

PERCEPTUAL ENHANCEMENTS FOR AN
INTEROPERABLE FS-1016 CELP
SPEECH CODER

By

GUENTER ALBAN DANNORITZER

Diplom Ingenieur Electrical Engineering

Fachhochschule Dieburg

Dieburg, Germany

1996

Submitted to the Faculty of the
Graduate College of the
Oklahoma State University
in partial fulfillment of
the requirements for
the Degree of
MASTER OF SCIENCE
July, 1998

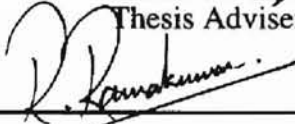
OKLAHOMA STATE UNIVERSITY

PERCEPTUAL ENHANCEMENTS FOR AN
INTEROPERABLE FS-1016 CELP
SPEECH CODER

Thesis Approved:



Thesis Adviser







Dean of the Graduate College

ACKNOWLEDGMENTS

The purpose of this study was to determine perceptual enhancements for an interoperable FS-1016 CELP coder. This study would not have been possible without the support and guidance of a number of people. I would like to take this opportunity to acknowledge and thank some of those people.

First, I would like to thank my adviser, Dr. Keith A. Teague, for his guidance, support and friendship throughout my studies at Oklahoma State University. I would especially like to thank Dr. Teague for offering me a position to work with him in the field of speech processing. I am very appreciative of this time with Dr. Teague and the people I met through this working experience. It was a very enjoyable and educational period in my life.

Additionally, I wish to thank the other members of my advisory committee, Dr. Scheets and Dr. Ramakumar, for the support they have given me during the last two years, as well as the School of Electrical and Computer Engineering.

I am also deeply grateful for the support and understanding given to me by my fiancé, my parents, and my brother and his family.

Finally, I would like to thank the Fulbright commission and the Institute of International Education for financing and supporting my stay here in the United States.

TABLE OF CONTENTS

Chapter	Page
I INTRODUCTION	1
Purpose	1
Thesis Outline	2
II BACKGROUND	4
Analysis-by-Synthesis Speech Coder	4
Noise Shaping	10
Auditory Masking	13
III PROPOSED SOLUTIONS	24
Improved Adaptive Codebook Search	25
Improved Stochastic Codebook Search	28
IV IMPROVED ADAPTIVE CODEBOOK SEARCH	30
Harmonic Filter	30
Results	33
V IMPROVED STOCHASTIC CODEBOOK SEARCH	40
Calculating the Masking Threshold	40
Time Domain Aliasing Cancellation (TDAC)	45
Noise Shaping with Masking Threshold versus LPC Spectrum	56
Postfilter Effect on Masked Regions of the Speech	63
Results	65
VI CONCLUSION	68
Discussion of the Results	68
Future Research	70
REFERENCES	73

LIST OF TABLES

Table	Page
1. Quiet environment DRT and DAM evaluation of the harmonic filter	34
2. Office environment DRT and DAM evaluation of the harmonic filter	35
3. MOS scores	38
4. Predicted MOS scores in quiet and office environment, based on the DAM results	39
5. Redundancy and voice quality due to raised discard threshold	55
6. Speech quality depending on spectral shape and amount of the added noise . . .	59
7. Attenuation introduced by the postfilter in masked regions of a 22 second long speech file	65

LIST OF FIGURES

Figure	Page
1. Analysis-by-Synthesis Structure	5
2. FS-1016 CELP Coder	6
3. Creating the excitation from the adaptive codebook: a.) The delay D is greater than the subframe size L ; b.) for $D < L$ the D sample long vector is repeated to fit L	7
4. Codebook search for the FS-1016 CELP coder	9
5. [KOND94] Original envelope of the speech and weighting filter functions as parameter of γ	12
6. [ZWICK90] Threshold in quiet, with age as parameter	16
7. [ZWICK90] Level of test tone just masked by critical band wide noise with 60 dB <i>SPL</i> , and center frequencies 0.25, 1, and 4 kHz; the dashed horizontal line shows the level of the noise; the other dashed line is the threshold in quiet	17
8. [ZWICK90] Level of a test tone just masked by critical band wide noise; the center frequency of the noise is 1 kHz with the different levels L_{CB}	18
9. [ZWICK90] Level of a test tone masked by 1 kHz tone of different levels	19
10. [ZWICK90] Level of a test tone masked by ten harmonics of a 200 Hz tone; the levels of the harmonics are given as a parameter	20
11. Implementation of the harmonic filter in the FS-1016 CELP coder; refer to Figure 2	27
12. Harmonic filter used to modify the speech for the ACB and the SCB search	30

13.	a) Plot of the word “figure” without background noise; b) synthesized by CELP 3.3; c) synthesized by the coder with the harmonic filter	36
14.	a) Plot of the word “figure”; b) delay values per subframe of the word “figure” determined by CELP 3.3; c) delay values per subframe determined by the coder modified with the harmonic filter	37
15.	a) Plot of the word “figure”; b) gain values per subframe determined by CELP 3.3; c) gain values determined by the coder with harmonic filter	38
16.	Virtual excitation pattern for a 1000 Hz (8.6 Bark) tone with level 70 dB (solid) and 20 dB <i>SPL</i> (dashed)	42
17.	Noise masking contribution for component f ; intensity within the critical bandwidth Δf around f is summed, excluding component f and its three DFT coefficient neighbors above and below	44
18.	SSB Filter Bank Analyzer	47
19.	Window arrangement in the time domain for the TDAC	48
20.	SSB Filter Bank Synthesizer	50
21.	a) Plot of the word “figure”; dashed is the original signal and dotted the reconstructed after the TDAC transform; b) error between original and reconstructed speech	53
22.	a) Frequency spectrum (solid) of 256 samples from the word “figure” and masking threshold (dashed); b) frequency samples below the masking threshold (dashed) are attenuated by the amount of the masking threshold	54
23.	Adding noise signals with different spectral shapes to a speech signal	57
24.	a) Speech spectrum (solid) vs masking threshold (dashed); b) LPC spectrum (solid) vs combined LPC and postfilter spectrum (dashed)	64

NOMENCLATURE

ACB	Adaptive Codebook
CELP	Code Excited Linear Prediction
D	Delay value in samples
DAM	Diagnostic Acceptability Measure
DCT	Discrete Cosine Transform
DRT	Diagnostic Rhyme Test
DST	Discrete Sine Transform
DFT	Discrete Fourier Transform
FFT	Fast Fourier Transform
Δf_g	Critical Bandwidth
FS-1016	Federal Standard 1016
G.723	Low Rate Speech Coding Standard
H	Linear Predictive Coding Filter
K	Number of Frequency Channels
ITU	International Telecommunications Union
I_G	Critical Band Intensity
L	Level / Vector size
L_E	Excitation Level

L_G	Critical Band Level
LP	Linear Prediction/Predictor
LPC	Linear Prediction Coefficients
m	Frame number
M	Progress of window
MOS	Mean Opinion Score
MPEG	Motion Picture Expert Group
N	Dimension of matrix, degree of filter, length of DFT
P	Window length
SCB	Stochastic Codebook
SPL	Sound Pressure Level
SSB	Single Sideband
STU	Secure Telephone Unit
TDAC	Time Domain Aliasing Cancellation
W	Perceptual Weighting Filter, Postfilter
z	Variable for frequency on Bark scale

CHAPTER I

INTRODUCTION

Purpose

This thesis describes several ways to improve the speech quality of a Federal Standard 1016 (FS-1016) Code Excited Linear Prediction (CELP) coder. The required changes for the improvement keep the coder interoperable with Federal Standard 1016. Although there are numerous papers that already propose improvements to CELP type speech coders, most of them rely on changes which do not let the newly fitted coder interoperate with the original standard coder. This makes the improvements useless in areas where the coder is already widely deployed. One of these cases where the FS-1016 coder is currently used is the secure telephone unit for the United States Government. An improvement that complies with the original standard would allow the incorporation of the new technology in new or modified units while maintaining full compatibility.

Based on literature search, the improvements made were focused on changing the codebook search procedure. Two developments are introduced which independently bring better speech quality. They can be implemented together without interfering with one another. The first major focus in this thesis is on improving the adaptive codebook search by using a harmonic filter. The second major focus is on improving the stochastic codebook search by applying perceptual noise shaping based on an auditory model.

Thesis Outline

The remainder of this thesis details the development of improvements to the coder to achieve a better perceptual synthesized speech quality. A breakdown for the rest of this thesis is presented in the following paragraphs.

Chapter II provides the reader with the necessary background to Analysis-by-Synthesis coding with a focus on the FS-1016 CELP coder. The explanation is emphasized in areas of the coder which are more important to understanding the proposed solutions and the way they are implemented. One of these areas is the adaptive codebook and how it models excitation for periodic parts of the speech signal. This leads to how a codebook search is performed in the CELP coder. A part of the codebook search is perceptual weighting of the error signal. With this knowledge it will be obvious that noise shaping cannot be done sufficiently without considering the phenomena of auditory masking of the human ear.

Chapter III points out the proposed changes to the coder and elaborates why they comply with the standard. Chapter IV gives background about the changes in the adaptive codebook search and details the results for the implementation.

Chapter V focuses on the improvements made in the stochastic codebook search. These improvements rely on a psycho-acoustic model. The chapter begins with a background on auditory masking. The model is tested for proper identification of masked and unmasked regions of a speech signal. Then noise shaping abilities based on the masking threshold are compared with the conventional noise shaping used in the CELP coder. Based on that, a decision is made on the best way to introduce the masking model in the noise

shaping procedure of the coder. Further tests are performed to determine if the proposed solution will be interoperable with the standard coder.

Chapter VI provides the conclusion to this work. Finally, some suggestions for future research are given.

CHAPTER II

BACKGROUND

Analysis-by-Synthesis Speech Coder

The goal of speech coding is to represent high quality speech with the least amount of data. Its basics [DELL87] can be traced back to the 1930's [ANDR84] with the first all-electrical speech synthesizer by Dudley et al. [DELL87].

Low rate speech coding deals with data rate below 8 kb/s and is used in areas where limited channel capacity is available. Such areas are secure telephone units or cellular phones, just to name a few [RAB94]. This chapter gives a brief description of a special group of low rate speech coders, the Analysis-by-Synthesis coder. The focus will be on one specific type of coder, the FS-1016 CELP [NCSO91] coder. Another emphasis will be on perceptual noise shaping used in the coder to improve the speech quality. This will lead to a discussion dealing with the ability of the ear to mask certain sounds in the presence of other sounds, called auditory masking. Necessary terminology that accompanies this topic and which will be important for later use in this thesis will be explained.

Analysis-by-Synthesis coding [KRO88] [KLEIJ95] [KOND94] is a novel way to achieve fairly good speech quality with data rates below 8 kb/s. In an iterative process a synthesized signal is matched to the original speech, based on some error criteria. Figure 1 shows the structure of an Analysis-by-Synthesis coder. Typically, the mean squared error is used for the matching of the synthesized speech \hat{S} to the original speech S .

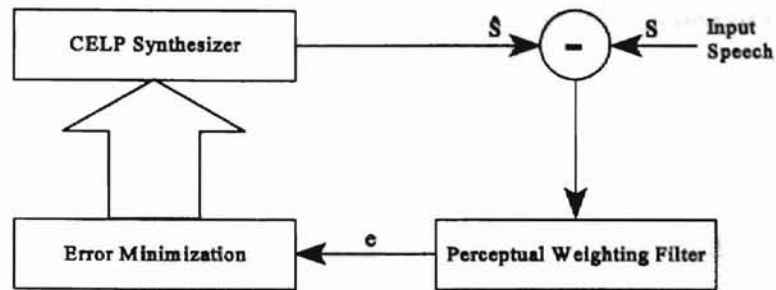


Figure 1. Analysis-by-Synthesis Structure

The benefits of this scheme are that it checks whether the coding procedure is operating efficiently, and it reduces the propagation of errors over several frames. The error criterion is modified by a perceptual weighting filter. This is a way to shape the noise in the synthesized speech and make it less audible. Perceptual weighting will be covered in more detail later.

Analysis-by-Synthesis coding started out being used in conjunction with linear predictive (LP) coders. It is common for linear predictive coders to have the speech signal divided into the vocal tract response and the excitation of the vocal tract. The vocal tract response is modeled by an adaptive filter which is determined through linear prediction. There are several ways to construct the excitation signal which leads to different coder types.

With Multi-Pulse excitation [ATAL82], the optimum excitation vector is specified by a small set of pulses with different amplitudes located at non-uniformly spaced intervals. The restriction of keeping the pulses uniformly spaced leads to the Regular Pulse Excitation Coder [KROO86]. Separating the excitation vectors according to the type of speech,

whether voiced or unvoiced, and storing many different vectors in one or more codebooks, results in the most popular of these linear predictive coders, the Codebook Excited Linear Prediction, or CELP, coder [SCHRO85].

Among many variations of the CELP coder, one is standardized in Federal Standard 1016 (FS-1016) [CAMP91] [NCSO91] [NCS92].

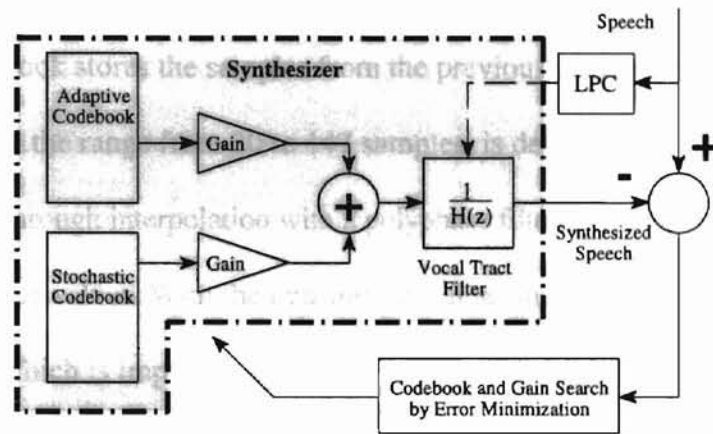


Figure 2. FS-1016 CELP Coder

The FS-1016 coder is shown in Figure 2. The speech synthesizer is shown in the gray part of Figure 2. The processing is done by splitting the speech into 240 sample long frames. A frame is further divided into 4 subframes of 60 samples each. The linear prediction (LP) analysis is done once per frame, and the codebook search is performed for every subframe. As mentioned above, the excitation signal can be divided into two categories. During unvoiced speech it is more noise-like and during voiced speech it is more periodic. This leads to the concept of two codebooks: one that stores codewords to model a noise-like excitation, called the stochastic codebook (SCB), and one that models the periodic excitation, called the adaptive codebook (ACB). The adaptive codebook's basic

principle can be explained with a quote from Machiavelli [MACH79, vol. III ch 43]. The English translation is from the speech processing book edited by Kleijn [KLEIJ95, p.592].

“AS A RULE, PRUDENT MEN CLAIM---NEITHER BY ACCIDENT NOR WITH OUT MERIT--- THAT SOMEONE WHO WANTS TO FORESEE WHAT IS TO COME SHOULD CONSIDER WHAT HAS BEEN: AS ALL THINGS ON EARTH, AT ANY TIME, FIND THEIR OWN COUNTERPART IN THE PAST TIMES.”

The codebook stores the samples from the previous excitation. For every subframe the delay value, in the range from 20 to 147 samples, is determined. There are 128 integer delay values and through interpolation with a polyphase filter, this can be further subdivided into 128 non-integer values. With the non-integer values the delay can be determined up to 1/4 of a sample, which is important for high pitch speakers.

Figure 3 shows how the excitation is constructed from the adaptive codebook. D_{min} and D_{max} refer to the delay range from 20 to 147 samples, respectively. Referring to Figure

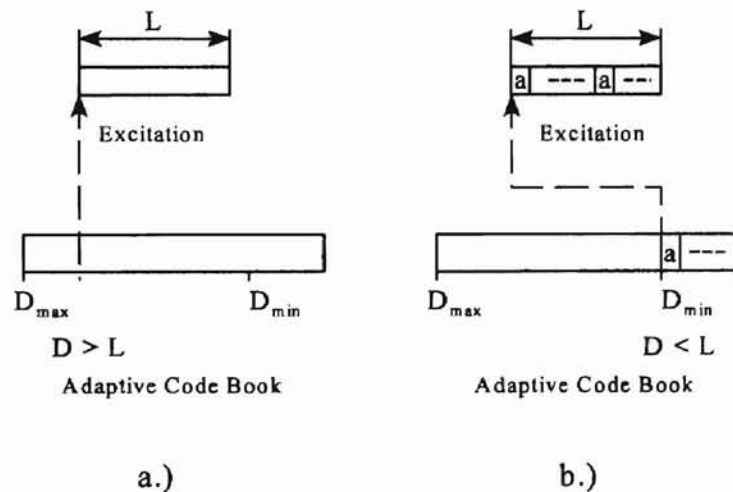


Figure 3. Creating the excitation from the adaptive codebook:
a.) The delay value D is greater than the subframe size L ;
b.) for $D < L$ the D sample long vector is repeated to fit L

3a, for delay values D greater than the subframe length L , the samples from D until $D-L$ are used for the excitation. In Figure 3b it is shown that the excitation for delay values that are smaller than the subframe size, the D samples are repeated until they add to L .

When the excitation signal is determined, which is done after the adaptive and the stochastic codebook search, the adaptive codebook is updated with that excitation.

With the stochastic codebook the noise-like excitation is formed. The stochastic codebook contains, in 512 codewords, samples of a zero-mean, unit-variance, white Gaussian sequence. The codewords are stored as ternary values.

Two amplifiers take care of the amount of periodic and noise-like excitation. For the stochastic codebook the amplifier ranges from -1330 to 1330 in 32 non-uniform steps [NCS92]. The gain for the adaptive codebook ranges from -0.993 to 1.991 in 32 non-uniform steps. For voiced frames the contribution of the stochastic codebook is decreased by attenuating its gain. This method introduced by Shoham [SHOH91] improves the subjective speech quality of the coder by reducing roughness and quantization noise in voiced areas.

The constructed excitation signal is fed to the linear predictive filter which models the vocal tract response. By trying all combinations of entrances in the two codebooks, an error value for each codebook entry is calculated. The two entries with the smallest error values are used to synthesize the speech signal.

The actual implementation of the codebook search is done in a slightly different way to reduce the amount of calculation and is shown in Figure 4 [NCS 92].

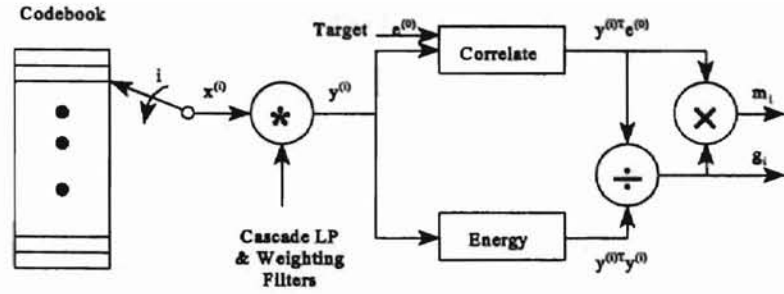


Figure 4. Codebook search for the FS-1016 CELP coder

In this form, minimizing the error equals maximizing the match score m_i . The codebook search is performed sequentially, first for the adaptive and then for the stochastic codebook. The two searches are basically identical. To find the maximum match score for one codebook, a so called target signal $e^{(0)}$ is compared with all filtered codewords $y^{(i)}$. This calculation is shown in equation (2.1).

$$m_i = \frac{(y^{(i)T}e^{(0)})^2}{y^{(i)T}y^{(i)}} \quad (2.1)$$

The optimal gain value for the codebook entry is calculated by equation (2.2).

$$g_i = \frac{y^{(i)T}e^{(0)}}{y^{(i)T}y^{(i)}} \quad (2.2)$$

To find the filtered codeword $y^{(i)}$, the codebook entry $x^{(i)}$ is convolved with the impulse response of the vocal tract filter H and the perceptual weighting filter W to form $y^{(i)}$, as shown in (2.3).

$$y^{(i)} = WHx^{(i)} \quad (2.3)$$

H and W are lower triangular matrices whose columns contain the truncated impulse response of the vocal tract filter and the weighting filter, respectively. This fact limits the type of filter, especially for the perceptual weighting filter, as the impulse response has to decrease sufficiently within the 60 sample size of the matrix to avoid excessive truncation error.

The search for the optimum codeword of the stochastic codebook uses the most calculation power of the whole coding process of the CELP coder.

To build the target $e^{(0)}$, the zero input response $\hat{S}^{(0)}$ of the vocal tract filter is subtracted from the speech S and convolved with the weighting filter (2.4).

$$e^{(0)} = W(s - \hat{s}^{(0)}) - WHu \quad (2.4)$$

In this calculation, there is a difference in whether the target is built for the adaptive codebook search or for the stochastic codebook search. For the stochastic codebook search the contribution of the adaptive codebook, denoted by u in (2.4) and found in the first stage of the search has to be incorporated. Convolved with the impulse response of H and W , it is subtracted from the part $W(S - \hat{S}^{(0)})$ of the target. During the first stage, the adaptive codebook search, u is set to zero.

In the following section a closer look is presented on the way perceptual weighting is done.

Noise Shaping

The principle behind noise shaping is to form the noise level in the synthesized

speech so that it is less audible [KOND94] [SEN94]. One popular, but not totally satisfactory way that is often found with Analysis-by-Synthesis coders, is the use of a perceptual weighting filter [ATAL79]. This filter modifies the error minimization in a way that the noise level is shaped according to the speech spectrum. The weighting filter is given by equation (2.5).

$$W(z) = \frac{H(z)}{H(z/\gamma)}, \quad 0 \leq \gamma \leq 1 \quad (2.5)$$

$H(z)$ is the vocal tract filter determined by linear prediction. The weighting filter is used to filter the error signal after the comparison of the original speech with the synthesized speech. The effect of the factor γ is to broaden the bandwidth of the formant frequencies by Δf , given by (2.6).

$$\Delta f = -\frac{f_s}{\pi} \ln \gamma \quad (\text{Hz}) \quad (2.6)$$

Here f_s is the sampling frequency in Hertz, used to digitize the speech for the coder. In Figure 5 it can be seen that the weighting filter de-emphasizes the frequency regions corresponding to the formants. With decreasing value of γ , more noise is allocated in the formant regions leaving less noise in the formant valleys, which improves the speech quality perceptually.

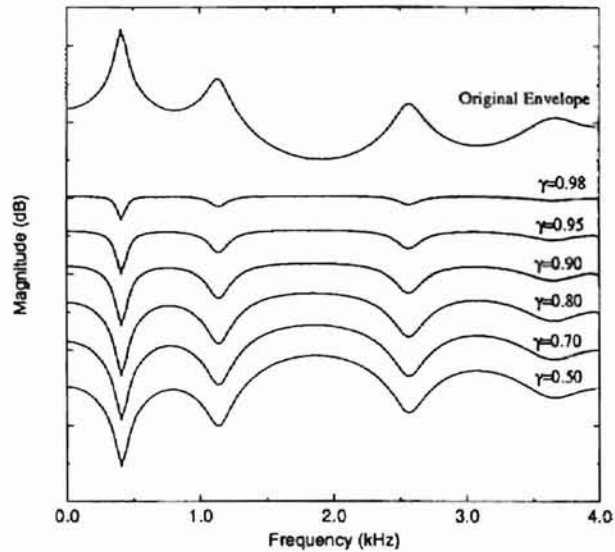


Figure 5. [KOND94] Original envelope of the speech and weighting filter functions as parameter of γ

Even though this process improves the quality of the synthesized speech perceptually, it does not account for adapting the synthesized signal to the final receiver, which is the human ear. It is recognized [SCHRO79] [ZWICK91] that by matching the coding to the hearing process, a better speech quality can be achieved. Investigations in auditory masking properties [SEN94] have shown that it is not enough to keep the noise below the signal level at all frequencies to guarantee that only the higher level signal will be perceived. While it is not the goal to eliminate the perceived noise at these bit rates, Sen [SEN94] hypothesized that the use of an auditory model to compute the masking level and replace the traditional weighting filter, would better match the coding to the hearing process. This would result in significantly better quality of the synthesized speech. This idea leads to the subject of auditory masking which will be explained next.

Auditory Masking

Masking is the phenomenon of reduced audibility of a signal due to the presence of a second so called masker signal [ZWICK90]. Most information about sound perception is based on psycho-acoustic experiments, and some results important for the use of auditory masking in speech coding will be presented in the following paragraphs.

Masking can be distinguished in its effects in the frequency domain and in the time domain. In the time domain three ways of masking are known, depending on what time they occur relative to the masker signal. Pre-masking occurs during a very short time before the masker signal appears and is rather unimportant to be used in coding techniques. The period while the masker is switched on is called simultaneous masking. Even though the masker is switched off, it has for a short time a decaying masking effect, which is called post-masking. Of these three, the simultaneous masking is the most effective one to be incorporated into speech or audio coding techniques. For simplicity, the masking effects in the frequency domain can be seen as constant in the period of simultaneous masking. Before explaining the results of several frequency domain masking tests, the concept of critical bands is introduced.

Critical Bands

The idea of critical bands was first proposed by Fletcher [FLET72] and an explanation can be found in Zwicker [ZWICK90] and Sen [SEN94]. For uniformly distributed noise, only a part around a test tone contributes to mask that test tone. Masking is achieved when the acoustic power of the tone and of the noise lying in the critical band

around the tone are the same. Noise outside the critical band does not contribute to masking. The critical bandwidth for tones has been determined by listening tests [ZWICK90]. From the data of these tests several equations are derived for calculating the critical bandwidth for a certain frequency. Equation (2.7) shows one given by Davidson [DAVI90] where f is the center frequency of the tone and Δf_g is the critical bandwidth.

$$\Delta f_g = 6.23f_{Hz}^2 + 93.4f_{Hz} + 28.5 \quad (2.7)$$

For frequencies up to 500 Hz the critical bandwidth can be assumed constant and is about 100 Hz. Above 500 Hz up to 13 kHz, it is approximately 20% of the center frequency.

The critical band concept [ZWICK90] is so important for describing hearing sensations that a unit was defined, leading to the so called critical band rate scale. This scale is based on our hearing system's feature that analyzes a broad spectrum in parts that correspond to critical bands. The audible frequency range to 16 kHz can be divided into 24 critical bands. The unit of the critical band rate scale is "Bark¹", which is represented by the variable z . Here again as with the critical bandwidth there are several equations given to calculate the critical band rate scale. Sen [SEN94] compared these and how they change the result when calculating a masking function. He came to the conclusion that the choice of the function did not cause large fluctuations of the masking threshold. As a result he used Terhardt's [TERH79] model which is defined in equation (2.8).

¹ The unit "Bark" was introduced in memory of Barkhausen, a scientist who introduced the "phon", a value describing loudness level.

$$z_{Bark} = 13.3 \arctan \left[0.75 \left(\frac{f_{Hz}}{1000} \right) \right] \quad \text{for } 0 \leq f_{Hz} \leq 4000 \text{ Hz} \quad (2.8)$$

With this equation the critical band rate z is calculated for the frequency component f_{Hz} .

These absolute values for z do not mean that the critical band exists only between these values. It should rather be seen as being able to shift over the frequency range. For one example is the critical band rate between 1 and 2 Bark which is the frequency range between 100 and 200 Hz. Having a tone at 190 Hz does not mean that the effect of a uniformly masking noise on that tone ends at 200 Hz. Rather the noise within the critical bandwidth of that tone masks it. The scale is used to display hearing characteristics, for example, as in plots.

Before going into the explanation of different types of masking effects in the frequency domain, some explanation about sound physiology is needed. Sounds are usually described in terms of time varying sound pressure. The unit of sound pressure is the Pascal (Pa). For psycho-acoustics, relevant sound pressures range from 10^{-5} Pa to 10^2 Pa. To cope with such a broad range the sound pressure level, L , is used. Sound pressure and sound pressure level are related by the equation (2.9).

$$L = 20 \log \frac{p}{p_0} \text{ dB} \quad (2.9)$$

The reference value of the sound pressure p is standardized to $p_0 = 20 \mu\text{Pa}$. Sometimes the abbreviation *SPL* instead of L is used to avoid confusion when used in

combination with levels referring to voltage or energy.

Threshold in Quiet

The threshold in quiet represents as a function of frequency the sound pressure level of a just audible tone. Figure 6 shows the threshold in quiet. It has a dynamic range from about 70 dB in the audible frequency range. Starting out with a threshold of about 70 dB sound pressure level at low frequencies around 200 Hz, the ear has its highest sensitivity at frequencies around 2 to 5 kHz. For higher frequencies the threshold rises again and reaches a limit at 16 to 18 kHz above which no audible sensation occurs [ZWICK90].

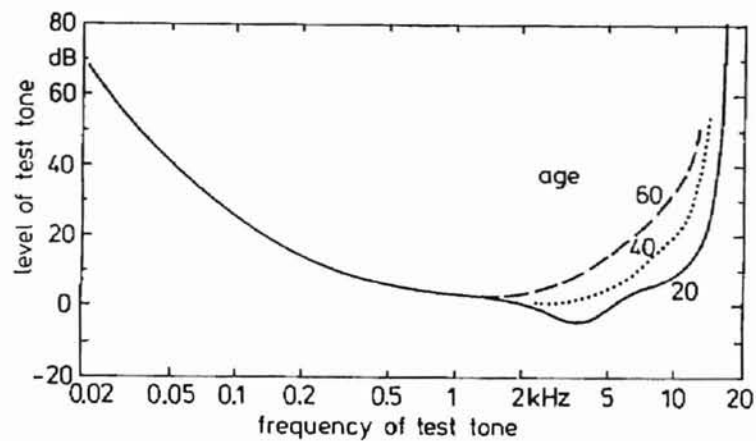


Figure 6. [ZWICK90] Threshold in quiet, with age as parameter.

This upper limit is dependent on the age and the amount of exposure to loud sound environments of the listener. Figure 6 shows the age dependency by the dotted and the dashed lines.

Pure Tone Masked by Narrow-Band Noise

In this experiment the noise bandwidth is equal to or smaller than the critical bandwidth, which will be referred to as “*critical band wide noise*.” For the center frequencies of 250 Hz, 1 kHz, and 4 kHz a critical band wide noise with a level of 60 dB is used.

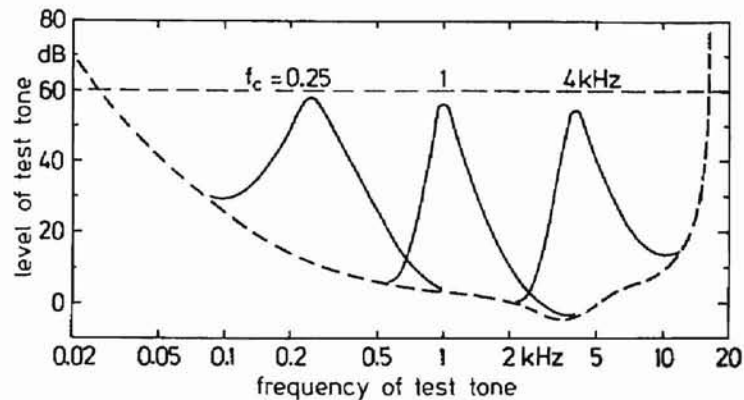


Figure 7. [ZWICK90] Level of test tone just masked by critical band wide noise with 60 dB *SPL*, and center frequencies 0.25, 1, and 4 kHz; the dashed horizontal line refers to the level of the noise; the other dashed line is the threshold in quiet

Figure 7 shows the threshold of the pure tone masked by the noise. Lets take for example the noise around the 250 Hz center frequency. The critical band width is here, as mentioned earlier, around 100 Hz. A tone is swept from the lower frequencies through that critical band wide noise to higher frequencies. For every frequency value the level of the tone is increased until it is audible. The maximum inaudible level is the plotted solid line in Figure 7. The horizontal dashed line is to display the level of the noise. It is meant to display the difference from the maximum masking threshold to the noise level and is helpful to notice the following observation.

The maximum masking threshold decreases with increasing frequency. For the 250 Hz center frequency, the maximum masking threshold is 58 dB [ZWICK90], 57 dB for the 1 kHz center frequency, and only 55 dB for the 4 kHz center frequency. This will be used later when the masking threshold is calculated. Another feature to point out is that the shape of the masking thresholds at 1 kHz and 4 kHz are similar in broadness. The shape at 250 Hz is much broader than the shapes at higher frequencies. From lower frequencies the masking threshold increases until reaching the maximum value at the center frequency of the narrow band noise. From there the threshold decreases, flatter than its increase. The increase is about 100 dB per octave [ZWICK90].

In another experiment related to the one above, the level dependency of the critical band wide noise on the masking threshold is investigated. Figure 8 shows the masking threshold of a test tone just masked by a critical band wide noise. The noise is centered at 1 kHz. The slope of the masking threshold below the center frequency is independent from the level of the noise. In contrast to that, the masking threshold slope above the center

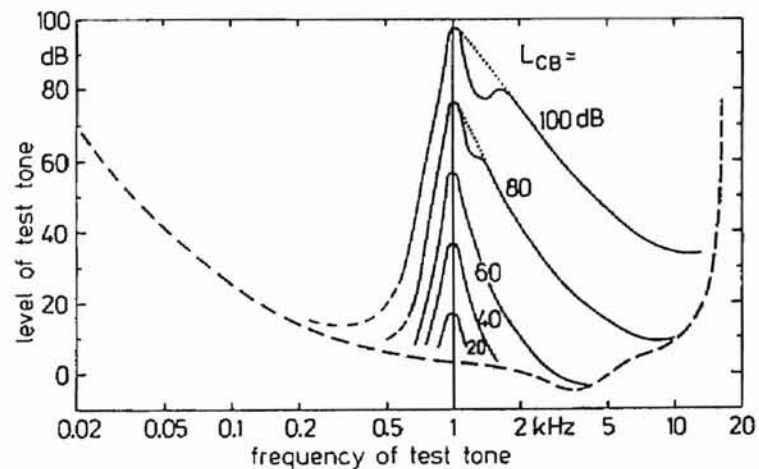


Figure 8. [ZWICK90] Level of a test tone just masked by critical band wide noise; the center frequency of the noise is 1 kHz with the different levels L_{CB}

frequency depends on the level of the noise. This effect is important for later when the masking threshold is calculated.

The dips for masker levels of 80 and 100 dB result from nonlinear effects in our hearing system which leads to interaction between the noise and the test tone [ZWICK90]. The solid line of the dips are the actual measured level and the plot is corrected to represent the actual behavior, shown by the dotted line [ZWICK90].

Pure Tone Masked by Tone

This experiment is similar to the one above, with the difference that instead of using a critical band wide noise this time a tone is used as a masker. Comparing the results of this experiment shown in Figure 10 with the previous one that has the narrow band noise as the masker yields an interesting aspect. Masking a tone with another tone gives a masking threshold where both slopes depend on the level of the masker signal. The lower slope of the masking threshold becomes flatter with decreasing level of the masker tone. For low levels, the slope is even flatter than the upper slope. At a masker level of 40 dB both slopes

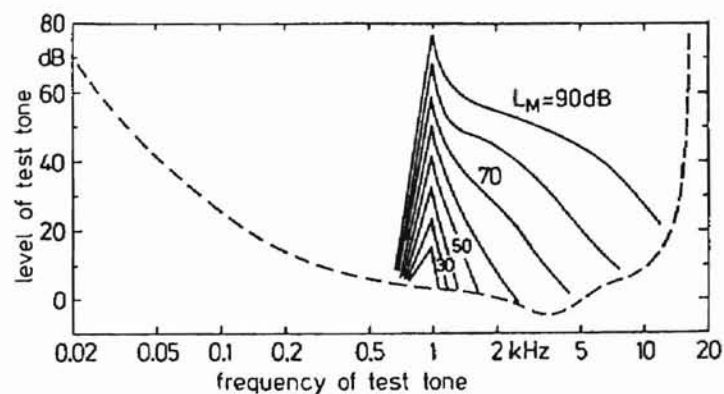


Figure 9. [ZWICK90] Level of a test tone masked by 1 kHz tone of different levels

are symmetrical. Another noticeable difference is that the maximum of the masking threshold is more pronounced than in the noise masking case.

The maximum of the masking threshold is lower than with the narrow band noise as masker. The curve in the area of the 1 kHz tone is estimated due to distortion by the difference tone. This difference between the maximum masking threshold of a narrow band noise versus that of a tone as a masker will be explained more in detail later and has an effect on how the overall masking threshold for a signal is calculated.

Pure Tone Masked by Complex Tones

After these more theoretical focused tests, a more practical one is explained that shows a scenario found in every day life. It is the masking of a tone by a complex tone.

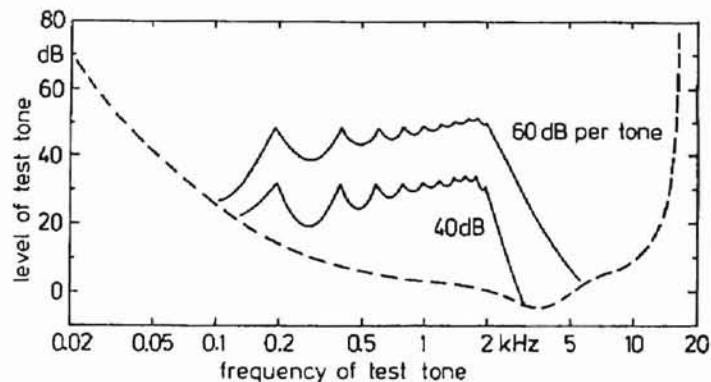


Figure 10. [ZWICK90] Level of a test tone masked by ten harmonics of a 200 Hz tone; the levels of the harmonics are given as a parameter

The complex tone is comprised of a 200 Hz fundamental frequency and nine higher harmonics with the same amplitude and random in phase. Figure 10 shows the result. The harmonics appear spread further for lower frequencies because of the logarithmic scale. For

higher frequencies the valleys between the harmonics become smaller and the maximum masking threshold even increases. After the last harmonic, the masking threshold decreases with a slope that depends on the level of the masker. Towards higher frequencies more and more harmonics fall into a critical band. This leads to the effect of masking a tone by a tone which turns over into masking a tone by a narrow band noise. Zwicker [ZWICK90] found through experimentation that the masking effect of a critical band wide noise can be approximated very well by five tones.

Asymmetry of Masking

So far the effects of masking a tone have been explained. In speech coding the primary goal in dealing with the effect of masking is to make noise inaudible that was introduced by the coder. Hellman [HEL72] determined that noise is more effective as a masker than a tone. This effect is called the asymmetry of masking and was already noticed in Zwicker's experiments, described above. The masking effect of a 90 dB *SPL* tone at 1 kHz is the same as a 72 dB *SPL* of critical band wide noise with a center frequency of 1 kHz. Through further investigation he found out that decreasing the bandwidth of the noise below the critical bandwidth improves the symmetry, but at high intensities a difference of close to 20 dB remains. Masking ceases when the effective energy of the masking signal and the signal that is masked are the same.

Zwicker's Model of Excitation

After considering masking in terms of tones and band limited noise Zwicker

[ZWICK90] developed a model to calculate based on the previous results the masking threshold. The first step to that is the calculation of the so called excitation. This term excitation is not to be mistaken with the excitation used in the speech coder to excite the LP filter.

To approximate the frequency selectivity of our hearing system, the frequency domain signal is divided in terms of critical bands. Sound intensity I within a critical band is used to calculate the critical band intensity I_G according to equation (2.10).

$$I_G(z) = \int_{z^{-0.5}}^{z^{+0.5}} \frac{dI}{dz} dz \quad (2.10)$$

The critical band level is calculated with respect to $I_0 = 10^{-12} \text{ W/m}^2$ by (2.11).

$$L_G = 10 \log \frac{I_G}{I_0} \text{ dB} \quad (2.11)$$

Calculating the critical band level is comparable to shifting a one Bark wide window continuously over the critical band scale. Integrating the intensity over that window results in the critical band intensity for the center frequency of the window. To better represent the discrimination performed by the hearing system, the infinitely steep slopes of the Bark wide window are replaced by the upper and lower slopes of the masking threshold for sinusoidal tones masked by narrow band noise, shown in Figure 8. Smoothing the slopes, leads to the excitation level L_E . From here, the masking threshold is calculated by adding the threshold in quiet and a so called masking index. This index is about -2 dB at low frequencies and

increases to -6 dB at higher frequencies. The masking index results from the effect shown in the experiment of masking a tone by narrow band noise in Figure 7. The maximal masking threshold decreased in that experiment with increasing frequency.

Building on this background in the next chapter, two solutions are proposed that each lead to an improvement in perceptual speech quality of the FS-1016 CELP coder.

CHAPTER III

PROPOSED SOLUTIONS

Before discussing the proposed solutions to improve the coder, let's talk about what it means for changes to be interoperable with the FS-1016 standard [NCSO91]. The coder transforms the speech on a frame by frame basis into parameters that describe the speech. For the CELP coder those parameters are the adaptive and stochastic codebook indices, the adaptive and stochastic codeword gain, and the vocal tract response. Federal Standard 1016 describes how these parameters are to be transmitted. To make changes to the coder and keep the altered coder interoperable with the standard, there are two points that have to be kept:

- Maintain the existing parameters and their representation.
- Change only the non-synthesizer part of the analyzer or parts of the synthesizer that are not part of the analyzer.

While maintaining the parameters and their representation, either the analyzer or the synthesizer can be changed in any way with the only restriction that the synthesized speech sounds like the input speech to the coder. This would allow changes to either the whole analyzer or the synthesizer, which is not the goal of this work.

Point two restricts us to making algorithmic changes to the present coder. Recall that the synthesizer is part of the analyzer. The non-synthesizer parts of the analyzer are the LP analysis and parts of the Analysis-by-Synthesis loop, such as the perceptual weighting filter and the search procedure for the codebook indices and gains. When changes are made

to these parts, the synthesizer can still deal with them. On the other side, there can be changes made to the parts of the synthesizer that are not in the synthesizer of the analyzer, which would be the post-filter for example. The solutions proposed here aim to improve the search procedure for the codebooks and the gains in such a way that the codewords chosen for the excitation lead to perceptually better synthesized speech than the codewords picked by the standard coder.

Improved Adaptive Codebook Search

The primary goal of the adaptive codebook search is to find an excitation vector that models the periodic part in the synthesized speech. The delay value determined by the adaptive codebook search is related to the pitch frequency of the speaker. One problem with the adaptive codebook, as with many other pitch predictors, is that incorrect delay values are often selected. Especially common is that the found value is double or half of the actual value. Delay errors lead to audible distortion in the synthesized speech, often characterized by excessive roughness.

One way to improve the adaptive codebook search is to apply a comb filter and thus emphasize the periodic structure of the speech. The adaptive codebook search is then more likely to select the right value of delay. There are several ways to add this comb filter to improve the search procedure.

Two approaches that have appeared in the literature or in coding standards such as ITU standard G. 723, which is more recent than FS-1016, are harmonic noise weighting and harmonic filtering. Harmonic noise weighting [MANO95] [KWON97] [GERSO92] refers

to the process of adding to the conventional perceptual weighting filter explained in Chapter II a comb filter, which is made adjustable to the delay value of the applied frame of speech. This implements weighting not only in terms of formant peaks and formant valleys as explained in Chapter II, but also in terms of harmonic peaks and valleys. The filter is invoked for voiced speech frames and more noise is placed in areas of harmonic peaks than is put in the areas of harmonic valleys.

Harmonic filtering [WANG90] refers to the process of enhancing the harmonic structure either of the synthesized speech, the excitation signal, or the input speech.

Both methods have in common the use of a comb filter of the form given in equation (3.1).

$$P(z) = 1 + \beta z^{-D} \quad (3.1)$$

The value for β determines how strongly the filter is invoked. β is updated for each frame based on the processed speech. Here D is the delay value which for the CELP coder is in the range from 20 to 147 samples. This results in an impulse response for the comb filter that depends on the delay value D and ranges from 20 up to 147 samples. When incorporating this filter as a harmonic weighting filter in the CELP coder, it would be added when filtering the codebook entry given in equation (3.2) and (3.3).

$$y^{(i)} = PWHx^{(i)} \quad (3.2)$$

The harmonic weighting filter is denoted by P which is a lower triangular matrix of its impulse response.

$$e^{(0)} = PW(s - \hat{s}^{(0)}) - PWHu \quad (3.3)$$

As mentioned in Chapter II, the convolution of the codebook vectors with the LP filter and the weighting filter is truncated to 60 values. This discounts the effect of the truncated part to the present frame [BERO84]. Truncating the impulse response after the comb filter operation of equation (3.1) would only work for low delay values.

This fact led to favor a harmonic filter to improve the adaptive codebook search.

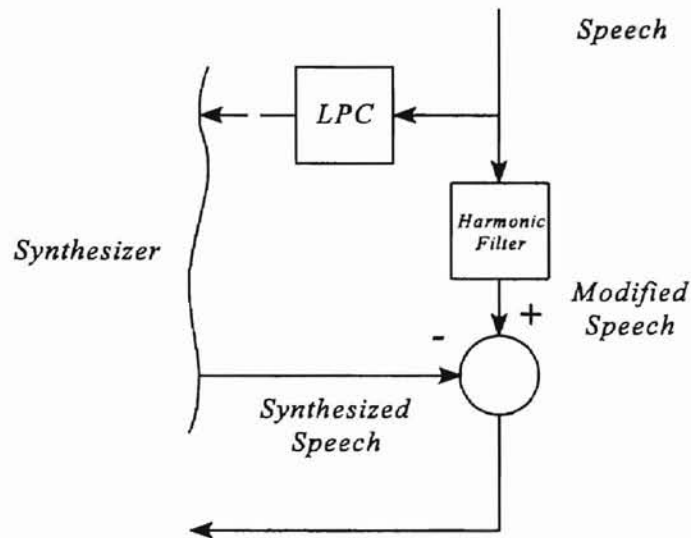


Figure 11. Implementation of the harmonic filter in the FS-1016 CELP coder; refer to Figure 2

Figure 11 shows where this harmonic filter is implemented in the CELP coder. The input speech for the LP calculation is kept unchanged and the harmonic modification is only performed on the reference speech for the adaptive and the stochastic codebook searches. The main effect of the harmonic emphasis will be on the adaptive codebook search, but for easier implementation it will be applied during both searches. Chapter IV describes the

implementation and discusses the results that the filter brings to the coder in terms of improved estimation of delay values and subsequent improvement in speech quality as produced by FS-1016 CELP.

Improved Stochastic Codebook Search

Chapter II led the way to investigate further the effect of an auditory masking model to improve the noise shaping performed by the CELP coder. After the FS-1016 standard was published, several papers addressing the use of an auditory masking model in CELP type speech coders were published. Auditory masking models are already in practical use in subband coders [VELD89] for audio signals. One standardized coder which uses auditory masking to enhance performance, for example, is the MPEG audio coder [AMBI97].

Montagna et al. [MONT91] described a modification for the CELP coder where the transfer function of the perceptual weighting filter is not based on the inverse vocal tract response, but rather on the auditory masking function of the ear. After determining the auditory masking function for a frame of speech, which is performed in the frequency domain, the coefficients for the perceptual weighting filter are determined. Here again a limiting factor is the truncated impulse response for the codebook search as mentioned above.

Sen [SEN93] [SEN94] followed a different approach by performing the stochastic codebook search in the frequency domain. After calculating the auditory masking function for the frame of speech, the error minimization is performed in the frequency domain. Here the problem with the truncated impulse response is bypassed, and the convolution in the

time domain for the codebook search becomes a multiplication in the frequency domain. Sen also discussed in more detail the calculation of a suitable masking function.

In Chapter V a closer look at the auditory masking model described by Sen [SEN94] is presented. Its functionality is tested first by introducing noise in the masked areas of a speech signal. A subband coding technique called "Time Domain Aliasing Cancellation" (TDAC) is used to transfer the speech signal into the frequency domain. After the masking threshold of the noise is calculated, noise is added by distorting the frequency samples that are below the masking threshold. Then the samples are transferred back in the time domain. The reconstructed speech signal should be perceptually indistinguishable from the original speech. The TDAC technique is used because of its ability to transfer speech into the frequency domain and back in the time domain without adding distortion [PRIN86]. This feature will be demonstrated.

In a second test, the masking model will be compared with the present noise shaping technique used in the CELP coder that is based on the vocal tract response. Two noise signals are created both with equal energy. One is shaped in the frequency domain according to the masking threshold. The other is shaped according to the vocal tract response. These two noise signals are added to the speech signal and the difference in perceiving the noise is investigated.

Which of the above mentioned methods is more suitable for implementation into the coder will be based on the results of this investigation.

CHAPTER IV

IMPROVED ADAPTIVE CODEBOOK SEARCH

Harmonic Filter

This chapter explains the implementation of the harmonic filter and its test results, which was introduced in Chapter III. The harmonic filter is applied to the input speech for the adaptive and the stochastic codebook searches. The major improvement however is to the adaptive codebook search by increasing the accuracy of the delay value calculation. The filter operation is performed on the speech S in equation (2.4).

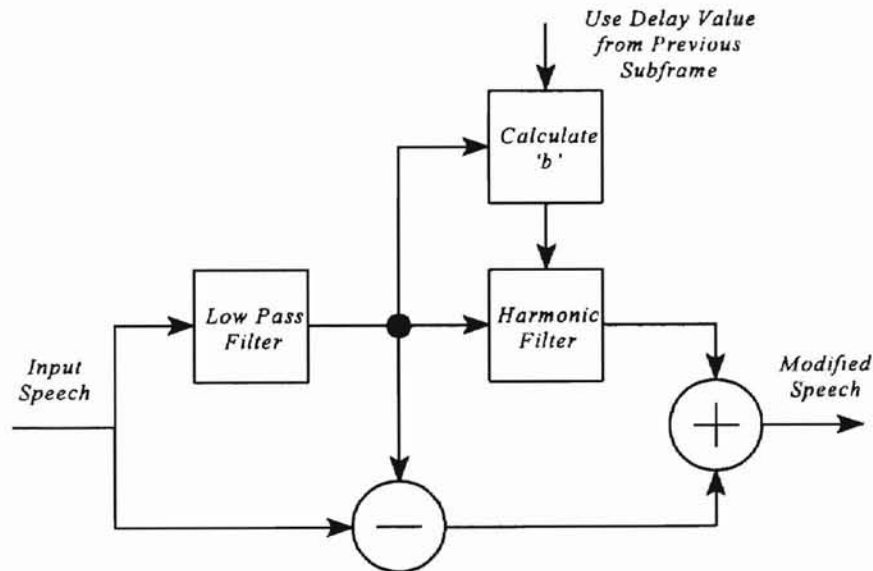


Figure 12. Harmonic filter used to modify the speech for the ACB and the SCB search

Figure 12 shows the way the speech is modified. Performing the harmonic filter operation only to the lowpass filtered speech was chosen because only integer values of the

delay were used to adjust the filter to the speech.

The CELP coder determines the delay during the adaptive codebook search up to a fraction of a sample. As the harmonic filter is adjusted only by integer values of the delay the filter will be off at higher frequencies of the speech and introduce distortion. To avoid this distortion the filter is applied only to the lowpass part of the speech. The modified speech is reconstructed by adding the low and the highpass contributions. The following paragraphs describe the individual blocks in detail.

A lowpass filter of the form (4.1) separates the lowpass part of the speech.

$$H_{LP}(z) = \frac{1}{2}(1 + z^{-1}) \quad (4.1)$$

Subtracting the lowpass signal from the input speech yields the highpass signal shown in (4.2).

$$H_{HP}(z) = 1 - H_{LP}(z) \quad (4.2)$$

To improve the harmonic structure of the lowpass filtered speech signal a comb filter is used.

$$P(z) = \frac{1}{1 + b\beta}(1 + b\beta z^{-D}) \quad (4.3)$$

In (4.3) D is the delay of the previous subframe, β is a constant, determined empirically to be 0.09. The term $1/(1 + b\beta)$ normalizes the filter and keeps the overall gain at one.

The value b is an adaptive coefficient which is calculated new for every subframe.

This is achieved by a two step calculation. First, a value b_{calc} is determined by finding the cross correlation between the speech signal $s(n)$ of the present subframe and that of the subframe past in time by D as shown in (4.4). Here again D is the delay value of the previous subframe. The value b_{calc} is limited to the range $[-1 \ 1]$. Values above or below that are set to the respective maximum value of that range.

$$b_{calc} = \frac{\sum_{n=0}^{59} s(n)s(n-D)}{\sum_{n=0}^{59} s(n-D)s(n-D)}, \quad 20 \leq D \leq 147 \quad (4.4)$$

In the second step, b is calculated by considering the value of the previous subframe. The b value for the present subframe never exceeds +/- 40% of its previous value. The calculation is shown in (4.5). This limit was again determined heuristically for the FS-1016 coder. By properly smoothing the b value, a rapid on and off switching of the harmonic filter is avoided. Without the smoothing, artifacts can be introduced in the synthesized speech.

$$b = \begin{cases} b_{calc}, & 1.4b_{previous} \leq b_{calc} \leq \frac{b_{previous}}{1.4} \\ 1.4b_{previous}, & b_{calc} > 1.4b_{previous} \\ \frac{b_{previous}}{1.4}, & b_{calc} < \frac{b_{previous}}{1.4} \end{cases} \quad (4.5)$$

When the subframe is an unvoiced area, b_{calc} is a very small value because the delay

value is any value and the cross correlation will be small. Thus the harmonic filter effect is automatically decreased during unvoiced areas. In a voiced area the cross correlation yields a greater value and the filter has an effect on the speech.

Using the delay value of the previous subframe has an interesting effect. In an unvoiced area the filter is switched off. When a transition occurs and the delay value derived from the adaptive codebook represents the harmonic structure in the speech, the harmonic filter will be invoked and amplify the harmonic structure corresponding to the delay value. Even though this effect occurs actually one subframe later than the actual voiced part in the speech occurred, overall it has a smoothing effect on the delay values determined by the adaptive codebook.

After the improvement of the harmonic structure, the speech is constructed by adding the modified lowpass and the unchanged highpass part of the speech. This harmonically enhanced speech is then used as reference in the Analysis-by-Synthesis loop of the coder.

Results

In December 1997 the FS-1016 CELP coder fitted with the harmonic filter underwent Diagnostic Acceptability Measure (DAM) [KLEIJ95 pp 477-478] and Diagnostic Rhyme Test (DRT) [KLEIJ95 p 481]. The results are given in Table 1 and 2.

With a DAM test the speech quality is evaluated in terms of sixteen different scales, which range each from 0 to 100 points. These scales belong to one of the three categories signal quality, background quality, and overall quality. A weighted average of all 16 scales

forms the final score, again in the range from 0 to 100, that characterizes the speech coder under test. The test is performed by specially trained listeners. This is not the case for the DRT test which is performed by naive listeners. Its results detail the intelligibility of the coder under test. The listener is presented one out of two words that only differ in the leading consonant. In addition, the difference is restricted to one out of eight distinctive phonetic features. A score is calculated for each feature, stating how many words were responded to correctly as a percentage of one feature. The overall score is again the weighed average of the feature scores.

For each of the DAM and the DRT tests a set of six speakers with and without background noise are tested. Office environment means that on the test files, in addition to the speaker, background noise found in an office environment can be heard. Table 1 shows the test results where the speaker was in a quiet environment. In Table 2 the results from the office environment are given.

The first row gives the score of the coder fitted with the harmonic filter. In the second row are the scores for the standard coder, and the third row gives the difference in score due to the harmonic filter. A positive value indicates that the coder with harmonic

Table 1. Quiet environment DRT and DAM evaluation of the harmonic filter

	DRT		DAM	
	<i>Score</i>	<i>Standard Error</i>	<i>Score</i>	<i>Standard Error</i>
<i>Harmonic</i>	91.9	0.54	70.7	1.1
<i>CELP 3.3</i>	91.1	0.77	69.8	0.6
<i>Difference</i>	0.8		0.9	

filter achieved a better score over the standard coder. The score fluctuates about the standard error around the score value.

In the quiet environment the coder with harmonic filter improved by 0.8 point in the DRT score over the standard coder. The improvement of the harmonic filter is greater than the standard error, which means that it is significant.

For the DAM score the harmonic filter achieved an improvement of 0.9 points, but here the standard error of the harmonic score is so big that the improvement can result from the scattering of the data.

In the following Table 2 the results for the office environment are listed. The improvements of the coder fitted with the harmonic filter were here more dramatic as in the quiet environment. Both, the DRT and the DAM score improved over the standard error, making obvious the general improvement achieved by incorporating the harmonic filter in the CELP coder. We can conclude that the harmonic filter provides more improvement when low level background noise is present. This is possibly a result of the CELP codebook search being easily degraded in the presence of noise.

Table 2. Office environment DRT and DAM evaluation of the harmonic filter

	DRT		DAM	
	<i>Score</i>	<i>Standard Error</i>	<i>Score</i>	<i>Standard Error</i>
<i>Harmonic</i>	90.5	0.56	59.1	0.9
<i>CELP 3.3</i>	89.0	0.82	57.9	1.2
<i>Difference</i>	1.5		1.2	

These improvements can be shown in some plots which are given below. In Figure 13 the word “figure” spoken in a quiet environment is plotted. The top is the original speech, followed by the CELP coded speech and finally the coded speech with CELP improved by the harmonic filter.

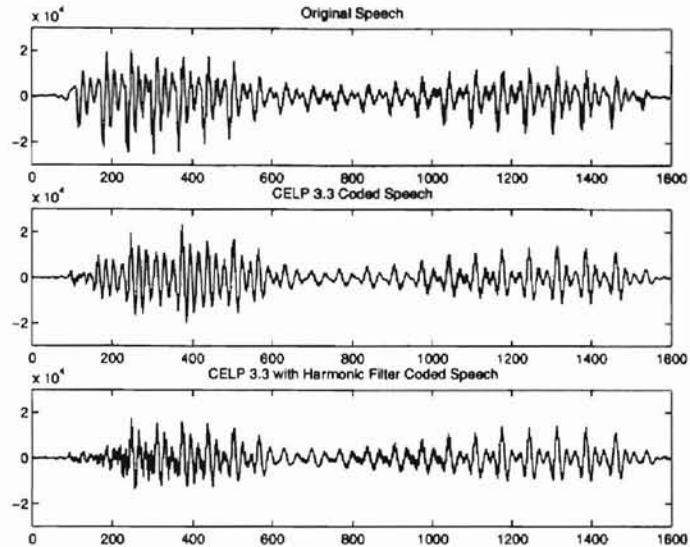


Figure 13. a) Plot of the word “figure” without background noise; b) synthesized by CELP 3.3; c) synthesized by the coder with the harmonic filter

Compared with original delay values in the top plot, the CELP coder picks the wrong delay value up to sample 600. The third plot shows that the harmonic filter improves the way the adaptive codebook search is performed.

In the bottom plot of the word “figure” the onset of the word seems to contain more noise than the original one which can not be found in the synthesized speech of the standard coder. This kind of noise was only found at this specific place of the speech file, and the previous test scores show that it doesn’t seem to degrade the speech significantly.

Figure 14 compares the determined delay values. The top plot shows again the

original speech signal. Each plotted value in the middle and the bottom plot counts for a delay value of one subframe. The third plot shows that with the harmonic filter in the CELP coder the delay values change much smoother than they do with the standard coder, plotted

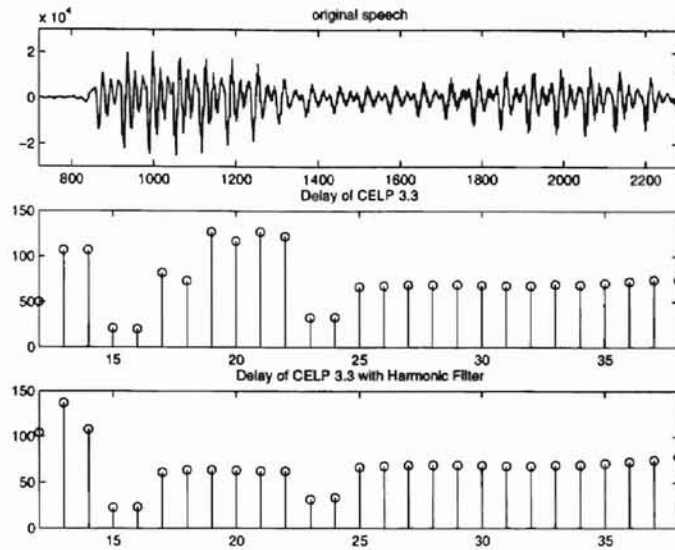


Figure 14. a) Plot of the word “figure”; b) delay values per subframe of the word “figure” determined by CELP 3.3; c) delay values per subframe determined by the coder modified with the harmonic filter

in the middle. Estimating the delay in the original speech shows about three maximum values in the range from sample 1000 to 1200, yielding a delay of $200/3 = 66.67$ samples. From subframe 16 to 22 the standard coder picks a value approximately twice the estimate. With the harmonic filter the right delay value is picked as shown in the bottom plot.

The next Figure 15 compares the gain associated with the delay values. From subframe 14 to 20 the sign of the gain changes much less for the coder with harmonic filter, compared with the standard coder.

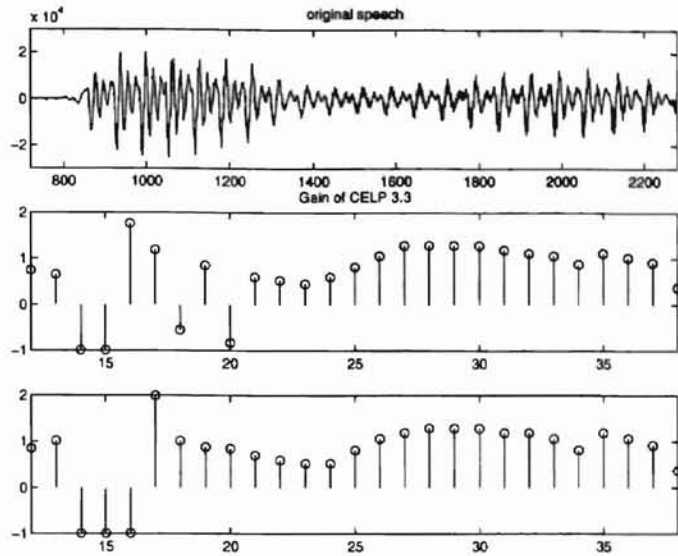


Figure 15. a) Plot of the word “figure”; b) gain values per subframe determined by CELP 3.3; c) gain values determined by the coder with harmonic filter

Another listening test often performed to categorize the quality of a speech coding system is the absolute category rating test [KLEIJ95 p 482]. It is performed with naive listeners who are presented with samples of processed speech material. They are asked to give a rating using a 5 point scale as given in Table 3. The average of all votes obtained for a particular system represents the mean opinion score (MOS).

Table 3. MOS scores

<i>Description</i>	<i>Rating</i>
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Along with the results of the DAM test comes a prediction of the skilled listener who performed the test, what the MOS score would be when a naive listener would perform an absolute category rating. In Table 4 these predicted MOS scores are given.

Table 4. Predicted MOS scores in quiet and office environment, based on the DAM results

	Predicted MOS in Quiet		Predicted MOS in Office	
	<i>Score</i>	<i>Standard Error</i>	<i>Score</i>	<i>Standard Error</i>
<i>Harmonic</i>	3.93	0.11	3.03	0.08
<i>CELP 3.3</i>	3.84	0.09	2.95	0.11
<i>Difference</i>	0.09		0.08	

For both environments the coder fitted with the harmonic filter improved in the score, but again the standard error is in the range of the change.

CHAPTER V

IMPROVED STOCHASTIC CODEBOOK SEARCH

Calculating the Masking Threshold

To benefit from the masking abilities of the ear in order to improve the quality of speech coders, we must identify the masked and unmasked regions of the speech. This is done by deploying an auditory masking model that is based on psycho-acoustic test results from Chapter II. With this model a masking threshold is calculated for the signal. Spectral values above the threshold are declared audible and spectral values below the threshold are declared to be masked. Carrying out the noise shaping in terms of masked and unmasked regions of the speech has been shown to increase the quality of speech coders [ZWICK91] [SEN94].

In Chapter II, Zwicker's model of excitation was introduced. Here again a reminder that the term "excitation" comes from the field of auditory physiology and is not to be mistaken for the excitation of the speech coder that is fed into the LP filter. Terhardt [TERH79] built on Zwicker's work and published a model that incorporated the asymmetry of masking, investigated by Hellman [HEL72]. Sen [SEN94] carried on these results and derived a masking model suitable for a speech coder.

In the pages that follow, the calculation of this masking model is explained. Next, a test is performed to demonstrate the proper operation of the model. We then compare the noise masking abilities of the conventional weighting filter in the CELP coder with noise masking that deploys a masking model. Based on these results an evaluation of the two

methods named in Chapter III is made.

A first step in calculating a masking threshold is to represent the speech signal into the frequency domain. Here it is important to consider the spectral resolution necessary to calculate the masking threshold. For the excitation calculation [ZWICK90] it is important to have a frequency resolution at least equal to a critical band. Another factor is to avoid leakage over the critical bands [SEN94]. With a sampling rate of 8 kHz used for the CELP coder, a 256 sample DFT provides a resolution of 31.25 Hz. At low frequencies this gives three samples per critical band, which is a sufficient resolution [ZWICK90]. To prevent leakage Sen [SEN94] suggests the use of a Hamming or a Hanning window. The bandwidth for both [OPP89] is $8\pi/P$, where P is the window length. With the 256 point DFT and a sample rate of 8 kHz this leads to a bandwidth of 125 Hz. For the Hanning window the peak side lobe is -31 dB and for the Hamming -41 dB below the maximum of the window's main lobe [OPP89].

The masking threshold is calculated by first determining the excitation. With respect to the psycho-acoustic masking results described in Chapter II, the first step of calculating the excitation signal is by assuming that every spectral sample is a tonal component. A virtual excitation pattern is associated with every sample in the frequency spectrum.

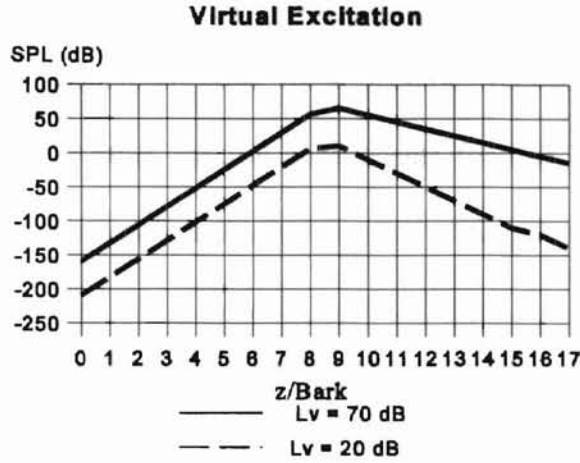


Figure 16. Virtual excitation pattern for a 1000 Hz (8.6 Bark) tone with level 70 dB (solid) and 20 dB *SPL* (dashed)

Figure 16 shows the virtual excitation pattern for a 1 kHz tone. The solid line represents the excitation pattern by a tone of 70 dB *SPL*, and the dashed line represents the pattern of a 20 dB *SPL* tone. The pattern is calculated by equation (5.1),

$$L_{E_v}(z) = L_v + s_v(z - z_v), \quad (5.1)$$

where L_v is the level of the tone v that creates the excitation, and L_{E_v} is the excitation level for the specific component due to component v . The upper and lower slopes s_v are given by (5.2).

$$s_v = \begin{cases} -24 - \frac{230}{f_v} + 0.2L_v \text{ dB/Bark} & z > z_v \\ 27 \text{ dB/Bark} & z \leq z_v \end{cases} \quad (5.2)$$

The lower slope is constant, whereas the upper slope depends on the level of the

tone.

To calculate the excitation value for a specific frequency value f , the virtual excitation values of all the other spectral components for that specific component are summed. Converting them into an amplitude value and adding them according to equation (5.3) accomplishes this.

$$L_E(f) = 10 \log \sum_{v=1, v \neq f}^N 10^{\frac{L_{E_v}(f)}{20}} \quad (5.3)$$

Summing the amplitude values is only an approximation, as there is no precise equation to calculate the masking threshold due to several masker signals [SEN94]. This calculation is the first of three steps to determine the excitation of a speech signal. The second step is based on Hellman's results concerning the asymmetry of masking. A noise signal masks better than a tone even when both have equal intensity. In the following way this is incorporated into the model and automatically accounts for whether the signal is more noise-like or more tonal like. In the case of speech this would be the distinction between unvoiced and voiced speech.

To calculate the noise masking effect on a spectral component f , the critical bandwidth of that component is determined. In the next step the intensity within that critical bandwidth is summed, excluding the three neighbors of the 256 point DFT above and below component f as well as component f itself. This calculation is performed for every frequency value in the spectrum and is denoted by equation (5.4).

$$N(f) = 10 \log \sum_{\substack{n = z(f) - 0.5 \\ n = f \pm (1, 2, 3)}}^{z(f) + 0.5} I_n \quad (5.4)$$

Figure 17 displays this calculation. For component f the critical bandwidth is Δf . The three neighbors are excluded as well as component f .

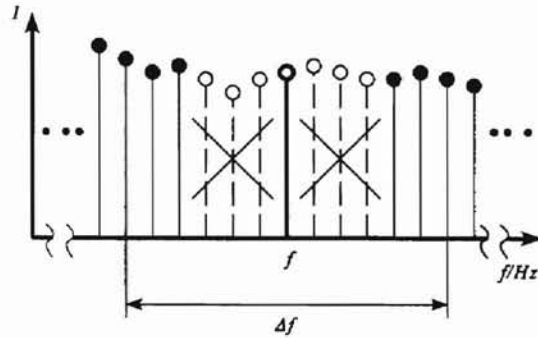


Figure 17. Noise masking contribution for component f ; intensity within the critical bandwidth Δf around f is summed, excluding component f and its three DFT coefficient neighbors above and below

When the signal is more noise-like, the spectrum is flat and this calculation contributes more to the excitation than for signals that have a harmonic structure. Due to the harmonic structure the spectrum changes more, and the above calculation does not add as much contribution to the overall excitation as for a noise-like signal.

The last contribution to the overall excitation is the threshold in quiet. It is approximated with equation (5.5) and added to the two excitation patterns from above.

$$L_{TH}(f) = 3.64 \left(\frac{f}{1000} \right)^{-0.8} - 6.5 e^{-0.6 \left(\frac{f}{1000} - 3.3 \right)^2} + 10^{-3} \left(\frac{f}{1000} \right)^4 \quad (5.5)$$

The total excitation calculation is described by equation (5.6).

$$L_E(f) = 10 \log \left(\sum_{v=1, v \neq f}^N 10^{\frac{L_{E_v}(f)}{20}} + \sum_{\substack{n=z(f)+0.5 \\ n=z(f)-0.5 \\ n \neq f \pm (1,2,3)}} I_n + 10^{\frac{L_{M(f)}}{10}} \right) \quad (5.6)$$

As mentioned in Chapter II, the actual masking threshold is about 2 to 6 dB below this excitation value, depending on the frequency. However for testing this masking model, the excitation pattern is used as the masking threshold.

A first test of the masking model will be described in the next section.

Time Domain Aliasing Cancellation (TDAC)

To show the effectiveness of an auditory masking model, the model is tested with unprocessed speech. The excitation is calculated according to the just described model and used as the masking threshold. The result of this test will show how applicable the described masking model is and whether it is possible to use the calculated excitation as the masking threshold.

The speech is first transformed into the frequency domain. After determining the masked areas of the speech, they are attenuated by the amount of masking threshold at the specific frequency. Then the speech is transformed back into the time domain and, by informal listening tests, the accuracy of the masking model is determined.

For the transform, a method called time domain aliasing cancellation (TDAC) [PRIN86] is used. The core of the method is a critically sampled single sideband filter bank. Figure 18 shows the structure of the analyzer.

Due to the critical sampling in the frequency domain, aliasing is introduced which is then canceled out in the time domain by the synthesizer. The method allows the signal to be perfectly reconstructed. The following discussion of this transform will focus on utilizing the DFT in calculating the analyzer and the synthesizer of the filter bank. For further details of TDAC the reader is referred to Princen's and Bradley's paper [PRIN86].

Sen [SEN94] derived from the structure in Figure 18 the following equation (5.7) for the analyzer,

$$\begin{aligned}
 X_k(m) = & (-1)^{mk} \cos\left(\frac{m\pi}{2}\right) \sum_{r=0}^{P-1} x_m(r) h(P-1-r) \cos\left(\frac{2\pi k}{K}(r+n_0)\right) \\
 & + (-1)^{mk} \sin\left(\frac{m\pi}{2}\right) \sum_{r=0}^{P-1} x_m(r) h(P-1-r) \sin\left(\frac{2\pi k}{K}(r+n_0)\right),
 \end{aligned} \tag{5.7}$$

where P is the length of the time domain window $h(n)$ which will be restricted to K , the number of frequency channels. $M=K/2$ is the progress of the window in the time domain. This means the window length in the time domain is equal to the number of frequency channels, and consecutive windows are overlapped by 50%. $X_k(m)$ is the k^{th} subband signal of the window number m . $r = n - mM$ is the window taken out of the time axis of the overall signal and shifted back to the origin. To achieve perfect reconstruction it follows [PRIN86] that $2n_0$ has to be $M+1$.

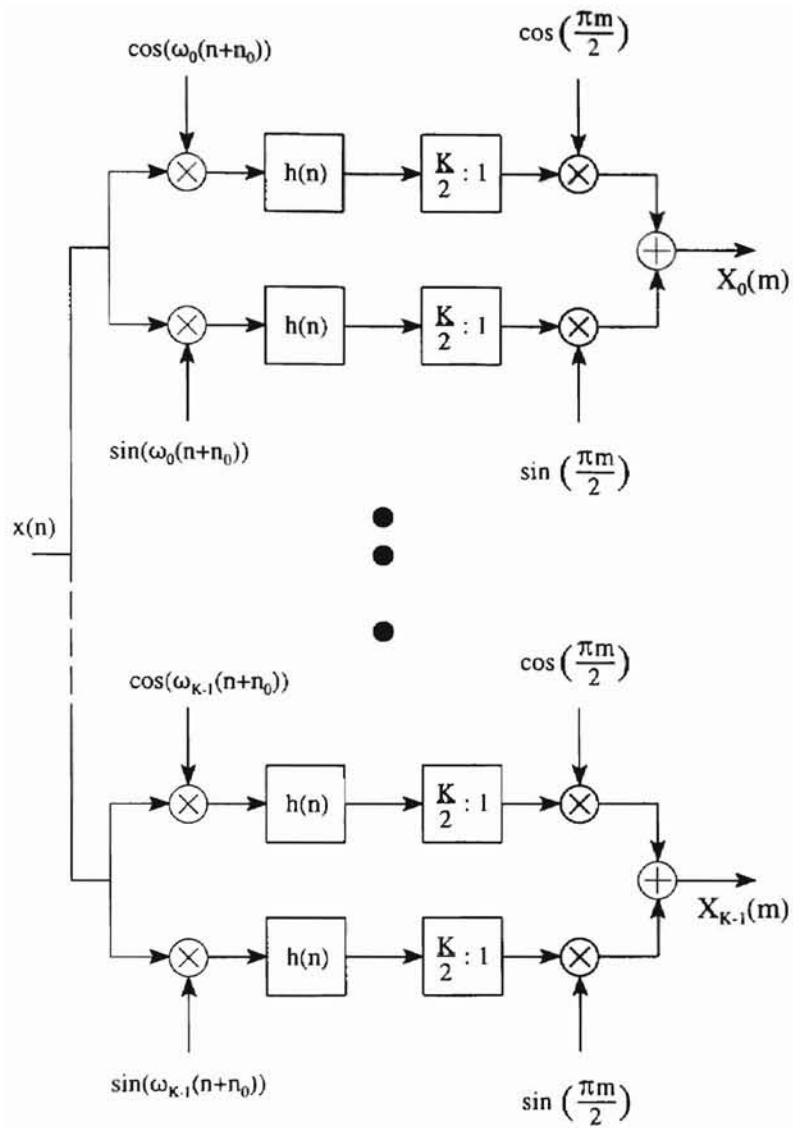


Figure 18. SSB Filter Bank Analyzer

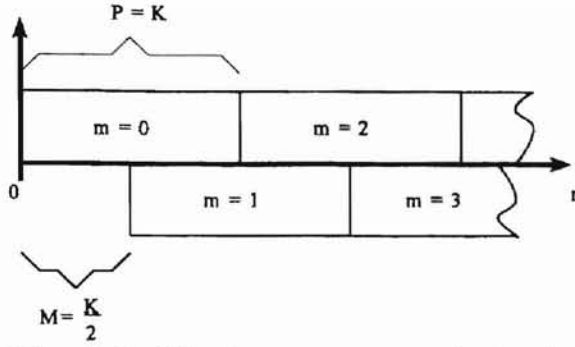


Figure 19. Window arrangement in the time domain for the TDAC

Figure 19 displays the structure of the overlapped windows and the variables that go along with them.

To take advantage of the DFT, two terms $X_C(k)$ and $X_S(k)$ are defined according to equation (5.8).

$$\begin{aligned}
 X_C(k) &= \sum_{r=0}^{P-1} x_m(r) h(P-1-r) \cos\left(\frac{2\pi kr}{K} + \frac{2\pi kn_0}{K}\right) \\
 X_S(k) &= \sum_{r=0}^{P-1} x_m(r) h(P-1-r) \sin\left(\frac{2\pi kr}{K} + \frac{2\pi kn_0}{K}\right)
 \end{aligned} \tag{5.8}$$

Multiplying (5.8b) by j and subtracting it from (5.8a) leads to equation (5.9) which introduces the DFT.

$$\begin{aligned}
 X_C(k) - jX_S(k) &= \sum_{r=0}^{P-1} x_m(r) h(P-1-r) e^{-j\left(\frac{2\pi kr}{K} + \frac{2\pi kn_0}{K}\right)} \\
 &= e^{-j\frac{2\pi kn_0}{K}} \text{DFT}\{x_m(r) h(P-1-r)\}
 \end{aligned} \tag{5.9}$$

Let's go back to equation (5.7). One of the two terms $\sin(m\pi/2)$ or $\cos(m\pi/2)$ cancels out, depending on whether m is even or odd. By taking this into account and plugging (5.9) into equation (5.7), gives the two equations (5.10) for even m and (5.11) for odd m .

$$\begin{aligned}
X_k(m_{\text{even}}) &= \sum_{r=0}^{P-1} x_m(r) h(P-1-r) \cos\left(\frac{2\pi kr}{K} + \frac{2\pi kn_0}{K}\right) \\
&= \text{Re}\{X_C(k) - jX_S(k)\} \\
&= \text{Re}\left\{e^{-j\frac{2\pi kn_0}{K}} \text{DFT}\{x_m(r) h(P-1-r)\}\right\}
\end{aligned} \tag{5.10}$$

For even m the term $(-1)^{mk}$ is always one and the first line in (5.10) is just the real part of equation (5.9).

$$\begin{aligned}
X_k(m_{\text{odd}}) &= (-1)^{mk} \sum_{r=0}^{P-1} x_m(r) h(P-1-r) \sin\left(\frac{2\pi kr}{K} + \frac{2\pi kn_0}{K}\right) \\
&= (-1)^{mk} (-1) \text{Im}\{X_C(k) - jX_S(k)\} \\
&= (-1)^{mk} (-1) \text{Im}\left\{e^{-j\frac{2\pi kn_0}{K}} \text{DFT}\{x_m(r) h(P-1-r)\}\right\}
\end{aligned} \tag{5.11}$$

When m is odd the value of $(-1)^{mk}$ depends on k whether it is one or minus one, so the term has to be considered here. The first line of (5.11) is the imaginary part of equation (5.9) multiplied by the $(-1)^{mk}$ term.

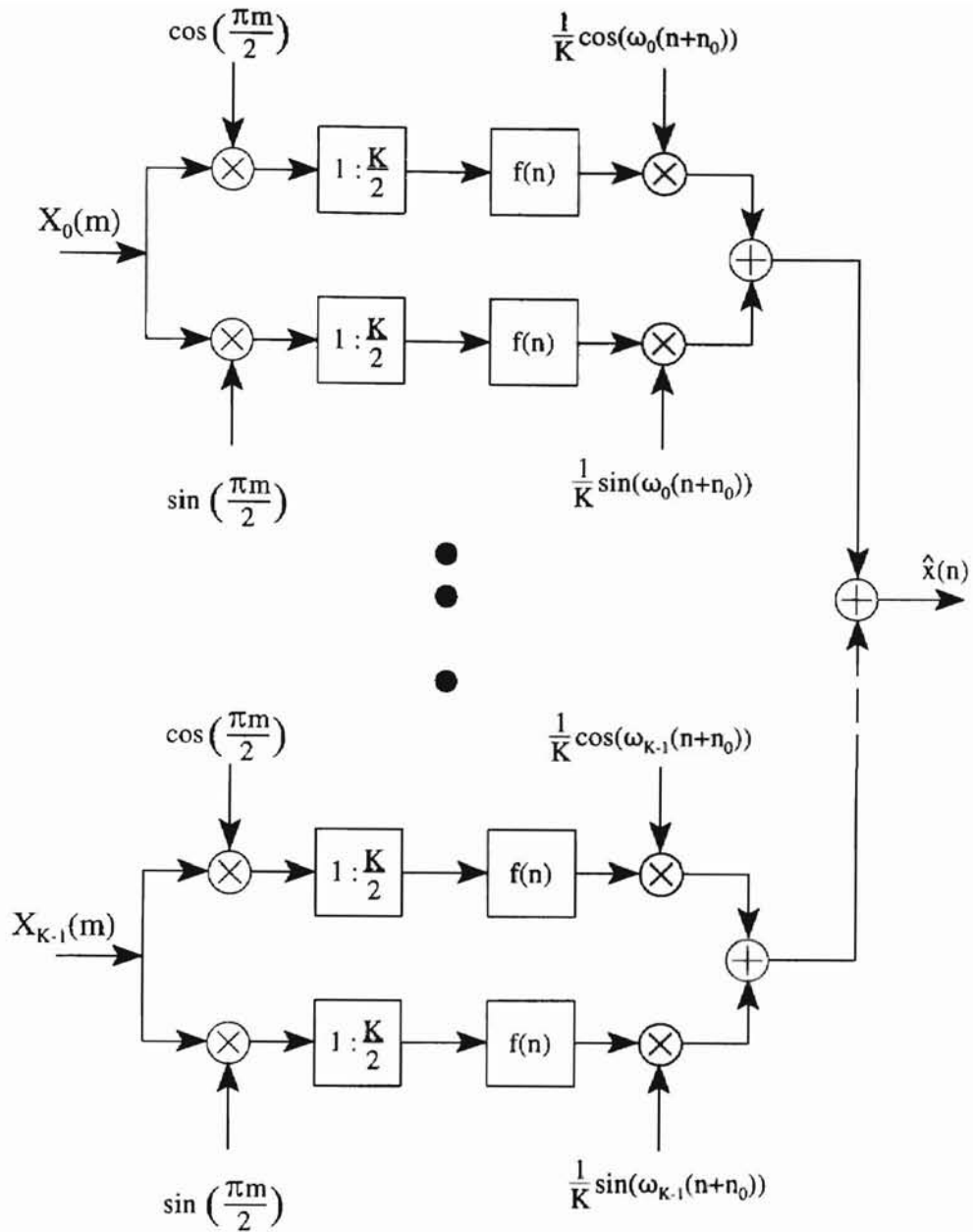


Figure 20. SSB Filter Bank Synthesizer

For the reconstruction of the signal in the synthesizer a similar procedure is followed by using the DFT. Figure 20 shows the synthesizer of the filter bank.

The signal $y_m(r)$ is the K time samples for the time frame m of the overall

synthesized signal, defined according to (5.12) [SEN94].

$$\begin{aligned}
 y_m(r) = & \cos\left(\frac{m\pi}{2}\right) \frac{1}{K} \sum_{k=0}^{K-1} (-1)^{mk} X_k(m) \cos\left(\frac{2\pi k}{K}(r+n_0)\right) \\
 & + \sin\left(\frac{m\pi}{2}\right) \frac{1}{K} \sum_{k=0}^{K-1} (-1)^{mk} X_k(m) \sin\left(\frac{2\pi k}{K}(r+n_0)\right)
 \end{aligned} \tag{5.12}$$

Let us first consider how to use the DFT to calculate this signal and then how to put these individual frames m together to achieve the overall time domain signal $\hat{x}(n)$. Again two variables, $x_C(r)$ and $x_S(r)$, are defined according to equation (5.13).

$$\begin{aligned}
 x_C(r) &= \sum_{k=0}^{K-1} (-1)^{mk} X_k(m) \cos\frac{2\pi k}{K}(r+n_0) \\
 x_S(r) &= \sum_{k=0}^{K-1} (-1)^{mk} X_k(m) \sin\frac{2\pi k}{K}(r+n_0)
 \end{aligned} \tag{5.13}$$

Remember that $X_k(m)$ is a real signal. Thus, by subtracting j times $x_S(r)$ from $x_C(r)$ we get the following equation (5.14) which allows the use of the DFT.

$$\begin{aligned}
 x_C(r) - jx_S(r) &= \sum_{k=0}^{K-1} (-1)^{mk} X_k(m) e^{-j\frac{2\pi k}{K}(r+n_0)} \\
 &= \text{DFT}\left\{(-1)^{mk} X_k(m) e^{-j\frac{2\pi kn_0}{K}}\right\}
 \end{aligned} \tag{5.14}$$

When we consider that in equation (5.13) one of the two terms $\sin(m\pi/2)$ or $\cos(m\pi/2)$ becomes zero depending on whether m is even or odd and we substitute equation (5.13) into (5.14) we get for even m the equation (5.15) and for odd m (5.16).

$$\begin{aligned}
y_{m_{\text{even}}}(r) &= \frac{1}{K} \sum_{k=0}^{K-1} X_k(m) \cos\left(\frac{2\pi k}{K}(r+n_0)\right) \\
&= \text{Re}\{x_c(r) - jx_s(r)\} \\
&= \text{Re}\left\{ \text{DFT}\left\{ X_k(m) e^{-j\frac{2\pi kn_0}{K}} \right\} \right\}
\end{aligned} \tag{5.15}$$

Here again, as in the analyzer, the $(-1)^{mk}$ term is one for even m and can be neglected in (5.15). For odd m it depends on the channel k and it has to be incorporated in (5.16).

$$\begin{aligned}
y_{m_{\text{odd}}}(r) &= \frac{1}{K} \sum_{k=0}^{K-1} X_k(m) \sin\left(\frac{2\pi k}{K}(r+n_0)\right) \\
&= (-1) \text{Im}\{x_c(r) - jx_s(r)\} \\
&= (-1) \text{Im}\left\{ \text{DFT}\left\{ (-1)^{mk} X_k e^{-j\frac{2\pi kn_0}{K}} \right\} \right\}
\end{aligned} \tag{5.16}$$

To obtain the overall synthesized signal $\hat{x}(n)$, the time frames y_m have to be overlapped and added. For perfect reconstruction the windows $h(n)$ in the analyzer and $f(n)$ in the synthesizer have to satisfy the following condition (5.17) [PRIN86].

$$\begin{aligned}
f(n) &= h(n) \\
f^2(r+M) + f^2(r) &= 2 \quad r = 0 \dots M-1
\end{aligned} \tag{5.17}$$

Choosing a sinusoidal window leads to a constant overlap demanded by (5.17) but the value is "1" in contrast to "2". This is only an amplitude factor which can be corrected by multiplying the sinusoidal window function by $\sqrt{2}$. Equation (5.18) shows the window function.

$$h(n) = f(n) = \sin\left(\frac{\pi}{K}n\right) \quad n = 0..K-1 \quad (5.18)$$

To test the perfect reconstruction of the transform, a speech signal is transformed into the frequency domain and back into the time domain. Figure 21a shows the two plots. The original speech is plotted dashed and the reconstructed speech is plotted over that with a dotted line.

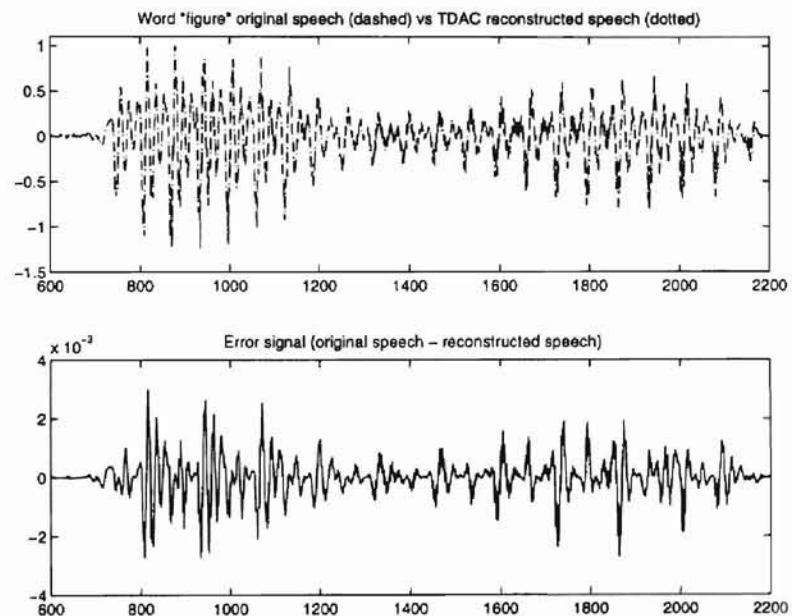


Figure 21. a) Plot of the word “figure”; dashed is the original signal and dotted the reconstructed after the TDAC transform; b) error between original and reconstructed speech

There is barely a difference noticeable. In Figure 21b the error between the two signals is plotted. The maximum error for that plot is approximately 0.3%, which is

neglectable small.

Next, the filter bank is used to test the masking model. The speech signal is transformed into the frequency domain, and the masking threshold is calculated. Figure 22a shows the frequency plot of a 256 sample wide window of the speech signal shown in Figure 21.

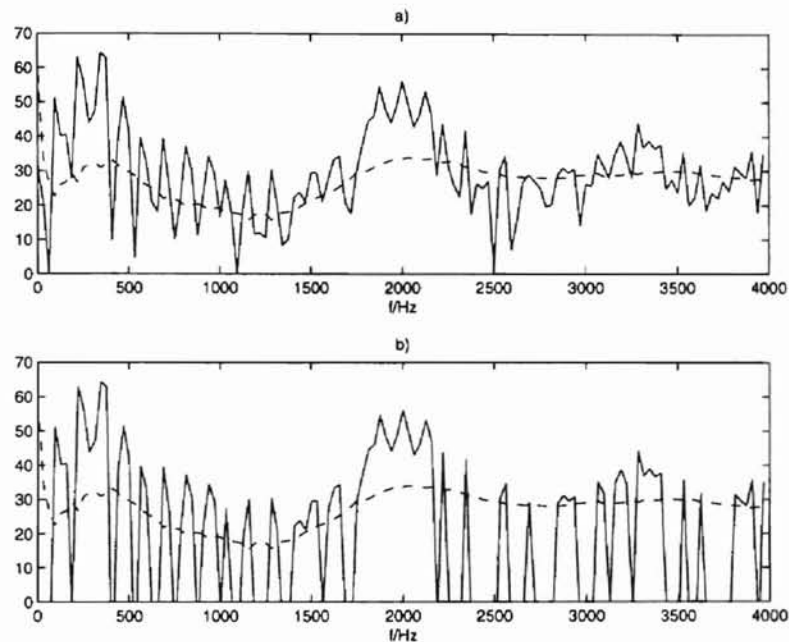


Figure 22. a) Frequency spectrum (solid) of 256 samples from the word “figure” and masking threshold (dashed); b) frequency samples below the masking threshold (dashed) are attenuated by the amount of the masking threshold

The dashed line shows the masking threshold for this frequency spectrum. Next, all the frequency samples that are below the masking threshold are attenuated by that specific masking threshold value. This is shown in Figure 22b. The speech is then transformed back into the time domain.

A 22 second long speech file with female and male speakers was processed according to the just mentioned procedure. In a further step the masking threshold was raised by 5, 10, and 15 dB over the calculated threshold. This was done to determine whether the calculated threshold is already the upper limit or whether it is lower than the actual threshold of the ear. The attenuation procedure was repeated with each raised value. Table 5 shows the redundancy and voice quality due to the raised discard threshold.

Table 5. Redundancy and voice quality due to raised discard threshold

<i>Offset (dB)</i>	<i>Percentage of Samples below Discard Threshold</i>			<i>Subjective Quality</i>
	<i>Min</i>	<i>Average</i>	<i>Max</i>	
0	4	58	96	Transparent
5	5	72	98	Slight Tonal Artifacts
10	11	84	99	Strong Tonal Artifacts
15	21	91	99	Artifacts + Degraded Speech

Remember that for this test the excitation was used as the masking threshold. The actual masking threshold is 2 to 6 dB below that, depending on the frequency. As one can see from Table 5, with the excitation used as the masking threshold on average there were 58% of the frequency samples below the threshold and the speech quality was transparent. The minimum number of samples that was discarded during one frequency transformation was 4% and the maximum was 96%. The value for the maximum was determined without considering the silent parts of the speech where usually 100% of the samples were discarded. With raising the masking threshold by 5 dB, slight tonal artifacts occurred. These

artifacts were separate to the speech and appeared to be a background signal. After raising the threshold by 10 dB, the artifacts became stronger but were still separated from the speech. With a 15 dB raised threshold, the speech started to become degraded.

What this test demonstrates is that the masking model is applicable the way it was calculated and that it is possible to use the calculated excitation as a masking threshold. This simplification of the masking threshold calculation will be used in the following tests. Another result is that on average over 50% of the frequency samples of a speech signal are perceptually unimportant. This test does not show how noise shaping based on an auditory masking model would perform in comparison to the conventional noise shaping used in the CELP coder. This is what we will investigate next.

Noise Shaping with Masking Threshold versus LPC Spectrum

We want to compare the effectiveness of shaping the noise according to the masking threshold versus shaping the noise according to the LPC spectrum, like it is done in the CELP coder. Therefore, noise signals are created with a defined amplitude spectrum. This is done by equation (5.19).

$$\begin{aligned} X(k) &= A(k) \cos(\theta) \\ &+ jA(k) \sin(\theta) \quad \theta = -\pi \dots \pi \end{aligned} \quad (5.19)$$

The noise signal is formed based on the given amplitude spectrum $A(k)$. Theta in equation (5.19) is a uniform distributed random variable with zero mean in the range from

$-\pi$ to π . To transform the frequency spectrum into a real time domain signal the following symmetry properties must hold [OPP89].

$$\begin{aligned} \text{Re}\{X(k)\} &= \text{Re}\{X(N-k)\} \\ \text{Im}\{X(k)\} &= -\text{Im}\{X(N-k)\} \end{aligned} \quad (5.20)$$

Three noise signals are created, differing in their amplitude spectrum. The created noise signals are added to the speech signal and the audibility of the noise is tested. The processing is shown in Figure 23.

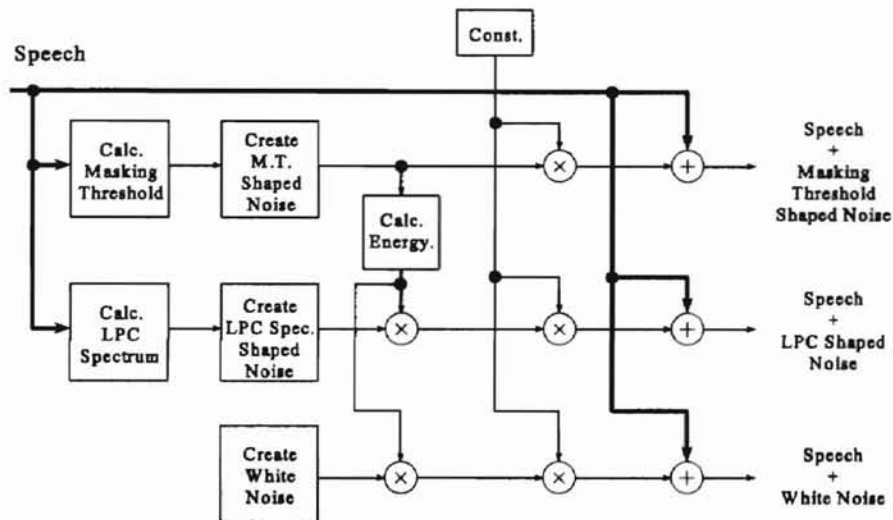


Figure 23. Adding noise signals with different spectral shapes to a speech signal

A speech signal's characteristics can be considered stationary over an interval of approximately 20 - 30 msec [KLEIJ95 p 11]. Thus the processing is done in time frames of 256 samples or 32 ms at a sample frequency of 8 kHz. The amplitude spectrum of two of the noise signals are adapted to the stationary characteristics of the speech signal. The third

is a white noise signal which is used to give a comparison of how the amount of added noise energy sounds without any shaping.

One noise signal's amplitude spectrum is shaped according to the calculated masking threshold of the processed frame. This is shown in the top part of Figure 23. After calculating the masking threshold, the noise is shaped according to the threshold. The energy of this created noise is calculated and the energies of the other two noise signals are normalized to that. The second noise signal's amplitude spectrum is shaped according to the LPC spectrum of the speech. A LP analysis is performed on the speech frame similar to the one performed in the CELP coder. From the LP coefficients (LPC) the spectrum is calculated and a noise signal is created based on that LPC spectrum.

After the energies are equalized to the energy of the masking threshold, their amplitudes are multiplied by a constant factor. Each noise signal is then added to the speech signal, creating three speech signals with the same energy of noise, but different shaped amplitude spectra.

The following Table 6 comprises the results of the experiment, which was performed with a 4 seconds duration speech file. The gain of the noise was raised from one to five times the amount that the masking threshold can bear. In the second column the total energy of the added noise is given, normalized to the 32,500 sample data of the speech file. To get an idea in which areas the noise energy of the performed test is compared to the speech coder, the same speech file was coded by the CELP coder and the energy of the error signal determined. Normalized to the number of samples, the energy is $2.96 (10^6)$. This shows that the noise added to the speech is below the actual noise added by the coder.

The first observation is again like the previous test with TDAC, that with a gain of one the masking threshold masks all the added noise and the speech sounds transparent to the original. However, the LPC shaped noise also results in transparent speech. This might be based on the fact that due to its nature the LPC shaped noise is correlated with the speech spectrum and so low energy noise is hidden by the speech.

Raising the amplitude of the noise signal results in more audible noise for the masking threshold shaped noise. The uniform noise compared to that is much more audible. This is due to the fact that with the masking threshold shaped noise up to that limit the noise

Table 6. Speech quality depending on spectral shape and amount of the added noise

Gain	Masking Energy/ Sample	Speech Quality of Noise Shaped According ..		
		Masking Threshold	LPC Spectrum	White Noise
1	2.26 (10^4)	Transparent	Transparent	Slight Audible Noise
2	9.03 (10^4)	Barely Audible Noise	Barely Muffled	Audible Noise
3	2.03 (10^5)	Slight Audible Noise	Slight Muffled	Strong Audible Noise
4	3.61 (10^5)	Audible Noise	Muffled	Very Strong Audible Noise
5	5.88 (10^5)	Strong Audible Noise	Very Muffled and Degraded Speech	Noise Partly Extinguishes Speech

is shaped optimally according to the hearing system. Noise that exceeds this threshold is spread uniformly above the masking threshold. In the sense of the hearing process this noise

exceeds the threshold in a minimum audible way. However, the noise shaped according to the LPC spectrum changes the speech in a different way. With increasing noise added, the noise itself becomes distorted and unlike white noise. Rather, the speech sounds more muffled and, with increased noise energy, distorted. A reason for that is the way the noise is shaped. By shaping it according to the LPC spectrum, the amplitude spectrum of the noise is correlated with that of the speech. This becomes obvious when listening to only the noise that was created by shaping it according to the LPC spectrum of the speech. It contains the speech information. By adding this noise to the speech, both signals fuse perceptually [KLEIJ95 p 428] and sound like speech with a buzzy characteristic.

In Chapter III, two ways were referred to how an auditory masking model can be incorporated into a speech coder. One approach by Montagna et al. [MONT91] proposed to adjust the transfer function of the perceptual weighting filter to the inverse of the masking threshold function. This would cause the coder to add more noise in frequency areas where the masking threshold is higher and less noise in areas where the threshold is lower. This scheme is similar to the test described above. The noise is shaped according to the masking threshold. As seen from the results, this is practical when the noise energy is in the range of the masking threshold energy. When the noise exceeds the energy, it becomes audible.

The second approach named in Chapter III was introduced by Sen [SEN94]. Here the stochastic codebook search is performed in the frequency domain. The search procedure is modified in a way that the error signal is only minimized in the unmasked areas of the speech.

Let's take the stochastic codebook search for example. The codebook search is

performed by searching for the closest match in the 60 dimensional vector space between the 512 codewords and the test vector. Due to the limited amount of codewords there might be no one perfect match. There are two ways to increase a match. One is to add more codewords to the codebook and the other is to reduce the dimension over which the error minimization is performed.

In an earlier test it was shown that on average about 50% of the frequency samples of a speech signal are perceptually unimportant. When the stochastic codebook search is performed in the frequency domain the error minimization can be adjusted in a way that the dimension of the codewords in the frequency domain is reduced depending on the masked regions of the speech. This allows the unmasked regions to be modeled better with the available codewords. In equation (5.21) and (5.22) the calculation of the gain g_k and the match score m_k in the frequency domain is shown.

$$m_k = \frac{\left(\sum_{i=1}^N \mathit{mask}(i) Y_k(i) T(i) \right)^2}{\sum_{i=1}^N Y_k^2(i)} \quad (5.21)$$

Note that the two equations are closely related to the time domain versions, equation (2.1) and (2.2), given in Chapter II.

$$g_k = \frac{\sum_{i=1}^N \mathit{mask}(i) Y_k(i) T(i)}{\sum_{i=1}^N Y_k^2(i)} \quad (5.22)$$

What is added is the masking threshold function *mask* which denotes a N dimensional vector, having the value one for unmasked and the value zero for masked regions of the speech spectrum. The LP filtered code Y is now a N dimensional frequency vector and the target is represented by T . The index k represents the codeword for which the gain and the match are calculated.

There is one drawback that comes along with this method. When the error is minimized only in the unmasked regions of the speech there is no guarantee for the error signal in the masked regions to be below the masking threshold. For this reason Sen [SEN94] introduced what he calls a “*maskfilter*” into the synthesizer that attenuates the masked regions of the synthesized speech. This maskfilter makes the direct implementation of the approach not interoperable with the standard coder as it demands for this specific maskfilter to be in every synthesizer that will be connected with this improved analyzer.

Comparing the two methods we can say that the first method relies solely on shaping the noise. The second method tries to move as much noise as possible in the masked region and then attenuates that noise by the maskfilter. From this point it seems that the second method is more beneficial than the first one, and we will spend more effort in exploring ways to make Sen’s proposed method interoperable with the FS-1016 standard.

There are two approaches that seem to be worthwhile for a closer look. One is the postfilter in the CELP coder which is based on bandwidth expanded LP coefficients. It would be interesting to know what attenuation effect this filter has on masked regions of the speech signal. Another point is to modify the search for the gain and match score in a way that the noise in the masked regions is introduced in a controlled way. This will be left for

future research and explained further in Chapter VI.

Postfilter Effect on Masked Regions of the Speech

The synthesizer of the CELP coder includes a postfilter that improves the quality of the synthesized speech. In this section we investigate the attenuation that is introduced by that filter to masked regions of the speech signal. The result may give an indication whether it is suitable to combine an analyzer with the improved codebook search proposed by Sen with a standard coder fitted with the conventional postfilter.

The postfilter of the CELP coder is given by equation (5.23),

$$W(z) = \frac{H(z/\beta)}{H(z/\alpha)} (1 - 0.5\mu z^{-1}), \quad \beta = 0.5, \alpha = 0.8, \quad (5.23)$$

where $H(z)$ is the transfer function of the LP filter. The coefficients α and β cause the transfer function to change in a way that the filter amplifies the formant regions of the speech and attenuates the formant valleys. This concept is shown in Figure 24b. The solid plot shows the LPC spectrum and the dashed plot the combined LP and postfilter spectrum. To avoid amplitude distortions of the spectrum caused by the overall lowpass characteristic of the postfilter, a highpass filter is added and made adaptive to the first reflection coefficient μ . The reflection coefficient is based on the LP coefficients and represents the tilt of the spectrum. The more spectral tilt is introduced by the lowpass characteristic of the LP filter, the more the highpass compensates for that.

To calculate the attenuation introduced by the postfilter, the transfer functions of the

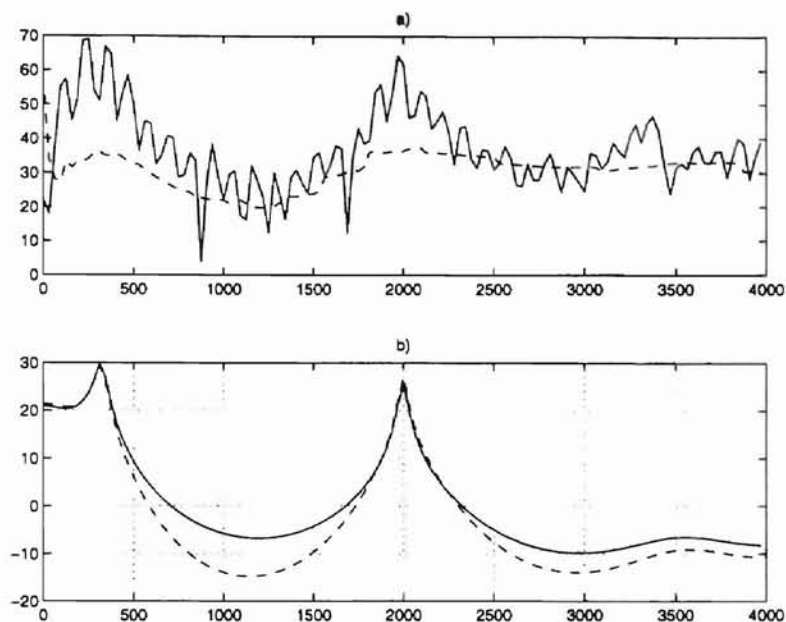


Figure 24. a) Speech spectrum (solid) vs masking threshold (dashed); b) LPC spectrum (solid) vs combined LPC and postfilter spectrum (dashed)

LP filter and of the postfilter are calculated. The energy of the postfilter transfer function is then equalized with the energy of the LP filter transfer function, shown in Figure 24b. Now the masking threshold is calculated, displayed as the dashed line in Figure 24a. The solid line is the speech spectrum. For the masked areas of the speech, the ratio of the LP filter and the postfilter is calculated and summed. Minimum and maximum values are stored and the average value is calculated from the summed ratios.

The test was performed on a 22 second long speech file containing male and female speakers. The results are shown in Table 7. On average the amplitude in the masked regions of the spectrum was attenuated approximately 2.5 dB. The maximum peak value for this 22 second of speech was around 11.9 dB below its original value. Unfortunately the postfilter did not only attenuate but also amplified spectral parts in the masked region up to 2.9 dB.

Table 7. Attenuation introduced by the postfilter in masked regions of a 22 second long speech file

<i>Minimal Value</i>	<i>Average Value</i>	<i>Maximum Value</i>
-2.9 dB	2.5 dB	11.9 dB

This comes from the fact that amplification and attenuation are determined in terms of formant peaks and formant valleys. The nature of the masking threshold is to raise slightly towards higher frequencies. However especially voiced LPC spectra have a spectral lowpass tilt. Around higher formant frequencies there might be spectral values that are amplified by the postfilter, but actually are below the masking threshold.

Results

In this chapter we explained the implementation of a published auditory masking model suitable for implementation in a speech coder. The next step was to demonstrate how effective the model is by utilizing a single sideband block filter method called time domain aliasing cancellation. The main benefit of this method is that it introduces an overall minimal distortion when the signal is transformed into the frequency domain and then back into the time domain.

After transforming the speech into the frequency domain, the masked and unmasked regions were identified by calculating the masking threshold. Masked regions of the speech were distorted by attenuating the respective frequency samples, and the speech was transformed back into the time domain. As a result on average over 50% of the frequency spectrum was declared masked, and the distorted speech sounded transparent.

Implementing one of the two methods described in Chapter III requires major changes to the coder. Therefore further tests were performed to make a decision on either method based on the gained results.

The first test was to compare the noise shaping abilities of the masking threshold with that of the conventional perceptual weighting filter used in the CELP coder. To compare that, noise signals were created with different spectral amplitude shapes. One noise signal's amplitude spectrum was shaped according to the masking threshold and a second one was shaped according to the LPC spectrum. To have a reference, a third white noise signal was created. All three noise signals were added to the speech creating three speech signals with the same overall noise energy, but with different shaped spectral noise. The result was that by adding noise energy that the masking threshold can bear both signals sounded transparent. In the third speech signal the white noise was audible. With raising the energy of the noise the masking threshold shaped noise became audible as noise, where as the LPC spectrum shaped noise distorted the speech in a way that it sounded muffled.

This result led to the conclusion that shaping the noise according to the masking threshold of the speech benefits in a major way when the energy of the noise is below the masking threshold. In low rate speech coding this is not the case and the introduced noise is much greater than the masking threshold. Therefore a combinational way of shaping and attenuating the noise seems to be the most beneficial one.

This poses a problem for keeping a modified FS-1016 CELP coder interoperable with the standard as it requires a change to the analyzer as well as adding a different postfilter to the synthesizer. As we saw in the previous test, the conventional postfilter alone

does not provide sufficient attenuation to suppress uncontrolled added noise in the masked regions of the speech. In some cases it even amplifies the signal in the masked regions which would be bad when the uncontrolled added noise is already high in this areas.

CHAPTER VI

CONCLUSION

Discussion of the Results

In this thesis we have discussed two possible speech quality improvements to the FS-1016 CELP coder that are interoperable with the standard. The first solution was to include a harmonic filter to improve the periodic structure of the speech. With the second solution noise was shaped in a perceptually better way by utilizing masking effects of the ear. This chapter briefly summarizes the results and points out interesting areas for future research that arose from this work.

The added harmonic filter was adapted to the CELP coder and values of the coefficients were determined empirically. It is included as part of the Analysis-by-Synthesis loop of the analyzer and thus makes the modified coder fully interoperable with the Federal Standard 1016. The major effect of the harmonic filter is to force the adaptive codebook search to choose more often the correct delay value in voiced speech areas. This results in improved speech quality and intelligibility which is documented by DAM and DRT test scores. The improvements are more pronounced in noisy speaker environments than in quiet ones. Very little additional computational overhead is added to the coder by this modification.

For the second improvement we implemented a recently published auditory model and determined its proper function. This was done by utilizing a block filter method, called the time domain aliasing cancellation, which introduces minimum distortion to the signal.

The masking model is suitable to use, yet its calculation adds a substantial amount of overhead when implemented in the coder.

We introduced two published methods that incorporated a masking model and they were applied to the CELP coder. In a further test with the masking model we compared the noise shaping abilities of the masking model with that of the perceptual weighting filter used in the CELP coder. As a result it was shown that when the energy of the noise signal is greater than the masking threshold, the noise becomes audible. Based on this result an evaluation of the two published methods was made. It was concluded that modifying the stochastic codebook search and determining the best codeword based on masked and unmasked regions of the reference speech would bring the most quality improvements. The published method had one characteristic that made it not interoperable with the standard coder. It added uncontrolled noise in the masked regions of the speech which had to be filtered by what was called a maskfilter. In a next step we investigated whether the postfilter used in the FS-1016 coder could be used to reduce this noise. The average attenuation which the postfilter brings to the masked regions of a speech signal was determined. It was found that the value was too low to have the postfilter take over the role of the maskfilter.

We can conclude from the test results that basing the stochastic codebook search on masked and unmasked regions of the reference speech will result in better speech quality instead of changing the transfer function of the perceptual weighting filter according to the inverse masking threshold. However, the success of the proposed method relies on a maskfilter that makes the modified coder not interoperable with the standard coder. In the following section we will present some ideas that might lead to changes that make the

solution interoperable with the standard.

Finally, the two suggested methods of changing the adaptive and the stochastic codebook search can be combined. With the harmonic filter the input speech signal to the codebook searches is harmonically enhanced. For the improved stochastic codebook search this reference signal is then transformed into the frequency domain, thus the methods do not interfere with each other.

Future Research

The results of the presented work suggest a variety of additional research ideas. In particular, the results of the improved stochastic codebook search left the question open whether the method will be interoperable with Federal Standard 1016. This possibility, along with ideas for improving the harmonic filter, are the subject of the following section.

Improved Adaptive Codebook Search

The advantage of the harmonic filter to automatically switch itself on in voiced areas and off in unvoiced areas brings one disadvantage. As the decision is based on the delay value of the previous subframe, the decision will be always one subframe late. Additional delay is introduced as a result of the smoothing of the filter coefficient. Voiced onsets are known to be modeled poorly by the FS-1016 coder. A more complex smoothing of the filter coefficient which allows the filter to be invoked quicker during voiced onsets could improve these areas. Since intelligibility is known to be related to the accuracy with which the

leading consonant is represented, and therefore improvement here could enhance the intelligibility of the coder. The first reflection coefficient gives a measure for voiced and unvoiced areas and is available in advance to the adaptive codebook search.

Improved Stochastic Codebook Search

For the stochastic codebook search there is one major point to be determined and that is whether the proposed method can be made interoperable with the FS-1016 standard CELP coder. The need for the maskfilter arose due the fact that the calculation for the match score and the gain included a masked function that is unity in unmasked and zero in masked areas of the speech signal. This added uncontrolled noise in the masked areas as the error minimization was only performed for the unmasked areas of the speech.

One way to change this would be to set the masked function not to zero but to a value smaller than one. This value could be made adaptable. The transition areas from unmasked to masked speech could be given greater values than the middle of a masked area. This would extend the error minimization again in the masked areas of the speech but give less emphasis to masked areas. This is in a way similar to changing the perceptual weighting filter to the inverse masking function, but it gives more flexibility to the way in which this masking function is chosen.

Another way would be to perform the stochastic codebook search in two stages. In the first stage the masking function is used as proposed, but the winning codeword is not determined yet. Rather, a second search is performed over the best N winning codewords of the first search with a changed masked function. Now the masked areas are incorporated into the error minimization. Here again the error in the masked areas of the speech can be

introduced in a more controlled manner than without considering the masked areas at all.

Another worthwhile area to look at is the actual calculation of the masking threshold. This calculation is computationally complex. However, except at lower frequencies, the resulting threshold shows a very constant form over the spectrum. Even over several frames of time the form does not seem to change very much. It would be interesting to explore whether there is a simpler way to come from the speech signal to an approximation to the masking threshold without performing the full calculations defined by Sen.

REFERENCES

- [AMBI97] E. Ambikairajah, A.G. Davis, W.T.K. Wong, "Auditory Masking and MPEG-1 Audio Compression," *Electronics & Communication Engineering Journal*, IEE, vol. 9, issue 4, pp 165-175, 1997.
- [ANDR84] H. Andrews, "Speech Processing," *Computer*, pp 315-324, Oct 1984.
- [ATAL79] B. Atal, M. Schroeder, "Predictive Coding of Speech Signals and Subjective Error Criteria," *IEEE Trans. On ASSP*, pp 247-254, 1979.
- [ATAL82] B. Atal, J. Remde, "A new Model of LPC Excitation for Producing Natural Sounding Speech at Low Bit Rates," *Proc. of ICASSP*, pp 614-617, 1982.
- [BERO84] M. Berouti, H. Garten, P. Kabal, P Mermelstein, "Efficient Computation and Encoding of the Multipulse Excitation for LPC," *Proc. Int. Conference Acoustic Speech and Signal Processing*, 1984.
- [CAMP91] J. Campbell, Jr., T. Tremain, V. Welch, "The Federal Standard 1016 4800 bps CELP Voice Coder", *Digital Signal Processing 1*, pp 145-155, 1991.
- [DAVI90] G. Davidson, L. Fielder, M. Antill, "Low Complexity Transform Coder for Satellite Link Applications," *AES Convention*, Preprint 2966.
- [DELL87] J. R. Deller, J. G. Proakis, J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Prentice Hall, NJ, 1987.
- [FLET72] H. Fletcher, *Speech and Hearing in Communication*, R.E. Krieger Publishing, 1972.
- [GERSO92] I. A. Gerson, M. A. Jasiuk, "Techniques for Improving the Performance of CELP-type Speech Coders," *IEEE Journal on Selected Areas in Communications*, vol.1, no.5, pp 858-865, 1992.
- [HEL72] R. P. Hellman, "Asymmetry of Masking Between Noise and Tone," *Perception & Psychophysics*, vol. 11 (3), pp 241-246, 1972.
- [KLEIJ95] W. B. Kleijn, K. K. Paliwal, *Speech Coding and Synthesis*, Elsevier, 1995.
- [KOND94] A. M. Kondoz, *Digital Speech, Coding for Low Bit Rate Communications. Systems*, pp 141-144, Wiley 1994.

- [KROO86] P. Kroon, E. F. Deprettere, R. Sluyter, "Regular-pulse Excitation: A novel Approach to Effective and Efficient Multipulse Coding of Speech," *IEEE Trans. On ASSP*, pp 1054-1063, October 1986.
- [KROO88] P. Kroon, E. F. Deprettere, "A Class of Analysis-by-Synthesis Predictive Coders for High Quality Speech Coding at Rates Between 4.8 and 16 kbits/s", *IEEE Journal on Selected Areas in Communications*, vol 6, no.2, pp 353-363, Feb. 1988.
- [KWON97] C. H. Kwon, C. K. Un, "Low-Rate CELP Speech Coding Using an Improved Weighting Function", *IEEE ICASSP*, pp 743-746, 1997.
- [MACH79] N. Machiavelli, "Discorsi sopra la prima deca di Tito Livio," in *Il Principe e le opere politiche*, Milan (Italy): Aldo Garzanti Editore, second ed., 1979. Appeared first in 1521.
- [MANO95] K. Mano, T. Mariya, S. Miki, H. Ohmuro, K. Ikeda, J. Ikedo, "Design of a Pitch Synchronous Innovation CELP Coder for Mobile Communicatins," *IEEE Journal on Selected Areas in Communications*, vol. 13, no. 1, pp 31-41, January 1995.
- [MONT91] R. Drogo de Iacovo, R. Montagna, "Some Experiments in Perceptual Masking of Quantizing Noise in Analysis-By-Synthesis Speech Coders," *Eurospeech*, pp 825-828, Genova, Italy 1991.
- [NCS92] "Details to Assist in Implementation of Federal Standard 1016 CELP", *Office of the Manager National Communications System*, Arlington Virginia, 1992.
- [NCSO91] "Telecommunications: Analog to Digital Conversion of Radio Voice by 4,800 Bit/Second Code Excited Linear Prediction (CELP)," *Federal Standard 1016*, National Communications System Office of Technology & Standards, published by General Services Administration Office of Information Resource Management, February 1991.
- [OPP89] A. V. Oppenheim, R. W. Schaffer, *Discrete-Time Signal Processing*, Prentice Hall, New Jersey, 1989.
- [PRIN86] J. Princen, A. Bradley, "Analysis/Synthesis Filter Bank Design Based on Time Domain Aliasing Cancellation," *IEEE Transactions on Acoustic, Speech, and Signal Processing*, vol. ASSP-34 no. 5 pp 1153-1161, Oct. 1986.

- [RAB94] L. Rabiner, "Applications of Voice Processing to Telecommunications," *Proc. IEEE*, vol. 82, no.2, pp 315-324, Feb. 1994.
- [SEN93] D. Sen, W. Holmes, "PERCELP - Perceptually Enhanced Random Codebook Excited Linear Prediction," *IEEE Workshop on Speech Coding for Telecommunications*, pp 101-102, Quebec, Canada 1993.
- [SEN94] D. Sen, *Perceptual Enhancement of Low Rate Speech Coders*, Ph.D thesis, B.E., University of New South Wales, Australia 1994.
- [SCHRO79] M. Schroeder, B. Atal, J. Hall, "Optimizing Digital Speech Coders by Exploiting Masking Properties of the Human Ear," *J. Acoust. Soc. Am.* vol. 66 no. 6, pp 1647-1652, 1979.
- [SCHRO85] M. Schroeder, B. Atal, "Code Excited Linear Prediction: High Quality Speech at Low Bit Rates," *Proc. of ICASSP*, pp 937-940, 1985.
- [SHOH91] Y. Shoham, "Constrained-Stochastic Excitation Coding of Speech at 4.8 kb/s," In *Advances in Speech Coding* (B. Atal, V. Cuperman and A. Gersho, Eds.) Kluwer Academic Publishers, Boston, pp 339-348, 1991.
- [TERH79] E. Terhardt, "Calculating Virtual Pitch," *Hearing Research*, 1, pp 155-182, 1979.
- [VELD89] R. N. J. Veldhuis, M. Breeuwer, R. van der Waal, "Subband Coding of Digital Audio Signals Without Loss of Quality," *IEEE ICASSP 1989*, pp 2009-2012, 1989.
- [WANG90] S. Wang, A. Gersho, "Improved Excitation for Phonetically-Segmented VXC Speech Coding Below 4 kb/s," *Proc. IEEE Global Telecomm. Conf.*, pp 946-950, 1990.
- [ZWICK90] E. Zwicker, H. Fastl, *Psychoacoustics - Facts and Models*, Springer Verlag, 1990.
- [ZWICK91] E. Zwicker, U. T. Zwicker, "Audio Engineering and Psychoacoustics: Matching Signals to the Final Receiver, the Human Auditory System," *Journal Audio Engineering Society*, vol 39 no 3, pp 115-125, March 1991.

VITA

Guenter Alban Dannoritzer

Candidate for the Degree of

Master of Science

Thesis: PERCEPTUAL ENHANCEMENTS FOR AN INTEROPERABLE FS-1016
CELP SPEECH CODER

Major Field: Electrical Engineering

Biographical:

Personal Data: Born in Hanau, Germany, February 11, 1969, the son of Franz and Elli Dannoritzer.

Education: Graduated from Kreuzburg School, Hainburg, Germany in June 1985; had vocational training at AEG Company in Seligenstadt, Germany and received a skilled working degree as an electrical technician in February 1989; served in the German Armed Forces from 1989 to 1991; received a degree from Ludwig Geissler Technical School, Hanau, Germany in June 1992; graduated with a "Diplom Ingenieur" from Fachhochschule, Dieburg, Germany, in July 1996. Completed the requirements of the Master of Science degree with a major in Electrical Engineering at Oklahoma State University, Stillwater, Oklahoma, in July 1998.

Experience: Internship in a development department for antennas and digital television from February to August 1995. Research Assistant, Department of Electrical and Computer Engineering, Oklahoma State University, January 1997 to present.

Professional Memberships: IEEE

Honors: Received a Fulbright Scholarship in the academic years 1996/97 and 1997/98.