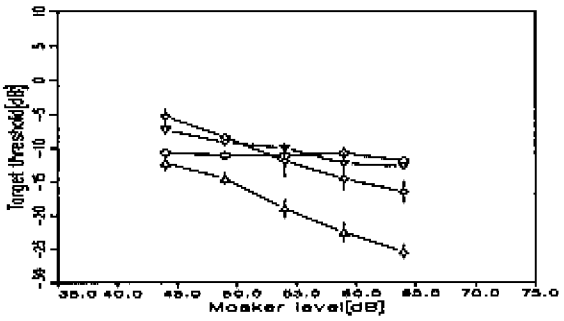


Figure 6.10: Thresholds of noise bands in spectrally-flat, zero-phase maskers as a function of masker level. The center frequency of the noise band is the parameter: 4000 Hz (△), 2000 Hz (◇), 1000 Hz (▽) and 500 Hz (○).

the level effects and isolate the effect of temporal envelope dispersion. To do so, we first determine the local spectral levels at 1, 2, and 4 kHz for maskers with spectral slopes of -3 and -6 dB/oct. The thresholds in a spectrally-flat masker for these spectral levels are then calculated by interpolating the data in Fig. 6.10. These calculated values are finally compared to the measured thresholds in the sloped maskers (see Table I).

It can be seen from Table I that the thresholds of the noise targets for the two spectrally-tilted maskers are higher than in the spectrally-flat masker at all three frequencies. For example, the threshold difference

Slope\Freq.	1 kHz	2 kHz	4 kHz
-3 dB/oct	4.4 dB	6.1 dB	9.1 dB
-6 dB/oct	3.5 dB	5.0 dB	7.2 dB

Table I: The differences between the thresholds of narrowband noise targets in spectrally-flat maskers and in two spectrally-tilted maskers. The spectrally-flat masker and the spectrally-tilted maskers have the same spectral level at the center frequencies of noise bands. The threshold differences are referred to an overall level of 64 dB for the maskers with spectral slopes of -3 and -6 dB/oct.

between a noise target at 2 kHz in a spectrally-flat and a spectrally-tilted masker with a slope of -3 dB/oct is 6.1 dB. This threshold difference can only be caused by the difference in envelope modulation resulting from the spectral tilt of the masker. For the masker with a spectral slope of -6 dB/oct, the response envelopes become even less modulated due to the low spectral levels at high frequencies. Threshold differences between maskers with spectral slopes of 0 and -6 dB/oct are therefore smaller than threshold differences between maskers with spectral slopes of 0 and -3 dB/oct.

6.6 Experiment 4

Phase difference limens (DLs) in spectrally-tilted harmonic complex sounds

We have shown in the previous experiments that thresholds of the noise bands at high frequencies are mainly determined by temporal analysis of the masker waveform and that this temporal analysis is strongly influenced by the fundamental frequency, the level, and the spectral shape of the maskers. Studies of auditory phase sensitivity for equal-amplitude harmonic complex sounds also showed that, especially at high frequencies, the phase-shifted component was detected by temporal analysis of the stimuli (Schroeder, 1959; Schroeder and Strube, 1986; Patterson, 1987; Moore and Glasberg, 1989). Typically, phase difference limens of a phase-shifted high harmonic in zero-phase complexes with equal amplitude harmonics were measured (Moore and Glasberg, 1989). These phase DLs indicate the smallest detectable changes of stimulus waveform by temporal analysis. Moreover, the phase DLs and the thresholds of the noise targets could be related because a phase-shifted harmonic is mathematically equivalent to adding another harmonic with the same frequency, but different amplitude and phase values. The precise quantitative relation is given by:

$$\begin{aligned} cos(2\pi nF_0t + \varphi) = {} & 2sin(\varphi/2)sin(2\pi nF_0t + \pi + \varphi/2) \\ & + cos(2\pi nF_0t) \end{aligned} \tag{6.4}$$

For convenience, this added component is called target and the original signal is called masker. The temporal analysis for detection of noise targets can therefore be further manifested by measuring phase DLs of each harmonic component of the masker. The study of phase sensitivity in complex tones is also of growing practical value in, for example,

speech coding and speech perception. Therefore, in this experiment, the measurement of phase sensitivity of individual harmonics is extended by using broadband harmonic complexes with different spectral slopes.

For a comparison with masking experiments, we also express the phase difference limen as threshold of the added component (amplitude ratio between the added component and the original harmonic component, expressed in decibels), i.e.

$$TD = 20log2sin(\varphi/2). \tag{6.5}$$

where φ takes values from 0 to 180 degree. A phase change of 180 degree is equivalent to a threshold value of 6 dB. Of course, this definition of threshold is meaningful only when the phase shift is detectable.

The maskers in this experiment were zero-phase harmonic complexes with three different fundamental frequencies of 100, 200 and 400 Hz and two spectral slopes of 0 and -6 dB/oct, which are the same as used in Experiment 3. The 100-Hz masker with two spectral slopes was presented at an overall sound pressure level of 70 dB. For the 200 and 400-Hz maskers, the harmonic components at the same frequencies were presented at the same amplitude as for the 100-Hz maskers with corresponding spectral slopes.

The threshold of the added component was determined by using the same procedure as we used before. In order to make the added component audible, the value of ϕ for the added component was initially set to 180 degree and the amplitude was additionally increased by 6 dB. If the threshold was greater than 6 dB, then the phase-shifted component was not detectable and the corresponding phase value was set to 180 degrees.

6.6.1 Results

The just noticeable phase-shifts (average for two subjects) are plotted in Figs. 6.11 (a) (c) and (e). These values of phase shift are also expressed as thresholds of the added components in the 100, 200 and 400-Hz maskers and plotted as a function of harmonic frequency respectively in Figs. 6.11 (b), (d) and (f) ($\triangle$ for 0 dB/oct and $\diamond$ for -6 dB/oct slopes).

One observes that for high-frequency harmonics, thresholds are generally low for the maskers with fundamental frequencies of 100 and 200 Hz and with a flat spectrum. For maskers with a fundamental frequency of 100 Hz, threshold differences between the two spectral slopes

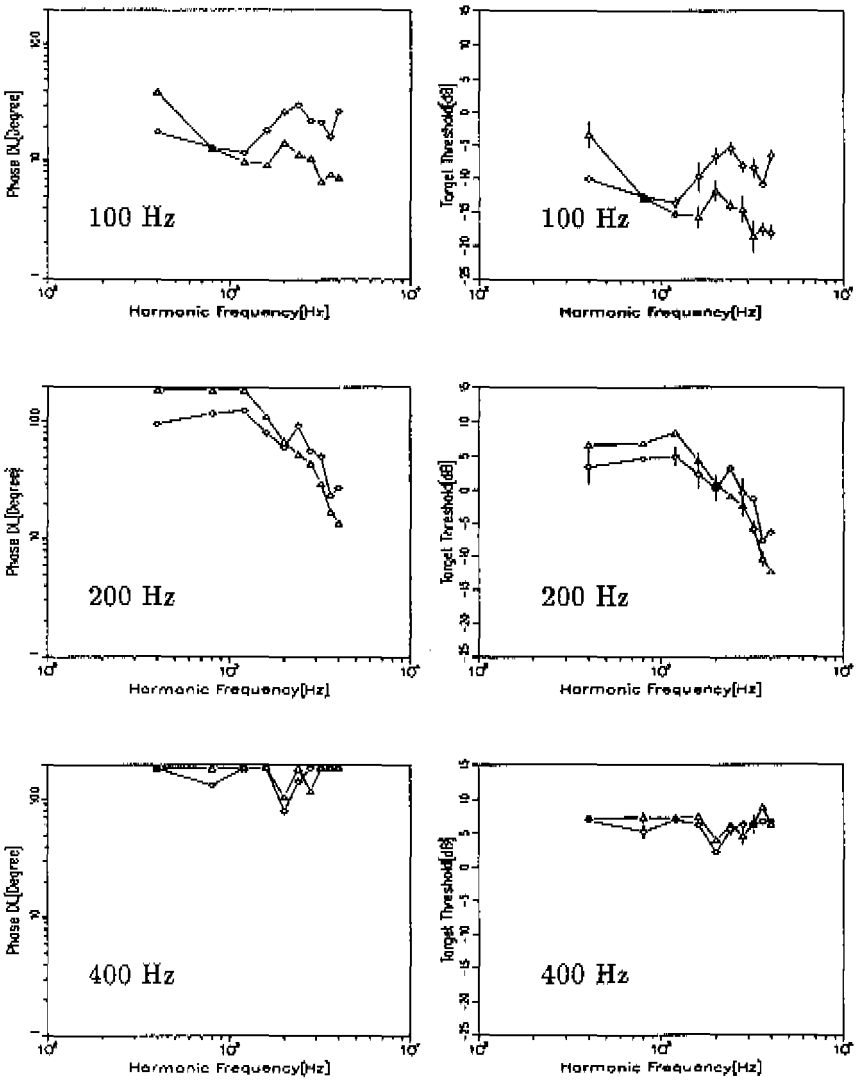


Figure 6.11: Left panels: Phase DL expressed in degree of phase shift in stimuli with spectral slopes of 0 dB/oct (△) and -6 dB/oct (◇). Right panels: Results from corresponding left panels transformed into threshold of the added component. Parameter is the fundamental frequency of the stimulus.

are large. For maskers with fundamental frequencies of 200 and 400 Hz, on the other hand, the thresholds are similar for the two spectral slopes. The phase shift is hardly detectable for any harmonic of the 400-Hz stimulus and for low-frequency harmonics of the 200-Hz stimulus with a spectral slope of -6 dB/oct.

6.6.2 Discussion

When the detection of phase-shifted harmonics is dependent on the temporal resolution of the auditory system, the phase difference limens must be accordingly dependent on the fundamental frequency and the spectral slope of the masker. For the 100-Hz stimuli, the phase DLs are the lowest and are mostly affected by the spectral slope. This is because the masker with a low fundamental frequency and a high spectral level results in a deeply-modulated temporal waveform. This phenomenon has also been observed in Experiment 3. The phase DLs for the phase-shifted components are therefore larger for the spectrally-tilted stimulus. They also generally increase with the fundamental frequency. For the 400-Hz stimuli with two spectral slopes, hardly any phase shifts are detectable. In this case, the valleys in the envelopes of responses for this masker become shallow and the added component becomes difficult to detect. In other words, the phase-shifted component is not resolved temporally.

The phase-shifted high harmonic appears to pop out of the complex and the cue for detection is the same as in a masking experiment (Duifhuis, 1970; Moore and Glasberg, 1989). As a result, the threshold of the added component decreases with the increase of harmonic number at high frequencies. Threshold patterns of the added components are therefore similar to masking patterns obtained in Experiment 3. For low-frequency components, on the other hand, the detection is mainly based on change of sound quality, such as a change of timbre or roughness of the sound (Plomp and Steeneken, 1969; Duifhuis, 1970; Patterson, 1987; Moore and Glasberg, 1989). In this case, the number of harmonics in a critical band is less than three and the interpretation of of a phase change as an addition of another component (see equation 6.4) becomes psychoacoustically less relevant. The threshold pattern of the added component is therefore different from the masking pattern obtained in Experiment 3 (see Fig. 6.11 for the 100-Hz masker). If the harmonic components are well resolved as for the 400-Hz masker, the phase shift is no longer detectable.

A direct comparison can be made with results from the phase DLs

obtained by Moore and Glasberg (1989). They measured the phase DL of each harmonic of zero-phase complex tones consisting of 20 equal-amplitude harmonics. They found that the phase DLs for the 10*th* harmonics of the 100-Hz and 200-Hz stimuli were about 10 degrees. Our phase DLs at this harmonic number are 10 degrees for the 100-Hz stimuli and about 68 degrees, however, for the 200-Hz stimuli due to low stimulus levels. In the measurements of Moore and Glasberg, the phase DLs increased for high-frequency harmonics. This result is obviously different from our measurements where the phase DLs decrease for higher component frequencies. The difference is explained by the different bandwidths of the complexes in the two studies. Moore and Glasberg used bandlimited complexes and the increase of the phase DL occurred at the upper spectral edge. In our experiment, the stimuli were wideband up to 7.8 kHz and measurements were obtained up to 4 kHz.

From Experiment 3, we expect that the phase DLs at high frequencies decrease with an increase of masker level. Our results indeed show that phase DLs for the 100-Hz stimuli with a low spectral level (due to the spectral slope of the stimuli) are large. This increase of phase DLs due to a low level is in agreement with observations by Patterson (1987) and Moore and Glasberg (1989). Since the detection of a phase shift for high-frequency harmonics is mainly based on the temporal analysis of the stimulus, phase difference limens for unresolved harmonics are totally equivalent to detection of an added component.

6.7 Experiment 5

Masking of narrowband noise by broadband harmonic complexes with different phase relations

The previous experiments have shown the importance of the temporal structure of complex-tone maskers in the masking of noise bands. In those experiments, all phases of the masker components were set to zero. The temporal structure of a harmonic masker, however, is very dependent on the phase relationships between its harmonic components. For this reason, the influence of phase on masking of narrowband noise is investigated in this experiment. From many possible phase relationships between the harmonic components of a broadband masker, three special phase relationships reported in the literature (Schroeder and Mehrgardt, 1982; Patterson, 1987) were chosen. One is a Schroeder-phase, another one is alternating-phase, and the third is the zero-phase condition that

has been used so far in all experiments. The reason to include Schroeder-phase maskers is that they have low peak factors of the temporal waveform. This contrasts with the zero-phase maskers which have high peak factors. Waveforms of the maskers with alternating phase are quasi periodic, with the quasi-period being half the period of the zero-phase masker.

In the first experiment, the detection of noise targets in two types of Schroeder-phase maskers (m_- and m_+) with a fundamental frequency of 100 Hz was investigated. The maskers were presented at sound pressure levels of 44 and 64 dB. The noise targets were centered at frequencies of 500, 1000, 2000 and 4000 Hz, with bandwidths of 100, 100, 200 and 400 Hz, respectively.

In the second experiment, the masker with a fundamental frequency of 100 Hz had an alternating-phase relationship for all harmonic components. The targets were narrowband noise signals of critical-band width, as used in Experiment 1 (see page 83). For convenience of comparison with the results from Experiment 1, the maskers were presented at a sound pressure level of 80 dB.

In the third experiment, we studied how an alternating-phase relationship below a certain frequency influences the detection of a high-frequency target. Maskers were computed in such a way that the harmonics below a certain frequency (transition frequency) had an alternating-phase relationship and zero phase above that frequency. The maskers had fundamental frequencies of 100, 200 and 400 Hz and were presented at 64 dB SPL. The target was a narrowband noise of 100 Hz width, centered at 5 kHz.

6.7.1 Results

Thresholds of noise targets in two Schroeder-phase maskers (m_- and m_+) were obtained for two subjects and the average thresholds are plotted in Fig. 6.12. In addition, thresholds in zero-phase maskers with the same levels are plotted for comparison. One can observe that there are significant threshold differences for the three types of maskers, but only at high center frequencies of the noise targets. For the maskers at 64 dB SPL, the threshold for targets at high frequencies is the lowest in the zero-phase masker and is the highest in the m_- masker. The difference is 16 dB. For the maskers at 44 dB SPL, the differences between the thresholds are reduced and the thresholds for the m_+ and m_- maskers are nearly identical.

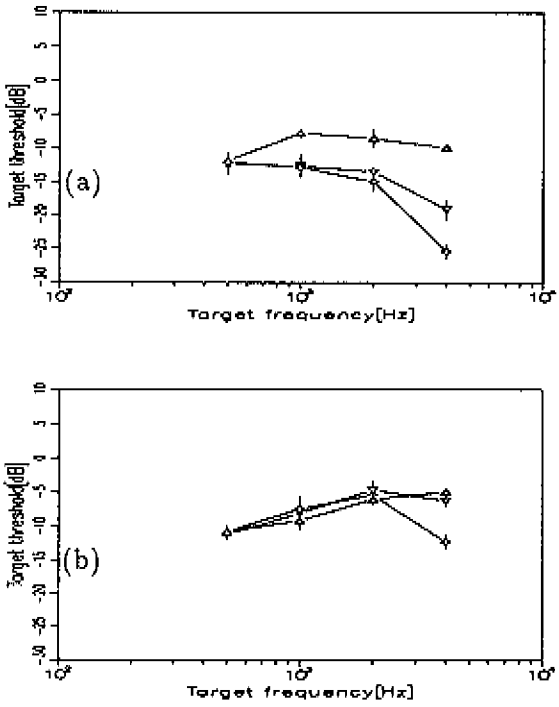


Figure 6.12: Thresholds of noise bands for three maskers with a fundamental of 100 Hz: the zero-phase masker (◇), the m_- masker (△) and the m_+ masker (▽). Panel (a) for masker level 64 dB and panel (b) for masker level 44 dB.

Results for the second experiment are obtained for three subjects and the average thresholds are shown in Fig. 6.13. For comparison, the thresholds of critical-band noise targets in zero-phase maskers with fundamental frequencies of 100 and 200 Hz are replotted from Fig. 6.3. Comparing all thresholds for low target frequencies, the thresholds for alternating-phase and zero-phase maskers are close if the maskers have the same fundamental. For high target frequencies, on the other hand, thresholds of the 100-Hz alternating-phase masker are similar to the 200-Hz zero-phase masker within some 5 dB.

For the third experiment, the thresholds of the narrowband noise signal, centered at 5 kHz, are obtained for five subjects, and the averages are presented in Fig. 6.14. The noise threshold is plotted as a function of the transition frequency below which the harmonics are in sine-cosine

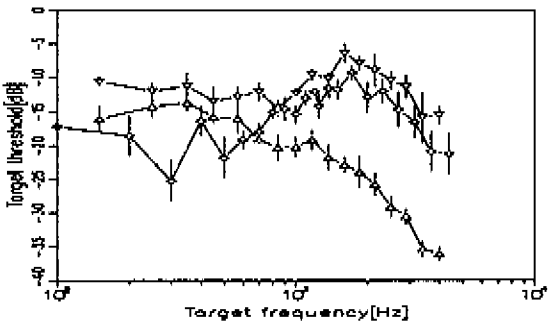


Figure 6.13: Thresholds of critical-band noise targets in a 100-Hz masker with an alternating phase for all harmonics (▽). Thresholds of the noise bands in a 100-Hz zero-phase masker (△) and in a 200-Hz zero-phase masker (◇) are plotted for comparison. Masker level is 80 dB SPL.

alternating phase. Thresholds for maskers with three fundamental frequencies of 100 (△), 200 (◇) and 400 Hz (▽) are plotted in the same panel. One can see that the noise threshold does not change for low transition frequencies for all three maskers. As the transition frequency increases, however, the threshold starts to increase at about 1.6 kHz for the 100-Hz masker and at about 3 kHz for the 200-Hz masker. The increment is about 15 dB in the 100-Hz masker and is about 4 dB in the 200-Hz masker. For the 400-Hz masker, a small decrease in threshold is observed for the two highest transition frequencies.

6.7.2 Discussion

Schroeder-phase maskers

As the detection of noise targets at high frequencies is predominantly determined by the temporal waveform of the masker, it is expected that the target threshold for the two types of Schroeder-phase maskers is higher than that for the zero-phase masker. Fig. 6.12 indeed shows that the noise targets masked by the zero-phase masker have the lowest thresholds.

The difference between thresholds for the two Schroeder-phase maskers is not manifested by the envelopes of their waveforms. One notices from panels (e) and (f) in Fig.6.2 that the two maskers are time-reversed versions of each other. It is of course reasonable to use the en-

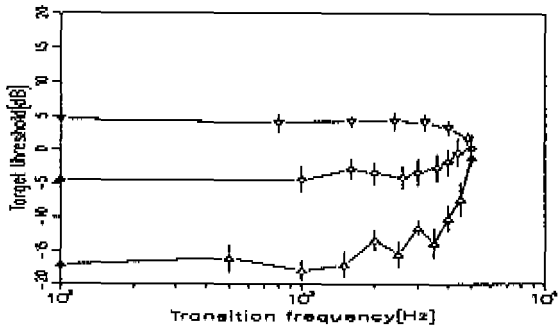


Figure 6.14: Thresholds of a noise band at 5 kHz in maskers with a frequency-dependent phase relationship. Below the transition frequency indicated at the abscissa, the components are in sine-cosine alternating phase. The components above the transition frequency are in zero phase. Parameter is the fundamental frequency of the maskers: 100 Hz (△), 200 Hz (◇) and 400 Hz (▽). Masker level is 64 dB SPL.

velopes of the cochlear-filter responses to the two maskers to explain the difference. The gamma-tone filter (Patterson and Moore, 1986) is used here to calculate the envelopes of the responses to the three maskers. Response waveforms at 4 kHz and their envelopes plotted on a decibel scale are shown in Fig. 6.15 (a-c), respectively. It turns out that the envelope (plotted on a decibel scale) for the zero-phase masker shows deeper and wider valleys than for the m_- and m_+ maskers. The envelopes for the two Schroeder-phase maskers, however, appear quite similar and they cannot convincingly account for the large threshold differences observed in the data.

An alternative approach is to choose a basilar-membrane (BM) model the properties of which are based on measurements of basilar-membrane motion and of neurophysiological responses (e.g., Schroeder, 1972; Allen, 1978; De Boer, 1980; Viergever, 1980; Strube, 1985). The cochlear model implemented by Strube (1985) provides a reasonably good fit of the phase response of the cochlear filter. Temporal envelopes derived from this filter have been shown to be able to explain the threshold differences between the two Schroeder-phase maskers (Strube, 1985; Smith et al., 1986; Kohlrausch, 1988). This BM model is therefore used to calculate the cochlear response to the three maskers. In the calculation of response envelopes, the parameters of the model are set

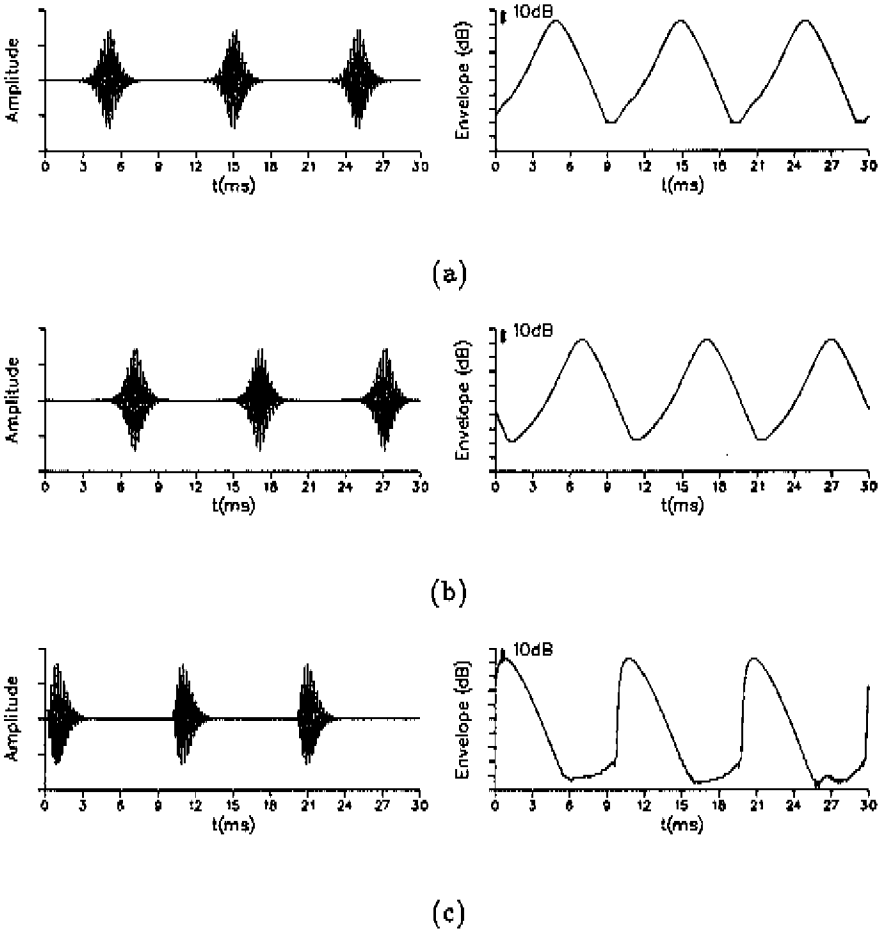


Figure 6.15: Left panels: Waveforms of gamma-tone filter responses to (a) the m_- masker, (b) the m_+ masker, and (c) the zero-phase masker, respectively. Right panels: Corresponding envelopes, plotted on a decibel scale. The resonance frequency of the filter is 4 kHz. The maskers are normalized to have the same RMS value.

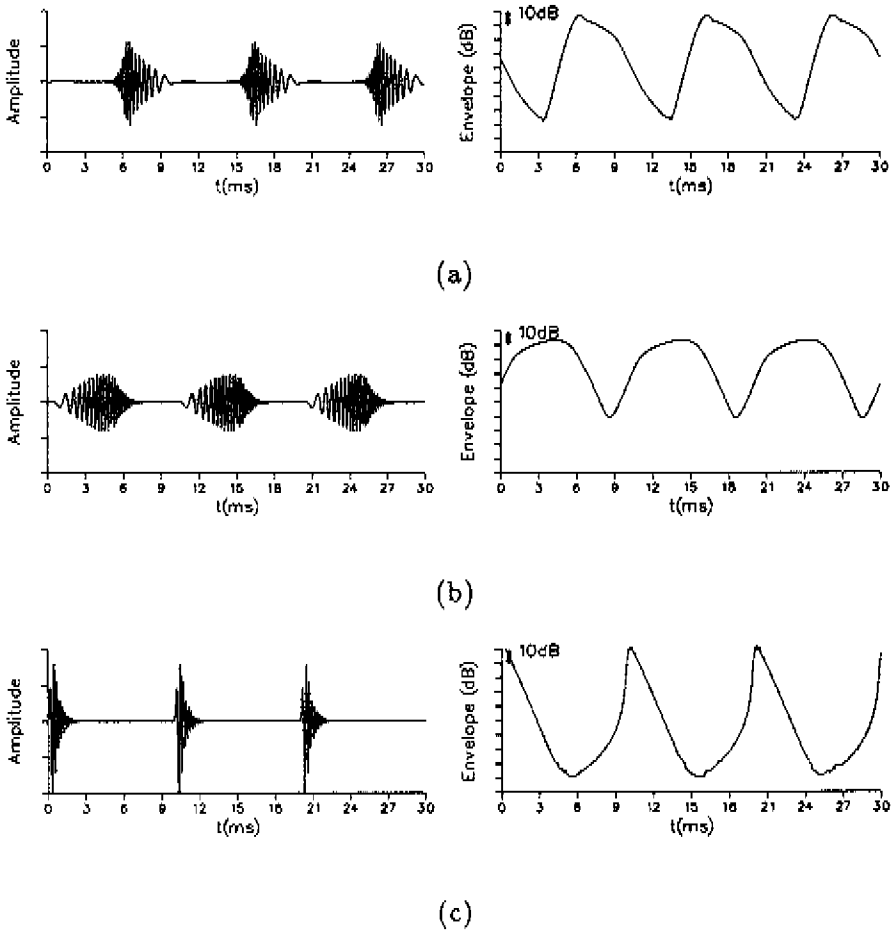


Figure 6.16: Left panels: Waveforms of basilar-membrane filter responses to (a) the m_+ masker, (b) the m_- masker, (c) the zero-phase masker. Right panels: Corresponding envelopes, plotted on a decibel scale. The resonance frequency of the filter is 4 kHz. The maskers are normalized to have the same RMS value.

according to Strube (1985). The parameter V_0/b_0 in the model (where V_0 is the friction coefficient per unit area associated with the basilar membrane motion and b_0 is the width of the basilar membrane at the stapes), however, is chosen to be 32000 mg/mm^3s, while Strube used a value of 16000 mg/mm^3s for most of his simulations. The BM model with a larger V_0/b_0 seems to provide a better match of the phase response to the experimental data (Viergever, 1980). The response waveforms at resonance frequency 4 kHz and their envelopes (plotted on a decibel scale) for the three maskers are shown in Fig. 6.16. From Fig.6.16, we can see that the valleys in the envelope for the zero-phase masker are the deepest and the widest whereas the valleys in the envelope for the m_- masker are the shallowest. The valleys in the envelope of the m_- masker are shallower and narrower than those for the m_+ masker. This is in line with the experimental finding at 4 kHz that the lowest target threshold is obtained with the zero-phase masker and the highest for the m_- masker, and that the threshold for the m_+ masker is lower than that for the m_- masker.

The results of this experiment are in agreement with masking experiments where the masker or the maskee contained some special frequency-dependent group delays. Firstly, to investigate the temporal resolution of the auditory system, Patterson and Green measured the simultaneous masking of tones by maskers generated as Huffman sequences (Patterson and Green, 1971). The Huffman sequences were actually the impulse responses of allpass filters and differed from each other by their group delays which were strongly frequency dependent. Two sets of group delays had the opposite sign and thus the waveforms of masker plus target were just time-reversed versions of each other. From the masking of pure tones we know that it is easier to mask a tone by a second tone of lower frequency than by one of higher frequency. It is therefore reasonable to consider that in the experiments of Patterson and Green the threshold differences for the two different maskers are predominantly determined by the difference of the frequency-dependent group delays of the maskers below the target frequency. Thus, the target thresholds are higher in the masker with a group delays increasing with an increase of frequency than with a group delays decreasing with an increase of frequency (Patterson and Green, 1971). Our results with two Schroeder-phase maskers and noise targets are in agreement with this result. In our case, the m_- masker has a group delay increasing with an increase frequency.

Secondly, a direct comparison can be made with the masking of pure tones by the two Schroeder-phase maskers (Smith et al.,

1986; Kohlrausch, 1988). The large threshold differences between the two Schroeder-phase maskers at high masker levels are, in principle, in agreement with the experimental finding by Smith et al. (1986) and Kohlrausch (1988). It is reasonable to suggest from these experiments that the group delays of the cochlear filters contribute to disperse or to concentrate the spectro-temporal distribution of the masker.

Although our results for the two Schroeder-phase maskers qualitatively agree with those from Kohlrausch's experiment, it is still difficult to compare the results between these two experiments quantitatively due to the difference in bandwidth and group delay function of the maskers. One notices in Fig. 6.12, on the one hand, that the threshold of the noise target at 1 kHz for the zero-phase masker with a fundamental frequency of 100 Hz is similar to that for the m_+ masker. In Kohlrausch's experiment, on the other hand, threshold of a pure tone at 1 kHz for the 100-Hz zero-phase masker was significantly higher than for the m_+ masker (Kohlrausch, 1988). This difference is caused mainly by the difference between group delay functions of the maskers used in the two experiments.

Alternating-phase maskers

The masking of the critical-band-noise targets by the masker with alternating-phase relationship for all harmonic components differs obviously at high frequencies from the masking patterns of the 100-Hz zero-phase masker (see Fig. 6.13). This is due to the fact that the alternating-phase manipulation introduces a secondary peak in the middle of a period. Therefore the 100-Hz masker with an alternating phase has a quasi-period of 5 ms and the responses of the cochlear filters at high frequencies are very similar to the responses for the 200-Hz zero-phase masker. Its masking pattern in the high-frequency region is therefore close to that obtained from the 200-Hz masker. This is clearly illustrated in Fig.6.17, where the responses of the basilar-membrane filter at 4 kHz to these three maskers are plotted. The valleys in the log-envelope plot for the 100-Hz zero-phase masker are much deeper than those for the 100-Hz masker with an alternating-phase relationship. On the other hand, the log-envelope plots for the 200-Hz zero-phase masker and the alternating-phase masker are similar. In the low-frequency region, the frequency resolution plays a dominant role in determining the masking thresholds. Therefore, the phase choice does not influence the masked threshold.

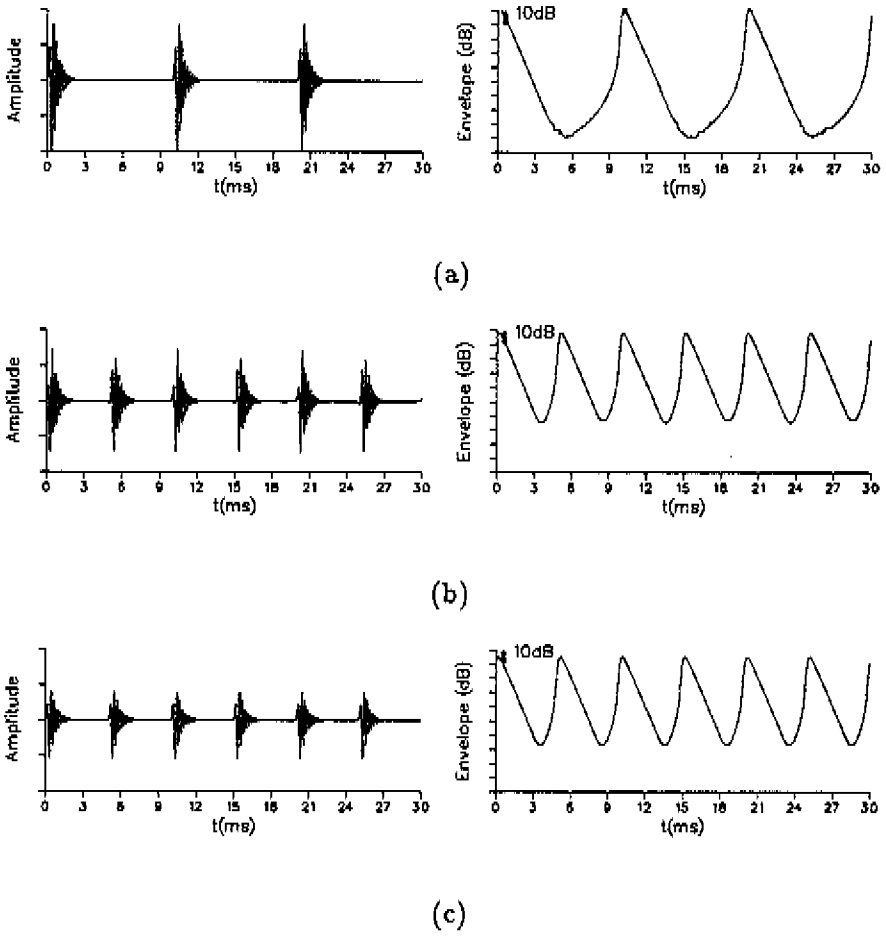


Figure 6.17: Left panels: Waveforms of basilar-membrane filter responses to (a) the 100-Hz zero-phase masker, (b) the 100-Hz masker with all harmonics in an alternating phase, (c) the 200-Hz zero-phase masker. Right panels: Corresponding envelopes, plotted on a decibel scale. The resonance frequency of the filter is 4 kHz. The maskers are normalized to have the same RMS value.

The effects of alternating-phase manipulation on the target thresholds are strongly dependent on the fundamental frequencies of the maskers. This is clearly demonstrated by the third experiment where the low-frequency components of the maskers are set to alternating phase and high-frequency components are set to zero phase. In this experiment, the detection of the noise band is predominantly determined by the critical-band filter at 5 kHz which has a bandwidth of approximately 900 Hz. Consequently, when only the low frequency harmonics of the maskers are set to alternating phase relationship, the threshold of the noise band remains constant with the lowest threshold for the 100-Hz masker and the highest for the 400-Hz masker. As the transition frequency increases to a certain region, the noise threshold for the 100-Hz and the 200-Hz masker increases. This is because the alternating phase of the harmonics results in a secondary peak between two successive pulses of the masker waveform and the height of the secondary peak increases when more harmonics are in alternating phase. Since the waveform for the 100-Hz masker has the strongest modulation, the effect of the secondary peak on the masked threshold is the strongest for this fundamental frequency. For the 400-Hz masker, the critical-band filter passes through only two or three harmonics and the valleys in the envelope of the filter response are hardly affected by the phase change. The threshold of the noise targets therefore remains rather constant for the 400-Hz masker.

6.8 Experiment 6

Masking of narrowband noise by synthetic vowel sounds

As an application to speech perception, it is desirable to show how masking patterns of speech sounds such as vowels can be understood in terms of evidence presented previously in this chapter. For instance, in Experiment 3, we have seen that the threshold decreases at target frequencies above 1 kHz when the masker level increases. Since spectra of the vowel sounds consist of several formant regions represented by peaks and valleys, the spectral level of a vowel masker is a function of frequency. It is therefore investigated in this experiment whether these differences in spectral levels of vowel maskers influence the thresholds of noise targets.

Vowel sounds were synthesized by using spectrally-flat zero-phase harmonic complexes with fundamental frequencies of 100 and 200 Hz as inputs to a linear-predictive-coding (LPC) filter. The formant fre-

quencies of the vowel were 680, 1110, 2347, 3202 and 4500 Hz. The targets were noise bands of critical-band width as used in Experiment 1 (see page 83). The targets passed through the same LPC filter as used for the vowel sounds and they therefore had locally the same spectral envelope as the vowel sound. The threshold of the target was then defined as the spectral level difference between the masker and the target, expressed in decibels. All vowel maskers were presented at a sound pressure level of 80 dB.

6.8.1 Results

The average thresholds for three subjects are plotted in Figs. 6.18 (a,b) for vowel fundamentals of 100 and 200 Hz respectively. The spectral envelope of the vowel masker is plotted in Fig. 6.18(c). One observes that there is no global decrease of threshold towards high frequencies as we have observed in Experiment 1 and 2 (see page 83). At low target frequencies, one sees from Fig. 6.18 (b) that the threshold function shows dips (at 300 and 500 Hz) and peaks (at 400 and 600 Hz) for the 200-Hz vowel masker, since masker harmonics are spectrally well resolved. By comparing Figs. 6.18 (a) and (c), it can be seen that threshold peaks correspond to spectral valleys and threshold valleys to spectral peaks of the vowel masker, which is a very counter-intuitive result.

6.8.2 Discussion

Vowel sounds have complex spectra and their masking patterns can vary in many ways. In principle, the (relative) threshold is rather constant as a function of frequency and globally less frequency-dependent than in a flat-spectrum zero-phase masker (Experiment 1, see page 83) This is because the spectrum of a vowel sound has in general a slope of -6dB/oct. As we have seen in Experiment 3, the threshold of the noise target increases as a result of the spectral slope of the masker. In Fig. 6.18(a) the valleys in the masking pattern are found to be located approximately at the formant frequencies, i.e. at the spectral peaks of the masker, and the peaks of the masking pattern are located at the spectral valleys of the masker. In particular, the threshold dips around frequencies 1110 and 2350 Hz correspond to the second and the third formant frequencies of the vowel masker. The threshold peak at about 1900 Hz corresponds to the spectral valley at 1900 Hz. These low thresholds at the formant frequencies are quite similar to the result of Experiment 3 where the

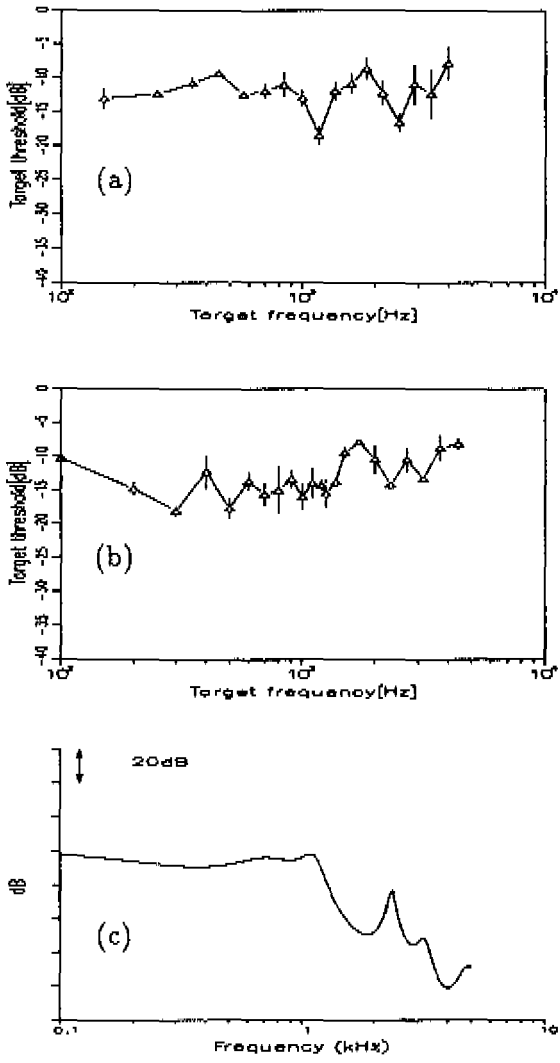


Figure 6.18: Thresholds of critical-band noise targets for vowel maskers as a function of center frequency of the noise band. (a) 100-Hz vowel masker. (b) 200-Hz vowel masker. (c) Spectral envelope of the vowel masker.

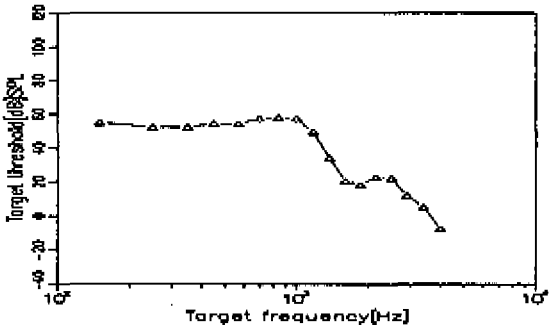


Figure 6.19: Thresholds of critical-band noise targets in the 100-Hz vowel masker, expressed in dB SPL.

thresholds of noise targets in the maskers with high levels have a lower spectral density ratio.

The finding that the masking pattern of a vowel sound is a blurred version of its physical spectrum has been attributed to the limited frequency resolution of the auditory system (Moore and Glasberg, 1983a; Tyler and Lindblom, 1982; Houtgast, 1974). The difference in decibels between peaks and valleys in the masking pattern is less than what is found in the physical spectrum of the masker. Our result suggests that among other reasons, the level effects due to the spectral peaks and valleys in the vowel masker can also attribute to the reduction of peak-valley difference in the masking pattern. The thresholds expressed in spectral density ratios in this experiment can be easily transformed to absolute target level by incorporating masker power density and target bandwidth. The so transformed threshold values are plotted in Fig.6.19. They show indeed that the masked threshold curve is a blurred version of the physical spectrum of the masker.

The masking pattern of the 200-Hz vowel masker of Fig. 6.18 (b) clearly reflects the spectral composition of the masker at low frequencies, and the formant regions are not well delineated. This is a consequence of the spectral resolution of the auditory system associated with a vowel with a high fundamental frequency. Only at high frequencies, threshold peaks occur at spectral valleys of the envelope of the vowel masker, at about 1900 and 3000 Hz.

In summary, our results suggest that the blurring of the vowel spectrum, which is observed in spectral masking patterns, should not be

attributed to the spectral resolution of the auditory system (Moore and Glasberg, 1983a). It rather seems that temporal resolution in combination with nonlinear level effects can quite well account for the observed reduction in peak-valley differences.

6.9 General discussion

In general, the auditory system is a nonlinear system consisting of a bank of bandpass filters with varying frequency resolution. Frequency resolution decreases and, accordingly, temporal resolution increases with an increase of the center frequency of the filter. Therefore, the detection of targets at low frequencies is mainly determined in the frequency domain by spectral analysis of the masker. At high frequencies, on the other hand, the detection of targets is predominantly determined by temporal analysis of the masker. The relative contributions of spectral and temporal analysis strongly depends on the fundamental frequency of the masker. The auditory system can therefore easily detect targets at high frequencies where envelopes of auditory-filter responses show deep valleys (Goldstein, 1967; Duifhuis, 1970; Duifhuis, 1973; Patterson, 1987). This has been shown by studies of masking period patterns in the literature (Duifhuis, 1970; Kohlrausch, 1988). Similarly, targets at low frequencies can be easily detected from maskers with high fundamental frequencies.

Moreover, the temporal resolution changes nonlinearly with masker level. Better resolution is associated with a higher masker level. Therefore, in deeply modulated maskers with high levels the targets appear to be detected easier (if the threshold is expressed as a spectral density ratio of target and masker) than they are at low masker levels. For maskers with a spectral slope of -6 dB/oct, in addition to the influence of waveform dispersion, the threshold of targets at high frequencies increases significantly due to the low spectral level associated with a strong spectral slope.

The auditory system is sensitive to the phase relationships among spectral components within a critical band. This phase sensitivity is mainly determined by the temporal analysis of the stimulus and therefore, is strongly dependent on the fundamental frequency and the spectral slope of the stimuli. This is a reason why only with low-fundamental stimuli, the flat-spectrum vowel (Schroeder and Strube, 1986) and the phase-vowel (Traunmüller, 1987) stimuli can show vowel-like qualities. Typically, phase spectra of these phase-vowels were similar

to natural vowels and the envelopes of the amplitude spectra were flat (Traunmüller, 1987). From our phase DLs measurements, we can expect that phase-vowels with fundamentals of 200 and 400 Hz will not have vowel qualities.

Although experiments in this chapter have been performed with rather artificial sounds, the results have also relevance for speech technology. For instance, the audibility of quantization noise is one of the main concerns in speech coding. In the perceptual weighting technique, which is mainly based on the masking of pure tones, the spectrum of the noise is shaped such that it is similar to the spectral envelope of the coded speech signal (Schroeder et al., 1979). If the spectral level of the noise is significantly lower than the spectral level of the speech sound, the quantization noise can be made inaudible. Our experiments showed, however, that the masking of noise targets by broadband harmonic complex sounds is mainly determined by local details of the masker in the frequency domain or in the time domain, and not primarily by global features of the maskers' spectra. Our results suggest that the weighting in the low-frequency region and for high-pitched sound should be associated with the harmonic structure of the speech signal. At high frequencies, the perceptual weighting should be associated with the temporal waveform of the speech signal. Although the masking threshold of the noise target at high frequencies is generally higher in vowel sounds than in pulse trains, the threshold can still be lower in the spectral peak region than in the valley region. In view of the continuously changing time-frequency structure of the sounds, a dynamic adaptation of perceptual weights could improve the quality of the low bit-rate coded speech, especially for coding of transient sounds.

Phase manipulation of speech signals will certainly influence their subjective sound quality, especially for transient sounds. Although the long-term spectrum of speech is usually preserved after phase manipulations, the speech waveform within each pitch period of vowel sounds is totally changed (Strube, 1982; Moriya and Honda, 1986; Quatieri et al., 1990). The auditory system is sensitive to this kind of waveform changes, especially at high frequencies, where good temporal resolution is retained. For example, a phase dispersion system has been designed to reduce the peak/RMS ratio of speech signal in broadcasting (Quatieri et al., 1990), where actual phase spectra of vowel sounds were replaced by phase spectra of upward frequency sweeps while amplitude spectra were left unchanged. Under broadcasting conditions, only a small loss of voice quality was reported (Quatieri et al., 1990). All these phase

manipulations are quite acceptable for voiced sounds. Although the masking experiments in this paper have suggested that the change of waveform is most detectable at high frequencies, this does not necessarily suggest the same for judgement of quality difference. Moreover, due to the global spectral tilt of vowel sounds, the phase changes at high frequencies become less noticeable. Our measurements of phase DLs for spectrally-tilted complex tones do show that the phase shift at high frequencies is quite detectable. We can expect that for real high-quality speech sound the phase of high harmonics will play an important role.

References

Allen, J.B. (**1978**). "Cochlear Model-1978," in *Models of the Auditory System and Related Signal Processing Techniques*, edited by M. Hoke and E. De Boer, Scand. Audiology suppl. **9**, 1-16.

Allen, J.B. (**1983**). "Magnitude and Phase-Frequency Response to Single tones in the Auditory Nerve," J. Acoust. Soc. Am. **73**, 2071-2093.

Atal, B.S. (**1988**). "Speech Coding and Human Speech Perception", In *Working Models of Human Perception* Edited by B.A.G.Elsendoorn and H. Bouma, (Academic Press, London), 101-126.

Békésy, G. von, (**1960**). *Experiment in Hearing*. (McGraw-Hill, New York).

Bos, C.E., and De Boer, E. (**1966**). "Masking and Discrimination," J. Acoust. Soc. Am. **39**, 708-715.

Charpentier, F., and Moulines, E. (**1989**). "Pitch-synchronous Waveform Processing Techniques for Text-to-speech Synthesis Using Diphones," Proceedings EUROSPEECH-89, **2**, 13–19.

De Boer, E. (**1980**). "A Cylindrical Cochlea Model: the Bridge Between Two and Three Dimensions", Hearing Res. **3**, 109-131.

Duifhuis, H. (**1970**). "Audibility of High Harmonics in a Periodic Pulse", J. Acoust. Soc. Am. **48**, 888-893.

Duifhuis, H. (**1971**). "Audibility of High Harmonics in a Periodic Pulse. II. Time Effect", J. Acoust. Soc. Am. **49**, 1155-1162.

Duifhuis, H. (**1973**). "Consequences of Peripheral frequency Selectivity for Nonsimultaneous Masking," J. Acoust. Soc. Am. **54**, 1471-1488.

Egan, J.P., and Hake, H.W. (**1950**). "On the Masking Pattern of a Simple Auditory Stimulus," J. Acoust. Soc. Am. **22**, 622-630.

Ehmer, R.H. (**1959**). "Masking Pattern of Tones," J. Acoust. Soc. Am. **31**, 1115-1120.

Fastl, H. (**1976**). "Temporal Masking Effects: I. Broad Band Noise Masker," Acustica **35**, 287-302.

Fastl, H. (**1976/1977**). "Temporal Masking Effects: II. Critical Band Noise Masker," Acustica **36**, 317-331.

Fastl, H. (**1979**). "Temporal Masking Effects: III. Pure Tone Masker," Acustica **43**, 282-294.

Fletcher, H. (**1929**). *Speech and Hearing*, (Van Nostrand, New York).

Fletcher, H. (**1940**). "Auditory Patterns," Rev. Mod. Phys. **12**, 47-56.

Goldstein, J.L. (**1967**). "Auditory Spectral Filtering and Monaural Phase Perception," J. Acoust. Soc. Am. **41**, 458-479.

Greenwood, D.D. (**1961**). "Critical Bandwidth and the Frequency Coordinates of the Basilar Membrane," J. Acoust. Soc. Am. **33**, 1344-1356.

Holmes, J.N. (**1973**). "The Influence of Glottal Waveform on the Naturalness of Speech From a Parallel Formant Synthesizer," IEEE Trans. on Audio and Electroacoustics, AU-**21**, No.3.

Houtgast, T. (**1974**). "Auditory Analysis of Vowel-like Sound," Acustica **31**, 320-324.

Houtsma, A.J.M., and Smurzynski, J. (**1990**). "Pitch Identification and Discrimination for Complex Tones with Many Harmonics," J. Acoust. Soc. Am. **81**, 305-310.

Jesteadt, W., Bacon, S.P., and Lehman, J.R. (**1982**). "Forward Masking as a Function of Frequency, Masker level, and Signal Delay," J. Acoust. Soc. Am. **71**, 951-963.

Johnston, J.D. (**1988**). "Transform Coding of Audio Signals Using Perceptual Noise Criteria," IEEE Journal on selected areas in communication, **6**, 314-323.

Kohlrausch, A. (**1988**). "Masking Patterns of Harmonic Complex Tone Maskers and the Role of the Inner Ear Transfer Function,", in *Basic Issues in Hearing*, edited by H. Duifhuis, J.W. Horst and H.P. Wit, (Academic Press, London), 339-350.

Kohlrausch, A. (**1992**), "Phase Effects in Masking related to Dispersion in the Inner Ear III. Masking Period Patterns of Short Targets," J. Acoust. Soc. Am. , (in preparation).

Krasner, M.A. (**1980**). "The Critical Band Coder–Digital Encoding of Speech Signals Based on the Perceptual Requirements of the Auditory System," Proc. Int. Conf. ICASSP, 327-331.

Levitt, H. (**1971**). "Transformed Up-down Method in Psychoacoustics," J. Acoust. Soc. Am. **49**, 467-477.

Markel, J.D., and Gray, A.H. (**1976**). *Linear Prediction*, (Springer-Verlag, Berlin).

Martens, J.P. (**1982**). "A New Threory for Multitone Masking," J. Acoust. Soc. Am. **72**, 397-405.

Moore, B.C.J., and Glasberg, B.R. (**1983a**). "Growth of Forward Masking for Sinusoidal and Noise Maskers as a Function of Signal Delay; Implications for Suppression in Noise," J. Acoust. Soc. Am. **73**, 1249-1259.

Moore, B.C.J., and Glasberg, B.R. (**1983b**) "Forward Masking Patterns for Harmonic Complex Tones," J. Acoust. Soc. Am. **73**, 1682-1685.

Moore, B.C.J., and Glasberg, B.R. (**1983c**). "Masking Patterns for Synthetic Vowels in Simultaneous and Forward Masking," J. Acoust. Soc. Am. **73**, 906-917.

Moore, B.C.J., and Glasberg, B.R. (**1989**). "Difference Limens for Phase in Normal and Hearing-impaired Subjects," J. Acoust. Soc. Am. **86**, 1351-1365.

Moriya T., and Honda, M. (**1986**). "Speech Coder Phase Equalization and Vector Quantization", Proc. Int. Conf. ICASSP, 1701-1704.

Patterson, R.D., and Moore, B.C.J. (**1986**). "Auditory Filters and Excitation Patterns as Representations of Frequency Resolution" in *Frequency Selectivity in Hearing*, edited by B. C.J. Moore, (Academic Press, London), 123-174.

Patterson, R.D. (**1987**). "A Pulse Ribbon Model of Monaural Phase Perception," J. Acoust. Soc. Am. **82**, 1560-1586.

Patterson, J. H., and Green, D.M. (**1971**). "Masking of Transient Signals Having Identical Energy Spectra," Audiology **10**, 85-96.

Pickles, J.O. (**1986**). "The Neurophysiological basis of frequency selectivity," in *Frequency Selectivity in Hearing*, edited by B.C.J. Moore, (Academic Press, London), 51-112.

Plomp, R., and Steeneken, H.J.M. (**1969**). "Effect of Phase on the Timbre of Complex Tones," J. Acoust. Soc. Am. **46**, 409-421.

Quatieri, T.F., Lynch, J.T., Malpass, M.L., McAulay, R.J., and Weinstein, C.J. (**1990**). "The VISTA Speech Enhancement System for AM Radio Broadcasting," *Final Technical Report*, Lincoln Lab., MIT, 29.

Rosenberg, A.E. (**1971**). "Effect of Glottal Pulse Shape on the Quality of Natural Vowels," J. Acoust. Soc. Am. **49**, 583-590.

Schafer, T.H., Gales, R.S., Shewmaker, C. A., Thompson, P.O. (**1950**). "The Frequency Selectivity of the Ear as Determined by Masking Experiments," J. Acoust. Soc. Am. **22**, 490-496.

Schöne, P.
von.(**1977**). "Nichtlinearitäten im Mithörschwellen-Tonheitsmuster von Sinustönen," Acustica **37**, 37-44.

Schöne, P. von.(**1979**). "Mithörschwellen-Tonheitsmuster Maskierender Sinustöne," Acustica **43**, 197-204.

Schroeder, M.R. (**1959**) "New Results Concerning Monaural Phase Sensitivity," J. Acoust. Soc. Am. **31**, 1579A.

Schroeder, M.R. (**1970**). "Synthesis of Low-Peak-Factor Signals and Binary Sequences with Low Autocorrelation," IEEE Trans Inf. Theory **16**, 85-89.

Schroeder, M. R., (**1972**) "An Integrable Model for the Basilar Membrane," J. Acoust. Soc. Am. **53**. 429-434.

Schroeder, M. R. (**1975**). "Models of Hearing," Proc. IEEE, **63**, 1333-1351.

Schroeder, M.R., and Mehrgardt, S. (**1982**). "Auditory Masking Phenomena in the Perception of Speech," in *The Representation of Speech in the Peripheral Auditory System*, edited by R. Carlson and B. Granström, (Elsevier, Amsterdam), 79-87.

Schroeder, M.R., and Strube, H.W. (**1986**). "Flat-Spectrum Speech," J. Acoust. Soc. Am. **79**, 1580-1583.

Schroeder, M.R., Atal, B.S., and Hall, J.L. (**1979**) "Optimizing Digital Speech Coder by Exploiting Masking Properties of the Human Ear," J. Acoust. Soc. Am. **66**, 1647-1652.

Small, A.M. (**1959**). "Pure-tone Masking," J. Acoust. Soc. Am. **31**, 1619-1625.

Smith, B.K., Sieben, U.K., Kohlrausch, A., and Schroeder, M.R. (**1986**). "Phase Effects in Masking Related to Dispersion in the Inner Ear," J. Acoust. Soc. Am. **80**, 1631-1637.

Strube, H.W. (**1982**). "How to Make an All-pass Filter with a Desired Impulse Response," IEEE Trans. Acoust., Speech, Signal Processing, ASSP **30**, 336-337.

Strube, H.W. (**1985**). "A Computationally Efficient Basilar-Membrane Model," Acustica **58**, 207-214.

Terhardt, E. (**1974**). "On the Perception of Periodic Sound Fluctuations (roughness)," Acustica **30**, 201-213.

Traunmüller, H. "Phase Vowel," in *The Psychophysics of Speech Perception*, edited by M.E.H. Schouten, (Martinus Nijhoff, Dordrecht), 377-384.

Tyler, R.S., and Lindblom, B. (**1982**). "Preliminary Study of Simultaneous-Masking and Pulsation-Threshold Patterns of Vowels," J. Acoust. Soc. Am. **71**, 220-224.

Veldhuis, R.N.J., Breeuwer, M., and Waal, R.V.D. (**1989**). "Subband Coding of Digital Audio Signals Without Loss of Quality," Proc. Int. Conf. ICASSP, 2009-2012.

Viergever, M. (**1980**). *Mechanics of the Inner Ear*, (Delf University Press, Delft).

Wegel, R.L., and Lane, C.E. (**1924**). "The Auditory Masking of One Pure Tone by Another and its Probable Relation to the Dynamics of the Inner Ear," Phys. Rev. **23**, 266-285.

Yates, G.K. (**1986**). "Frequency Selectivity in the Auditory Periphery," in *Frequency Selectivity in Hearing*, edited by B. C.J. Moore, (Academic Press, London), 1-50.

Zwicker, E. (**1970**). "Masking and Psychological Excitation as Consequences of the Ear's Frequency Analysis," in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G.F. Smoorenburg, (A.W. Sijthoff, Leiden), 376-396.

Zwicker, E. (**1976a**). "Masking Period Patterns of Harmonic Complex Tones," J. Acoust. Soc. Am. **2**, 429-439.

Zwicker, E. (**1976b**). "A Model for Predicting Masking-Period Patterns," Biol. Cybernetics **23**, 49-60.

Zwicker, E., and Fastl, H. (**1990**). *Psychoacoustics: Facts and Models*, (Springer-Verlag, Berlin).

Zwicker, E., and Terhardt, E. (**1980**), "Analytical Expressions for Critical-band Rate and Critical Bandwidth as a Function of Frequency," J. Acoust. Soc. Am. **68**, 1523-1525.

Zwicker, E., Flottorp, G., and Stevens, S.S. (**1957**). "Critical Band Width in Loudness Summation", J. Acoust. Soc. Am. **29**, 548-557.

Summary

SPEECH signal processing is a very important way of achieving suitable signal representations for speech communication. Psychophysical studies of speech representation, on the other hand, provides a research direction to deal with the perceptually important aspects of the speech signal in the signal processing. The main theme of this dissertation was devoted to the study of some psychophysical and signal-processing aspects of speech representation.

One speech analysis technique proposed in Chapter 2 is a robust linear-predictive coding (LPC) analysis using a short- time-energy (STE) weighting function. We derived a generalized STE-based LPC analysis under the linear-least-square criterion. The sample selection window or the weighting function in this algorithm were based on the short-time energy of the speech signal. Their effect was to over-weight the speech samples that fit the LPC model well and to down-weight the other samples. This novel LPC approach produces less deviating estimates of the formant frequencies than those obtained from the conventional LPC and is less sensitive to the values of the fundamental frequency. From the experimental observations, the STE-thresholded LPC solution is preferable to the sample-selective method based on two-step LPC analyses in terms of computation efficiency and robustness in the selection of speech samples and preferable to the STE-weighted LPC from the viewpoint of estimation accuracy.

The weighted LPC analysis was further developed in Chapter 3, where the relation between the covariance linear prediction (CLP) analysis of a frame of a speech signal and the CLP analysis of its subframes was established. The results of CLP analysis derived from a set of subframes were equivalent to those of a residual-weighted CLP analysis of the complete frame and the solutions of the residual-weighted CLP are the same as those of the generalized weighted average of subframe CLP. Those subframes which best reflect the filter model of speech production

can therefore be chosen to improve the accuracy of the estimate of the LPC parameters.

Another signal processing technique, proposed in Chapter 4, was to use a singular-value decomposition (SVD) technique to detect instants of glottal closure. The exact detection of such instants in a speech signal is a very important step in speech-coding and speech-manipulation applications. This new technique used the Frobenius norm of the linear-predictive matrix for the detection of instants of glottal closure. The sequential computation of the Frobenius norm of the matrix is reduced to just the addition of the sum of the squared entries of the last row of the matrix and the subtraction of the sum of the squared entries of the first row of the preceeding matrix. Therefore, the new method is computationally very attractive. As an additional benefit, the new method is less sensitive to noise.

Motivated by using the phase of the Fourier transform of speech to extract formant frequencies, we studied in Chapter 5 the conditions under which a signal is uniquely determined by its Fourier transform phase. If a speech signal cannot be uniquely determined by its phase (to a factor), then it is not possible to directly extract formant frequencies from its phase spectrum. We showed that uniqueness corresponds to the non-singularity of a matrix which can be formed from the finite length real sequence.

The study of auditory masking in Chapter 6 is intended to shed light on how the auditory system processes complex sounds. The auditory system can be considered as a nonlinear system consisting of a bank of bandpass filters with varying frequency resolution. Frequency resolution decreases and, accordingly, temporal resolution increases with an increase of center frequency of the filter. Therefore the detection of targets at low frequencies is mainly determined by spectral properties of maskers. At high frequencies, on the other hand, the detection of targets is predominantly determined by temporal behaviour of maskers. The relative contributions of spectral and temporal analysis strongly depends on the fundamental frequency of the masker.

The temporal resolution changes nonlinearly with masker level. Better resolution is associated with a higher masker level. Therefore, in deeply modulated maskers with high levels the targets appear to be detected easier (if the target threshold is expressed relative to the masker level) than they are at low masker levels. For maskers with a spectral slope of -6 dB/oct, in addition to the influence of waveform dispersion, the threshold of targets at high frequencies increases significantly due to

the low spectral level associated with a strong spectral slope.

From the masking experiments, insight is also gained into the discrimination thresholds for those speech sounds that result from from speech processing techniques. For instance, quantization noise in coded speech signal can be masked, if the spectral envelope of the noise is shaped properly by a perceptual weighting technique. The perceptual weighting technique so far is mainly based on the masking behaviour of pure tones. Masking behaviour of complex sounds, however, cannot be easily predicted from masking behaviour of pure tones. Our experiments showed that the masking of noise targets by harmonic complex sounds is mainly determined by local details of the masker in the frequency domain or in the time domain, and not primarily by global features of the maskers' spectra. We therefore suggest that the weighting in the low-frequency region and for high-pitched sound should be associated with the harmonic structure of the speech signal. At high frequencies, the perceptual weighting should be associated with the temporal waveform of the speech signal. In view of the continuously changing time-frequency structure of the sounds, a dynamic adaptation of perceptual weights should be used to improve the quality of low bit-rate coded speech, especially for coding of transient sounds.

Our measurements of phase difference limens of individual components in spectrally-tilted complex tones showed that the phase shift is easily detectable at high frequencies. Phase manipulations of speech signals will influence their subjective sound quality especially for transient sounds, because the auditory system is sensitive to this kind of waveform changes, especially at high frequencies. For real high-quality speech we have to take care of the phase relationships of high harmonics.

Samenvatting

HET toepassen van signaalbewerking op spraakgeluiden is een belangrijke manier om goede signaalrepresentaties te verkrijgen voor spraakcommunicatie. Psychofysisch onderzoek naar representaties van spraaksignalen laat zien welke aspecten van het bewerkte spraaksignaal perceptief van belang zijn. Het hoofdthema van dit proefschrift is onderzoek naar psychofysische en signaalbewerkingsaspecten van spraakrepresentatie.

Een van de spraakanalysetechnieken die wordt voorgesteld in Hoofdstuk 2, is een robuuste lineair-predictieve coderingsanalyse (LPC) waarbij een korte-termijn-energieweegfunctie (STE) wordt gebruikt. Een gegeneraliseerde STE-gebaseerde LPC-analyse met een lineair-kleinste-kwadraatmaatstaf is ontwikkeld. Het venster voor het selecteren van monsters of de weegfunctie voor dit algoritme is gebaseerd op de korte-termijn-energie van het spraaksignaal. Het effect hiervan is dat die spraaksignaalmonsters die goed voldoen aan het LPC-model zwaarder worden gewogen dan monsters die niet zo goed in dit model passen. Deze nieuwe LPC-benadering resulteert in kleinere afwijkingen in schattingen van formantfrequenties dan bij conventionele LPC en is ook minder gevoelig voor de waarde van de grondfrequentie. Proefondervindelijk is ook aangetoond dat de STE-gedrempelde LPC-oplossing de voorkeur verdient boven de selectieve-monster-methode wat betreft robuustheid ten aanzien van geselecteerde monsters en wat betreft doelmatigheid van berekeningen, en ook te verkiezen is boven de STE-gewogen LPC-methode met betrekking tot de nauwkeurigheid van geschatte waarden.

De gewogen LPC-analyse-methode is verder ontwikkeld in Hoofdstuk 3, waarin het verband tussen de covariantie-lineaire-predictieanalyse (CLP) van een fragment spraaksignaal en de CLP-analyses van de sub-fragmenten wordt aangetoond. Het resultaat van een CLP-analyse gebaseerd op een aantal sub-fragmenten, was gelijkwaardig aan het resultaat van een residu-gewogen CLP-analyse van het gehele fragment

en de oplossingen vanuit de residu-gewogen CLP bleken hetzelfde als de oplossingen verkregen uit het gegeneraliseerd-gewogen gemiddelde sub-fragment CLP. Die sub-fragmenten die het best passen bij het filtermodel van spraakproduktie, kunnen daarom worden geselecteerd om zo de schattingsnauwkeurigheid voor LPC-grootheden te verhogen.

Een andere signaalbewerkingstechniek, besproken in Hoofdstuk 4, bestaat uit het gebruik van singuliere-waarde-decompositie (SVD) om de momenten van het sluiten van de stembanden te bepalen. Het precies bepalen van deze momenten is een erg belangrijke stap bij het coderen of het manipuleren van spraakgeluiden. Deze nieuwe techniek gebruikt de Frobenius-norm van de lineaire voorspellingsmatrix voor het detecteren van de momenten van het sluiten van de stembanden. De sequentiële berekening van de Frobenius-norm van een matrix is teruggebracht tot het optellen van de som van de gekwadrateerde elementen in de laatste rij van de matrix, en tot het aftrekken van de som van de gekwadrateerde elementen in de eerste rij van de voorafgaande matrix. Deze nieuwe methode is daarom rekentechnisch erg aantrekkelijk. Bovendien blijkt de nieuwe methode minder gevoelig voor ruis.

Geïnspireerd door het gebruik van de fase bij de Fourier transform van spraak voor het bepalen van de formantfrequenties, is in Hoofdstuk 5 onderzocht onder welke voorwaarden een signaal op unieke wijze is bepaald door zijn Fourier-transformfase. Als een spraaksignaal niet op unieke wijze bepaald is door zijn fasefunctie, op een vermenigvuldigingsfactor na, dan is het ook niet mogelijk om op directe wijze de formantfrequenties uit het fasespectrum af te leiden. Er is aangetoond dat uniekheid overeenkomt met de niet-singulariteit van een matrix die gevormd kan worden uit de eindige reeks van de signaalmonsters.

Het onderzoek naar auditieve maskering, besproken in Hoofdstuk 5, is bedoeld om inzicht te krijgen in hoe ons gehoorsysteem samengestelde geluiden verwerkt. Het gehoorsysteem kan beschouwd worden als een niet-lineair systeem van banddoorlaatfilters met een verlopend oplossingsvermogen. Frequentie-oplossend vermogen neemt af en, daarmee samenhangend, tijdoplossend vermogen neemt toe naarmate de afstemfrequentie van het filter toeneemt. Daarom is het detecteren van een doelsignaal met lage frequentie grotendeels bepaald door spectrale eigenschappen van de maskeerder. Bij hoge doelsignaalfrequenties, daarentegen, wordt het detectieproces grotendeels bepaald door het temporele gedrag van de maskeerder. De relatieve bijdragen van spectrale en temporele analyse in het oor hangen sterk af van de grondfrequentie van de maskeerder.

Temporeel oplossend vermogen verandert op niet-lineaire wijze met het maskeerniveau. Een beter oplossend vermogen gaat samen met hogere maskeerniveaus. Bij sterk gemoduleerde maskeergeluiden op hoge geluidsniveaus, lijken doelsignalen gemakkelijker detecteerbaar dan bij maskeergeluiden op lage geluidsniveaus, zolang de detectiedrempel relatief wordt uitgedrukt met betrekking tot het geluidsniveau van de maskeerder. Voor maskeerders met een spectrale helling van -6 dB/octaaf neemt de zo uitgedrukte detectiedrempel voor hoog-frequente doelsignalen aanzienlijk toe als gevolg van het lage spectrale niveau van de maskeerder verbonden aan zo'n spectrale helling. Deze drempeltoename wordt verder versterkt door de invloed van golfvormdispersie.

Door het uitvoeren van maskeringsexperimenten is tevens inzicht verkregen in ons onderscheidingsvermogen voor spraakgeluiden die resulteren uit bepaalde spraakbewerkingstechnieken. Kwantiseringsruis bij een gecodeerd spraaksignaal kan bijvoorbeeld gemaskeerd worden als de spectrale omhullende van deze ruis op de juiste manier bepaald wordt door middel van een perceptieve weegtechniek. Deze weegtechniek is voornamelijk gebaseerd op maskeergedrag van ons oor voor zuivere sinustonen. Maskeergedrag voor samengestelde geluiden kan echter erg moeilijk voorspeld worden vanuit kennis over maskeergedrag voor zuivere tonen. Onze experimenten hebben aangetoond dat het maskeren van ruisachtige doelsignalen door harmonisch samengestelde tonen voornamelijk bepaald wordt door plaatselijke details van de maskeerder in ofwel het frequentiedomein ofwel het tijdsdomein, en niet zozeer wordt bepaald door globale eigenschappen van het spectrum van de maskeerder. Er wordt daarom voorgesteld dat de weging in het laagfrequente gebied voor geluiden met een hoge grondfrequentie geassocieerd wordt met de harmonische structuur van het spraaksignaal. In het hoog-frequente gebied behoort het perceptieve gewicht geassocieerd te worden met de golfvorm van het signaal in het tijdsdomein. Gezien de voortdurend veranderende tijd-frequentiestructuur van spraakgeluiden, moet een dynamische adaptatie van perceptieve gewichten worden aangewend ter bevordering van de kwaliteit van spraakgeluid dat is gecodeerd met een lage bit-stroom, vooral in het geval van transiente geluiden.

Metingen van onze gevoeligheid voor faseverschillen tussen individuele deeltonen van spectraal-gekantelde samengestelde geluiden hebben aangetoond, dat een faseverschuiving gemakkelijk detecteerbaar is bij hoge frequenties. Faseveranderingen in spraaksignalen beïnvloeden hun subjectieve geluidskwaliteit, in het bijzonder voor transiente geluiden,

omdat ons gehoorsysteem erg gevoelig is voor golfvormveranderingen, speciaal bij hoge frequenties. Om zeker te zijn van spraakgeluid met een hoge kwaliteit moet zorg gedragen worden voor het behoud van de juiste faserelaties, vooral tussen hoge-orde harmonischen.

Acknowledgements

I wish to express my appreciation to the colleagues of the Instituut voor Perceptie Onderzoek (IPO) who helped me in one way or another, from daily life to scientific research. In particular, I acknowledge Prof. dr. H. Bouma for providing me the opportunity to conduct my Ph.D. research at IPO. I also express my gratitude from the bottom of my heart to my promoters, Prof. dr. A.J.M. Houtsma, Prof. dr. Y. Kamp, and Dr. A. Kohlrausch, for their every effort to guide me through the painstaking stage of writing this thesis. I wish also to express my thanks to Ir. L.F. Willems for his kind help and constant guidance. Furthermore I acknowledge Prof. dr. R. Collier for his constructive advice. Finally, I also thank Ing. Theo de Jong for making weekend-life more enjoyable, both inside and outside IPO.

Stellingen

behorende bij het proefschrift
Psychophysical and Signal-processing Aspects of Speech Representation
van Changxue Ma

I

The relative importance of the Fourier transform phase and amplitude for the perception of a reconstructed signal is strongly dependent on the window size of the short-time Fourier analysis.

A. V. Oppenheim and J.S. Lim, "The importance of phase in signals, Proc. of the IEEE, Vol. 69, pp.529-541, 1981.

II

PIOLA (pitch inflected overlap and add) technique is very appropriate for manipulating the pitch of speech sounds and producing natural-sounding speech, because our hearing is relatively insensitive to the phase manipulation in speech signals.

L.L.M. Vogten, C. Ma, W.D.E. Verhelst and J.H. Eggen, "Pitch inflected overlap and add speech manipulation", European Patent: 91202044.3

III

A finite signal sequence $(x_n, n = 0, 1, 2, \ldots, N-1)$ is not uniquely determined by its Fourier phase if the polynomial $\sum_{n=0}^{N-1} x_n z^n$ contains a factor of $1 + \alpha z^{-1} + z^{-2}$, given $\alpha \subset R$.

This thesis, Chapter 5.

IV

In the coding of high-pitched speech sounds, the use of a long-term pitch predictor improves the subjective quality of the coded speech. This is due to that fact that in the optimization process, coding errors around the low-harmonics are downweighted and coding errors between the low-harmonics are overweighted.

S. Singhal and B.S. Atal, "Amplitude optimization and pitch prediction in multipulse coders", IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 37, pp. 317-327, 1989.

V

Models which functionally describe natural processes should work well, be structurally simple, and be computationally efficient. Therefore, there are always possibilities to modify and to improve a model.

VI

The estimation of the coefficients of linear-prediction-coding (LPC) of speech can be improved by the selection of speech samples which fit the LPC model well.

This thesis, Chapters 2, 3

VII

Omission and simplification help us to understand things, but one must first understand what to omit and how to simplify.

VIII

High-fidelity loudspeakers are designed to have a flat spectral response only in testing rooms, not in living rooms. It is a good solution to use an adaptive digital filter for the frequency equalization at a fixed position, but it is a bad practice to achieve the frequency equalization for time-varying positions.

IX

Although the Fast Fourier Transform (FFT) was discovered again and again after Gauss, it has been valued only after the invention of digital computers.

M.T. Heideman, D.H. Johnson and C. S. Burrus, "Gauss and the history of the Fast Fourier Transform", IEEE ASSP Magazine, Vol. 1, No. 4, pp.14-21, 1984.

X

Who does not honor his/her teachers lacks wisdom in spite of his/her knowledge.