Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Tokyo Institute of Technology

Source Excitation Generation for a HMM Based Synthesizer

Raphael Meyer

March 30, 2006

Diploma Thesis DA-2006-02
October 2005 – March 2006

Tutor:	Javier Latorre
Supervisors:	Prof. Sadaoki Furui
	Dr. Beat Pfister
Professor:	Prof. Lothar Thiele

Contents

Chapter 1

Introduction

In our research we focus on the development of a polyglot HMM based text to speech system. The HMM based synthesis uses a vocoder like analysis/synthesis technology. In the last stage of the synthesis the speech is generated by a MLSA filter using mel cepstral coefficients (MCC) as input and a simplified residual signal as excitation. This excitation is a generated pulse/noise signal,

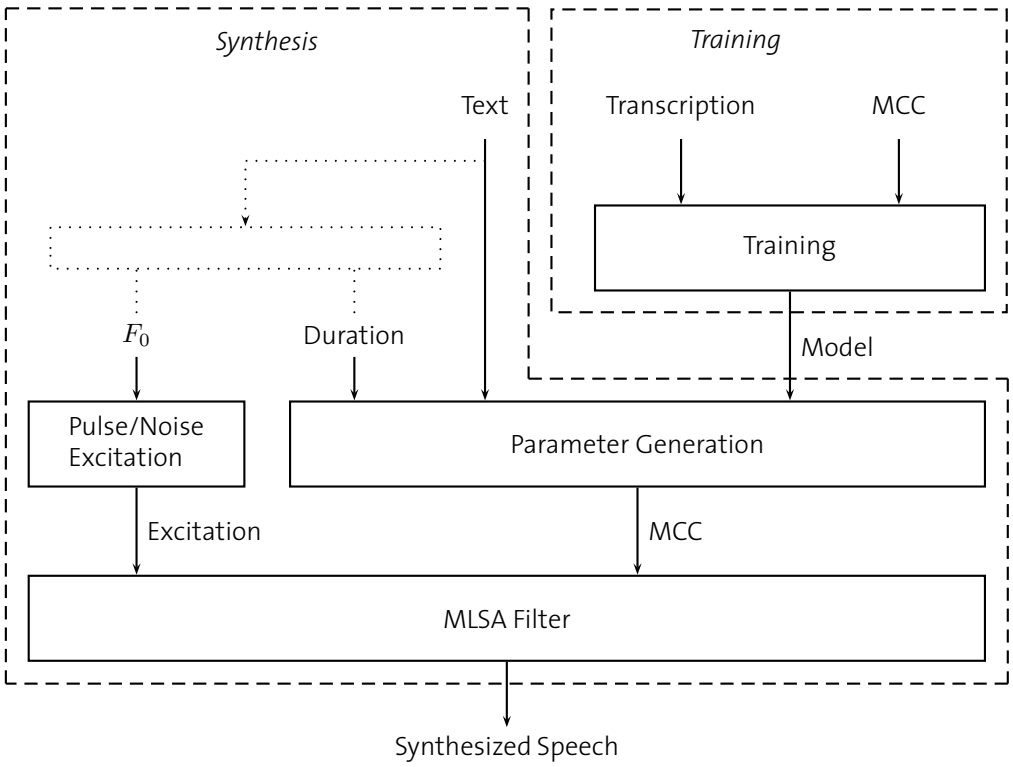


Figure 1.1: An overview of our HMM based TTS. F_0 and duration values may be generated from the input text using a prosodic model. But as this is not part of this project, we used values estimated from recorded data.

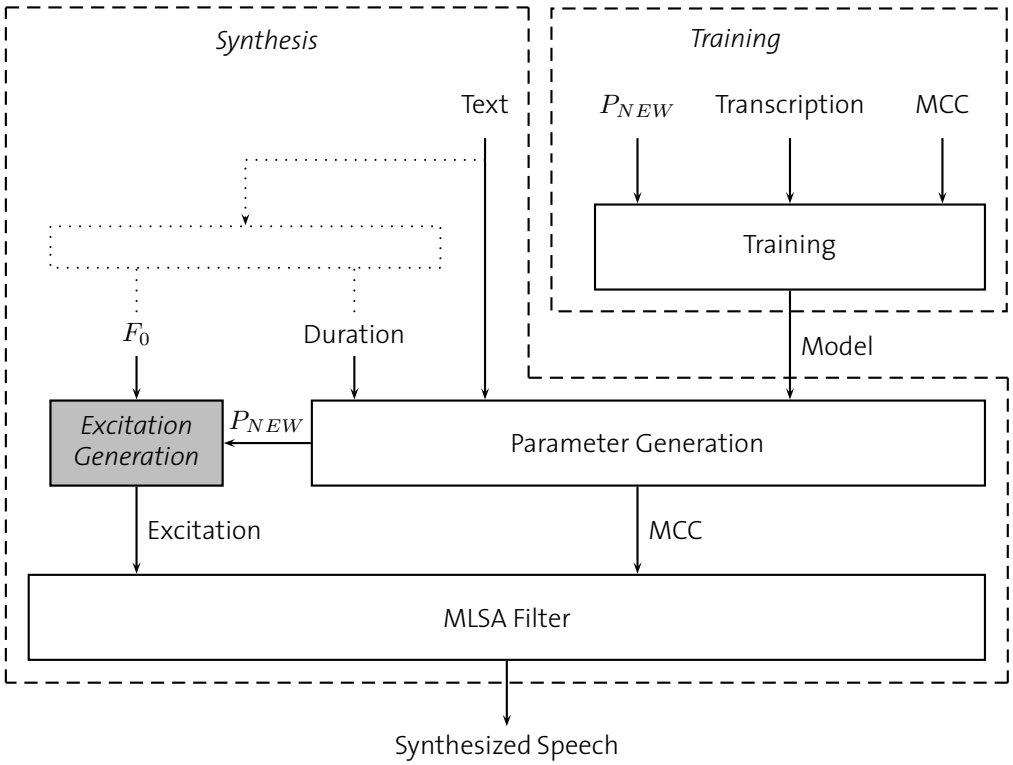


Figure 1.2: The task of this thesis is to replace the pulse/noise based excitation generation by a more sophisticated model. Therefore, in addition the MCC new parameters P_{NEW} for the source excitation generation may be introduced as well.

similar to the one in a LPC decoder. Although this simplification is useful, it results in a synthetic and buzzy sound quality of the synthesized speech. A simplified diagram of the system is shown in figure 1.1.

1.1 Task

The goal of this thesis is the study of a new model of source excitation generation that improves the quality of the speech synthesized with an HMM based synthesizer.

In a basic HMM based synthesizer, the source excitation is modeled as a sequence of pulses or noise signal. As a result, the synthesized speech usually has a very metallic sound. The idea to improve the speech quality consists in modeling the residuals excitation with different types of noises and pulse forms based on the analysis of the original residuals. In order to integrate this model into an HMM model, the different noises and pulse forms have to depend on a reduced number of parameters so that they can be easily generated.

A first approach to this problem was to use an excitation model such as MELP (mixed excitation linear prediction) standard. MELP is able to eliminate the synthetic buzz and has also

been used with some succes in HMM synthesis as shown by Yoshimura et al. [YTM+01]. However, mixed excitation produces some other effects on the quality of the synthesized speech that needed to be investigated.

In a second step, new parameters were introduced and tested. These new parameters were also incorporated into the HMMs. Figure 1.2 emphasizes the modifications to the old systems shown in figure 1.1.

Summarizing, the task at the beginning of this thesis were:

- Analysis of the MELP speech coding standard and the effects on the speech quality of its artifacts.
- Integration of the MELP parameters into an HMM based speech synthesizer.
- Development of a new set of parameters that produces a more robust estimation of the source excitation when included in an HMM speech synthesizer based on mel cepstral coefficients.
- Conducting of a subjective evaluation to compare the quality of the different source excitation models.

Chapter 2

Speech Coding

2.1 Linear Predictive Coding

The linear predictive coding (LPC) vocoder [IS68], [AH71] uses a fully parametric model to mimic human speech. In this approach, only the parameters of a speech model are transmitted and a decoder is used to regenerate speech with the same perceptual characteristics as the input speech waveform. Since periodic update of the model parameters requires fewer bits than direct representation of the speech signal, an LPC vocoder can operate at low bit rates. Still, the parametric representation in LPC coefficients preserves the critical information of a speech signal needed for other applications such as speech recognition.

Block diagrams of a LPC encoder and a decoder are shown in figure 2.1 and figure 2.2, respectively.

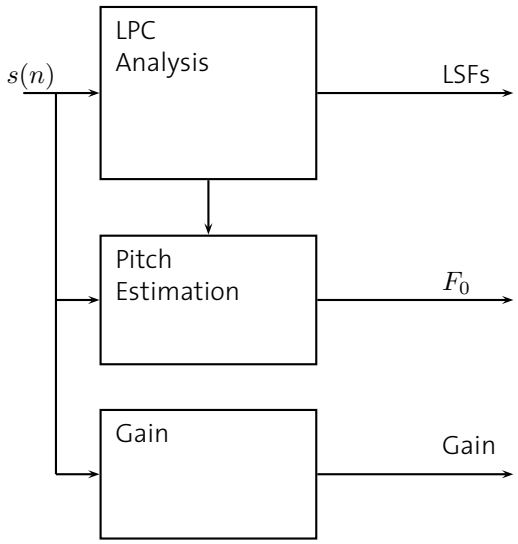


Figure 2.1: Block diagram of a LPC encoder.

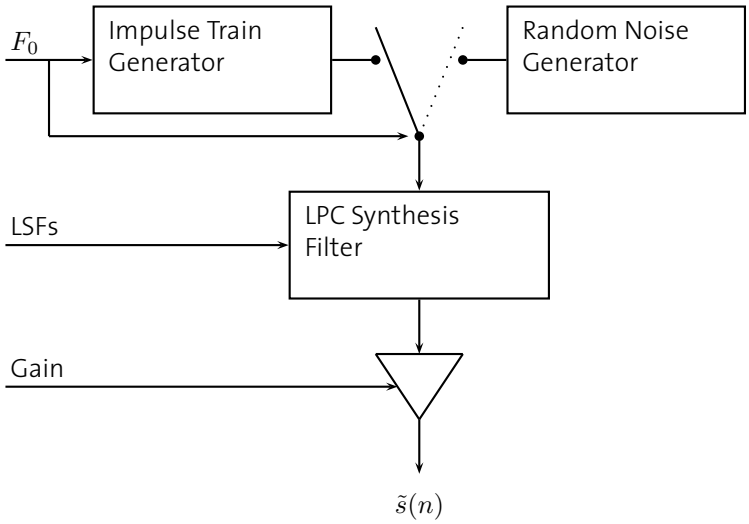


Figure 2.2: Block diagram of a typical LPC decoder.

The fact that consecutive samples of a speech signal are statistically dependent led to the idea of using linear prediction in speech coding. In a linear prediction, the nth sample of a signal $s(n)$ can be estimated by a weighted sum $\tilde{s}(n)$ of K preceding samples.

$$\tilde{s}(n) = -\sum_{K}^{k=1} a_k \, s(n-k) \tag{2.1}$$

Dividing the signal in overlapping frames of length N results in a system of N independent equations for every frame, where typically $N \gg K$. The coefficients a_k can be computed with autocorrelation based methods like the *Durbin* algorithm.

The prediction error is

$$\begin{aligned} e(n) &= s(n) - \tilde{s}(n) && (2.2)\\ &= s(n) + \sum_{k=1}^{K} a_k \, s(n-k) && (2.3)\\ &= \sum_{k=0}^{K} a_k \, s(n-k) \qquad (a_0 \equiv 1) && (2.4) \end{aligned}$$

and can be written in the z-transform as

$$\begin{aligned} E(z) &= S(z)\, A(z) && (2.5)\\ &= S(z)\, \frac{1}{H(z)}. && (2.6) \end{aligned}$$

The error is also called residual and corresponds to the glottal excitation. The original speech signal $S(z)$ can be restored from the error $E(z)$ using the synthesis filter $H(z)$:

$$S(z) = H(z)\,E(z). \tag{2.7}$$

In LPC speech coding the actual error $e(n)$ is replaced by a generated signal $\tilde{e}(n)$. This approximated residual is modelled by an impulse train based on the pitch F_0 for voiced speech or random noise for unvoiced speech. An example of such a signal is shown in figure 2.3(a). Therefore each frame of speech can be reduced to a set of LPC coefficients, the pitch value which is used for the voiced/unvoiced decision and the gain value. In the decoder the speech signal is reconstructed by a synthesizer based on a time varying all-pole filter.

LPC coefficients are very sensible to errors. Transmission errors and quantization errors result in a strong degradation of the quality of the synthesized speech. Therefore, the LPC coefficients are generally transformed to the much more robust line spectral frequencies (LSF) for transmission.

2.2 Mel Log Spectrum Approximation Filter

Imai et al. [ISF83] presented a mel log spectrum approximation (MLSA) filter to synthesize speech using mel cepstral coefficients (MCC). The advantage of MCC is that it represents spectra that have logarithmic frequency resolutions similar to the human ear which has a high resolution at low frequencies. For obtaining the MCC several methods have been proposed. In conjunction with HMM the most commonly used methods are based on the methods proposed by Fukuda et al. [FTKI92] and Imai et al. [Ima83].

2.3 Mixed Excitation LPC Vocoder Model

The major drawback of the simple model for the source excitation generation in the LPC vocoder is that the decoded speech sounds synthetic. To improve the perceptual quality of the speech encoded with the LPC model McCree [MB95] proposed a mixed excitation LPC vocoder model (MELP).

The main idea behind MELP is to split the speech signal in a set of frequency bands. For each band a separate excitation based on pulse train and noise is generated. This excitations are then combined into a single mixed excitation. Figure 2.3(b) shows an example of a mixed excitation signal.

Figures 2.4 and 2.5 show block diagrams of the MELP encoder and decoder. To achieve a more natural sound for the synthesized speech the MELP model has the following five additional features:

- Mixed pulse and noise excitation

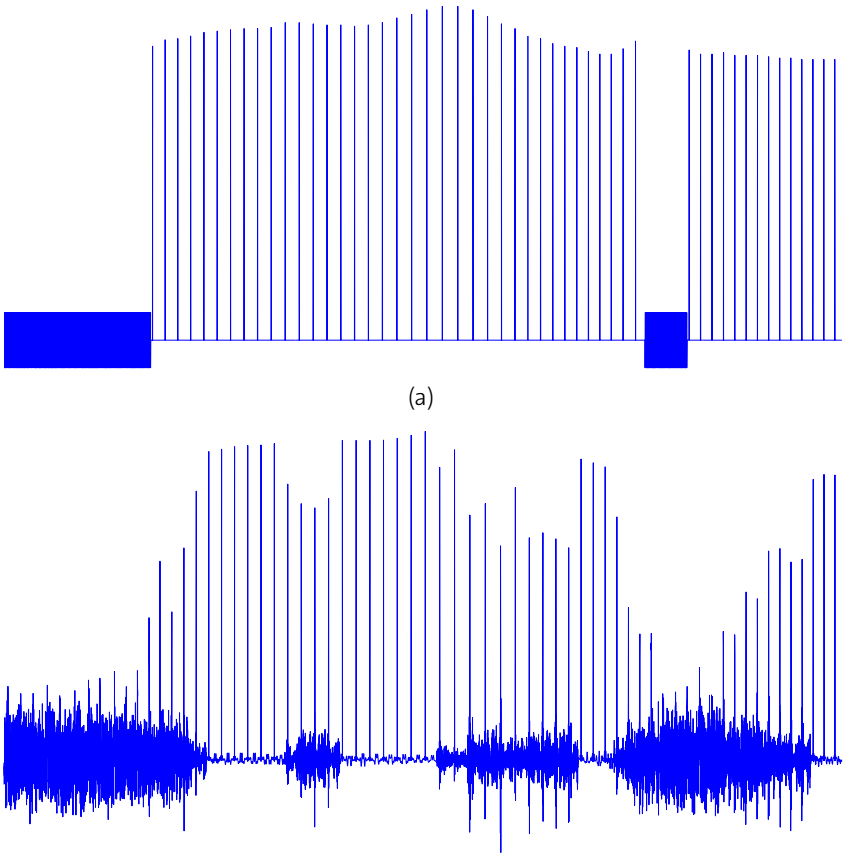


Figure 2.3: Pulse and noise based (a) and MELP based (b) excitation.

- Pulse Generation
  - Periodic or aperiodic pulses
  - Transmission of the Fourier magnitudes
- Adaptive spectral enhancement filter
- Pulse dispersion filter

2.3.1 Mixed Excitation

The most important feature of a MELP coder is the mixed pulse and noise excitation. Its primary effect is the reduction of the buzzy quality of the basic LPC vocoder.

The mixed excitation LPC encoder generates an excitation signal with different mixtures of pulse and noise in each of a number of frequency bands. The standard MELP vocoder uses five bands of 0–500 Hz, 500–1000 Hz, 1000–2000 Hz, 2000–3000 Hz and 3000–4000 Hz. The pulse

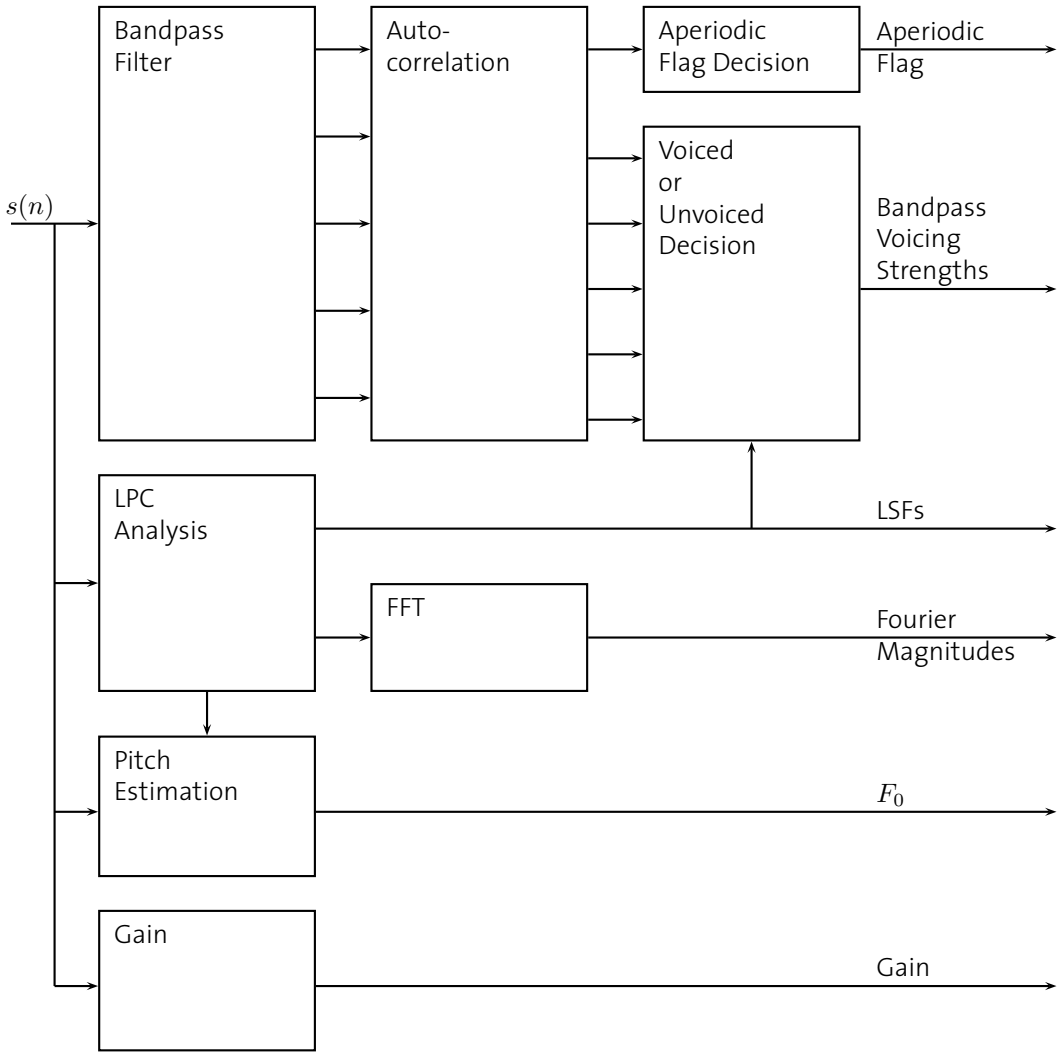


Figure 2.4: Block diagram of the MELP Encoder

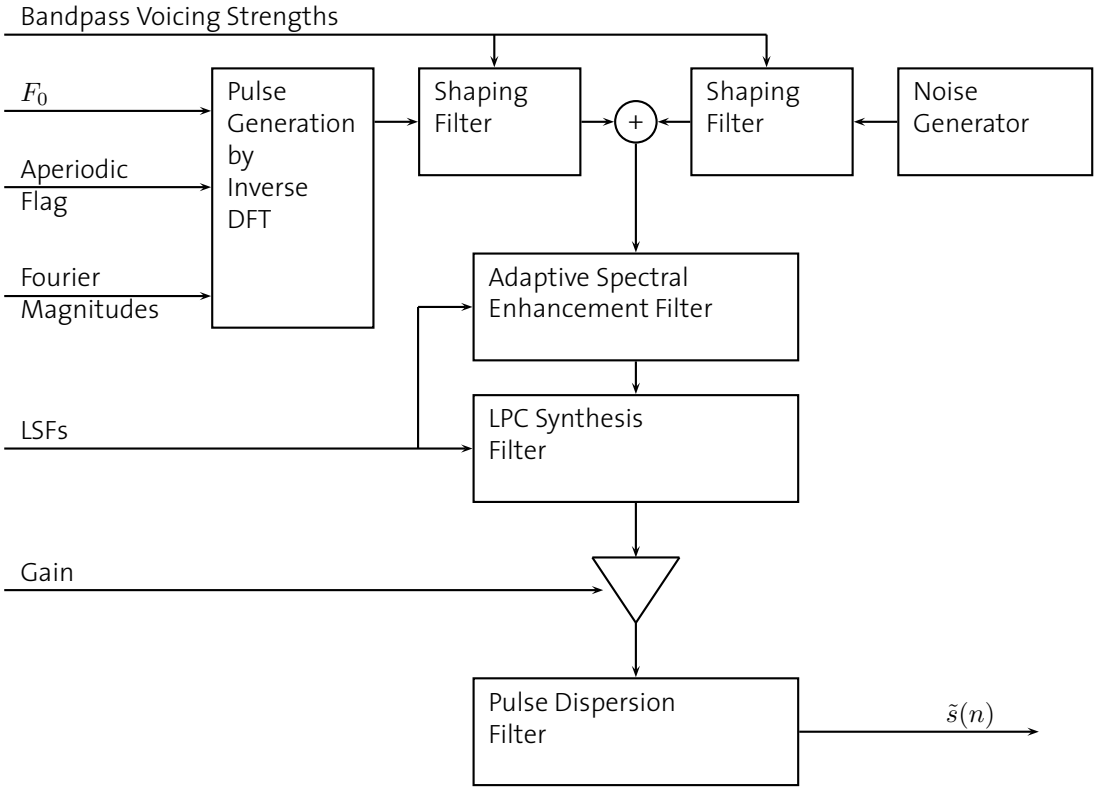


Figure 2.5: Block diagram of the MELP Decoder

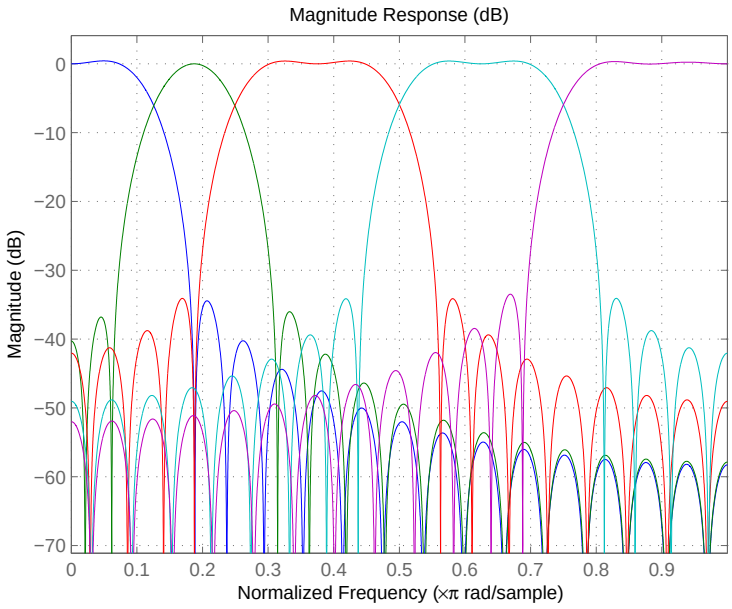


Figure 2.6: Magnitude response of the five shaping filters.

excitation $e_{pulse}(n), n \in [0, 1 \ldots T-1]$ is computed using an inverse discrete Fourier transform of one pitch period in length of a signal $M(k)$.

$$e_{pulse}(n) = \frac{1}{T} \sum_{k=0}^{T-1} M(k) e^{\frac{j2\pi nk}{T}} \tag{2.8}$$

$M(k)$ is a sequence of ones except for the first K terms that may be set according to the Fourier magnitudes as explained in subsection 2.3.2.2. The phases of $M(k)$ are set to zero. The noise is generated by a uniform random number generator and then normalized. The pulse train and noise signal are each passed through time-varying spectral shaping filters and then added together to give a fullband excitation. For each frame, the ith frequency shaping filter coefficients $c_{pulse}(i)$ and $c_{noise}(i)$ are generated by a weighted sum of the fixed coefficients $c_{bp}(i)$ of N bandpass filters.

$$c_{pulse}(i) = \sum_{j=1}^{N} w_j c_{bp}^{(j)}(i) \tag{2.9}$$

$$c_{noise}(i) = \sum_{j=1}^{N} (1 - w_j) c_{bp}^{(j)}(i) \tag{2.10}$$

In the standard MELP coder the bandpass voicing strengths are simplified to a voiced/unvoiced decision. Therefore the weights w_j are either 1 or 0. The bandpass filters are FIR filters of 32th order. The magnitude responses of these filters are shown in figure 2.6.

2.3.2 "Modifications" to the Pulse Generation

2.3.2.1 Periodic or Aperiodic Pulses

Another problem of standard LPC is a distortion called *tonal noise*. This distortion is introduced when periodicity is present in speech frames which are actually unvoiced. This is often encountered in the voicing transition region, especially for female speakers. In order to reduce this kind of distortion, the periodicity in the voiced excitation is destroyed by varying each pitch period length with a pulse position jitter uniformly distributed up to $\pm$ 25 %. This allows the synthesizer to mimic the erratic glottal pulses which are often encountered in voicing transitions or in vocal fry. However, this cannot be done for strongly voiced frames without introducing a hoarse quality. Therefore a control algorithm is used to determine when the jitter should be added. In the standard MELP coder a jitter will be added if the lowest bandpass voicing strength does not exceed a certain threshold.

2.3.2.2 Fourier Magnitudes

The Fourier magnitudes of the first ten harmonics of the residual signal are used to improve the quality of the synthesized speech. This is particularly effective for male speakers and when background noise is present. By using the Fourier series, the magnitudes of the selected harmonics of the fundamental pitch frequency are reproduced and as a result the performance of the speech production model at the perceptually important lower frequencies is improved. The first ten Fourier magnitudes are determined from the peaks of the Fourier transform of the prediction residual signal. These coefficients are used in the generation of the pitch pulse of the mixed excitation signal.

2.3.3 Adaptive Spectral Enhancement Filter

Another feature in the mixed excitation LPC vocoder model is adaptive spectral enhancement. This adaptive filter helps the bandpass filtered synthetic speech to match natural speech waveforms in the formant regions. Typical formant resonances usually do not completely decay in the time between pitch pulses in either natural or synthetic speech. However, the synthetic speech waveforms reach a lower valley between the peaks than natural speech waveforms do. This is probably caused by the inability of the poles in the LPC synthesis filter to reproduce the features of formant resonances in natural human speech.

The adaptive spectral enhancement filter provides a solution to the problem of matching formant waveforms. This adaptive pole/zero filter is widely used in CELP coders [CG87] to reduce quantization noise in between the formant frequencies. The poles are generated by a bandwidth expanded version of the LPC synthesis filter, with β equal to 0.8. Since this all-pole filter $A(\beta z^{-1})$ introduces a disturbing lowpass filtering effect by increasing the spectral tilt, a weaker all-zero filter $A(\alpha z^{-1})$ calculated with α equal to 0.5 is used to decrease the tilt of the overall filter without reducing the formant enhancement. In addition a simple first order FIR filter $(1 + \mu z^{-1})$ is used to further reduce the lowpass muffling effect.

Thus the transfer function of the enhancement filter $H_{asz}(z)$ is given by

$$H_{asz}(z) = \frac{A(\alpha z^{-1})}{A(\beta z^{-1})} \cdot (1 + \mu z^{-1}) \tag{2.11}$$

where $\alpha = 0.5p$ and $\beta = 0.8p$. The tilt coefficient μ is calculated as $\min(0.5 \cdot k_1, 0)$ multiplied by the signal probability p. The first reflection coefficient $k1$ is calculated from the decoded LSFs. The signal probability p is estimated by comparing the power in the current speech frame with a long-term estimate of the noise power. It is a linear ramp value between 0 and 1 corresponding to an estimated gain value between 12 dB and 30 dB. The signal probability is introduced to reduce the fluctuations caused by this adaptive filter when background noise is present [McC99].

2.3.4 Pulse Dispersion Filter

The pulse dispersion filter improves the match of bandpassed filtered speech waveforms in frequency bands which do not contain a formant resonance. At these frequencies, the synthesized speech often decays to a very small value between the pitch pulses. This is also true for frequencies near the higher formants, since these resonances decay significantly between excitation points, especially for the longer pitch periods of male speakers. In these cases, the bandpass filtered natural speech has a smaller peak-to-valley ratio than the synthesized speech. In natural speech, the excitation may not all be concentrated at the point in time corresponding to closure of the glottis. This additional excitation prevents the natural bandpass envelope from falling as low as the synthetic version. This could be due to a secondary excitation peak from the opening of the glottis or aspiration noise resulting from incomplete glottal closure.

The pulse dispersion filter is a fixed FIR filter, based on a spectrally flattened synthetic glottal pulse which introduces time-domain spread to the synthetic speech. A triangle pulse based on a typical male pitch period is used. The filter coefficients are generated by taking a DFT of the triangle pulse, setting the magnitudes to unity, and taking the inverse DFT.

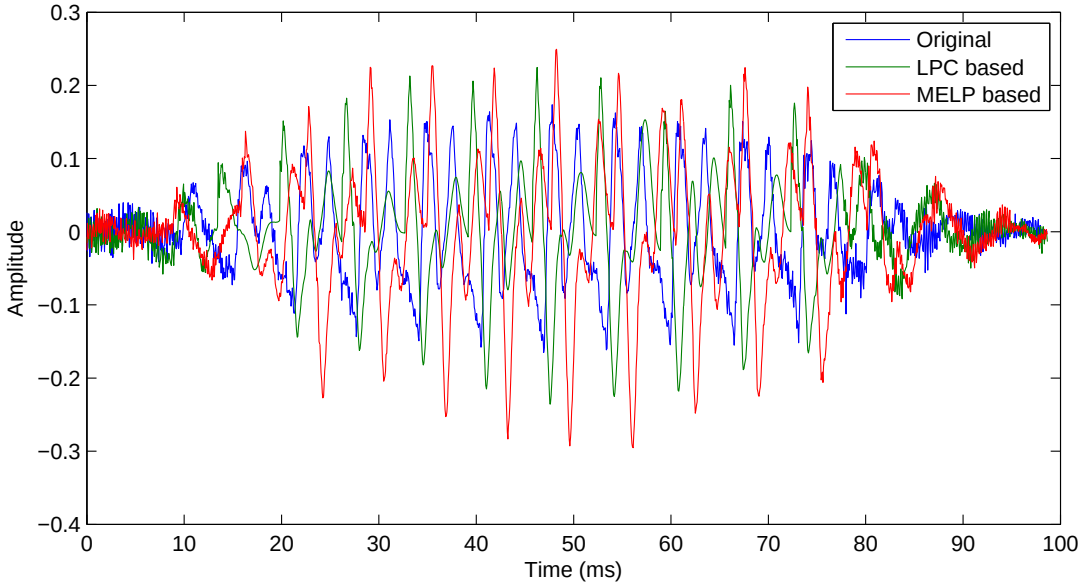
This filter decreases the peakiness of the synthesized band passed signal in frequencies away from the formants. This results in more natural sounding LPC speech output.

2.4 Wideband MELP Coder

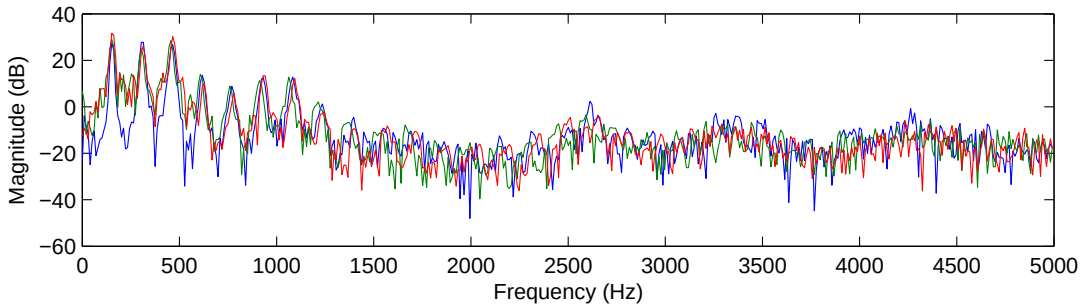
The original MELP coder proposed by McCree [MB95] was designed as a narrowband coder with low bit rate. However, we are interested in high quality rather than in low bit rates. Therefore we used the wideband MELP coder proposed by Lin [Lin00], [LKL00].

Because the standard MELP model has been altered to code the full speech band, i.e. 50–7000 Hz, the sampling rate has been changed from 8 kHz to 16 kHz. The frame period has been set to 11.25 ms (180 samples) which is half the duration used in the MELP Standard. The filter orders and cut-off frequencies have been adapted and the analysis windows have been resized to fit the higher sampling rate and shorter frame period. The $10\,th$ order LPC filter of the standard MELP coder used for narrow band speech signals is no longer adequate to model the spectral envelope of wideband speech signals. The LPC order order has been increased to 20. The pulse dispersion filter has been eliminated as the $20\,th$ order LPC synthesizer has been found experimentally [Lin00] to be sufficiently good in representing the envelope of the speech spectral envelope.

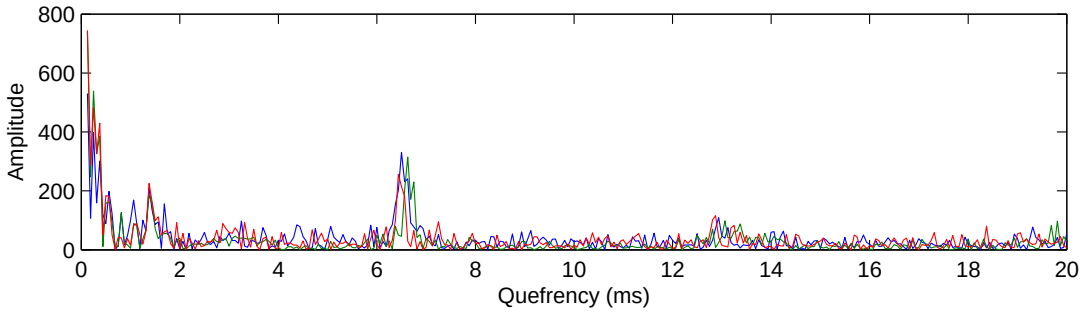
The MELP model was not only changed from a narrowband to wideband model, but the pitch estimation has also been modified to improve its accuracy. In the MELP standard the final pitch estimate is calculated from the original speech signal, a residual signal and a candidate pitch from an earlier stage in the analysis. The wideband MELP coder evaluates every pitch in set of values in the range of valid pitch values. The pitch that corresponds to the maximum normalized autocorrelation value among this set is selected as the final pitch estimate. Furthermore a pitch tracking



(a)



(b)



(c)

Figure 2.7: Waveform (a), spectrum (b) and cepstrum (c) for a voiced speech frame.

method has been added in order to preserve continuity of the pitch estimates between neighboring speech frames and to reduce sudden pitch changes that lead to perceptual distortions. If the pitch estimated in the current frame exceeds the range of the pitches estimated for the previous and the next frame by a certain percentage, the current pitch is set to the average of the pitch of the previous and the next frame.

In an informal listening test of preference [Lin00] the 8.4 kpbs wideband MELP coder performed equal to the 48 kbps ITU G.722 coder [Mai88]. Furthermore, it was even preferred at a 3:2 ratio over the 14.4 kpbs MPEG4 wideband CELP coder [Mot98].

Chapter 3

HMM Based Synthesis

The hidden Markov models (HMMs) are statistical models widely used to characterize the sequence of speech spectra. HMMs have successfully been applied to speech recognition systems. Based on these facts it was obvious that HMMs may also be useful in speech synthesis. Tokuda et al. [TMY+95] proposed an algorithm for speech parameter generation from HMMs using mel cepstral coefficients. Masuko et al. [MTKI96] proposed a new algorithm which includes delta and delta-delta parameters.

3.1 Hidden Markov Models

A hidden Markov model (HMM) is a finite state machine which generates a sequence of discrete time observations. At each time unit the HMM changes states according to the state transition probability distribution and then generates an observable output according to the output probability distribution of the current state. The challenge is to determine the hidden parameters from the observable outputs.

There are three different problems to solve with HMMs.

- The computation of the probability of a particular output sequence, given the model parameters, which can be solved by the *forward algorithm*.
- Calculate the most likely sequence of (hidden) states which could have generated a given output sequence given the model parameters. This can be found by using the *Viterbi algorithm*.
- Find the most likely set of state transition and output probabilities for a given output sequence. This can be solved by the *Baum-Welch algorithm*.

3.2 HMM Based Synthesis

In HMM based synthesis the first stage is to train a model. This is common to HMM based speech recognition. In an optional second stage the model is adapted to a specific speaker. Although this is optional, it is often either desired as a feature or done because it might help improving the quality of the synthesized speech. Finally speech is synthesized by extracting parameters out of the model given the text converted to a sequence of phonemes.

3.2.1 Training

The HMMs used in speech synthesis are often left-to-right models with no skip.

Initially, a set of monophone models is trained. These models are cloned to produce triphone models for all distinct triphones in the training data. The triphone models are then reestimated with the embedded version of the *Baum-Welch* algorithm. The states of the triphone HMMs are clustered using the furthest neighbor hierarchical clustering algorithm [YW94]. The output distributions in the same cluster are tied to reduce the number of parameters and to balance model complexity against the amount of available data. Tied triphone models are reestimated with the embedded training again.

3.2.2 Speaker Adaptation

In the speaker adaptation stage, initial model parameters, such as mean vectors of output distributions, are adapted to a target speaker using a small amount of adaptation data uttered by the target speaker. The initial model can be speaker dependent or independent, however for a speaker dependent model the speaker for the initial model has to be selected carefully to get an optimal result.

3.2.3 Synthesis

The text to be synthesized is transformed into a sequence of phonemes. According to this sequence of phonemes, triphone HMMs are concatenated to a model representing a whole sentence. For single mixture HMMs the speech parameters can be calculated out of this sentence HMM using the algorithm presented by Tokuda et al. [TKI95]. Tokuda et al. [TMY+95] later proposed an improved version of the algorithm that can generate parameters from multi mixture HMM.

Let $\mathbf{O} = \{\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_T\}$ be the vector sequence of speech parameters. Further, let $\mathbf{Q} = \{(q_1, i_1), (q_2, i_2), \ldots, (q_T, i_T)\}$ be the state sequence of an HMM λ where (q, i) indicates the ith mixture of state q. Assume that the vector of speech parameters $\mathbf{o}_t$ at frame t consists of the

static feature vector $\mathbf{c}_t$ and the dynamic feature vector $\Delta\mathbf{c}_t$.

$$\mathbf{o}_t = \{\mathbf{c}_t, \Delta\mathbf{c}_t\} \tag{3.1}$$

where

$$\mathbf{c}_t = \left[c^{(1)}, c^{(2)}, \ldots, c^{(M)}\right]^T \tag{3.2}$$

$$\Delta\mathbf{c}_t = \left[\Delta c^{(1)}, \Delta c^{(2)}, \ldots, \Delta c^{(M)}\right]^T \tag{3.3}$$

and $\Delta\mathbf{c}_t$ is defined as

$$\Delta\mathbf{c}_t = \sum_{i=-L}^{L} w_i \mathbf{c}_{t+i}. \tag{3.4}$$

The problem is to determine the parameter sequence $\mathbf{c} = \left[\mathbf{c}_1^T, \mathbf{c}_2^T, \ldots, \mathbf{c}_T^T\right]^T$ which maximizes

$$P\left[\mathbf{O}|\lambda\right] = \sum_{all\,\mathbf{Q}} P\left[\mathbf{Q}, \mathbf{O}|\lambda\right] \tag{3.5}$$

for a given HMM λ. However, since the problem is difficult to solve, the optimum sequence is considered in a similar manner of the Viterbi algorithm.

$$\bar{P}\left[\mathbf{O}|\lambda\right] = \max_{\mathbf{Q}} P\left[\mathbf{Q}, \mathbf{O}|\lambda\right] \tag{3.6}$$

has to be maximized with respect to $\mathbf{c}$. Since $\mathbf{Q}$ and $\mathbf{c}$ have to be determined simultaneously, dynamic programming methods cannot be used in contrast to the Viterbi algorithm. To solve this problem Tokuda et al. [TMY+95] proposed to maximize with respect to $\mathbf{c}$

$$\begin{aligned}\log P\left[\mathbf{O}|\mathbf{Q}, \lambda\right] =& \alpha \sum_{k=1}^{K} \log p_{q_k}(d_{q_k}) + \sum_{t=1}^{T} \log c_{q_t, i_t} \\ & - \frac{1}{2}\epsilon(\mathbf{c}) - \frac{1}{2}\log|\mathbf{U}| - \frac{3MT}{2}\log 2\pi\end{aligned} \tag{3.7}$$

where

$$\epsilon(\mathbf{c}) = (\mathbf{O} - \mu)^T \mathbf{U}^{-1} (\mathbf{O} - \mu) \tag{3.8}$$

$$= (\mathbf{W}\mathbf{c} - \mu)^T \mathbf{U}^{-1} (\mathbf{W}\mathbf{c} - \mu) \tag{3.9}$$

and

$$\mu = \left[\mu_{q_1,i_1}^T, \mu_{q_2,i_1}^T, \ldots, \mu_{q_T,i_T}^T\right]^T \tag{3.10}$$

$$\mathbf{U} = \mathrm{diag}\left[\mathbf{U}_{q_1,i_1}, \mathbf{U}_{q_2,i_2}, \ldots, \mathbf{U}_{q_T,i_T}\right] \tag{3.11}$$

$$\mathbf{W} = \left[\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_T\right]^T \tag{3.12}$$

$$\mathbf{w}_t = \left[\mathbf{w}_t^{(0)}, \mathbf{w}_t^{(1)}\right] \tag{3.13}$$

$$\mathbf{w}_t^{(0)} = [\mathbf{0}_{M\times M}, \ldots, \mathbf{0}_{M\times M}, \mathbf{I}_{M\times M}, \mathbf{0}_{M\times M}, \ldots, \mathbf{0}_{M\times M}]^T \tag{3.14}$$

$$\begin{aligned}\mathbf{w}_t^{(1)} = [&\mathbf{0}_{M\times M}, \ldots, \mathbf{0}_{M\times M},\\ &w_{-L}\mathbf{I}_{M\times M}, \ldots, w_0\mathbf{I}_{M\times M}, \ldots, w_L\mathbf{I}_{M\times M},\\ &\mathbf{0}_{M\times M}, \ldots, \mathbf{0}_{M\times M}]^T.\end{aligned} \tag{3.15}$$

By setting $\frac{\partial \log P[\mathbf{O}|\mathbf{q},\lambda]}{\partial \mathbf{c}} = \mathbf{0}_{TM\times TM}$ we obtain a set of equations

$$\mathbf{R}\mathbf{c} = \mathbf{r} \tag{3.16}$$

where

$$\mathbf{R} = \mathbf{W}^T\mathbf{U}^{-1}\mathbf{W} \tag{3.17}$$

$$\mathbf{r} = \mathbf{W}^T\mathbf{U}^{-1}\mu. \tag{3.18}$$

The iterative algorithm to solve the equations (3.16) is derived from the standard RLS algorithm.

In the case of multi mixture states, a sub-optimal sub-state sequence can be found by means of an iterative algorithm. The outline of the algorithm is as follows:

1. Initialization
   (a) Determine a initial sub-state sequence $\mathbf{Q}$.
   (b) For the initial sub-state sequence obtain $\mathbf{c}$, ϵ and $\mathbf{P}$ where $\mathbf{P} = \mathbf{R}^{-1}$.
2. Iteration
   (a) For $t = 1, 2, \ldots, T$
      i. For each possible sub-state at frame t obtain the value of $\log P[\mathbf{Q}, \mathbf{O}|\lambda]$ from equation (3.7)
      ii. Choose the best sub-state in the sense that $\log P[\mathbf{Q}, \mathbf{O}|\lambda]$ is most increased by the sub-state replacement.
   (b) Choose the best frame in the sense that $\log P[\mathbf{Q}, \mathbf{O}|\lambda]$ is most increased by the sub-state replacement.

(c) If $\log P[\mathbf{Q}, \mathbf{O}|\lambda]$ cannot be increased by the sub-state replacement at the best frame, stop the iteration.

(d) Replace the sub-state of the best frame and calculate $\mathbf{c}$, ϵ and $\mathbf{P}$ for the next iteration step.

(e) Go to 2a.

3.3 HMM Based Polyglot Synthesizer

A special application of HMM based synthesis is the polyglot HMM based TTS (text to speech) synthesis. We call a system polyglot if it can generate intelligible speech in several languages having the same voice identity.

We are using a HMM based TTS system to investigate various aspects for polyglot TTS synthesis [LIF05c], [LIF05b], [LIF05a]. Figure 3.1 illustrates the layout of such a system. Our approach consists in combining monolingual corpora from several speakers in different languages to train a language independent and speaker independent HMM based synthesizer [LIF05b]. Since in our method no human polyglot talent is required we can expand it to any number of languages we want. Furthermore, since no phone mapping is needed for the languages included in the mixture, the perceptual intelligibility and the level of foreign accent when synthesizing these languages is lower than with other methods based on phone mapping.

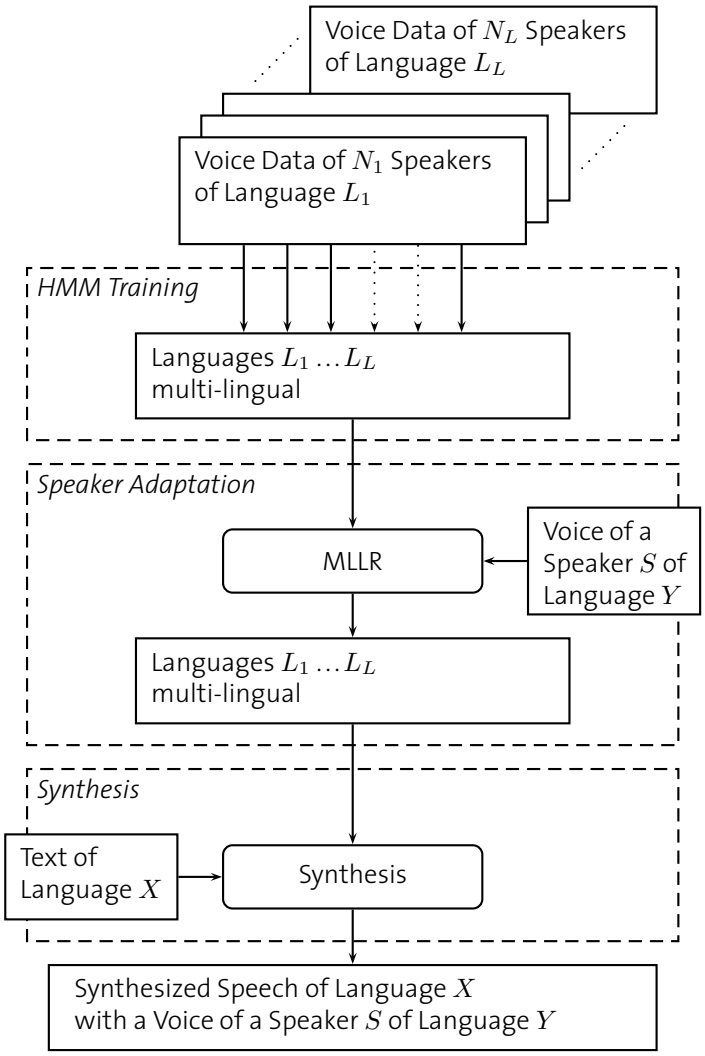


Figure 3.1: Overview of a possible layout of our polyglot HMM based TTS system. In this illustrative example the HMM is adapted to the voice of a speaker S of language $Y \in \{L_1 \dots L_L\}$. A given text in any language $X \in \{L_1 \dots L_L\}$ is synthesized with the same voice of speaker S.

Chapter 4

Distortion Measurement

The evaluation of speech quality is of critical importance in the field of speech coding. Not only is it necessary to have consistent subjective tests for the comparative assessments of alternative coders, it is also essential to have an objective distortion measure which, during the development phase, can give an immediate and reliable estimate of the anticipated perceptual quality of a particular coding algorithm.

We tried two different methods for distortion measurement. One is based on the cepstrum, while the other is based on the bark scale.

4.1 Cepstral based Distortion

The cepstral distortion $D^{(k)}$ for the kth segment can be calculated as

$$D^{(k)} = \frac{1}{N} \sum_{i=1}^{N} \left(M_x^{(k)}(i) - M_y^{(k)}(i) \right)^2 \tag{4.1}$$

where

$$M^{(k)} = IDFT(\log |DFT(s^{(k)})|). \tag{4.2}$$

The signal s is divided in overlapping segments of length N.

4.2 Bark Spectral Distortion

A psychoacoustically motivated measure is the bark spectral distortion (BSD) [WSG92]. The bark spectrum L reflects the ear's nonlinear transformations of frequency and amplitude, together with important aspects of its frequency analysis and spectral integration properties in response to complex sounds. Figure 4.2 shows examples of the bark spectrum.

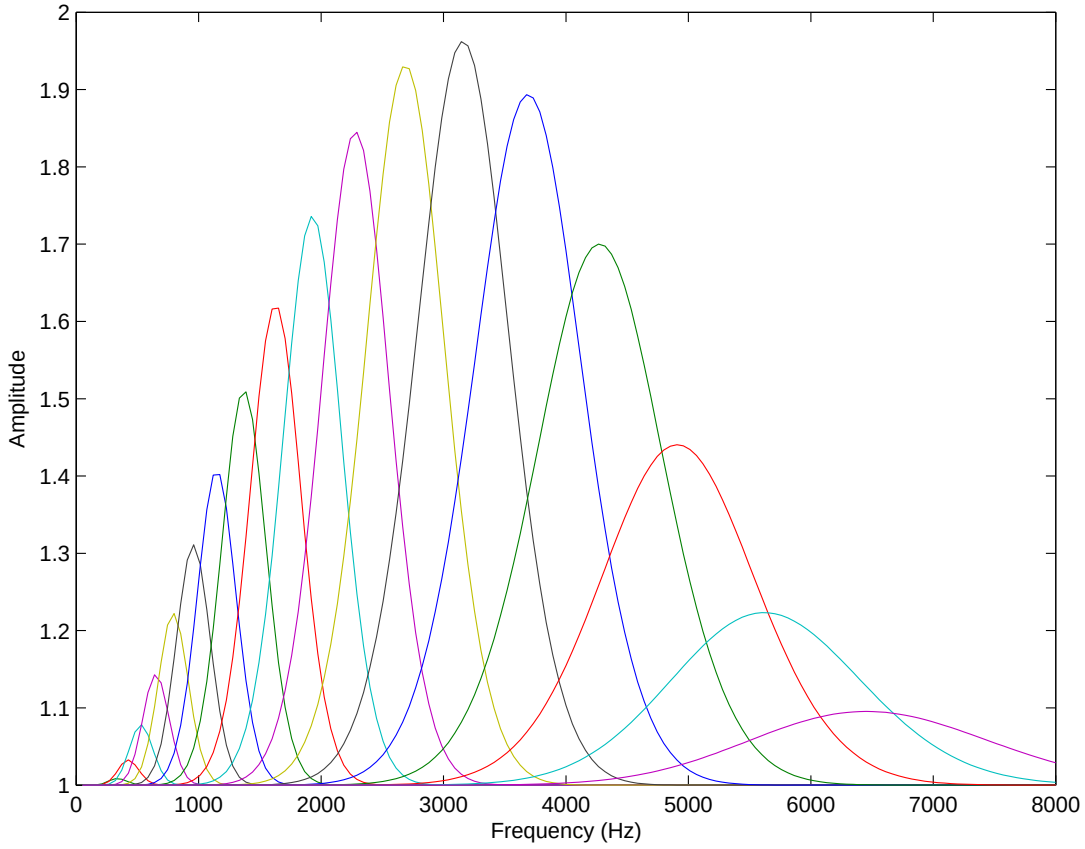


Figure 4.1: The nineteen weighting functions used to calculated the BSD. The functions are based on Gaussian distributions where the μs are set to the center frequencies of the corresponding critical band and the σs are set to the corresponding critical bandwidth.

For a value b in Bark the transformation to the value f in Hertz is defined as

$$f = 600 \sinh(\frac{(b + 0.5)}{6}) \tag{4.3}$$

To calculate the bark spectrum $L^{(k)}$ of the kth segment, intermediate coefficients $P^{(k)}$ are calculated first. A Discrete Fourier transform is applied to the kth segment of the signal. The coefficients from the DFT are then piecewise squared. The obtained values are then weighted by the functions shown in figure 4.1. The coefficients $P^{(k)}$ are finally transformed to the bark spectrum $L^{(k)}$ by

$$L^{(k)}(i) = \left(P^{(k)}(i)\right)^{\frac{1}{3}} \tag{4.4}$$

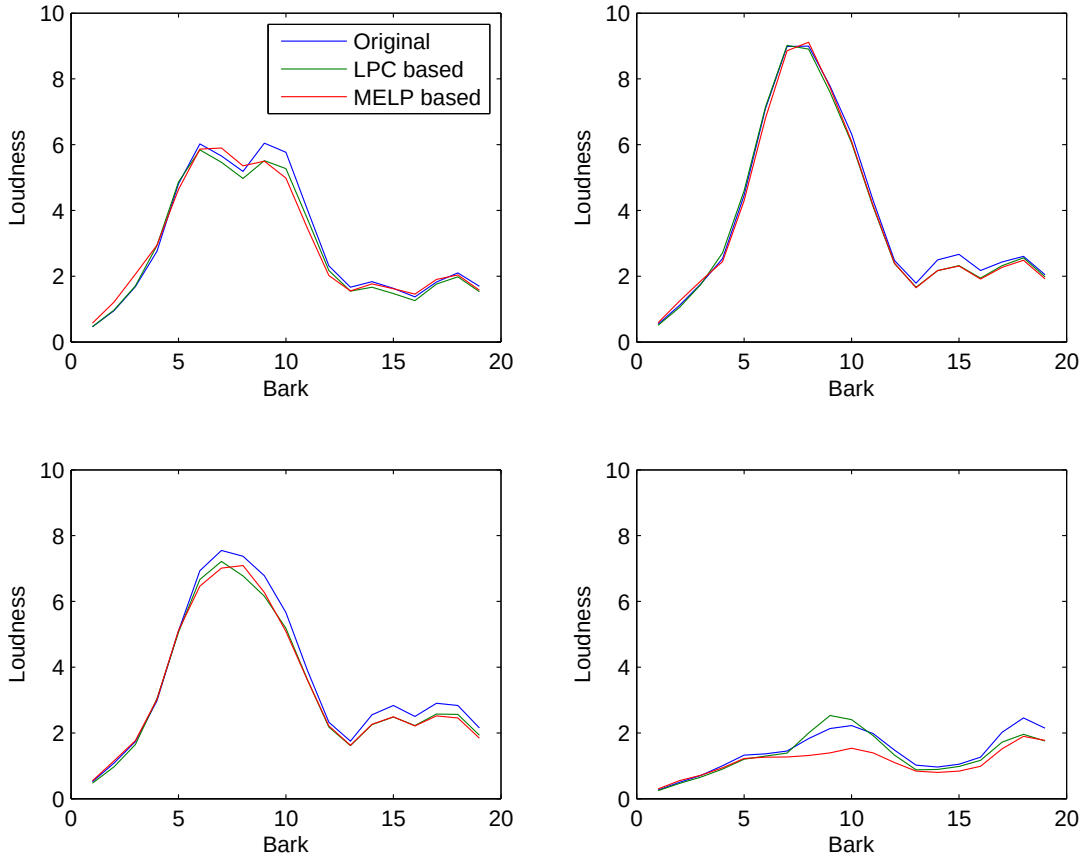


Figure 4.2: Bark spectra for four different segments of speech.

The BSD for the kth segment is given by

$$BSD^{(k)} = \sum_{i=1}^{N} \frac{\left[L_x^{(k)}(i) - L_y^{(k)}(i)\right]^2}{\left[L_x^{(k)}(i)\right]^2} \tag{4.5}$$

where N is the number of critical bands, $L_x^{(k)}$ the bark spectrum of the kth segment of original speech and $L_y^{(k)}$ the bark spectrum of the kth segment of coded speech. We used 19 critical bands.

A major drawback of the BSD is, that it only works for voiced speech. In the calculation of the BSD we only used voiced speech segments.

Chapter 5

Results

5.1 Vocoder

In a first step we worked with the MLSA filter on its own without considering the HMM. Therefore, the mel cepstral coefficients as well as the parameters for the source excitation generation have been extracted from recorded data.

5.1.1 Pulse Forms

In a first experiment we tried to find a function to replace the pulse. Previously, the pulse that we used in the pulse/noise based source excitation was a simple delta function (cf. figure 2.3(a)). Delta functions for pulses have the disadvantage that their DFT result in delta functions with high frequencies instead of reflecting the corresponding frequency of the pulse. This effect is illustrated in figure 5.1.

However, by using other functions for the pulse generation we had to recognize that the range for modifying the pulse is very narrow. Widening the pulse form resulted in smearing the synthesized speech, while a pulse form approaching the delta function resulted in sharpening the synthesized speech.

5.1.2 MELP

Several approaches have been proposed to eliminate the synthetic sound of LPC vocoder models. This includes e.g. CELP, RELP and MELP. MELP has also been used with some success in HMM based speech synthesis. Yoshimura et al. [YTM+01] proposed the idea to incorporate MELP parameters into the HMM based speech synthesis.

Figure 5.2 shows an example of a voicing transition from unvoiced to voiced speech. The MELP based excitation can model a smooth transition. In contrast, standard pulse/noise based excitation suddenly changes from unvoiced to voiced speech.

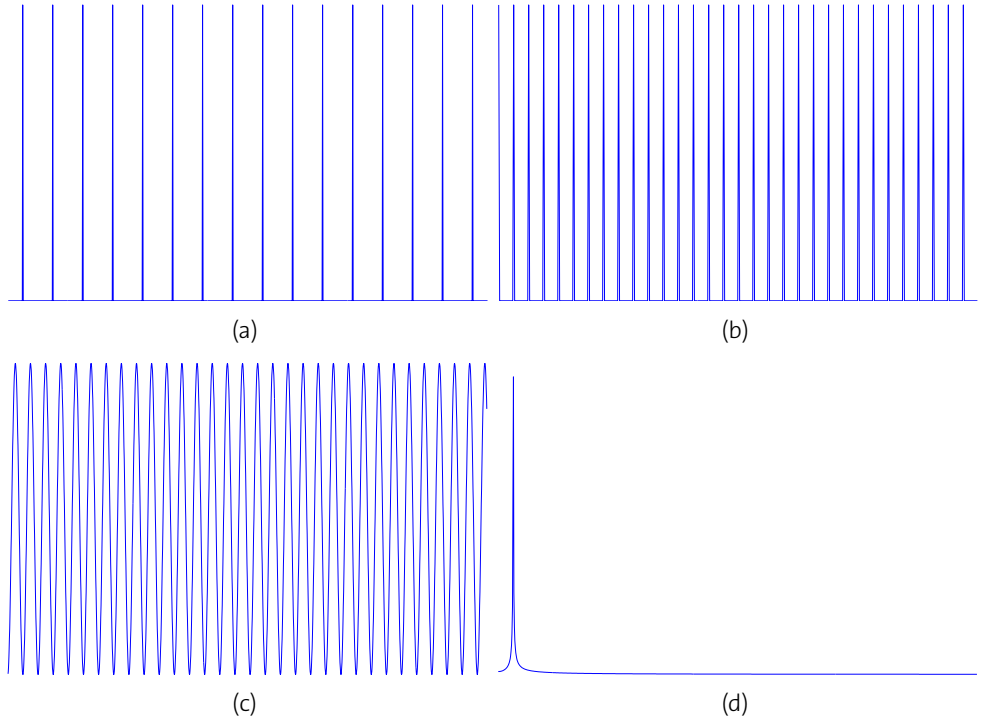


Figure 5.1: The DFT of a delta function (a) results in another delta function with a high frequency (b) instead of reflecting the corresponding frequency. In contrast, the DFT of a simple sine function (c) reflects its frequency (d).

In a setup as shown in figure 5.3 we extracted the source excitation from the standard MELP coder proposed by McCree [MB95]. We tried to modify this MELP coder to our needs. We experimented with different filter orders and analysis window sizes. A major focus was to increase the sampling rate of 8 kHz of the standard MELP coder which is far to low for our purposes. However, we decided to use the wideband MELP coder proposed by Lin [LKL00]. This MELP coder has already been optimized for a sampling rate of 16 kHz.

Figure 5.4 shows how we created the source excitation based on the wideband MELP coder. The MELP parameters were calculated by the analysis stage of the wideband MELP coder.

By using the source excitation from the MELP coder we could eliminate the synthetic buzz. The audio samples myi_a01.voc.wav that uses pulse/noise based excitation and myi_a01.vocm.wav that uses the source excitation from the wideband MELP coder illustrate the differences. However, the use of the source excitation from the wideband MELP coder together with mel cepstral coefficients for synthesis has a lowpass muffling effect, since we did not yet use a postfiltering technique.

It also introduces another distortion, which may be due to the fact, that the bandwidths in the wideband MELP coder are simply the bandwidths of the standard MELP coder multiplied by two. Therefore the higher frequency bands may be emphasized to much. Because of the logarithmic

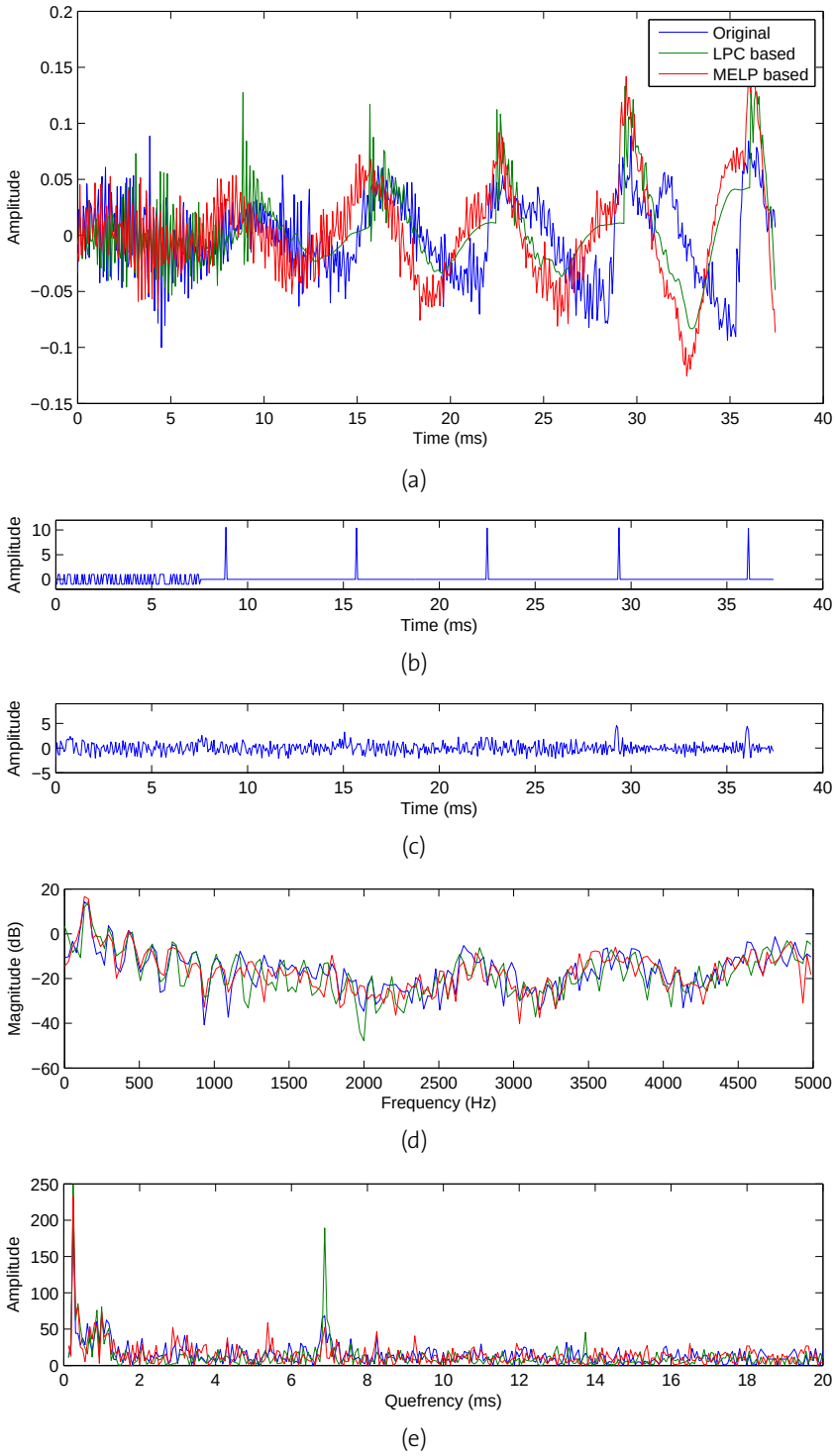


Figure 5.2: Waveform (a), spectrum (d) and cepstrum (e) for a speech frame with a transition from unvoiced to voiced speech. While the LPC based excitation (b) has a sudden change from voiced to unvoiced excitation, the MELP based excitation (c) can model a smoother transition.

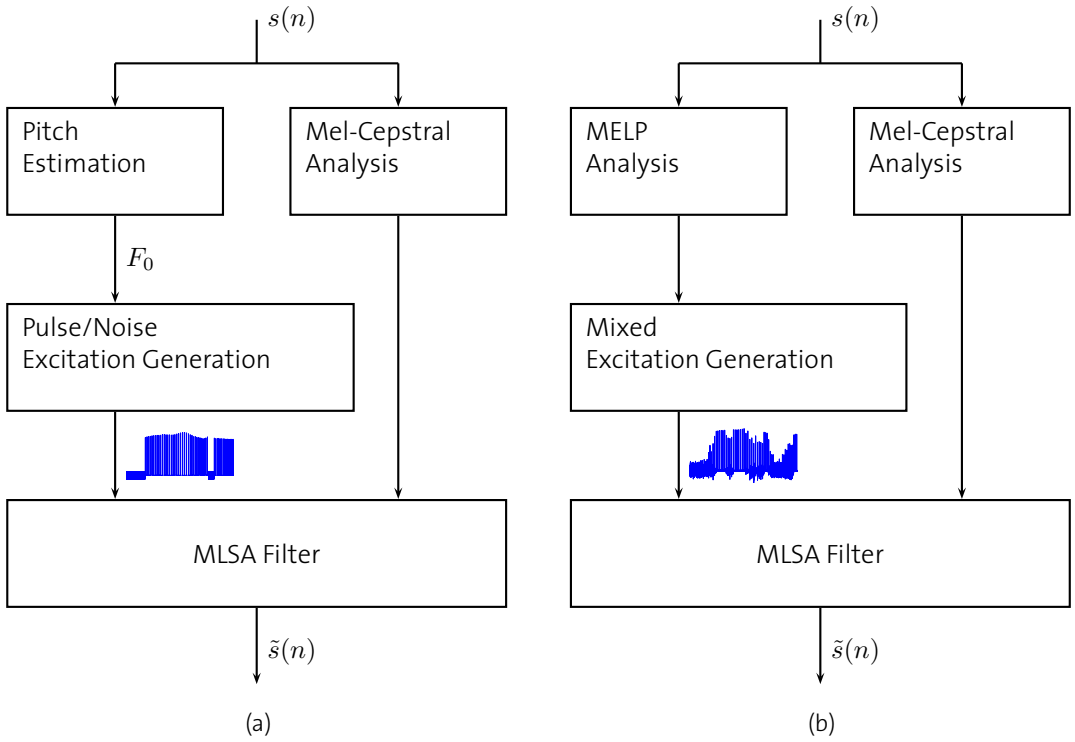


Figure 5.3: The setup used to evaluate LPC like source excitation (a) versus MELP based excitation (b).

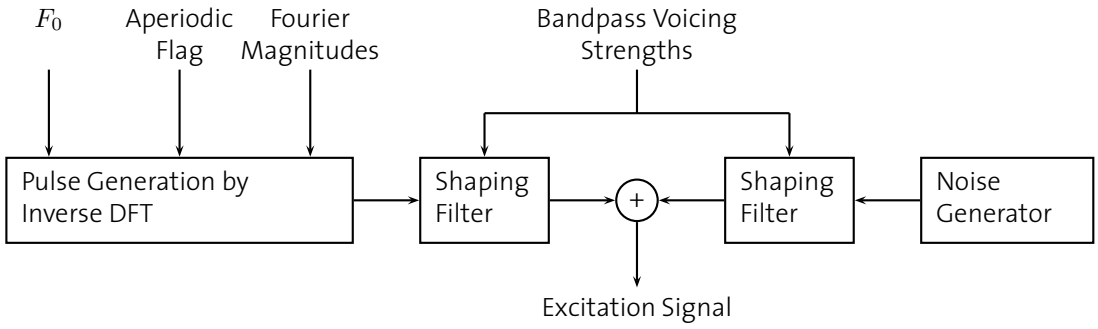


Figure 5.4: The Generation of the MELP based source excitation.

ATR (Male Japanese Speaker)		Average	Standard Deviation
MCC	Mixed Excitation	0.003732	0.000178
	Mixed Excitation (modified bands)	0.004375	0.000269
	Pulse/Noise Excitation	0.003316	0.000129
LSF	MELP coder	0.005846	0.000274
	MELP coder (modified bands)	0.006105	0.000282

Table 5.1: Average cepstral distortion for approxamately 38 minutes of speech data of a male Japanese speaker from the ATR speech corpus.

GlobalPhone (Male German Speaker)		Average	Standard Deviation
MCC	Mixed Excitation	0.003036	0.000107
	Mixed Excitation (modified bands)	0.003732	0.000165
	Pulse/Noise Excitation	0.002544	0.000114
LSF	MELP coder	0.004575	0.000311
	MELP coder (modified bands)	0.004894	0.000355

Table 5.2: Average cepstral distortion for approxamately 15 minutes of speech data of a male German speaker from the GlobalPhone speech corpus.

behavior of the human ear we experimented with mel scale sized frequency bands. However, we could not get a perceptual improvement of the quality by using such modified frequency bands. Although we think this two distortion effects to be perceptually far less disturbing than the buzzy sound of the pulse/noise excited synthesis, it has a considerable impact on distortion measurements.

For further investigations we made distortion measurements. We used the two different methods described in chapter 4. For this objective evaluation, we resynthesized the recorded speech data from one of the Japanese male speakers of the ATR speech corpus and a German male Speaker from the GlobalPhone speech corpus. The 500 utterances from the Japanese speaker summed up to about 38 minutes of audio data, the 162 utterances from the German male speaker summed up to about 15 minutes. The results are shown tables 5.1–5.4.

Even though the decoded output from the wideband MELP sounds very natural it got very bad results compared to the speech synthesized with the MLSA filter. A possible reason may be the

ATR (Male Japanese Speaker)		Average	Standard Deviation
MCC	Mixed Excitation	0.279637	0.169124
	Mixed Excitation (modified bands)	0.351270	0.184925
	Pulse/Noise Excitation	0.170400	0.181339
LSF	MELP coder	0.676615	0.244784
	MELP coder (modified bands)	0.676350	0.245344

Table 5.3: Average BSD for approximately 38 minutes of speech data of a male Japanese speaker from the ATR speech corpus.

GlobalPhone (Male German Speaker)		Average	Standard Deviation
MCC	Mixed Excitation	0.134708	0.041371
	Mixed Excitation (modified bands)	0.183611	0.053087
	Pulse/Noise Excitation	0.073733	0.029991
LSF	MELP coder	0.688848	0.223638
	MELP coder (modified bands)	0.689751	0.223065

Table 5.4: Average BSD for approximately 15 minutes of speech data of a male German speaker from the GlobalPhone speech corpus.

fact that improvement of the MELP coder is the better model for unvoiced speech and partially voiced speech. But distortion measurement methods have difficulties reflecting the human perception for unvoiced speech, or even fail as in the case of the BSD, and only work well for strong voiced speech segments. This result may suggest that MCCs are more appropriate for speech synthesis than LSFs.

A comparison of the mixed and the pulse/noise based excitation shows, that there is some distortion present in the mixed excitation. This is due to the lowpass muffling effect of the mixed excitation and another distortion that may be due to a too strong noise signal for certain bands.

We used two version of the wideband MELP coder in this objective evaluation. One had the original frequency bands proposed by Lin [LKL00] of 0–1000 Hz, 1000–2000 Hz, 2000–4000 Hz, 4000–6000 Hz and 6000–8000 Hz. In the other version we had modified the bands to have five bands of equal bandwidth on the mel scale. The bands were thus set 0–459 Hz, 459–1218 Hz, 1218–2475 Hz, 2475–4556 Hz and 4556–8000 Hz. It is interesting to note, that the modification of the frequency bands had almost now effect on the MELP coder, but significantly influenced the distortion measurements for speech synthesized with MCC and the mixed excitation signal from the MELP coder.

5.1.3 Simplified Mixed Excitation

During the process of building the HMM we implemented a simplified model for the source excitation generation. The most basic part of the MELP coder are the bandpass voicing strengths. We first built a very simple model that only used a pitch value and the bandpass voicing strengths. A possible weakness of the MELP coder may be that if a band is not consider voiced it is assumed as unvoiced. But this may not be true, the band may just be a weak powered band. In the MELP coder this is most probably compensated by using signal probabilities in the spectral enhancement filter and by using the gain values. However, in our simplified mixed excitation generation model this produced a strong distortion. To solve this problem we first tried different methods of weighting the noise signal depending on the bandpass voicing strengths. We then used the gain value from the MELP coder to scale the noise signal. By using this very simple method we got good results. The simplified source excitation generation model we built uses a pitch value,

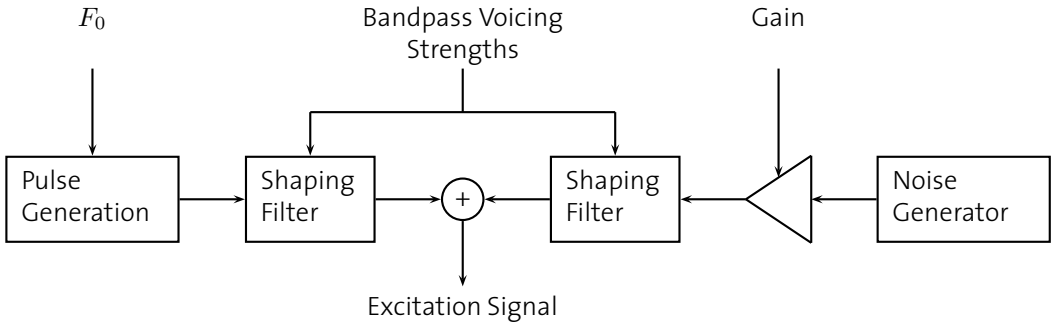


Figure 5.5: The model for the generation of the simplified mixed source excitation signal. It is only based on the pitch F_0, the bandpass voicing strengths and a gain value we used to scale the noise signal.

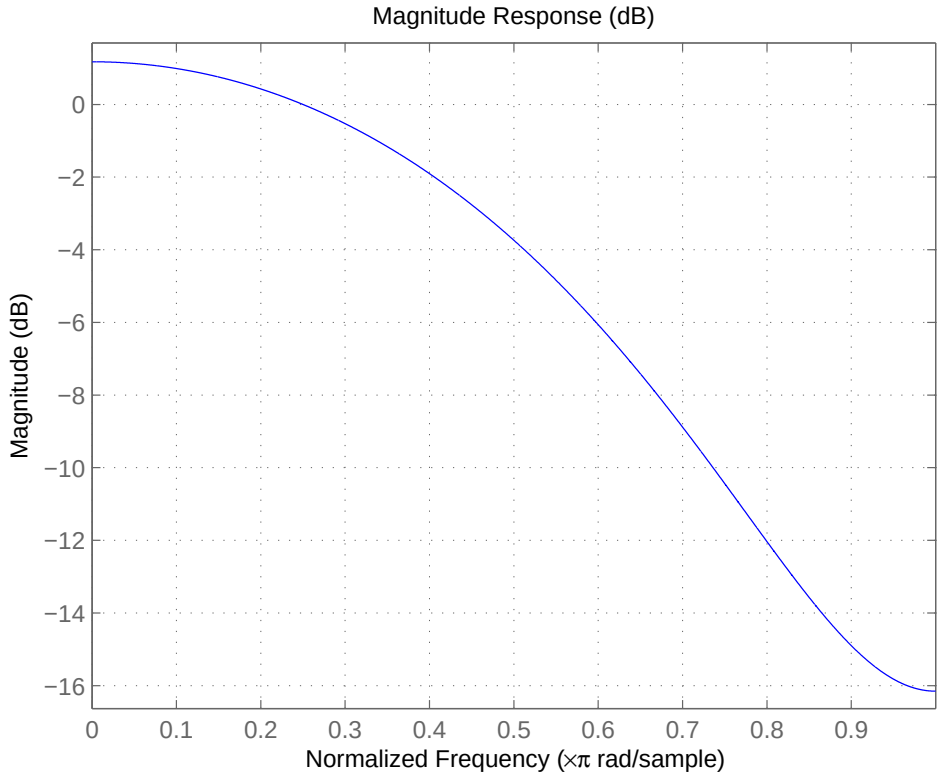


Figure 5.6: The magnitude response of the second order FIR filter we used to shape the white noise.

Parameter	Model M1	Model M2	Model M3
MCC	√		√
Bandpass Voicing Strenghts		√	√
Aperiodic Flag		√	√
Fourier Magnitudes		√	√
Gain		√	√

Table 5.5: The three different Models that were trained for evaluation.

the bandpass voicing strengths and a gain value. Figure 5.5 shows a diagram of this model.

We also tried to filter the white noise signal by a second order FIR shown in figure 5.6 to weaken the noise in higher frequencies. The function for this filter $H(z)$ is

$$H(z) = 0.2473 + 0.6503z^{-1} + 0.2473z^{-2}. \tag{5.1}$$

However, we could not find any difference in the synthesized speech.

A major advantage of this simplified mixed excitation generation model is, that it can be easily incorporated into HMM.

5.2 HMM

In the system we used before (see figure 1.1) the HMM was trained by mel cepstral and mel cepstral delta coefficients. Given the input text and duration information, the HMM was used to calculate mel cepstral coefficients. This coefficients were passed to a MLSA filter to synthesize the speech. This filter was excited with a pulse/noise signal according to the pitch information. For our tests we used pitch and duration information that are estimated from recorded data. In future, these will be generated from the text using a prosodic model.

5.2.1 HMM Training

We built our system mainly from components provided by the *Hidden Markov Model Toolkit* (HTK) [YOVW05]. An outline of the system for the training of the HMM is shown in figure 5.7.

We trained three different models shown in table 5.5 with the data of a male Japanese speaker from the ATR speech corpus. By setting the order of the cepstrum lifter to 20 the size of the feature vector containing the MCC is 21. For each frame there are five parameters for the bandpass voicing strengths, one for the aperiodic flag, ten Fourier magnitudes and two gain values.

We encountered some problems with the parameter generation from the HMM for large feature vectors. Numerical underflows caused the program that implements the algorithm explained in subsection 3.2.3 to produce parameter sequences of bad quality. Therefore the model M3 has not been used for further testing. In comparison with the samples myi_a01.m2.wav and

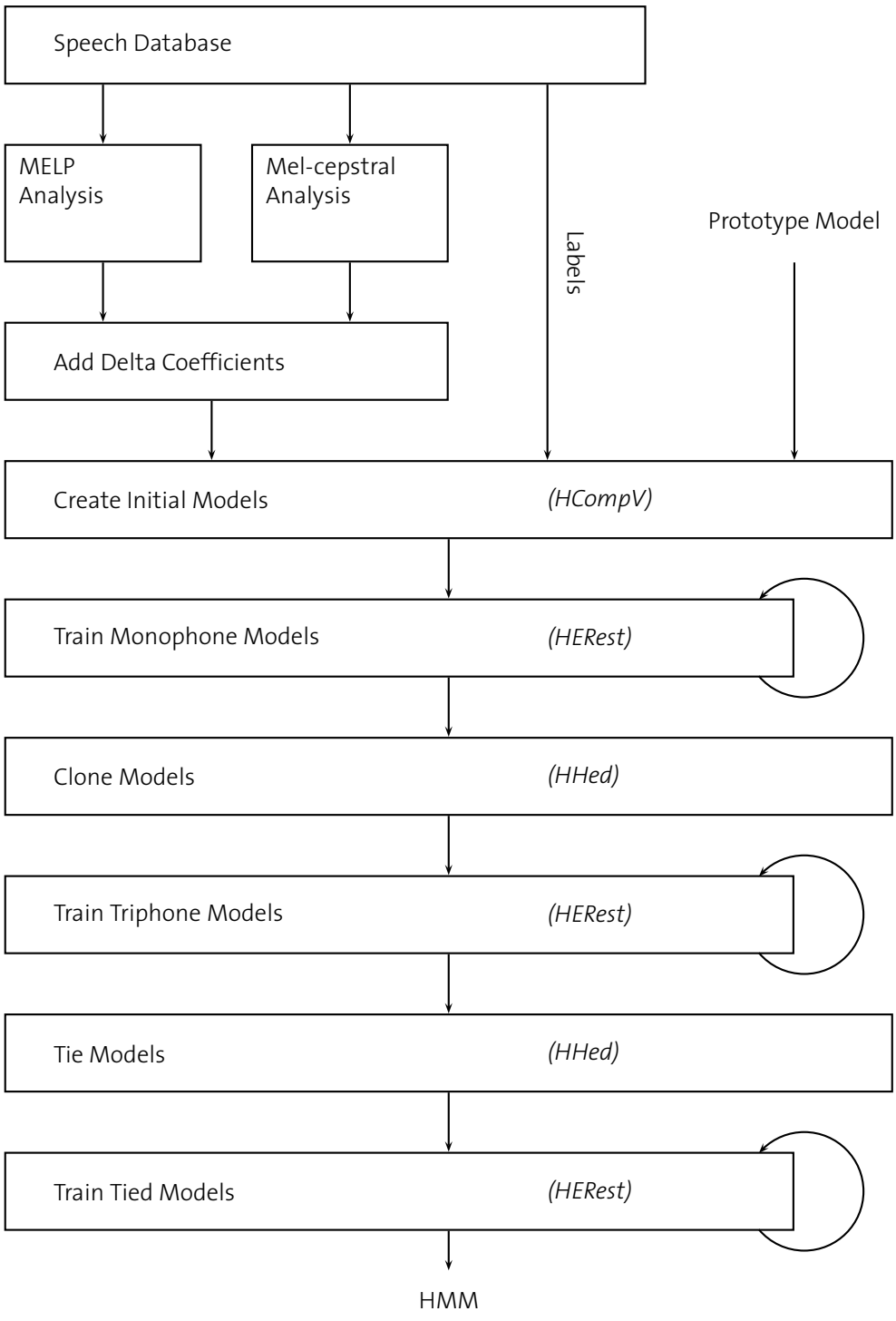


Figure 5.7: Overview of the training for the HMM based TTS system. The system is mainly built with components from HTK.

myi_a01.m2w.wav that were synthesized with parameters extracted from the models M1 and M2, the samples myi_a01.m1.wav and myi_a01.m1w.wav synthesized with parameters from the model M3 point out this degradation of quality.

For the further discussion the following three different types of synthesized speech are important:

Pulse/Noise Excitation The MCC are generated from model M1 and the speech is synthesized by using pulse/noise source excitation. An example is the sample myi_a01.old.wav.

Simplified Mixed Excitation The MCC are generated from model M1. By using the bandpass voicing strengths and gain values generated from model M2, the source excitation is created with the simplified mixed excitation generation model introduced in subsection 5.1.3. Sample myi_a01.m2.wav is such an example.

MELP Based Excitation The MCC are generated from model M1. The source excitation is created with the wideband MELP coder using the bandpass voicing strengths, aperiodic flag and Fourier magnitudes generated from model M2. An example of this kind is the audio sample myi_a01.m2w.wav.

By using the mixed excitation the synthetic buzz is eliminated. Another problem with the pulse/noise excitation in the HMM based TTS Synthesizer is that sometimes the generated MCC are poorly aligned with the pitch data. This results in a creaking distortion that occurs especially at the beginning of words, after pauses of the speech. Due to smoother modelling of the voicing transitions, this distortion can also be removed by using mixed excitation. However, the synthesized speech using pulse/noise excitation is sharper and cleaner. The speech that is synthesized by using mixed excitation suffers from a lowpass muffling effect and another light distortion, that is stronger for the MELP based excitation than for the simplified mixed excitation.

5.2.2 Subjective Test

We conducted an subjective listening test with nine subjects. Ten sentences that were not included in the training data were synthesized with pulse/noise excitation, MELP based excitation and simplified mixed excitation, respectively. Table A.2 lists the samples used in this test. For each of the 30 pairings the subject had to decide which of the two samples had better sound quality. After each decision the next pair was selected at random and then presented to the subject.

The result of the listening test were quite unexpected. Experts who knew the details about this research projects preferred the mixed excitation over the pulse/noise excitation. Because of the stronger distortion in the MELP based excitation they also preferred the simplified mixed excitation over the MELP based excitation. However, the subjects of the listening test preferred the pulse/noise excitation over the simplified mixed excitation by a 4:3 ratio. Anyway, a statistical analysis of the test result shows, that the error probability of this 4:3 ratio is higher than 10%.

Therefore the test result of this pairing is not reliable. The subjects were also indecisive about the preference of MELP based excitation and the pulse/noise excitation. The MELP based excitation was preferred over the simplified mixed excitation by a 2:1 ratio.

This unexpected result is subject to further investigations. The subjects may have considered a clear voice as a more important factor for quality than the naturalness of the voice. By using a postfilter similar to the pulse dispersion filter of the MELP coder or the postfilter proposed by Kishimoto et al.[KZT+02] we may obtain different results. For comparison, samples myi_a01.old.p.wav, myi_a01.m2.p.wav and myi_a01.m2w.p.wav have been postfiltered with the filter proposed by Kishimoto.

Chapter 6

Conclusion

The synthetic buzz of speech synthesized with a pulse/noise excitation can be eliminated by using mixed excitation. The creaking distortions due to poor alignment of the excitation and the MCC can also be removed by using mixed excitation. However, mixed excitation introduces a lowpass muffling effect.

Yoshimura et al. [YTM+01] proposed to use a mixed excitation model derived from the wideband MELP coder proposed by Lin et al. [LKL00], even though they are using mel cepstral coefficients for synthesis. But this MELP coder works on classical LPC coefficients transformed into LSFs for transmission. The additional features besides the mixed excitation are therefore mainly aimed to improve the quality of speech synthesized with LPC coefficients. Based on the results we got by using our simplified mixed excitation generation model, we also think that the basic principal of mixed excitation is a powerful option to improve the quality of a MLSA filter based synthesizer. However, we think that the model has first to be adapted to the characteristics of the synthesis with a MLSA filter.

Moreover, recent work by Pérez and Bonafonte [PB05] and Dinther et al. [vDVK05] on glottal pulse parameterization may also inspire new ideas for parametric residual generation.

The unexpected result of the subjective test has still to be analyzed. To further improve the quality of the synthesized speech it is also important to understand why one sample is preferred over another one.

Chapter 7

Further Work

7.1 Postfilter

The next step that has to be made is the evaluation of using a postfilter technique. In future evaluation we want to use the postfilter proposed by Kishimoto et al.[KZT+02] that is designed to improve the quality of speech synthesized by a MLSA filter.

7.2 Simplified Mixed Excitation

Our model for the simplified mixed excitation generation may be further developed into a model for mixed excitation generation adapted for MLSA filter synthesis. Also the analysis to retrieve the parameters from the audio data should be redesigned, rather then using the analysis of the wideband MELP coder. Other ideas include to have adaptive frequency bands, e.g. based on the cepstrum and pitch.

7.3 Speaker Adaption

The speaker adaptation is done with unconstrained maximum likelihood linear regression (MLLR) by using the HERest tool from HTK. The transformation matrices are obtained by solving a maximization problem using the Expectation-Maximization (EM) technique. Figure 7.1 shows a diagram of the speaker adaptation using HTK. As of version 3.3 which is the latest by now, the HTK [YOVW05] supports speaker adaptive training (SAT) with constrained maximum likelihood linear regression (CMLLR) transforms. This new feature should help to improve the speaker adaptation.

When it comes to polyglot synthesis, it is particularly difficult to get good performance for target speakers who's training data does not include the target language. By using CMLLR and SAT in our polyglot HMMs we hope to increase the similarity between the original and synthesized voiced of such a target speaker.

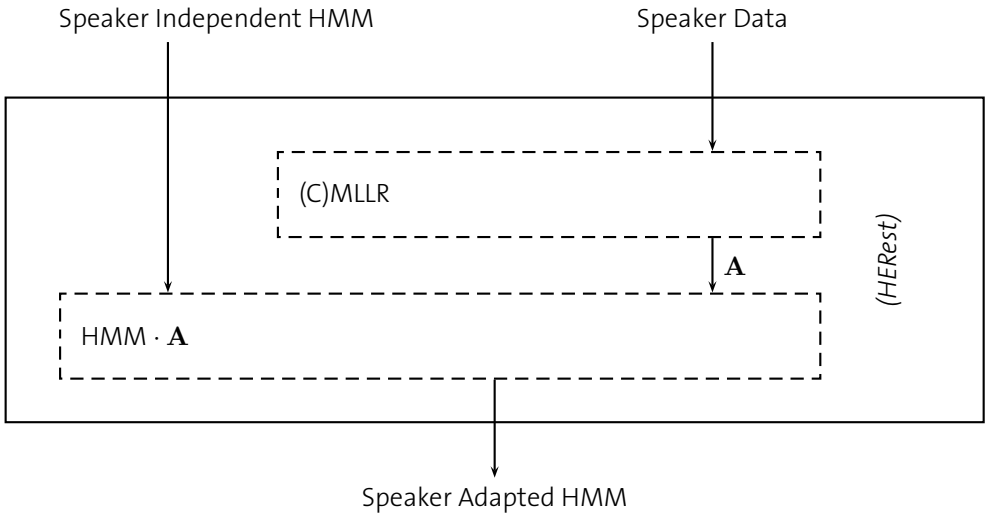


Figure 7.1: In speaker adaptation, HERest from the HTK first calculates an transformation $\mathbf{A}$ that is applied to the HMMs.

7.4 Polyglot HMM Based Synthesis

The mixed source excitation methods discussed in this thesis still need to be incorporated into our polyglot HMM based TTS synthesizer. Until now mixed excitation has only be tested with monolingual models.

Appendix A

Samples

The audio samples that are included in this thesis as listening examples all derive from the same utterance. It is a sample of a male Japanese speaker from the ATR speech corpus. The text of this sample reads

あらゆる 現実(げんじつ) をすべて 自分(じぶん) のほうへねじ 曲(ま) げたのだ。

The transcription in *kunreishiki* reads "*arayuru genjitu wo subete jibun no hou e nejimageta no da*."

Parameter Source	Method	Sample
	Original Audio Data	myi_a01.rec.wav
Recorded Data	Pulse/Noise	myi_a01.voc.wav
	Wideband Melp	myi_a01.vocm.wav
HMM (Models M1 and M2)	Pulse/Noise	myi_a01.old.wav
	Simplified Melp	myi_a01.m2.wav
	Wideband Melp	myi_a01.m2w.wav
HMM (Models M1 and M2)	Pulse/Noise (postfiltered)	myi_a01.old.p.wav
	Simplified Melp (postfiltered)	myi_a01.m2.p.wav
	Wideband Melp (postfiltered)	myi_a01.m2w.p.wav
HMM (Model M3)	Simplified Melp	myi_a01.m1.wav
	Wideband Melp	myi_a01.m1w.wav

Table A.1: Audio samples referenced in this thesis.

Pulse/Noise Excitation	Simplified Mixed Excitation	MELP based Excitation
myi_a34.old.wav	myi_a34.m2.wav	myi_a34.m2w.wav
myi_b43.old.wav	myi_b43.m2.wav	myi_b43.m2w.wav
myi_d30.old.wav	myi_d30.m2.wav	myi_d30.m2w.wav
myi_e16.old.wav	myi_e16.m2.wav	myi_e16.m2w.wav
myi_e26.old.wav	myi_e26.m2.wav	myi_e26.m2w.wav
myi_g04.old.wav	myi_g04.m2.wav	myi_g04.m2w.wav
myi_h44.old.wav	myi_h44.m2.wav	myi_h44.m2w.wav
myi_i11.old.wav	myi_i11.m2.wav	myi_i11.m2w.wav
myi_i36.old.wav	myi_i36.m2.wav	myi_i36.m2w.wav
myi_i41.old.wav	myi_i41.m2.wav	myi_i41.m2w.wav

Table A.2: This audio samples have been used in the informal listening test described in subsection 5.2.2. They were synthesized by parameters from the HMMs mentioned in subsection 5.2.1.

Appendix B

Tools

In the scope of this thesis we have used, modified and implemented several tools.

Tools that we have used:

HTK To build the HMM we used the *Hidden Markov Model Toolkit* (HTK) [YOVW05].

SPTK We used several tools from *Speech Signal Processing Toolkit* (SPTK), e.g. the MLSA filter *mlsadf*.

MATLAB MATLAB was often used to analyze synthesized speech and excitation signals, to design filters etc.

Tools that we have modified:

MELP First, we used the implementation by Texas Instruments, Inc. of the standard MELP coder as a basis for our experiments with the MELP model.

Wideband MELP Later, we worked with the wideband MELP coder by Lin [Lin00].

Tools that we have implemented:

Simplified Mixed Excitation Generator For a proof of concept we have implemented a simplified mixed excitation generator. This model turned out to be useful for further research.

Listening Test Interface To conduct the listening test we built a web interface in php using a mysql database. This helped us to have generated on the fly an individually randomized test for each subject.

Distortion Measurements The distortion measurements have been implemented as MATLAB scripts.

Data Handling We wrote several tools for data conversion, handling and manipulation, e.g. for the MELP delta parameter creation for HMM training.

List of Figures

List of Tables

Bibliography

[AH71] B. Atal and S. Hanauer. Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, 50(2):637–655, August 1971.

[CG87] Juin-Hwey Chen and Allen Gersho. Real-time vector apc speech coding at 4800 bps with adaptive postfiltering. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 2185–2188, Dallas, TX, USA, April 1987.

[FTKI92] Toshiaki Fukuda, Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai. An adaptive algorithm for mel-cepstral analysis of speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 137–140, San Francisco, CA, USA, March 1992.

[Ima83] Satoshi Imai. Cepstral analysis synthesis on the mel frequency scale. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 93–96, Boston, MA, USA, May 1983.

[IS68] F. Itakura and S. Saito. Analysis synthesis telephony based on the maximum like lihood method. In Y. Kohashi, editor, *6th International Conference on Acoustics*, pages C17–20, Tokyo, Japan, August 1968.

[ISF83] Satoshi Imai, K. Sumita, and C. Furuichi. Mel log spectrum approximation (MLSA) filter for speech synthesis. *Transactions of the IECE of Japan*, J66-A:122–129, February 1983.

[KZT+02] Y. Kishimoto, H. Zen, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. A postfiltering technique for HMM-based speech synthesis. In *Proceedings of Autumn Meeting of the Acoustical Society of Japan (ASJ)*, volume 1, pages 279–280, Akita, Japan, September 2002.

[LIF05a] Javier Latorre, Koji Iwano, and Sadaoki Furui. Cross-language synthesis with a polyglot synthesizer. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, pages 1477–1480, Lisbon, Portugal, September 2005.

[LIF05b] Javier Latorre, Koji Iwano, and Sadaoki Furui. Polyglot synthesis using a mixture of monolingual corpora. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages SP–L1.1, Philadelphia, PA, USA, March 2005.

[LIF05c] Javier Latorre, Koji Iwano, and Sadaoki Furui. Speaker adaptable multilingual synthesis. In *Symposium on Large-Scale Knowledge Resources (LKR2005)*, pages 235–238, Tokyo, Japan, March 2005.

[Lin00] Weiran Lin. Wideband speech coding based on mixed excitation. Master's thesis, School of Electrical & Electronic Engineering, Nanyang Technological University, 2000.

[LKL00] Weiran Lin, Soh Ngee Koh, and Xiao Lin. Mixed excitation linear prediction coding of wideband speech at 8 kbps. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume II, pages 1137–1140, Istanbul, Turkey, June 2000.

[Mai88] X. Maitre. 7 kHz audio coding within 64 kbit/s. *IEEE Journal on Selected Areas in Communications*, 6:283–298, February 1988.

[Mas02] Takashi Masuko. *HMM-Based Speech Synthesis and Its Applications*. PhD thesis, Tokyo Institute of Technology, November 2002.

[MB95] Alan V. McCree and Thomas P. Barnwell III. A mixed excitation LPC vocoder model for low bit rate speech coding. *IEEE Transactions on Speech and Audio Processing*, 3(4):242–250, 1995.

[McC99] Alan McCree. Adaptive filter and filtering method for low bit rate coding. United States Patent, October 1999. Patent number: 5,966,689.

[Mot98] Motion Picture Experts Group. *MPEG-4 Audio Final Committee Draft 14496-3*. International Organisation for Standardisation, 1998. ISO/IEC 14496-3.

[MTKI96] Takashi Masuko, Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai. Speech synthesis using HMMs with dynamic features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 389–392, Atlanta, GA, USA, May 1996.

[PB05] Javier Pérez and Antonio Bonafonte. Automatic voice-source parameterization of natural speech. In *Proceedings of the International Conference on Spoken Language Processing (Interspeech)*, pages 1065–1068, Lisboa, Portugal, September 2005.

[PHB04] Beat Pfister, Hans-Peter Hutter, and René Beutler. Sprachverarbeitung I. Skript zur Vorlesung. TIK, ETH Zürich, 2004.

[PHB05] Beat Pfister, Hans-Peter Hutter, and René Beutler. Sprachverarbeitung II. Skript zur Vorlesung. TIK, ETH Zürich, 2005.

[TKI95] Keiichi Tokuda, Takao Kobayashi, and Satoshi Imai. Speech parameter generation from HMM using dynamic features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 660–663, Detroit, MI, USA, May 1995.

[TMY+95] Keiichi Tokuda, Takashi Masuko, Tetsuya Yamada, Takao Kobayashi, and Satoshi Imai. An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 1, pages 757–760, Madrid, Spain, September 1995.

[vDVK05] R. van Dinther, R. N. J. Veldhuis, and A. Kohlrausch. Perceptual aspects of glottal-pulse parameter variations. *Speech Communication*, 46:95–112, May 2005.

[WSG92] Shihua Wang, Andrew Sekey, and Allen Gersho. An objective measure for predicting subjective quality of speech coders. *IEEE Journal on Selected Areas in Communications*, 10(5):819–829, June 1992.

[YOVW05] Steve Young, Dave Ollason, Valtcho Valtchev, and Phil Woodland. *The HTK Book (for HTK Version 3.3)*. Cambridge University Engineering Department, April 2005.

[YTM+01] Takayoshi Yoshimura, Keiichi Tokuda, Takashi Masuko, Takao Kobayashi, and Tadashi Kitamura. Mixed excitation for HMM-based speech synthesis. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, volume 3, pages 2263–2266, Aalborg, Denmark, September 2001.

[YW94] Steve J. Young and Phil C. Woodland. State clustering in hidden markov model based continuous speech recognition. *Computer Speech and Language*, 8:369–383, 1994.