Manu Airaksinen

Analysis/Synthesis Comparison of Vocoders Utilized in Statistical Parametric Speech Synthesis

School of Electrical Engineering

Thesis submitted for examination for the degree of Master of Science in Technology.

Espoo 19.11.2012

Thesis supervisor:

Prof. Paavo Alku

Thesis instructor:

M.Sc. (Tech.) Tuomo Raitio

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Manu Airaksinen

Title: Analysis/Synthesis Comparison of Vocoders Utilized in Statistical Parametric Speech Synthesis

Date: 19.11.2012 Language: English Number of pages:8+113

Department of Signal Processing and Acoustics

Professorship: Acoustics and audio signal processing Code: S-89

Supervisor: Prof. Paavo Alku

Instructor: M.Sc. (Tech.) Tuomo Raitio

This thesis presents a literature study followed by an experimental part on the state-of-the-art vocoders utilized in statistical parametric speech synthesis. In the experimental part, the analysis/synthesis properties of three selected vocoders (GlottHMM, STRAIGHT and Harmonic/Stochastic Model) are examined. The performed tests were the analysis of vocoder parameter distributions, statistical testing on the effect of emotions to the vocoder parameter distributions, and a subjective listening test evaluating the vocoders' relative analysis/synthesis quality.

The results indicate that the STRAIGHT vocoder has the most Gaussian parameter distributions and most robust synthesis quality, whereas the GlottHMM vocoder has the most emotion sensitive parameters and best but unreliable synthesis quality. The HSM vocoder's LSF parameters were found to be more Gaussian than the GlottHMM vocoder's LSF parameters. HSM was found to be sensitive to noise, and it scored the lowest score on the subjective listening test.

Keywords: vocoder, speech synthesis, HMM, vocoder parametrization, analysis/synthesis, statistical distribution, GlottHMM, STRAIGHT, HSM

AALTO-YLIOPISTO
SÄHKÖTEKNIIKAN KORKEAKOULU

DIPLOMITYÖN
TIIVISTELMÄ

Tekijä: Manu Airaksinen

Työn nimi: Tilastollisessa parametrisessa puhesynteesissä käytettyjen vokooderien analyysi-synteesi-vertailu

Päivämäärä: 19.11.2012 Kieli: Englanti Sivumäärä:8+113

Department of Signal Processing and Acoustics

Professuuri: Akustiikka ja äänenkäsittely Koodi: S-89

Valvoja: Prof. Paavo Alku

Ohjaaja: DI Tuomo Raitio

Tässä työssä esitetään kirjallisuuskatsaus ja kokeellinen osio tilastollisessa parametrisessa puhesynteesissä käytetyistä vokoodereista. Kokeellisessa osassa kolmen valitun vokooderin (GlottHMM, STRAIGHT ja Harmonic/Stochastic Model) analyysi-synteesi -ominaisuuksia tarkastellaan usealla tavalla. Suoritetut kokeet olivat vokooderiparametrien tilastollisten jakaumien analysointi, puheen tunnetilan tilastollinen vaikutus vokooderiparametrien jakaumiin sekä subjektiivinen kuuntelukoe jolla mitattiin vokooderien suhteellista analyysi-synteesi -laatua.

Tulokset osoittavat että STRAIGHT-vokooderi omaa eniten Gaussiset parametrijakaumat ja tasaisimman synteesilaadun. GlottHMM-vokooderin parametrit osoittivat suurinta herkkyyttä puheen tunnetilan funktiona ja vokooderi sai parhaan, mutta laadultaan vaihtelevan kuuntelukoetuloksen. HSM-vokooderin LSF-parametrien havaittiin olevan Gaussisempia kuin GlottHMM-vokooderin LSF parametrit, mutta vokooderin havaittiin kärsivän kohinaherkkyydestä, ja se sai huonoimman kuuntelukoetuloksen.

Avainsanat: vokooderi, puhesynteesi, HMM, vokooderiparametri, analyysi-synteesi, tilastollinen jakauma, GlottHMM, STRAIGHT, HSM

Preface

This thesis work has been performed as a part of the Simple4All project, funded by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287678. The work was done at the Department of Signal Processing and Acoustics at Aalto University School of Electrical Engineering.

I would like to give special thanks to my supervisor, professor Paavo Alku for the great opportunity to work on this project, and to my instructor Tuomo Raitio for providing me with excellent guidance and motivation. Additional thanks are given to Hannu Pulakka for the help with the subjective listening test, Marko Takanen for the help with the statistical testing methods, and to Emma Jokinen and Henna Tahvanainen for helping me with various small problems along the way. I would like to give collective thanks to the Laboratory of Acoustics and Audio Signal Processing for having such an awesome and inspiring working environment.

On a personal level, I would like to thank my fiancée Anna for her continuous support and patience, as well as my parents who have always been there for me. This thesis is dedicated to my son Aaro.

Otaniemi, 19.11.2012

Manu Airaksinen

Contents

Abbreviations

AbS	Analysis by Synthesis
ANOVA	Analysis of Variance
AP	Aperiodicity coefficient of STRAIGHT vocoder
ASR	Automatic Speech Recognition
CCR	Comparison Category Rating
CELP	Code Excited Linear Prediction (codec)
CLT	Closed-loop training
CWT	Continuous Wavelet Transform
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DSM	Deterministic plus Strochastic Model (vocoder)
$F0$	Fundamental frequency
FFT	Fast Fourier Transform (algorithm)
FIR	Finite Impulse Response
GCI	Glottal Closure Instant
GMM	Gaussian Mixture Model
GPF	Glottal Post-filtering (vocoder)
GSS	Glottal-Spectral Separation (vocoder)
HMM	Hidden Markov Model
HNM	Harmonic plus Noise Model (vocoder)
HNR	Harmonic-to-Noise Ratio
HSM	Harmonic/Stochastic Model
HTS	HMM-based speech synthesis system
IAIF	Iterative Adaptive Inverse Filtering (algorithm)
IIR	Infinite Impulse Response
LF-model	Liljencrants-Fant model
LPC	Linear Predictive Coding
LSF	Line Spectral Frequency
LSP	Line Spectral Pair
MBE	Multiband Excitation (vocoder)
ME	Mixed Excitation (vocoder)
MELP	Mixed Excitation Linear Prediction (codec)
MFCC	Mel-Frequency Cepstral Coefficient
MGC	Mel-Generalized Cepstrum
MGLSA	Mel-Generalized Log Spectrum Approximation (filter)
MLSA	Mel Log Spectrum Approximation (filter)
MOS	Mean Opinion Score
MT	Machine Translation
PCA	Principal Component Analysis
PSOLA	Pitch-Synchronous Overlap-Add (algorithm)
PSRC	Pitch-Synchronous Residual Codebook (vocoder)
SPTK	Speech Signal Processing Toolkit
STFT	Short-Time Fourier Transform

STRAIGHT	Speech Transformation and Representation using Adaptive Interpolation of weiGHT spectrum (vocoder)
SWLP	Stabilized Weighted Linear Prediction
TBE	Two-Band Excitation (vocoder)
TTS	Text-to-Speech
VT	Vocal Tract

1 Introduction

In the field of text-to-speech (TTS) synthesis, the traditional *unit selection synthesis* has been lately challenged both in quality and utility by *statistical parametric speech synthesis*. Unit selection synthesis is based on the concatenation of prerecorded waveform snippets, which at best results in natural sounding synthetic speech. However, the problem of this method is that the sufficient modeling the sound-space outside the range of the database is difficult, and of poor quality. A solution for this problem if offered by statistical parametric speech synthesis, also referred to as *Hidden Markov Model* (HMM)-based speech synthesis, where the speech signal is compressed into analysis parameters, which are used in phoneme specific statistical models.

Statistical parametric speech synthesis offers many advantages compared to the traditional speech synthesis methods, such as a wider obtainable sound-space with significantly lower memory and processing requirements. One of the largest individual problems of statistical parametric speech synthesis is the conservation of quality of the speech signal when it is analyzed into parameters and then *synthesized* back in to a speech waveform. Especially for older methods, the obtained synthesis quality is inadequate. More sophisticated *vocoders* (analysis/synthesis algorithms) have been developed to amend this problem, and their ultimate goal is to provide natural-sounding synthesis waveforms.

Even though the state-of-the-art vocoders have been successfully implemented and used in the framework of HMM-based speech synthesis, their vocoder parameter distributions have not been studied or documented. Because the parameter distributions are usually modeled as a single Gaussian distribution per synthesis context, it could be valuable to know the suitability of each vocoder parameter type for such modeling. This information could lead to new refinements in the vocoder parameter types. Also, comparative studies concerning the vocoders' synthesis quality are hard to come by as well as studies that evaluate the analysis/synthesis quality of the vocoders, which can be thought of as the optimal quality that a vocoder can achieve.

The goal of this thesis is to conduct a literature study on the state-of-the-art vocoders utilized in statistical parametric speech synthesis, and based on it to select a small number of prominent vocoders from different methodological backgrounds for the experimental section. The experimental section consists of the analysis of the vocoder parameters' statistical distributions and a subjective listening test measuring the relative analysis/synthesis quality of the vocoders.

The thesis is structured as follows: Chapter 2 presents basic information about speech production, analysis, and synthesis. Chapter 3 presents the basic framework of statistical parametric synthesis, the HMM-based speech synthesis model. Chapter 4 contains the literature study on the state-of-the-art vocoders utilized in HMM-based speech synthesis. Chapter 5 presents the first section of the experimental part of the thesis, the statistical analysis of vocoder parameters, and the second section, the subjective listening test, is presented in Chapter 6. Overall discussion about the obtained results as well final conclusions of the thesis are presented in Chapter 7.

2 Speech Modeling

The understanding of the physical and mathematical models of speech production is essential for speech synthesis. In this chapter, the basics of human speech production, and its mathematical modeling in the context of digital filters is presented.

2.1 Speech Production

Speech is a form of human communication, where a sequence of sounds is used as a symbolic representation of information. The arrangement of these sounds is unique for each language, but since they are all produced by the *human vocal system*, their building blocks are the same for all languages. These building blocks are called *phonemes*, and their study and classification is called *phonetics*.

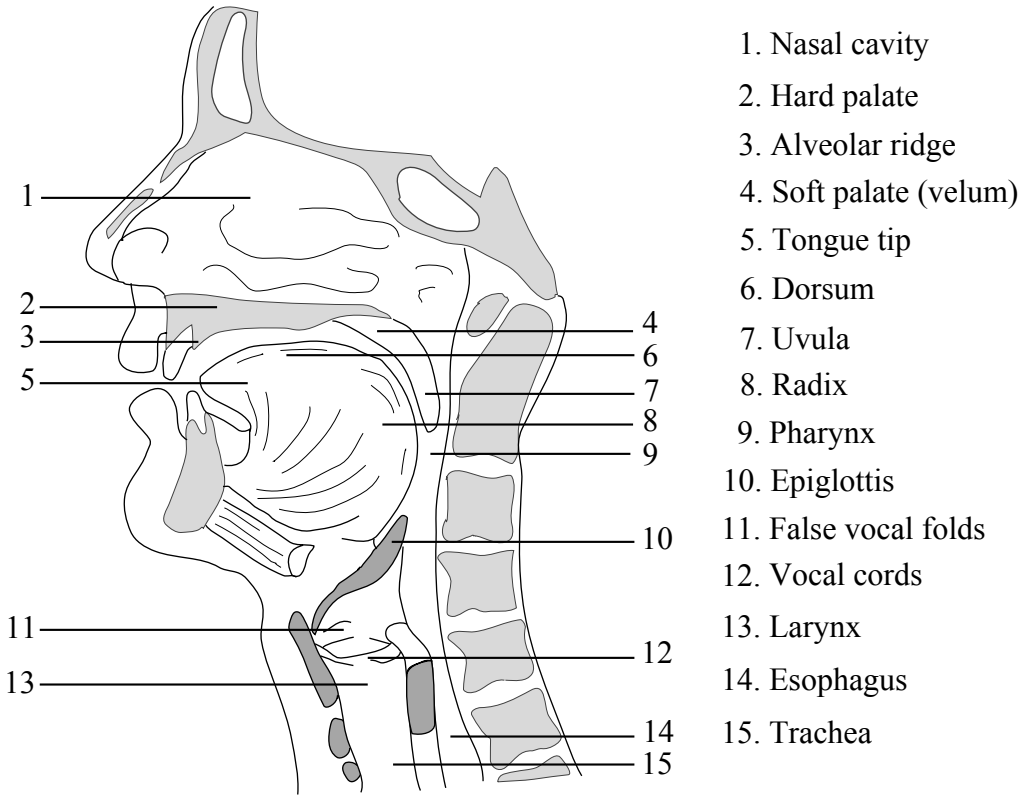


Figure 2.1: The human speech production system. [40]

The human vocal system along with its important features are illustrated in Figure 2.1. The *vocal tract* is the space ranging from the opening between the vocal cords (*glottis*) to the lips. The vocal tract can be divided into the *pharynx* (from the *esophagus* to the mouth) and the *oral cavity*. For an average male, the length of the vocal tract is about 17 cm, and the cross-sectional area varies from zero to about 20 cm^2 (depending of the positions of the tongue, lips, jaw, and velum) [61]. The *velum* is also used to control the acoustical coupling of the *nasal tract* to the vocal

tract. The sub-glottal system composed of the lungs, bronchi and trachea serves as a source of energy for the production of speech.

The sounds in speech can be divided into three categories according to their mode of excitation: *Voiced*, *unvoiced*, and *plosive* sounds. Voiced sounds are produced by blowing air through the glottis while the tension of the vocal cords is adjusted so that they vibrate in a relaxation oscillation. This produces quasi-periodic pulses of air flow which excite the vocal tract. Unvoiced sounds are generated by causing a constriction somewhere along the vocal tract and forcing air through it at a velocity that produces turbulence. This produces a noise source to excite the vocal tract. Plosive sounds are generated by making a complete closure somewhere along the vocal tract, and building up pressure behind it. When the closure is opened, the pressure is released as a burst of air-flow excites the vocal tract.

Given the excitation of the produced sound, the vocal and nasal tracts act as resonance tubes of non-uniform cross-sectional area similarly to organ pipes or wind instruments. The spectrum of the sound is shaped by the frequency selectivity of the tubes. The resonance frequencies of the vocal tract tube are called *formant frequencies* or *formants*, which are the most common spectral characteristics. *Anti-formants* are additional spectral characteristics, which are formed in nasal sounds (*nasals*), which are voiced sounds where the oral cavity is completely constricted at some point, and the air flows only through the nasal tract. In these sounds the oral cavity acts as an resonant cavity that traps acoustic energy at certain frequencies.

Different sounds are formed by varying the shape of the vocal tract (the positions of the formants), meaning that the two main characteristics of each sound used in human speech are the excitation and the spectral properties of the vocal tract.

The final property of human speech production is the transfer of the acoustical energy from the lips to the surrounding air. Because of the mismatch in acoustical impedance between the boundaries, the transfer of energy is not ideal especially in lower frequencies. Effectively this means that the so-called *lip radiation* effect acts as a high-pass filter to the outgoing sound.

2.2 Source-filter Theory

Various mathematical and physical models have been suggested for the modeling of speech. The most accurate models are based on direct physical modeling of the human speech production system, but their analytic solutions are too complex for most applications. A functional trade-off between the complexity and accuracy of the used model is the use of a *terminal analog* model, where speech is modeled as a linear system where the output has desired speech-like properties when controlled by a set of parameters that are somehow related to the process of speech production. Ideally, the output of the model is equivalent to the physical model, but the inner structure does not mimic the physical structure of speech production. For discrete-time digital systems, this means that the spectral properties are represented as digital filters, and the excitation is represented as a digital signal. This is known as the *source-filter model*, and it is discussed in further detail in Section 2.2.1.

The speech signal varies in time, thus meaning that the parameters of the used

model must also vary in time. A common assumption is that the properties of the excitation and vocal tract remain fixed for periods of 10-20 ms. This allows for the processing of the speech signal in consecutive *frames* that are assumed to be stationary.

2.2.1 Source-filter Model

The source-filter model used in the digital modeling of speech is a simple concept: the excitation signal $u_G(n)$ is used to excite a linear system $v(n)$ representing the vocal tract, and the product of their convolution is the produced speech-like signal $u_L(n)$:

$$u_L(n) = u_G(n) \circledast v(n), \tag{2.1}$$

or in z domain:

$$U_L(z) = U_G(z)V(z). \tag{2.2}$$

The vocal tract transfer function $V(z)$ can be sufficiently modeled as an all-pole filter of the form

$$V(z) = \frac{G}{1 - \sum_{k=1}^{p} \alpha_k z^{-k}}, \tag{2.3}$$

where G and $\{\alpha_k\}$ are dependent upon the properties of the vocal tract [61]. Because poles can be used only to model resonances, the modeling of the anti-formants in nasal sounds (which would require also zeros) is a problem for the all-pole representation. However, nasal sounds are relatively rare in most languages, and they can be adequately modeled by increasing the amount of poles [61].

The model for the excitation signal $u_G(n)$ is dependent on the type of the desired excitation: A quasi-periodic waveform is required for voiced sounds, and a random noise waveform is required for unvoiced and plosive sounds. A convenient representation for voiced speech is given by a system where an impulse train generated according to the *fundamental frequency*, $F0$, of the speech is used to excite a linear system $G(z)$ that is used to model the characteristics of the glottal pulse. The noise excitation is simply generated as white Gaussian noise, whose amplitude is scaled to the desired level.

The final part of the complete source-filter model is the inclusion of the lip radiation effect. As discussed in Section 2.2, the lip radiation effect acts as an high-pass filter to the outgoing sound. A sufficient approximation of this effect is achieved by using a first order differentiator $L(z)$:

$$L(z) = L_0(1 - z^{-1}), \tag{2.4}$$

where L_0 is a constant.

The block diagram for the complete source-filter model is presented in Figure 2.2. The output speech $S(z)$ can be represented as:

$$S(z) = E(z)G(z)V(z)L(z), \tag{2.5}$$

where $E(z)$ is the impulse/noise excitation signal, and $G(z) = 1$ for unvoiced sounds.

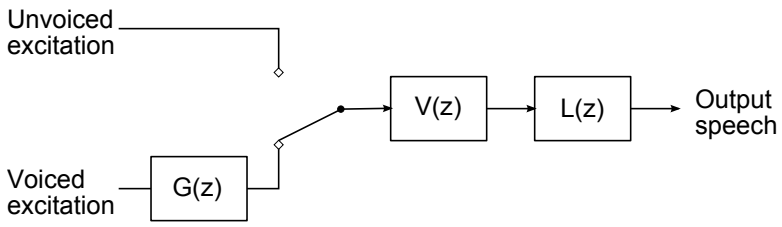


Figure 2.2: The block diagram for the complete source-filter model of speech production

The spectral properties $G(z)$, $V(z)$ and $L(z)$ of the linear system of Figure 2.2 are difficult to estimate separately with satisfactory accuracy. Because of that, it is usually convenient and useful to combine them into a single all-pole system $H(z)$:

$$H(z) = G(z)V(z)L(z) \tag{2.6}$$

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

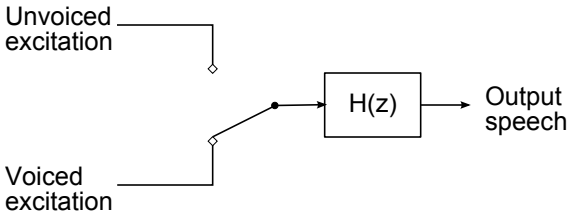


Figure 2.3: The block diagram for the unified source-filter model of speech production

Using this expression, the block diagram of the model can be presented as in Figure 2.3. This representation is the basis of almost every vocoding method discussed in this thesis, as well as the exact model used in the simple impulse excitation vocoder (see Section 4.1). More precisely, Figure 2.3 represents the *synthesis* block diagram of the proposed speech model. For the method to be usable, the parameters used by the synthesis method ($F0$, spectral envelope $H(z)$, and the excitation gains A_V and A_N) need to be estimated in some fashion. In statistical parametric text-to-speech synthesis, the parameter tracks are generated according to the text input as described in Section 3, but to be able to train the system, labeled training data from real speech signals is needed. The estimation of the *vocoder parameters* from a natural speech signal is called the *analysis phase*, and it can be seen as the inverse operation of the synthesis phase (which was derived according to the human speech production theory): the input to the system is a speech signal waveform, and the output is a number of vocoder parameters used by the source-filter model. A model of human speech with distinct analysis and synthesis procedures is called a *vocoder*, which is short for voice coder. However, it is notable that the analysis and synthesis operations of the source-filter model use simplified models, which makes the processes *lossy* and not perfectly invertible.

2.2.2 Source-filter Residual

For the single filter source-filter model of speech of Section 2.2.1, the analysis phase consists usually of the estimation of the all-pole spectral envelope $H(z)$, and the gain and $F0$ of the excitation. However, if $H(z)$ is known, the ideal excitation signal, or *residual*, $E(z)$ can be computed by filtering the original signal $S(z)$ with the inverse filter, or *analysis filter*, of the *synthesis filter* $H(z)$:

$$E(z) = \frac{1}{H(z)} S(z) \tag{2.8}$$

If $H(z)$ is an all-pole IIR filter, then $\frac{1}{H(z)}$ is a FIR filter with only zeros. Together with $H(z)$, the residual signal can be used to perfectly reconstruct the original signal in the synthesis phase. An example of the residual signal of the Finnish vowel [a] is presented in Figure 2.4(b). The residual signal is computed using an 18th-order LPC analysis filter estimation (see Section 2.3) of $H(z)$. It can be seen that the residual signal differs from the simplified binary pulse excitation (Figure 2.4(c)) in two main ways: The pulses in the signal are dispersed (not pure delta pulses), and the excitation has additive noise.

The spectrum of the residual signal can be seen in Figure 2.5(b) along with the spectrum of the original signal (Figure 2.5(a)) and the binary pulse signal (Figure 2.5(c)). The spectral envelope of the original signal has been flattened in the residual signal, but the harmonic structure has remained intact with its imperfections. Compared to the spectrum of the binary pulse, the residual signal's harmonic structure attenuates at higher frequencies.

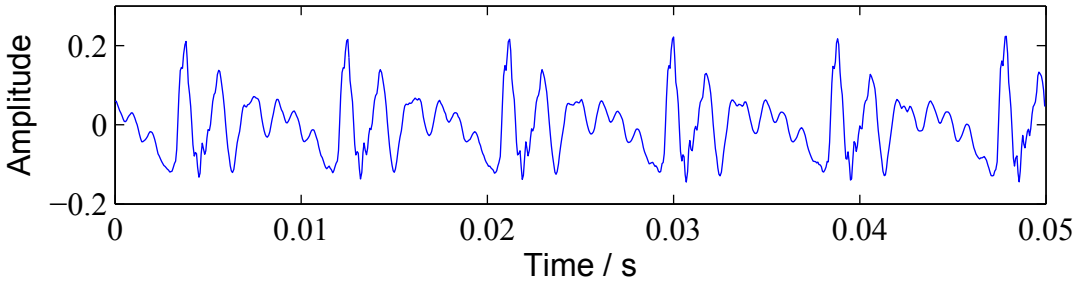
The residual signal is used as such in the field of *speech coding*, but it cannot be directly used in statistical parametric speech synthesis, because essentially it would require the statistical modeling of a waveform, not generalizable parameters. However, many methods have been developed to model the residual signal with generalizable parameters, and many of them are utilized in the state-of-the-art vocoders discussed in Section 4.

2.3 Linear Predictive Coding

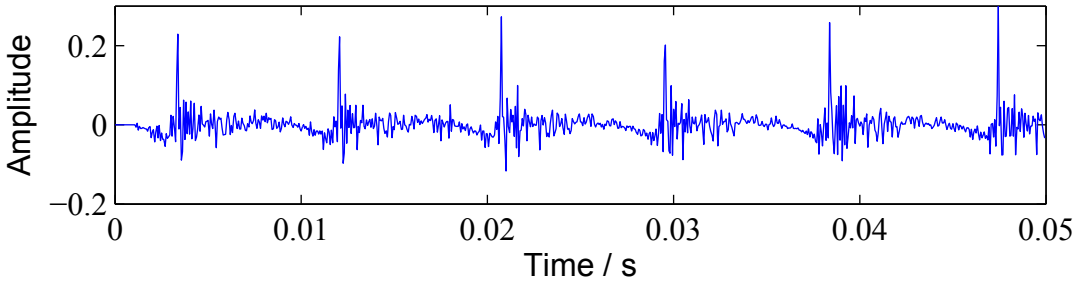
Linear Predictive Coding (LPC) of speech is one of the most powerful speech analysis techniques. Its main property is its computationally efficient ability to extract sufficiently accurate estimates of the spectral envelope $H(z)$ in the form of an all-pole filter. Due to this property LPC is a useful technique for estimating many basic speech parameters such as formants and spectra, and for low bit rate coding.

In the framework of statistical parametric speech synthesis, LPC is one of the main methods used to extract the filter parameters of the source-filter model of speech production (see Section 2.2.1). If the spectral envelope filter $H(z)$ is modeled as an all-pole filter of the form

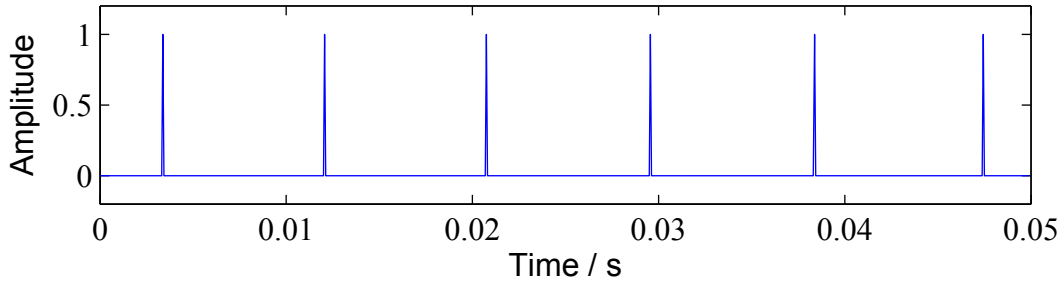
$$H(z) = \frac{S(z)}{E(z)} = \frac{G}{1 - \sum_{k=1}^{p} a_k z^{-k}}, \tag{2.9}$$



(a) The original waveform.



(b) The residual signal obtained with LPC analysis.



(c) The impulse excitation signal.

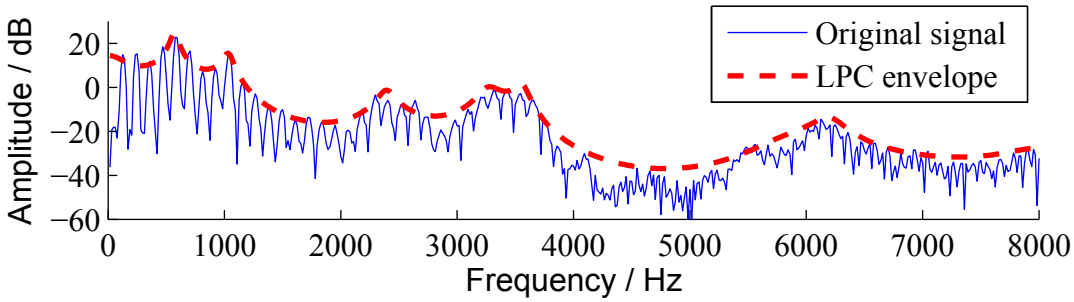
Figure 2.4: The Finnish vowel [a] with its LPC residual (excitation) signal and its impulse excitation signal.

the gain parameter G, and the filter coefficients $\{a_k\}$ can be obtained using linear predictive analysis. This means that only the excitation parameters (at the simplest form only the $F0$) need to be estimated additionally. The block diagram of a simple LPC vocoder is presented in Figure 2.6. The block diagram is identical to the combined block diagram of the source-filter model presented in Figure 2.3, with the addition of the gain parameter.

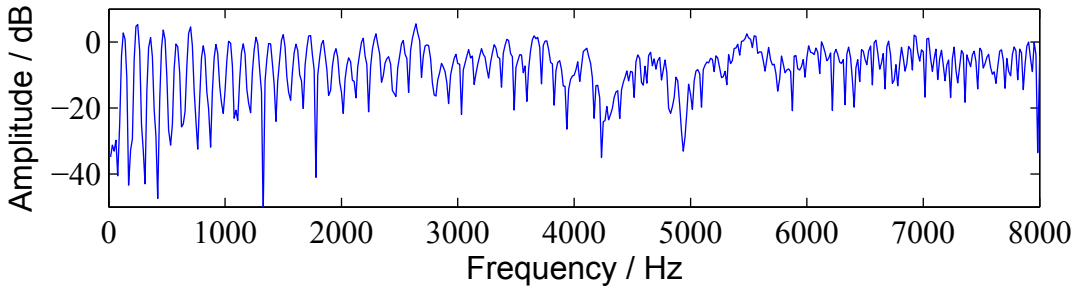
In the time domain, the simplified model of speech production (see Section 2.2.1) that uses the filter representation of Equation 2.9 becomes

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) \; + \; Gu(n), \tag{2.10}$$

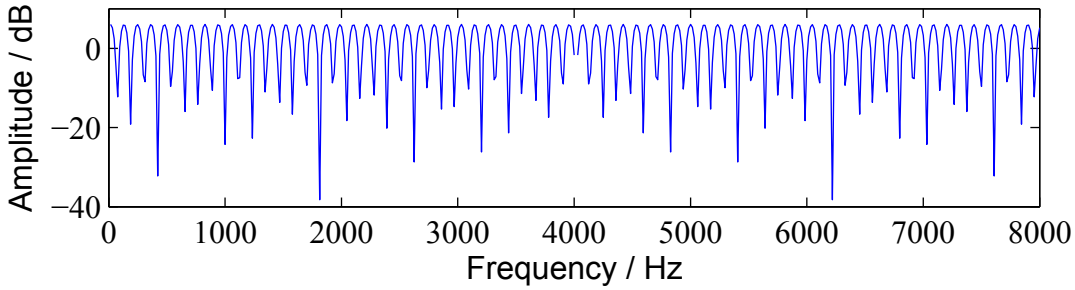
where $u(n)$ is the excitation signal.



(a) The spectrum of the original signal with its LPC envelope.



(b) The spectrum of the residual signal



(c) The spectrum of the impulse excitation signal

Figure 2.5: The spectra of the signals

2.3.1 Linear Predictive Analysis

The basic idea of LPC is the prediction of the value of the next signal sample based on a linear combination of p previous samples. A linear predictor of order p, with

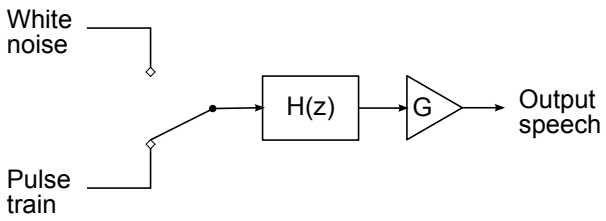


Figure 2.6: The block diagram for the LPC vocoder

prediction coefficients $\{\alpha_k\}$, is defined as a system whose output is

$$\tilde{s}(n) = \sum_{k=1}^{p} \alpha_k s(n-k) \tag{2.11}$$

A necessary feature to assess the functionality of the predictor is the *prediction error* $e(n)$, given by

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^{p} \alpha_k s(n-k) \tag{2.12}$$

Equation 2.12 can be presented in the z-domain as

$$E(z) = S(z)A(z), \tag{2.13}$$

where $E(z)$ is the z-transform of $e(n)$, $S(z)$ is the z-transform of $s(n)$, and $A(z)$ is the z-transform of the prediction error filter given by

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k} \tag{2.14}$$

Upon further inspection of Equations 2.10 and 2.12, it can be seen that if the model is exactly accurate for the speech signal, and if $\{a_k\} = \{\alpha_k\}$, then $e(n) = Gu(n)$. Thus, $A(z)$ becomes the inverse filter of the system $H(z)$ of Equation 2.9:

$$H(z) = \frac{G}{A(z)} \tag{2.15}$$

The main problem of linear predictive analysis thus becomes the estimation of the predictor coefficients $\{\alpha_k\}$ so that the prediction error $e(n)$ is minimized under some criterion. The mean squared error is by far the most utilized optimization criterion. The coefficients $\{\alpha_k\}$ that minimize the mean squared error are assumed to be the parameters of the system function $H(z)$ of Equation 2.9.

The squared prediction error E_n in a short-time frame $s_n(m)$ starting at sample n is defined as

$$E_n = \sum_m e_n^2(m) \tag{2.16}$$

$$= \sum_m (s_n(m) - \tilde{s}_n(m))^2 \tag{2.17}$$

$$= \sum_m \left(s_n(m) - \sum_{k=1}^{p} \alpha_k s_n(m-k) \right)^2, \tag{2.18}$$

where $s_n(m) = s(n+m)$. For the time being, the summation in the above equations is left unspecified, because the selection of the summation range affects the solution of the problem. The coefficients that minimize the error can be found by setting $\partial E_n / \partial \alpha_i = 0, i = 1, 2, \ldots, p$, thus obtaining the equations

$$\sum_m s_n(m-i)s_n(m) = \sum_{k=1}^{p} \alpha_k \sum_m s_n(m-i)s_n(m-k), \quad 1 \le i \le p \tag{2.19}$$

Equation 2.19 can be simplified by defining

$$\phi_n(i,k) = \sum_m s_n(m-i)s_n(m-k) \tag{2.20}$$

which leads to

$$\sum_{k=1}^{p} \alpha_k \phi_n(i,k) = \phi_n(i,0), \quad 1 \leq i \leq p \tag{2.21}$$

Equations 2.19 and 2.21 are sets of p equations with p unknowns, which means that they can be solved explicitly for the predictor coefficients $\{\alpha_k\}$ that minimize the prediction error $e(n)$. To solve the optimal coefficients, the values of $\phi_n(i,k)$ must be computed for $1 \leq i \leq p$ and $0 \leq k \leq p$.

The computation of $\phi_n(i,k)$ requires the summation interval m to be defined (and finite). Depending of the selected interval, two major approaches to the computation of the LPC coefficients have been developed: the *autocorrelation method* and the *covariance method*. The autocorrelation method assumes that the waveform segment $s_n(m)$ is zero outside the interval $0 \leq m \leq N-1$ (N is the window length):

$$s_n(m) = s(m+n)w(m), \tag{2.22}$$

where $w(m)$ is a finite length window function that is zero outside the interval $0 \leq m \leq N-1$. For the summation in Equation 2.16 this means that for a predictor of order p, the prediction error will be non-zero over the interval $0 \leq m \leq N-1+p$:

$$E_n = \sum_{m=0}^{N+p-1} e_n^2(m) \tag{2.23}$$

The covariance method assumes instead that the summation interval is fixed over the length of the frame interval over which the mean squared error is computed:

$$E_n = \sum_{m=0}^{N-1} e_n^2(m) \tag{2.24}$$

The autocorrelation method is the most popular method of the two in speech analysis/synthesis, because it will guaranteedly produce a stable all-pole filter [61]. Thus only the autocorrelation method will be described in more detail in this thesis.

When the summation limits of Equation 2.23 are used in the computation of $\phi_n(i,k)$, Equation 2.20 becomes

$$\phi_n(i,k) = \sum_{m=0}^{N+p-1} s_n(m-i)s_n(m-k), \quad \begin{matrix} 1 \leq i \leq p \\ 0 \leq k \leq p \end{matrix} \tag{2.25}$$

which can be expressed as

$$\phi_n(i,k) = \sum_{m=0}^{N-1-(i-k)} s_n(m)s_n(m+i-k), \quad \begin{matrix} 1 \leq i \leq p \\ 0 \leq k \leq p \end{matrix} \tag{2.26}$$

In this form $\phi_n(i,k)$ is identical to the short-time autocorrelation function $R_n(l)$ evaluated for $l = i - k$:

$$\phi_n(i,k) = R_n(i-k) \tag{2.27}$$

$$R_n(l) = \sum_{m=0}^{N-1-l} s_n(m)s_n(m+l) \tag{2.28}$$

Since $R_n(l)$ is an even function [61], the final form of $\phi_n(i,k)$ can be expressed as:

$$\phi_n(i,k) = R_n(|i-k|), \quad \begin{matrix} 1 \leq i \leq p \\ 0 \leq k \leq p \end{matrix} \tag{2.29}$$

Thus, Equation 2.21 can be expressed as:

$$\sum_{k=1}^{p} \alpha_k R_n(|i-k|) = R_n(i), \quad 1 \leq i \leq p \tag{2.30}$$

This set of equations can be expressed in matrix form as

$$\bar{\boldsymbol{R}}\bar{\alpha} = \bar{r} \tag{2.31}$$

or

$$\begin{bmatrix} R_n(0) & R_n(1) & R_n(2) & \cdots & R_n(p-1) \\ R_n(1) & R_n(0) & R_n(1) & \cdots & R_n(p-2) \\ R_n(2) & R_n(1) & R_n(0) & \cdots & R_n(p-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R_n(p-1) & R_n(p-2) & R_n(p-3) & \cdots & R_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ R_n(3) \\ \vdots \\ R_n(p) \end{bmatrix} \tag{2.32}$$

which can be solved by inverting the matrix $\bar{\boldsymbol{R}}$ and computing $\bar{\alpha} = \bar{\boldsymbol{R}}^{-1}\bar{r}$. Since the matrix $\bar{\boldsymbol{R}}$ is a symmetric Toeplitz matrix, the computation of the solution to Equation 2.32 can be done efficiently by utilizing the Durbin recursion [61].

After the predictor coefficients $\{\alpha_k\}$ have been obtained, the gain parameter G can be computed by (for the derivation, see [61]):

$$G^2 = R_n(0) - \sum_{k=1}^{p} \alpha_k R_n(k) \tag{2.33}$$

An illustration of the power of LPC analysis can be seen in Figure 2.5(a) of Section 2.2.2. The order p of the analysis coefficients determines the detail at which the LPC spectral envelope hooks to the original spectrum: a low order analysis hooks only to the strongest formants, and a high order analysis hooks also to the harmonic structure of the spectrum. To capture a good estimate of the spectral envelope of a speech segment, a rule of thumb is to select p so that it is the value of the sampling rate F_s in kHz added to a small integer so that the resulting integer is even. For example, for 16 kHz samples, a good predictor order is $p = 16 + 2 = 18$.

2.3.2 Line Spectral Pairs

The LPC predictor coefficients $\{\alpha_k\}$ can be used to efficiently model the vocal tract spectral envelope (see Section 2.3.1), but they are not robust in terms of quantization or statistical modeling: Even though the autocorrelation method guarantees a stable synthesis filter, a small error in the coefficient values may cause the synthesis filter to become unstable. Many methods have been proposed for the robust representation of the LPC coefficients, such as reflection coefficients or log area ratios [61]. One of the most prominent methods of presenting the LPC data is the *Line Spectral Frequency* (LSF) representation, which are the roots of the *Line Spectral Pair* (LSP) polynomials [66].

The LSP polynomials for the LP analysis filter $A(z) = 1+\sum_{k=1}^{p} a_k z^{-k}$ are defined as:

$$P(z) = A(z) + z^{-(p+1)}A(z^{-1}), \tag{2.34}$$

and

$$Q(z) = A(z) - z^{-(p+1)}A(z^{-1}). \tag{2.35}$$

Inspection of $P(z)$ and $Q(z)$ shows that $P(z)$ is a symmetric polynomial, $Q(z)$ is an anti-symmetric polynomial, and

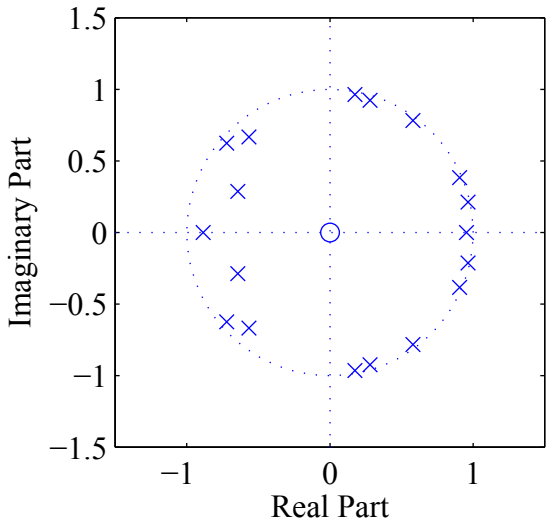
$$A(z) = \frac{1}{2}[P(z) + Q(z)] \tag{2.36}$$

$P(z)$ and $Q(z)$ are reported to have the following three properties when the LPC is estimated with the autocorrelation method [66]:

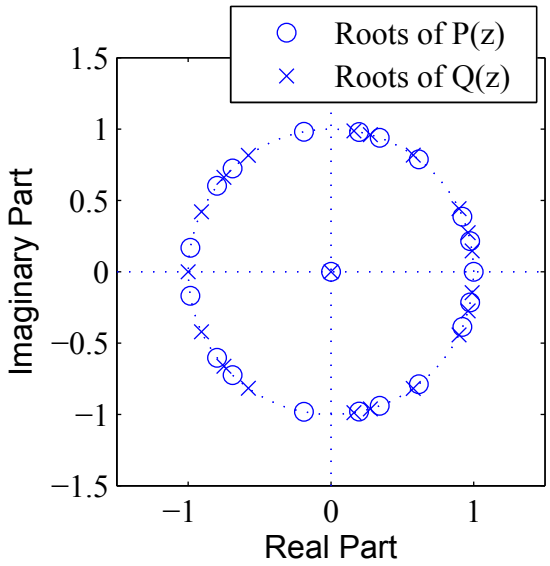
1. All zeros of $P(z)$ and $Q(z)$ are on the unit circle;
2. Zeros of $P(z)$ and $Q(z)$ are interlaced with each other; and
3. Zeros of $P(z)$ and $Q(z)$ do not overlap.

If these three properties are kept, the minimum phase property (and stability) of $A(z)$ is preserved after quantization of the zeros of $P(z)$ and $Q(z)$. Furthermore, since zeros of the LSP polynomials are located on the unit circle, their locations are easily determined by their angular frequency, or Line Spectral Frequency (LSF), which allows them to be represented as one real number instead of a complex number. The locations of the zeros can be used to reconstruct the polynomials $P(z)$ and $Q(z)$. If the order m of the LP analysis is even (which holds true in most practical applications), then $P(z)$ has a trivial zero at $z = -1$ and $Q(z)$ has a trivial zero at $z = 1$. In addition, the zeros of the LSP polynomials are scattered symmetrically with respect to the real axis, thus eliminating the need to report the LSF of the other half (top or bottom) of the unit circle. An illustration of the zeros of LSP polynomials on the unit circle is shown in Figure 2.7 (b).

The LSF parameters have been shown to have robust and top-notch performance with respect to the interpolation of LPC parameters, which is a desired quality in the field of speech synthesis and manipulation [59].



(a) The poles of the original LPC representation, A(z).



(b) The roots of the LSP polynomials P(z) and Q(z).

Figure 2.7: The poles and roots of the different LPC representations in the z-domain.

2.4 Cepstrum

Along with LPC, the cepstral analysis of speech is one of the most prominent methods for the extraction of the spectral envelope. The cepstrum is a *homomorphic transformation*, where a convolution $x(n) = x_1(n) \circledast x_2(x)$ is converted into a sum $\hat{x}(n) = \hat{x}_1(n) + \hat{x}_2(n)$. In the cepstral model, the speech signal is assumed to be a convolution of two components: The vocal tract system (including the lip radiation effect and the spectral envelope characteristics of the glottal pulse) $h(n)$, and the

excitation signal $e(n)$:

$$s(n) = e(n) \circledast h(n) \tag{2.37}$$

The excitation signal $e(n)$ can be considered as a high-frequency component of the speech spectrum, where as the spectral envelope $h(n)$ can be considered as a low-frequency component. Thus, if the convolution can be converted into a sum, a form of filtering can be done to separate these two components.

The *real* cepstrum $c_s(n)$ of a signal $s(n)$ is defined as the *inverse Fourier transform of the logarithm of the magnitude of the Fourier transform of the signal*:

$$c_s(n) = \mathcal{F}^{-1}\{\log|\mathcal{F}\{s(n)\}|\}, \tag{2.38}$$

where $\mathcal{F}\{\ \}$ and $\mathcal{F}^{-1}\{\ \}$ denote the discrete Fourier transform (DFT) and the inverse DFT operations respectively.

The homomorphic properties of this transformation can be seen, when we input $s(n) = e(n) \circledast h(n)$ into the system:

$$\mathcal{F}\{e(n) \circledast h(n)\} = E(e^{j\omega})H(e^{j\omega}) \tag{2.39}$$

$$\log|E(e^{j\omega})H(e^{j\omega})| = \log|E(e^{j\omega})| + \log|H(e^{j\omega})| \tag{2.40}$$

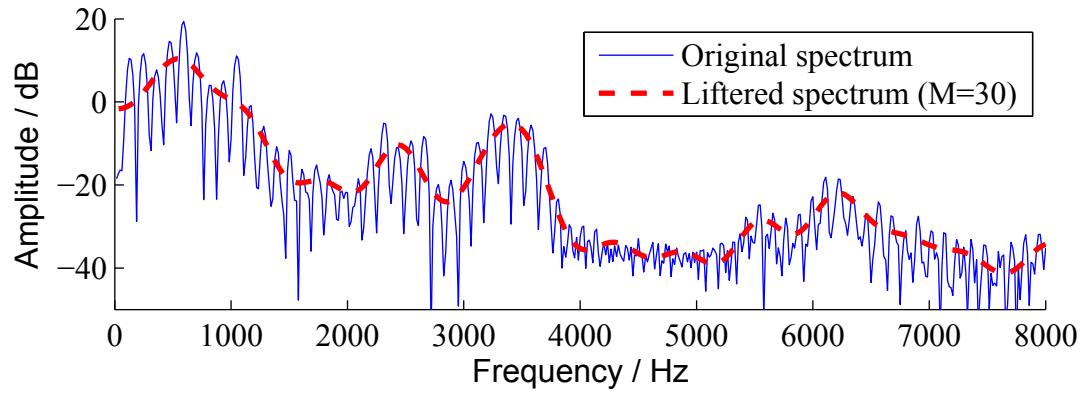
$$\begin{aligned} \mathcal{F}^{-1}\{\log|E(e^{j\omega})| + \log|H(e^{j\omega})|\} &= \mathcal{F}^{-1}\{\log|E(e^{j\omega})|\} + \mathcal{F}^{-1}\{\log|H(e^{j\omega})|\} \\ &= c_e(n) + c_h(n) \\ &= c_s(n) \end{aligned} \tag{2.41}$$

The use of the real cepstrum discards the phase information from the signal representation, which is acceptable for the determination of the minimum phase representation of the spectral envelope. If the phase information is needed to be preserved, the *complex* cepstrum can be used. The difference between the real and complex cepstra is that the real cepstrum takes the real logarithm of the magnitude spectrum, whereas the complex cepstrum takes the complex logarithm of the full spectrum. The complex logarithm is defined as:

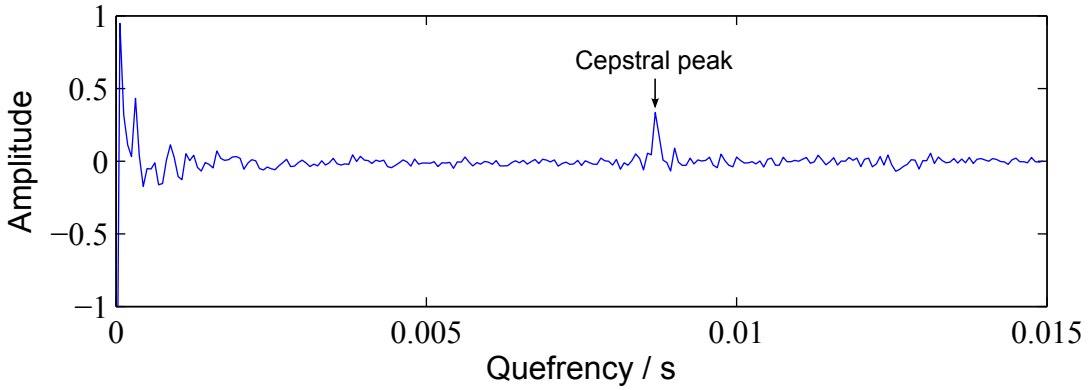
$$\log_c[X(e^{j\omega})] = \log|X(e^{j\omega})| + j\arg[X(e^{j\omega})] \tag{2.42}$$

In the context of this thesis, the cepstrum is assumed to be real.

The properties of the cepstral domain signal $c(n)$ can be thought of as a "spectrum of the spectrum": The first coefficient, $c(0)$, represents the energy of the signal, and for $n \geq 1$, $c(n)$ represents the magnitude of sinusoidal components of *quefrency* n in the spectrum. This means that assuming $|H(e^{j\omega})|$ has low-frequency and $|E(e^{j\omega})|$ has high-frequency fluctuations in the spectrum, $c_h(n)$ has relatively high coefficient values for small values of n, whereas $c_e(n)$ has most of its energy concentrated at the high values of n. Since $c_e(n)$ and $c_h(n)$ have their energies concentrated in different areas of the cepstrum, they can be separated from each other by filtering, or *liftering*, the cepstrum at a cut-off quefrency of n_0 simply by setting the samples above or below n_0 as zero.



(a) The spectrum of the vowel [a] and its cepstral envelope.



(b) The real cepstrum of the signal. The cepstral peak is located at $\tau = 0.008375\text{s} \Rightarrow F0 \approx 119.4\text{Hz}$.

Figure 2.8: The cepstral analysis of the Finnish vowel [a].

The impulse response $h(n)$ of the minimum phase spectral envelope of the signal can be obtained by utilizing the inverse cepstral transform to the liftered estimate of $c_h(n)$:

$$h(n) = \mathcal{F}^{-1}\{e^{\mathcal{F}\{c_h(n)\}}\} \tag{2.43}$$

The excitation component $c_e(n)$ of the cepstrum has a peak at a location corresponding to the fundamental period $F0$ (in samples) for voiced speech. For unvoiced speech, $c_e(n)$ is more noisy, and it does not contain the pitch peak. Cepstral analysis is demonstrated in Figure 2.8, where Figure 2.8(a) depicts the spectral envelope obtained by cepstral liftering, and Figure 2.8(b) depicts the real cepstrum of the analyzed signal.

The block diagram of a simple cepstral vocoder is presented in Figure 2.9: The impulse response of the spectral envelope filter is calculated from the cepstral coefficients, and it is convolved with either the impulse train or noise excitation for voice and unvoiced speech, respectively.

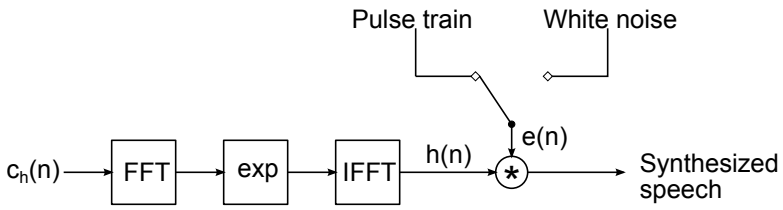


Figure 2.9: The block diagram of a simple cepstral vocoder

2.4.1 Mel-frequency Cepstrum

The problem with the conventional cepstral representation of speech is that compared to the LPC method, it requires a relatively high number of cepstral coefficients to model the spectral envelope sufficiently, and much of the information is not psychoacoustically relevant [34]. A solution for this problem has been to use *frequency warping* on the signal spectrum so that the warped spectrum emphasizes the psychoacoustic properties of human hearing. The psychoacoustic frequency-domain characteristics of the human hearing system can be modeled by the *Bark* and *mel*-scales, out of which the mel-scale has become the established norm for cepstral representation. The cepstrum computed from a mel-weighted spectrum is called as the mel-cepstrum. The mel-scale is defined as [58]:

$$m = 2595 \log_{10}(1 + \frac{f}{700}) = 1127 \log(1 + \frac{f}{700}), \tag{2.44}$$

where m is the mel-scaled value of frequency f.

With the use of discrete digital signals, the conversion to the mel-scale is usually implemented using an overlapping, log-energy normalized filter bank (Figure 2.10) corresponding to the mel-scale. The cepstrum obtained from the mel-warped spectrum is called the *mel-frequency cepstrum* (MFC), and its coefficients are called mel-frequency cepstral coefficients (MFCC). The computation of the MFCCs is commonly carried as follows [34]:

1. The signal is windowed to obtain the short-time signals
2. The DFT power spectrum of the signal is computed
3. The power spectrum is ran through the mel filter bank, and its log-energy is computed
4. The *Discrete Cosine Transform* (DCT) of the resulting power spectrum is computed to obtain the MFCCs

The MFCCs give a real valued, compact representation of the psychoacoustically interesting parts of the speech spectrum. One of their strongest features is that they are highly uncorrelated (meaning that they have a nearly diagonal covariance matrix), which makes them excellent for statistical modeling with diagonal Gaussian Mixture Models (GMMs) [34]. MFCCs are commonly utilized for example in automatic speech recognition, speaker recognition and speech synthesis.

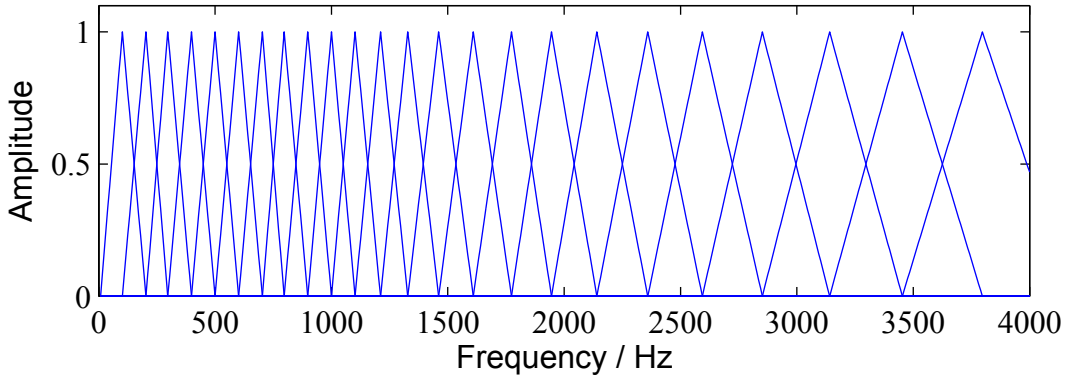


Figure 2.10: The mel-frequency filter bank for an 8 kHz signal.

2.4.2 Mel Log Spectrum Approximation Filter

The mel-frequency cepstral coefficients (MFCCs) give a good and compact representation for the psychoacoustically relevant spectral envelope information, but since the representation is on the mel-scale, the coefficients can not be used as such in the cepstral vocoder introduced in Section 2.4; frequency unwarping is needed. Additionally, the simple cepstral vocoder requires the computation of the impulse response of the spectral envelope, which is not computationally efficient. To overcome these problems, in 1983 Imai [37] proposed the *Mel Log Spectrum Approximation* (MLSA) filter, which obtains the spectral filter coefficients directly from the MFCCs by the means of a linear transformation.

MFCCs $c(m)$ represent the unwarped speech spectrum $H(z)$ by the following relation:

$$\begin{aligned} H(z) &= \exp F(z) \\ &= \exp \sum_{m=0}^{M} c(m)\tilde{z}^{-m}, \end{aligned} \tag{2.45}$$

where

$$\tilde{z}^{-1} = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \qquad |\alpha| < 1 \tag{2.46}$$

is an all-pass function which represents the mel-warped frequency characteristics, when α is a coefficient corresponding to the mel-scale (for example $\alpha = 0.35$ for 10 kHz sampling rate).

Since the filter of Equation 2.45 is not fractional, it is not realizable. In the MLSA filter, modified Padé approximation is utilized to obtain a rational transfer function approximation of the exponential type transfer function [37]. The (L, L)th order modified Padé approximation $R_L(w)$ for the exponential function $\exp(w)$ is given by

$$R_L(w) = P_L(w)/P_L(-w), \tag{2.47}$$

$$P_L(w) = 1 + p_{L,1}w(1 + p_{L,2}w(\cdots(1 + p_{L,L-1}w(1 + p_{L,L}))\cdots), \tag{2.48}$$

$$p_{L,l} = \lambda_{L,l}(L - l + 1)/(2L - l + 1), \qquad \lambda_{L,l} \approx 1 \tag{2.49}$$

If $F(z)$ is written as

$$F(z) = b_\alpha(0) + z^{-1} \sum_{m=1}^{M+1} b_\alpha(m) \tilde{z}^{-(m-1)}, \tag{2.50}$$

where

$$\begin{aligned} b_\alpha(M+1) &= \alpha c(M) && (2.51)\\ b_\alpha(m) &= c(m) + \alpha(c(m-1) - b_\alpha(m+1)), \qquad M \le m \le 2 && (2.52)\\ b_\alpha(1) &= c(1) - \alpha b_\alpha(2))/(1-\alpha) && (2.53)\\ b_\alpha(0) &= c(0) - \alpha b_\alpha(1) && (2.54) \end{aligned}$$

and if we let

$$\begin{aligned} F^{(0)}(z) &= b_\alpha(0) && (2.55)\\ F^{(1)}(z) &= z^{-1} b_\alpha(1) && (2.56)\\ F^{(2)}(z) &= z^{-1}(b_\alpha(2)\tilde{z}^{-1} + b_\alpha(3)\tilde{z}^{-2}) && (2.57)\\ F^{(3)}(z) &= z^{-1}(b_\alpha(4)\tilde{z}^{-3} + \ldots + b_\alpha(7)\tilde{z}^{-6}) && (2.58)\\ F^{(4)}(z) &= z^{-1}(b_\alpha(8)z^{-7} + \ldots + b_\alpha(M+1)z^{-M}) && (2.59) \end{aligned}$$

then the Padé approximation of $H(z)$ can be written as

$$H(z) = \exp(b_\alpha(0)) \prod_{k=1}^{4} R_3(F^{(k)}(\tilde{z})), \tag{2.60}$$

which is a rational function. The resulting filter is an IIR filter with poles and zeros, and its stability is guaranteed [37]. This means that the MLSA filter can model spectral valleys (generated by zeros in the transfer function) more efficiently than the LPC method.

The mel-cepstral analysis scheme was improved by Fukada et al. in 1992 [27], where an adaptive algorithm was proposed for the determination of optimal MFCCs, and an efficient digital filter structure for the MLSA analysis filter was presented.

3 Statistical Parametric Speech Synthesis

Statistical parametric speech synthesis has been a subject of growing research interest in the past decade. Together with the more traditional unit selection synthesis, statistical parametric speech synthesis can be seen at the moment as the other serious contender of the state-of-the-art speech synthesis methods. The main instance of statistical parametric speech synthesis techniques is Hidden Markov Model (HMM)-based speech synthesis. The HMM-based Speech Synthesis System (HTS) system [33], developed by researchers at the Nagoya Institute of Technology (Nitech), is the most prominent tool for HMM-based speech synthesis.

3.1 HMM-based Speech Synthesis

The main advantages of statistical parametric speech synthesis over unit selection synthesis are that it is more flexible, adaptable, and the size of the speech database required to construct a good quality, large sound space of synthetic speech is much smaller. Unit selection synthesis [76] can produce natural sounding speech when the generated speech is similar in phonetic contexts as the database samples. However, crude errors are made if the phonetic contexts required are not present in the database. Due to the massive number of possible arrangements of phonetic units and contexts, a database consisting of waveforms for each possible case is virtually impossible to obtain.

In statistical parametric speech synthesis, the speech waveform is broken down into a parametric representation, which is statistically modeled to obtain an average model of the parameter space. This representation enables modeling more accurately prosodic transitions which are not present at the training database. The statistical representation also enables transforming voice characteristics, speaking styles, and emotions, by mimicking voices (adaptation), mixing voices (interpolation), producing voices (eigenvoices), and controlling voices (multiple regression). Also, multilingual support is easy to implement. [76]

The drawbacks and challenges of statistical parametric speech synthesis are the vocoder quality, which is not on par with the pure waveforms of unit selection synthesis, the accuracy of the HMM-based acoustic modeling, which does not exactly model the real speech waveform, and the problem of over-smoothing of the HMM-generated parameter trajectories. In this thesis, only the vocoder model is considered. [76]

HMM-based speech synthesis has become the most popular technique of statistical parametric speech synthesis mainly because most of the theory and algorithms behind HMM-based automatic speech recognition (ASR), which has been studied extensively, can be used also in HMM-based synthesis. The HMM parameters λ are estimated from a set of training data by maximizing the likelihood of the data given the model parameters:

$$\hat{\lambda} = \arg\max_{\lambda} \{p(\mathbf{O}|\mathcal{W}, \lambda)\}, \tag{3.1}$$

where $\mathbf{O}$ is a set of training data, and $\mathcal{W}$ is a set of word sequences corresponding to $\mathbf{O}$. Speech parameters, $\mathbf{o}$, for a sequence to be synthesized, w, are then generated

by maximizing the output probabilities as

$$\hat{\mathbf{o}} = \arg\max_{\mathbf{o}} \{p(\mathbf{o}|w, \hat{\lambda})\}. \tag{3.2}$$

The block diagram for the HMM-based speech synthesis system is presented in Figure 3.1. It consists of the training part (upper half) and the synthesis part (lower half). In the training part, the labeled speech signal is analyzed in frames using the vocoder of choice into excitation parameters and spectral parameters, which make up the *feature vector* of each frame. The feature vectors are used in the acoustical modeling with HMMs of each phonetical unit that is used. The most common phonetical units used are *monophones*, which are essentially context independent phonemes.

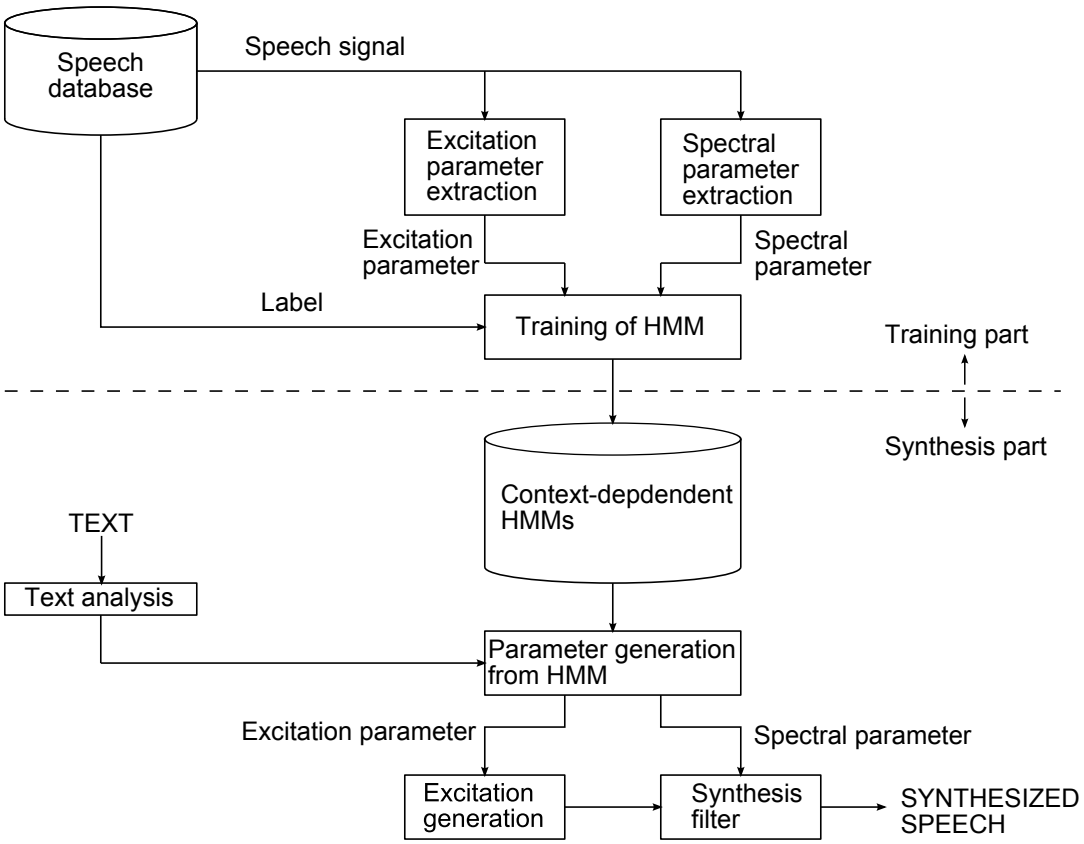


Figure 3.1: The block diagram of the HMM-based speech synthesis system

In the synthesis part, feature vectors are generated from the context-dependent HMMs according to the labels given by the text input. The feature vectors are input to the synthesis part of the vocoder of choice to construct the synthetic speech. Essentially Equations 3.1 and 3.2 describe the functions of the HMM training from the feature vectors, and the HMM feature vector generation from the text input, respectively.

3.2 Hidden Markov Models

Hidden Markov Models (HMMs) are statistical models used for modeling various kinds of sequential data. The speech signal is a good example of sequential data, where the next sample is statistically dependent from the present sample. For example, in English 'h' is likely to follow 't' but not 'x'. In the case of HMMs, the first-order Markov model is used, which means that the observed state q at time $t+1$ is only dependent in the state at time t:

$$P(q_{t+1} = S_j | q_t = S_i, q_{t-1} = S_k, \ldots) = P(q_{t+1} = S_j | q_t = S_i) \tag{3.3}$$

An HMM can be considered as a finite state system with discrete output states $S_1, S_2, \ldots, S_N$, which emit output values $\{v_1, v_2, \ldots, v_M\}$ within their individual probability distributions $\mathbf{B} = [b_j(m)]$, where

$$b_j(m) = P(O_t = v_m | q_t = S_j), \tag{3.4}$$

is the observation probability that we observe v_m in state S_j. The states are connected to each other via transition probabilities $\mathbf{A} = [a_{ij}]$, where

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i) \tag{3.5}$$

is the probability that the system will transfer from state S_i to S_j for the next output sample. $\mathbf{A}$ is assumed to be independent of time. In the general case, the transition matrix $\mathbf{A}$ has transitional probabilities between states, but in speech modeling a left-to-right topology is used, meaning that the system transitions successively from state n to state $n+1$, or stays in the same state, until the end of the system is reached. This is illustrated in Figure 3.2.

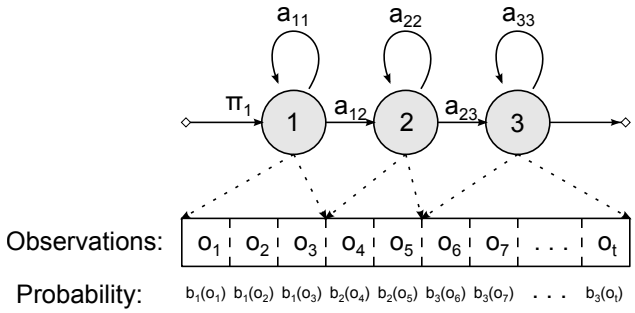


Figure 3.2: The block diagram of a 3-state left-to-right HMM

In HMMs, the state sequence $\mathbf{Q} = \{q_1, q_2, \ldots, q_t\}$ is hidden, and the observation sequence $\mathbf{O} = \{O_1, O_2, \ldots, O_t\}$ is generated by the state sequence. This means that there are multiple state sequences $\mathbf{Q}$ that could have generated the same observation sequence $\mathbf{O}$, but with different probabilities. The final component determining the HMM are the initial state probabilities $\mathbf{\Pi} = [\pi_i]$, where

$$\pi_i = P(q_1 = S_i) \tag{3.6}$$

The parameter set λ of an HMM is thus defined as $\lambda = (\mathbf{A}, \mathbf{B}, \mathbf{\Pi})$.

Given a number of sequences of observations, there are three basic problems concerning HMMs [6]:

1. Given a model λ, the evaluation of the probability of any given observation sequence $\mathbf{O}$, $P(\mathbf{O}|\lambda)$.

2. Given a model λ and an observation sequence $\mathbf{O}$, finding of the state sequence $\mathbf{Q}$, which has the highest probability of generating $\mathbf{O}$, $\mathbf{Q}^* = \arg\max_{\mathbf{Q}} P(\mathbf{Q}|\mathbf{O}, \lambda)$.

3. Given a training set of observation sequences, $\chi = \{O^k\}_k$, the learning of the model that maximizes the probability of generating χ, $\lambda^* = \arg\max_\lambda P(\chi|\lambda)$.

Problem 1 can be solved using the Forward-Backward alogrithm [6], problem 2 can be solved using the Viterbi algorithm [6], and problem 3 can be solved using the Baum-Welch algorithm [6].

3.3 Applications of HMM-based Speech Synthesis

HMM-based speech synthesis has multiple notable applications: First of all, HMM-based speech synthesis can be used as a plain text-to-speech synthesis system especially in applications that have a limited capacity of memory (such as mobile phones). These applications include for example screenreaders, telephone services, e-book readers, car navigators, and basic voice communication aids for people with disabilities. However, these applications are usually considered as an optional extra, and none of them are a ground-breaking must-have feature.

The new and emerging applications that are enabled by HMM-based speech synthesis are looking to be much more promising, as their focus is on the *voice*, not the text. These applications include voice cloning, voice reconstruction, personalized speech-to-speech translation, articulatory-controllable speech synthesis, and noise-adaptive speech synthesis. The underlying technology behind all of these applications is the possibility of statistical adaptation of the HMMs via a linear transform based on (in comparison) small amounts of adaptation data.

Voice cloning can be described as automatically creating synthetic voices from a relatively small amount of data. The system is trained on an average model consisting of multiple speakers, and then it is adapted to the target voice by using a linear transformation on the target vectors. This can be used in creating celebrity voices, or more importantly, in so-called voice banking. The idea of voice banking is that a person can record a relatively short segment of their speech, and if they happen to lose their voice for medical reasons, they can regain their old voice by the means of speech synthesis that resembles their own voice. [72], [73]

Voice reconstruction has the same goal as voice banking: to give people that have a degenerated ability to speak the ability to speak clearly with their own voice via synthesis. However, the problem usually is that people whose speaking ability has degenerated have no clear recordings of their speech for use in voice banking. Thus, the problem is to distinguish speaker-characteristic features from the disordered characteristics, and replace the disordered characteristics with natural models. [15]

Personalized speech-to-speech translation can be considered as one of the ultimate goals in speech technology: It is a system that takes in an input in language X, recognizes it with ASR, translates the ASR output using Machine Translation

(MT) into language Y, and synthesizes the MT output with a TTS system that has been adapted to the speaker voice. HMM-based speech synthesis enables the use of cross-lingual speaker adaptation that is used to adapt the synthesis in other languages.

4 Analysis/Synthesis Methods

Many different Analysis/Synthesis vocoders have been developed to be applied with HMM-based speech synthesis. They can be categorized into four different categories: Mixed excitation vocoders, Residual modelling vocoders, Glottal source modelling vocoders, and Sinusoidal modelling vocoders.

4.1 Impulse Excitation vocoder

The most basic vocoder used in statistical parametric speech synthesis is essentially the unified source-filter model introduced in Section 2.2.1: The speech signal is divided into source and filter parameters. The source signal is modeled as a pulse train for voiced segments, and as white Gaussian noise for unvoiced segments.

Analysis

The analysis phase of the impulse excitation vocoder is very straightforward: The $F0$ of each frame is estimated for the excitation parameter, and the spectral envelope is estimated using a cepstral or LPC based method. The most popular representation for the spectral envelope are the MFCCs (see Section 2.4.1), which are used for example in the baseline HTS system [33]. The analysis vector of the impulse excitation vocoder is depicted in Table 4.1.

Table 4.1: The analysis vector of the impulse excitation vocoder, where p denotes the order number of the spectral analysis.

Excitation parameters	$1 \times F0$
Spectral parameters	$p \times$ MFCC or LSF

Synthesis

The synthesis block diagram for the impulse excitation vocoder is presented in Figure 4.1. The excitation generation is based on the $F0$ parameter of the feature vector: For voiced frames, the excitation is generated as an impulse train where the pulses are separated by the length of the pitch period, and unvoiced excitation is generated as white Gaussian noise. The excitation is fed into the time varying synthesis filter, which is implemented as the MLSA filter (see Section 2.4.2) for the MFCC representation of the spectrum.

The impulse excitation vocoder is very simple to implement, but it produces a "buzzy" speech quality due to the unnaturally strong harmonic structure in the excitation spectrum. Also, the binary pulse/noise representation is unable to correctly model the speech sounds which are characterized by a combination of periodic and noise components, such as voiced fricatives. Despite its limitations, the simplicity of the impulse excitation vocoder has made it historically significant for the development of statistical parametric speech synthesis. Currently, the impulse excitation

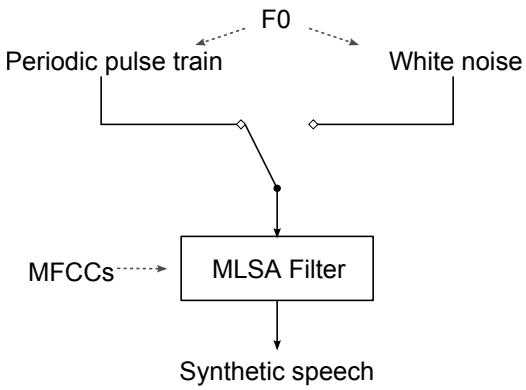


Figure 4.1: The synthesis block diagram of the impulse excitation vocoder.

vocoder is mainly used as a benchmark when testing the quality of more advanced vocoders.

4.2 Multi-Band Mixed Excitation Vocoders

Mixed Multi-Band Excitation (MBE) vocoders use additional parameters with the F0 value to generate a more accurate excitation signal (to reduce buzziness) than the impulse excitation vocoder. The common trait for all of these methods is that the parameters are extracted in a uniform way without case-specific adaptation.

The following three methods are introduced in this section: The Mixed Excitation (ME) vocoder, STRAIGHT vocoder, and the Harmonic plus Noise Model (HNM)-based Two-Band Excitation (TBE) vocoder.

4.2.1 Mixed Excitation

Proposed by Yoshimura et al. in 2001 [74], the Mixed Excitation vocoder for HMM speech synthesis was the first advanced vocoding method to improve the vocoder quality of HMM-based speech synthesis. This method is the HMM-adjusted version of the Mixed Excitation Linear Prediction (MELP) low bit rate speech codec, originally proposed by McCree et al. in 1995 [55]. The main difference between the HMM-adjusted and the original version is that the HMM-adjusted version uses MFCC coefficients instead of LPC coefficients for the representation of the spectral envelope. Also, the method called "Adaptive Spectral Enhancement" [55] used in the MELP codec is omitted from the HMM-adjusted version.

The main idea of the Mixed Excitation vocoder is based on the observations of Makhoul et al. [53] on the spectral characteristics of the LPC residual: The residual was found to have different degrees of periodicity and noise in different frequency bands. If the residual is modeled completely periodically (with a pulse train), the resulting voice will sound "buzzy". Similarly, if the residual is modeled completely with noise, the resulting voice will sound "hissy". With a correct combination of periodic and noise components in the excitation (residual), the synthesized speech will show a great increase in quality.

Analysis

During the analysis phase, the Mixed Excitation vocoder first determines the F0 as well as the mel-cepstral coefficients of each analysis frame. Next, the degree of voicing is estimated for each time frame under five sub-bands of the speech signal. For 16 kHz sample rate, the sub-bands are 0-1, 1-2, 2-4, 4-6, and 6-8 kHz. For each sub-band, the degree of voicing is estimated using Equation (4.1):

$$c_t = \frac{\sum_{n=0}^{N-1} s_n s_{n+t}}{\sqrt{\sum_{n=0}^{N-1} s_n s_n \sum_{n=0}^{N-1} s_{n+t} s_{n+t}}} \tag{4.1}$$

where t is the estimated pitch lag (inverse of F0 in samples), N is the window length in samples, and s_n is the nth sample of the processed window s. Equation (4.1) is in fact a normalized autocorrelation function of the frame, and because it is calculated with the pitch lag, it measures the highest order of periodicity within the frame.

Finally, a residual signal of the frame is obtained via inverse filtering using the MLSA (see Section 2.4.2) analysis filter, and the Fourier magnitudes of ten first harmonics are obtained from it.

The analysis feature vector extracted for each frame then consists of one logF0 coefficient, p mel-cepstral coefficients, five bandpass voicing strengths, and 10 Fourier magnitudes (see Table 4.2). Thus the total length of the feature vector becomes 40 for $p = 24$. For the purpose of HMM-training, each coefficient's delta and delta-delta coefficients are also calculated, but this is not needed in pure analysis/synthesis operation.

Table 4.2: The analysis vector of the Mixed Excitation vocoder.

Excitation parameters	$1 \times F0$
	$5 \times$ bandpass voicing
	$10 \times$ fourier magnitude
Spectral parameters	$p \times$ MFCC

Synthesis

Speech is synthesized from the analysis vectors according to the block diagram in Figure 4.2. First, the voiced and unvoiced parts of the excitation are generated separately. The voiced part is obtained by generating a periodic pulse train corresponding to the F0 value of each frame, with the spectral characteristics of the ten first Fourier magnitudes. If the voicing strength is weak, position jittering is applied to the pulses, where they are randomly shifted within ±25% of the original position.

Bandpass filters are formed for both the periodic and noise excitation, and the bands are weighted according to the values of their respective bandpass voicing

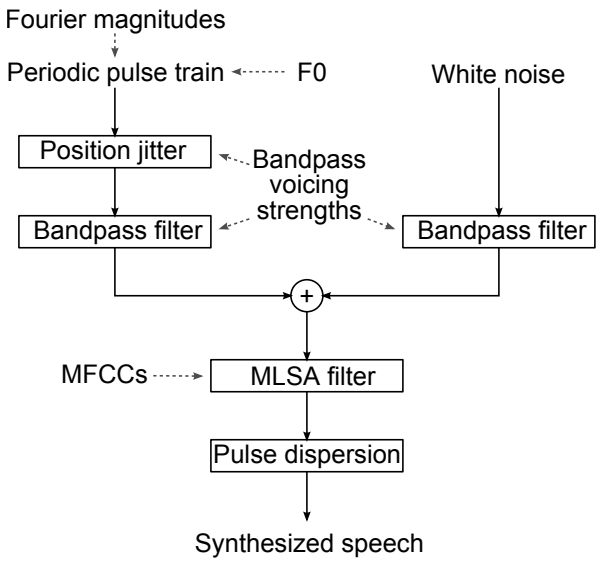


Figure 4.2: The synthesis block diagram of the Mixed Excitation vocoder.

strengths. Both excitations are filtered through their respective bandpass filters, and then they are added together to form the mixed excitation signal. The mixed excitation signal is filtered through the MLSA synthesis filter, forming the synthesized speech.

Final post-processing is done by filtering the frame with a pulse dispersion filter, which is a static FIR filter obtained from a spectrally whitened triangle pulse approximation of an average glottal pulse. This method has been shown to significantly enhance the quality of the synthesized speech [74], [55].

The quality of the synthesized speech of the ME vocoder was evaluated by a pair comparison test between the proposed system and the simple pulse train excitation vocoder. The ME vocoder was found to be preferred by a ratio of 9 to 1 over the impulse excitation vocoder, which means a great increase in quality.

4.2.2 STRAIGHT

STRAIGHT (Speech Transformation and Representation using Adaptive Interpolation of weiGHT spectrum) is the most established of the more sophisticated vocoding methods. Originally proposed by Kawahara in 1997 [41], it has gone through extensive research and development [45], [43], [42], [46], and it is often the main reference to which other vocoders in HMM based synthesis are compared [8], [10], [52], [64], [21].

STRAIGHT was first designed as a tool for speech transformation and accurate spectral envelope representation. Original STRAIGHT parameters are represented as Fourier transform magnitudes and aperiodicity measurements corresponding to them. They can not be used in HMM synthesis due to their high dimensionality. To overcome this problem, Zen et al. proposed an HMM-modified version of STRAIGHT [75], where the spectral envelope is represented as mel-frequency cepstral coefficients, and the corresponding aperiodicity measurements are averaged

over five sub-bands of frequency.

Analysis

The main idea behind STRAIGHT is the extraction of a smoothed spectral envelope, which minimizes the effect of periodicity interference in the analysis frames. This means that the STRAIGHT spectral envelope is essentially independent of the speech excitation, which is a great feature with respect to speech transformation.

The extraction of the spectral envelope is carried as follows: First, the signal is windowed using two complementary F0-adaptive windows that have equivalent temporal and spectral resolution. The windows w_p and w_c are based on the product of a Gaussian component and the 2nd order cardinal B-spline function (convolution of two 1st order spline functions/square pulses), and they are given by:

$$w_p(t) = e^{-\pi(t/t_0)^2} \circledast h(t/t_0), \tag{4.2}$$

$$w_c(t) = w_p(t) sin(\pi \frac{t}{t_0}), \tag{4.3}$$

where t is the time index, t_0 is the time of the fundamental period, and $h(t)$ is the 2nd order cardinal B-spline function given by:

$$h(t) = \begin{cases} 1 - |t|, & \text{if } |t| < 1, \\ 0, & \text{otherwise} \end{cases} \tag{4.4}$$

The length of the window functions can be seen to be two times the length of the fundamental period, which has the property of smoothing the temporal structure of the spectrogram. Smoothing in the frequency domain is acquired by the use of the 2nd order cardinal B-spline function, which essentially robustly interpolates the space between the magnitude spectrum samples (the Fourier transform of the signal is convolved with the Fourier transform of the window function). The complimentary window function w_c is sinusoidally modulated so that the spectrogram produces maxima there where the original spectrogram (acquired with window function w_p) has holes. [45]

Next, the original and complimentary magnitude spectrograms $P_o(\omega, t)$ and $P_c(\omega, t)$ are calculated using the window functions w_p and w_c respectively. The spectrograms are combined into the final spectrogram $P_r(\omega, t)$ by

$$P_r(\omega, t) = \sqrt{P_o^2(\omega, t) + \xi P_c^2(\omega, t)}, \tag{4.5}$$

where ξ is a blending factor that minimizes the temporal variation of the resultant spectrogram. Numerical search has provided an estimate value of $\xi = 0.13655$. [45]

The problem of this method is over-smoothing that is caused by the interaction of the convolution of the Gaussian and the 2nd order cardinal B-spline component of the window function. The proposed solution for this is the use of a quasi-optimal smoothing function $h(t)$ that consists of three 2nd order cardinal B-spline functions.

The aperiodicity measurements estimate the amount of harmonic information in relation to non-harmonic information in the signal. Ideally this is done by warping each frame according to the phase of its fundamental component, which makes the warped signal have a regular harmonic structure [44], and then calculating the ratios between lower and upper spectral envelopes S_L and S_U for each sample. The upper spectral envelope has the spectral peaks connected to each other, and the lower spectral envelope has the spectral valleys connected to each other.

In practice, the unwarped aperiodicity measures are obtained by performing a table lookup operation of the lower-upper ratio from a database of known aperiodicity measurements. After that, its weighted average in relation to the speech power spectrum is calculated to give the final aperiodicity measurement:

$$P_{AP}(\omega) = \frac{\int w_{ERB}(\lambda;\omega)|S(\lambda)|^2\Gamma(\frac{|S_L|^2}{|S_U|^2})d\lambda}{\int w_{ERB}(\lambda;\omega)|S(\lambda)|^2 d\lambda} \tag{4.6}$$

where w_{ERB} is a simplified auditory filter shape for smoothing the power spectrum at center frequency ω, $|S(\lambda)|^2$ is the speech power spectrum, and $\Gamma(\ )$ is the table lookup operation.

Finally, STRAIGHT uses a specific pitch extraction algorithm (PDA) called TEMPO (Time-domain Excitation extractor using Minimum Pertubatin Operator) [41], [45], [44] to extract the fundamental frequency trajectory of the target sample. The method is based on the concept of *instantaneous frequency* (first time-derivative of the instantaneous phase) [9], and it uses the following nearly harmonic model for the representation of speech:

$$x(t) = \sum_{k=1}^{N} a_k(t) cos\left(\int_0^t (k\omega_0(\tau) + \omega_k(\tau))d\tau + \phi_k(0)\right), \tag{4.7}$$

where $a_k(t)$ represents a slowly changing instantaneous amplitude, $\omega_0(\tau)$ is the instantaneous frequency, $\omega_k(\tau)$ is a slowly varying FM component of the kth harmonic, and $\phi_k(0)$ is the instantaneous phase.

The instantaneous frequency is extracted from the signal by the means of an analyzing continuous wavelet transform, which has the smallest amount of AM and FM properties at the fundamental frequency. This observation is used in the measure of "fundamentalness" over the frequency range, and the frequency with the highest fundamentalness is selected as the instantaneous frequency. The continuous wavelet transform (CWT) is defined as:

$$D(t,\tau_c) = |\tau_c|^{-\frac{1}{2}} \int_{-\infty}^{\infty} s(t)\Psi^{\star}\left(\frac{t-u}{\tau_c}\right) du, \tag{4.8}$$

where $\Psi^{\star}(t)$ is the complex conjugate of a wavelet function and τ_c is a scale factor of the wavelet. The wavelet used in the TEMPO algorithm is a based on a Gabor function, and it is defined by:

$$\Psi(t) = g(t-1/4) - g(t+1/4) \tag{4.9}$$

$$g(t) = e^{-\pi\left(\frac{t}{\eta}\right)^2} e^{-j2\pi t}, \tag{4.10}$$

where $\eta > 1$ is a parameter representing the frequency resolution of the wavelet transfer function. The fundamentalness measure $M(t, \tau_c)$ is defined as:

$$\begin{aligned} M(t,\tau_c) &= -\log\left[\int_\Omega \left(\frac{d|D|}{du}\right)^2 du\right] + \log\left[\int_\Omega |D|^2 du\right] \\ &\quad -\log\left[\int_\Omega \left(\frac{d^2 arg(D)}{du^2}\right)^2 du\right] + \log\Omega(\tau_c) + 2\log\tau_c, \end{aligned} \tag{4.11}$$

where $\Omega(\tau_c)$ is an integration interval set proportional to the size of the corresponding analyzing wavelet.

As stated in the introduction of the method, the HMM-adapted version of STRAIGHT transforms the STRAIGHT spectrum into a mel-frequency cepstral representation for the purpose of statistical modeling. The aperiodicity measurements are also transformed into a compressed representation. The original way is to average them over sub-bands of frequency, but recently Cotescu et al. [14] have proposed alternative representation methods, out of which a mel-frequency cepstral representation has provided the best results.

The acquired analysis vector for STRAIGHT (Table 4.3) thus consists of the F0 value, five aperiodicity coefficients and 20-40 spectral MFC coefficients.

Table 4.3: The analysis vector of the STRAIGHT vocoder, where p denotes the order number of the spectral analysis.

Excitation parameters	$1 \times F0$
	$5 \times$ aperiodicity measure
Spectral parameters	$p \times$ STRAIGHT MFCC

Synthesis

STRAIGHT synthesis is done in frame-by-frame basis by creating a mixed excitation signal of the length of two pulse periods based on the F0 and aperiodicity measurements. The harmonic pulse train is all-pass filtered with a randomized group-delay filter, which reduces the buzziness of the resultant synthesis. The acquired mixed excitation signal is convolved with the minimum phase MLSA filter derived from the frame's spectral MFCCs. Finally, the Pitch-Synchronous Overlap-Add (PSOLA) algorithm [56] is applied to the synthesized frames to get the final signal. The synthesis process is illustrated in Figure 4.3.

The components for the mixed excitation are generated by sub-band filtering the voiced (impulse train) and unvoiced (white Gaussian noise) parts separately in

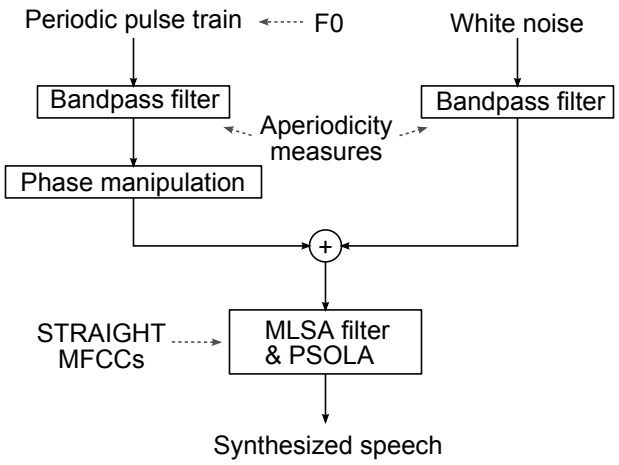


Figure 4.3: The synthesis block diagram of the STRAIGHT vocoder.

the frequency domain. The stepwise bandpass filters used are determined by the aperiodicity coefficients so that the resultant sub-bands will have the same average lower-to-upper envelope ratio as the respective aperiodicity coefficient.

After the sub-band weighting, the pulse train component is all-pass filtered with $\Phi(\omega)$ to adjust the phase characteristics of the excitation. $\Phi(\omega)$ is obtained by a group delay design, where the target group delay function $d_4(\omega)$ is calculated by

$$d_4(\omega) = \frac{d_g x(\omega)}{\sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi} |x(\omega)|^2 d\omega}} \tag{4.12}$$

$$x(\omega) = \rho(\omega) F^{-1}(W_s(\tau) N(\tau)) \tag{4.13}$$

$$W_s(\tau) = |\tau| e^{-\pi(\tau/\tau_{bw})^2} \tag{4.14}$$

where d_g is the desired spread of the target group delay function, $N(\tau)$ is an initial random group delay function made from white Gaussian noise, $W_s(\tau)$ is a weighting function in the spatial frequency domain, and $\rho(\omega)$ is a frequency-weighting function used to control temporal energy spread in each frequency region. The excitation phase characteristic $\Phi(\omega)$ is obtained by integrating $d_4(\omega)$. [41]

The synthesis quality of STRAIGHT is significantly better than the simple pulse train excitation vocoder, with a MOS around 3, where as the simple excitation vocoder has a MOS around 2.

4.2.3 Harmonic plus Noise Model

The Harmonic plus Noise Model (HNM) was originally proposed by Stylianou [69], [70] for the purpose of concatenative speech synthesis, and it has been the basis for various implementations for the use in statistical parametric speech synthesis [31], [47]. The most notable implementation is the Two-Band Excitation (TBE) vocoder proposed by Kim et al. in 2006 [47].

The original HNM is very similar to the original Harmonic/Stochastic Model (HSM) introduced in Section 4.5.2. The speech model used by the HNM is characterized by a sum of a harmonic part and a noise part that are separated by a cut-off frequency called the *maximum voiced frequency* F_m. Stylianou argues in [70] that even though the F_m assumption is not valid from a speech production point of view, it is useful in a perception point of view, as it leads to a simple model with high-quality synthesis. Also, Kim et al. [47] argue that the determination of the exact cut-off point of the mixed excitation is more important for the quality than the approximation of voicings in few fixed frequency bands which is used in the Mixed Excitation vocoder (Section 4.2.1).

The harmonic part $s_h(t)$ is modeled as the sum of harmonic sinusoids up to F_m:

$$s_h(t) = \sum_{k=-L(t)}^{L(t)} A_k(t) e^{jk\omega_0(t)}, \tag{4.15}$$

where $A_k(t)$ is a complex number representing the amplitude and phase of the kth sinusoidal component, $L(t)$ denotes the number of harmonics used in the harmonic part, and ω_0 is the fundamental angular frequency.

The noise part $s_n(t)$ is modeled as white Gaussian noise filtered by an autoregressive filter:

$$s_n(t) = e(t)[h(\tau, t) \circledast b(t)], \tag{4.16}$$

where $e(t)$ is a gain term, $h(\tau, t)$ is the time-varying autoregressive model, and $b(t)$ is white Gaussian noise.

The synthetic signal $\hat{s}(t)$ is obtained by

$$\hat{s}(t) = s_h(t) + s_n(t). \tag{4.17}$$

It is notable that the original HNM uses the amplitudes and phases of sinusoids for the modeling of the harmonic part, which is problematic for statistical parametric speech synthesis because of two reasons: First, the feature vector becomes very long, which makes the statistical modeling more difficult. Second, the statistical modeling of phase components is not possible with current methods. The implementations of the HNM for statistical parametric speech synthesis circumvent these problems. In the next part, the TBE model is discussed in more detail.

Analysis

The maximum voiced frequency F_m of the HNM is the key concept in the Two-Band Excitation model of Kim et al. [47]. In TBE, the HNM is simplified such that the F_m denotes the cut-off frequency for a mixed excitation that is generated by a low-pass filtered pulse train and high-pass filtered white Gaussian noise. Thus the parameters to be extracted by the TBE analysis are: F0, MFCC or LSF coefficients, and F_m (see Table 4.4).

There are various proposed methods for the estimation of the F_m: The original method of HNM, the initial TBE method of Kim and Hahn, and the refined TBE method of Han [30].

Table 4.4: The analysis vector of the HNM/TBE vocoder, where p denotes the order number of the spectral analysis..

Excitation parameters	$1 \times F0$
	$1 \times F_M$
Spectral parameters	$p \times$ MFCC or LSF

The original method used in the HNM of Stylianou is a harmonic test, where each harmonic peak's amplitude is compared to the cumulative amplitude of its surrounding spectral valleys. If the ratio between the two is above a certain heuristic limit, the harmonic component is considered voiced. Otherwise it is marked as unvoiced. This test is done for each harmonic component, and then the resultant voicing vector is filtered by a three-point median smoothing filter. F_m is selected as the highest frequency that has a non-zero value in the voicing vector.

The original way of estimating the F_m is problematic because of the heuristic limit value, which makes the method unrobust. Therefore Kim et al. proposed a new estimation method in their TBE vocoder. Their technique is based on the normalized autocorrelation coefficients of high-pass filtered frame segments. A high-pass filter with cut-off frequency f is denoted as h^f_{HPF}, and the high-pass filtered signal frame is denoted as

$$s^f_{HB}(n) = h^f_{HPF} \circledast s(n). \tag{4.18}$$

The normalized autocorrelation coefficient at the estimated pitch lag τ (in samples) is

$$R^f_{n,HB}(\tau) = \frac{\sum_{n=0}^{N-1} s^f_{HB}(n) s^f_{HB}(n+\tau)}{\sqrt{\sum_{n=0}^{N-1} \left\{ s^f_{LB}(n) \right\}^2 \sum_{n=0}^{N-1} \left\{ s^f_{HB}(n+\tau) \right\}^2}}, \tag{4.19}$$

where N is the analysis window size.

If the cut-off frequency f is higher than the real F_m, the normalized autocorrelation coefficient will be close to zero. Similarly, if the filter cut-off frequency is smaller than the real maximum voiced frequency, the normalized autocorrelation coefficient will be close to one. Thus, the normalized autocorrelation coefficient is evaluated with varying high-pass filter cut-off frequencies f ranging over the entire bandwidth. The highest cut-off frequency that satisfies $R^f_{n,HB}(\tau) > 0.5$ is selected as the estimate for the maximum voiced frequency F_m.

A refinement to the original TBE F_m estimation method was presented by Han et al. in 2009 [30]. In this method, the original TBE method is used to obtain an

initial estimate for the maximum voiced frequency, and then an analysis-by-synthesis scheme is applied to refine the estimate by minimizing spectral distortion.

In the analysis-by-synthesis scheme, the excitation is generated with the initial F_m and synthesized by the MLSA filter (see Synthesis section). Next, the spectral distortion is measured by the symmetric Kullback-Leibler distance between the normalized power spectra of adjacent sub-frames near the frame boundary:

$$D_{SKL} = \sum_{k=0}^{N-1} \left(S_i(k) - S_{i+1}(k) \log \frac{S_i(k)}{S_{i+1}(k)} \right), \tag{4.20}$$

where $S_i(k)$ is the normalized power spectrum in a sub-frame, i is the frequency index and k is the frequency bin index.

This analysis is applied to candidate F_ms around the initial estimate, and the F_m value that gives the least amount of spectral distortion according to Equation (4.20) is selected as the optimal maximum voiced frequency.

Synthesis

The HNM/TBE synthesis scheme is very similar to the synthesis scheme of the Mixed Excitation vocoder. The synthesis block diagram can be seen in Figure 4.4.

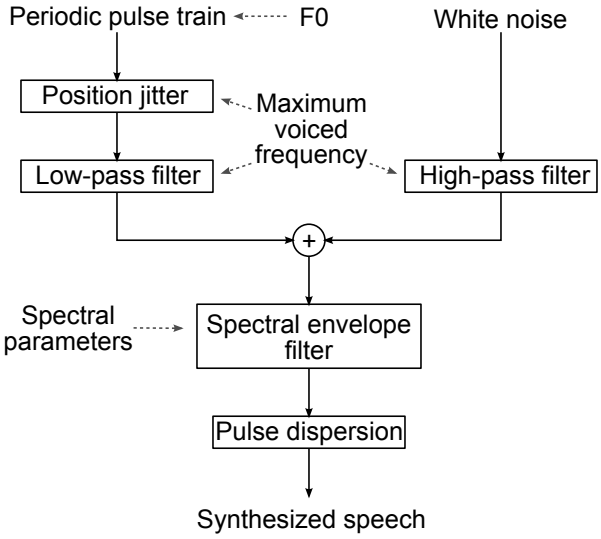


Figure 4.4: The synthesis block diagram of the HNM/TBE vocoder.

Mixed excitation is generated so that the voiced pulse train generated according to $F0$ is low-pass filtered according to the maximum voiced frequency F_m, and the unvoiced white Gaussian noise is high-pass filtered with the complement filter of the voiced part. The two parts are summed to obtain the mixed excitation signal. Also, the methods of pulse position jittering and pulse dispersion filtering are applied similar to the Mixed Excitation vocoder (see Section 4.2.1).

Finally, the mixed excitation signal is filtered with the MLSA filter derived from the mel-frequency cepstral coefficients to obtain the synthesized frame. The synthesis quality of the HNM/TBE method is on par with the Mixed Excitation vocoder,

and it uses significantly fewer excitation parameters. This makes the method viable in applications where memory and processing power are limited, such as mobile phones.

4.3 Residual Modeling Vocoders

Residual modeling vocoders use statistical optimization procedures to produce the optimal excitation signals based on the desired criteria. In practice this means that the methods try to recreate the exact waveform of the residual signal obtained in the analysis phase.

This approach has many advantages, because the residual signal contains much more information about the source than just the voicing and the pitch. For example, the residual contains phase information and non-linear effects which are not represented by the voicing and pitch parameters. [10]

The following two methods are presented in this section: The Closed-loop Training vocoder and the Pitch-synchronous Residual Codebook vocoder.

4.3.1 Closed-Loop Training

The excitation approach for HMM-based speech synthesis based on the Closed-Loop Training (CLT) procedure [3] was proposed by Maia in 2007 [52]. The main idea is the determination of optimal voiced and unvoiced filters for the excitation generation of each HMM state to maximize the likelihood of the synthesized excitation in comparison to the original excitation. This makes the method technically not a pure analysis/synthesis method, because the analysis phase does not break the signal frames into feature vectors, from which the signal could be synthesized again. Instead, the feature vectors contain only the $F0$ and MFCC information, and the determined filters are fixed based on the HMM states. To use this method for analysis/synthesis, each frame's HMM-states should be known. Labeled training data from the same database as the HMM training data is required for the training of the HMM-state dependent filters.

Analysis

The analysis phase of the CLT method for HMM-based speech synthesis is in fact a hybrid of analysis and statistical training procedures. Each frame is analyzed for its $F0$ and MFCC information, but the main property of the method, the training of the excitation filters $H_v(z)$ and $H_u(z)$, is not done frame-by-frame but for each HMM state.

The training is done in an Analysis-by-Synthesis (AbS) scheme illustrated in Figure 4.5. The AbS scheme is similar to the low bit-rate Code Excited Linear Prediction (CELP) codec used in speech coding [65]. In the AbS scheme, the voiced excitation (filtered pulse train) signal $v(n)$ is subtracted from the target excitation (residual) signal $e(n)$ to obtain the unvoiced excitation signal $u(n)$, which is filtered with the inverse unvoiced filter $\frac{1}{H_u(z)}$ to obtain the white noise error signal $w(n)$.

The goal is to minimize the error signal by adjusting the filter coefficients of $H_v(z)$ and $H_u(z)$, as well as the form of the pulse train $t(n)$. This is done in an iterative fashion, where an improved estimation of the filter coefficients is obtained with the help of the pulse train estimation and vice versa.

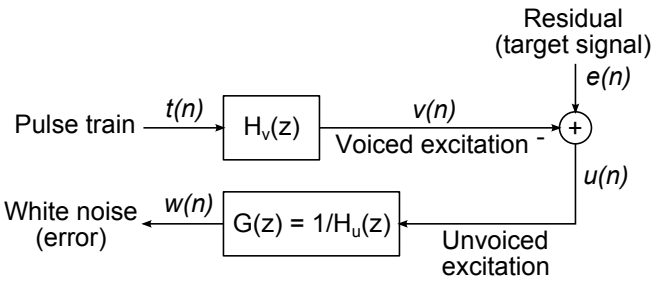


Figure 4.5: The block diagram of the AbS scheme.

The filters $H_v(z)$ and $H_u(z)$ are represented as M-order FIR and L-order IIR filters respectively:

$$H_v(z) = \sum_{l=-M/2}^{M/2} h(l) z^{-l}, \tag{4.21}$$

$$H_u(z) = \frac{1}{G(z)} = \frac{K}{1 - \sum_{l=1}^{L} g(l) z^{-l}}, \tag{4.22}$$

where K is the gain coefficient for the unvoiced filter.

The filter determination is done in a way that maximizes the likelihood of the target excitation signal $e(n)$ given the excitation model comprising of $H_v(z)$, $H_u(z)$, and pulse train $t(n)$. The likelihood of the excitation vector $\bar{\mathbf{e}} = [e(0) \cdots e(N-1)]^T$, given the voiced excitation vector $\bar{\mathbf{v}} = [v(0) \cdots v(N-1)]^T$ and matrix $\mathbf{G}$, is

$$P[\bar{\mathbf{e}}|\bar{\mathbf{v}}, \mathbf{G}] = \frac{1}{\sqrt{(2\pi)^N |\mathbf{G}^T\mathbf{G}|^{-1}}} e^{-\frac{1}{2}[\bar{\mathbf{e}}-\bar{\mathbf{v}}]^T \mathbf{G}^T \mathbf{G} [\bar{\mathbf{e}}-\bar{\mathbf{v}}]}, \tag{4.23}$$

where N is the whole database length in samples, and $\mathbf{G} = [\bar{\mathbf{g}}_0 \cdots \bar{\mathbf{g}}_{N-1}]$ is an $N \times (N+L)$ matrix containing the overall impulse response of the inverse unvoiced filter $G(z)$, which satisfies the equation

$$\bar{\mathbf{w}} = \mathbf{G}\bar{\mathbf{u}}. \tag{4.24}$$

Each column $\bar{\mathbf{g}}_j$ is obtained by

$$\bar{\mathbf{g}}_j = [0 \cdots 0 \;\; 1/K_s \;\; g_s(1)/K_s \cdots g_s(L)/K_s \;\; 0 \cdots 0]^T, \tag{4.25}$$

where there are respectively j and $(N+L-j)$ zeros before and after the inverse unvoiced filter coefficients $\{1/K_s, g_s(1)/K_s, \ldots, g_s(L)/K_s\}$. The index $s = \{1, \ldots, S\}$ indicates the HMM state in which the jth database sample belongs to.

Similarly, considering the HMM state dependency of the filters, the overall voiced excitation vector $\bar{\mathbf{v}}$ can be given by

$$\bar{\mathbf{v}} = \sum_{s=1}^{S} \mathbf{A}_s \bar{\mathbf{h}}_{v,s}, \tag{4.26}$$

where $\bar{\mathbf{h}}_{v,s} = [h_{v,s}(-M/2) \dots h_{v,s}(M/2)]^T$ is the impulse response vector of the voiced filter for HMM state s and $\mathbf{A}_s$ is the overall pulse train matrix where only the pulse train positions belonging to state s are non-zero.

The likelihood of the residual vector $\bar{\mathbf{e}}$ can be presented under the conditions of $H_v(z)$, $H_u(z)$, and $t(n)$ by placing Equation (4.26) into Equation (4.23). To solve for the maximum likelihood in relation to the voiced filter $H_v(z)$, the first partial derivative of log-likelihood function with respect to the voiced filter vector $\bar{\mathbf{h}}_{v,s}$ is set to zero and solved for $\bar{\mathbf{h}}_{v,s}$:

$$\frac{\partial \log P[\bar{\mathbf{e}}|H_v(z), H_u(z), t(n)]}{\partial \bar{\mathbf{h}}_{v,s}} = 0 \tag{4.27}$$

This results in

$$\bar{\mathbf{h}}_{v,s} = [\mathbf{A}_s^T \mathbf{G}^T \mathbf{G} \mathbf{A}_s]^{-1} \mathbf{A}_s^T \mathbf{G}^T \mathbf{G} \left[\bar{\mathbf{e}} - \sum_{l=1, l \neq s}^{S} \mathbf{A}_l \bar{\mathbf{h}}_{v,l} \right], \tag{4.28}$$

which is also the least-squares solution to the problem.

The unvoiced filter $H_u(z)$ can be expressed as the autoregressive spectral estimation (that is, for example, LPC with the autocorrelation method) of $u(n)$. For the proof, see [52]. The mean autocorrelation function is estimated for each HMM state, and LPC analysis using the autocorrelation method is applied to each state with the acquired autocorrelation estimates. The obtained LPC analysis coefficients are the filter coefficients of the inverse unvoiced filter $G(z) = 1/H_u(z)$, and the minimum phase all-pole LPC synthesis filter is the estimate for the unvoiced filter $H_u(z)$.

The optimization of the position and amplitude of the pulses is a necessary part of the filter training process to obtain accurate filter estimates. Pulse optimization is conducted so that the mean squared error (MSE) $\epsilon = \frac{1}{N} \bar{\mathbf{w}}^T \bar{\mathbf{w}}$ of Figure 4.6 is minimized while keeping the excitation filters $H_v(z)$ and $H_u(z)$ constant.

The obtained solution for the pulse train amplitudes a_i is [52]:

$$a_i = \frac{\bar{\mathbf{h}}_{gi}^T \left[\bar{\mathbf{e}}_g - \sum_{j=1, j \neq i}^{Z} a_j \bar{\mathbf{h}}_{gj} \right]}{\bar{\mathbf{h}}_{gi}^T \bar{\mathbf{h}}_{gi}}, \tag{4.29}$$

where $\bar{\mathbf{h}}_{gj}$ is the impulse response vector of the combined filters $H_v(z)$ and $H_u(z)$ with a delay of j samples, and Z is the number of pulses in the optimization process.

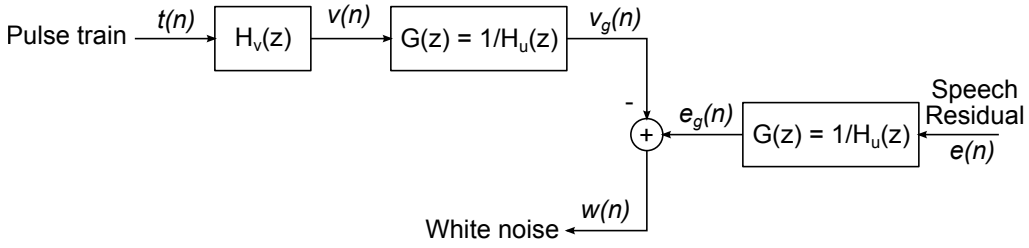


Figure 4.6: The block diagram of the pulse optimization scheme.

The best position p_i for each pulse is obtained by:

$$p_i = \arg \max_{p_i=1,\ldots,N} \frac{\left[\bar{\mathbf{h}}_{gi}^T \left(\bar{\mathbf{e}}_g - \sum_{j=1,j\neq i}^{Z} a_j \bar{\mathbf{h}}_{gj}\right)\right]^2}{\bar{\mathbf{h}}_{gi}^T \bar{\mathbf{h}}_{gi}} \tag{4.30}$$

The iterative algorithm used in the method first initially estimates the pulse amplitudes and positions according to Equations (4.29) and (4.30). Next, the voiced filters are estimated according to Equation (4.28), and the unvoiced filters are estimated using LPC analysis. Finally, the pulse train amplitudes and positions are adjusted according to the estimated filters. These steps are iterated until the algorithm converges or the maximum amount of iterations is reached.

Synthesis

The synthesis phase of Maia's HMM-adjusted CLT vocoder is illustrated in Figure 4.7:

First, the F0 and MFC coefficients as well as HMM state durations are generated from the trained HMMs. Next, the voiced and unvoiced filters H_v and H_u are determined according to each HMM-state. A periodic pulse train and white Gaussian noise are generated according to the frame-by-frame generated trajectory, and they are filtered with their respective filters. Finally, the voiced and unvoiced excitations are combined to form the final mixed excitation, which is filtered with the MLSA filter derived from the MFC coefficients. It is notable that the F0 and MFCC values vary frame-by-frame, but the mixed excitation filters vary according to the HMM states.

The synthesis quality of the method was evaluated in a subjective listening test by comparing it to a simple pulse excitation vocoder (see Section 4.1) and the HMM-adjusted STRAIGHT vocoder (see Section 4.2.2) [52]. The test type was pair comparison test, and the results were: Proposed system 60%, STRAIGHT 58.3%, and simple excitation 31.7%.

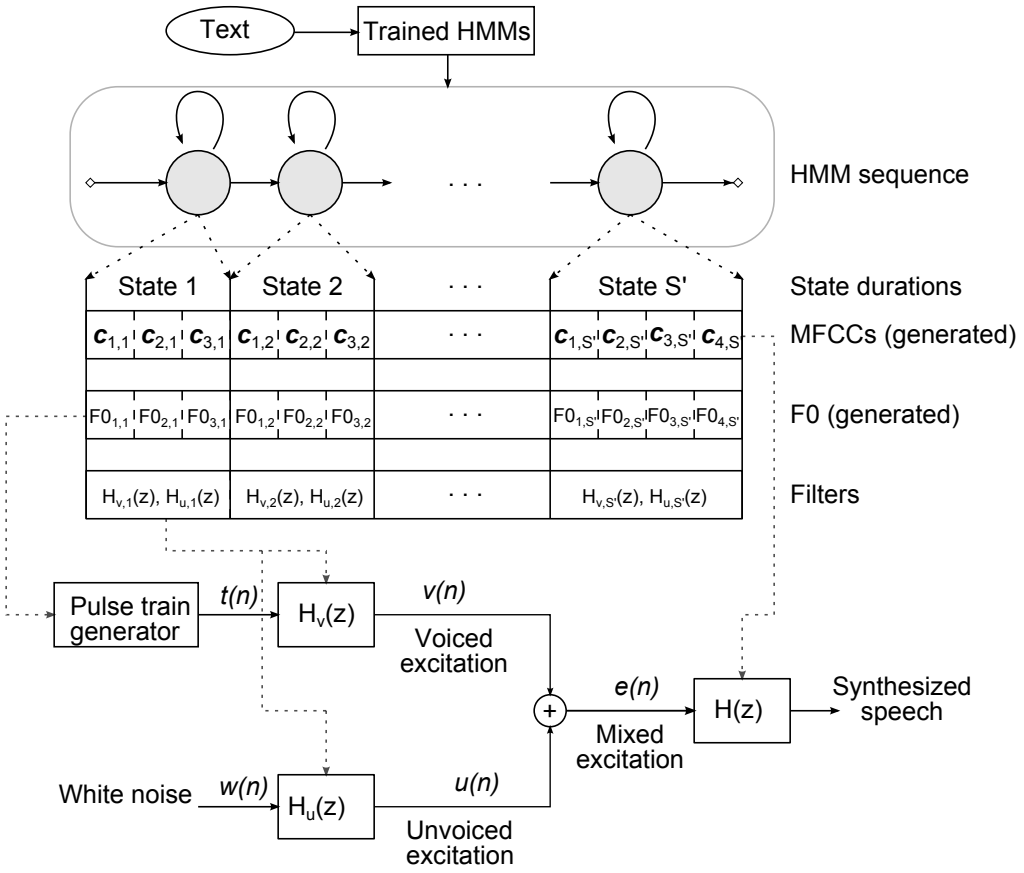


Figure 4.7: The block diagram of the CLT synthesis scheme.

4.3.2 Pitch-synchronous Residual Codebook

The vocoding method proposed by Drugman et al. [22] is based around the philosophy that the best way to reduce buzziness in vocoded speech is to use parts of the *real* excitation signal. In the proposed method, a pitch-synchronous (PS) residual codebook of typical excitation frames is constructed. In synthesis they are overlap-added with modifications to produce the synthetic excitation signal. The codebook frames used in the synthesis are determined by selecting the frame from the codebook that has the closest euclidean distance (in terms of compressed coefficients) from the target frame. The Pitch-Synchronous Residual Codebook vocoder for HMM-based speech synthesis can thus be considered as a modified PS-CELP codec [29].

Analysis

The analysis phase of the PSRC vocoder can be divided into a preliminary and main analysis phases. In the preliminary phase, the residual codebook is constructed from a speech database, and in the main phase the actual speech-to-be-analyzed is analyzed.

The construction of the residual codebook is carried out as follows: First, the training database is filtered with an analysis filter derived from mel-generalized cepstral (MGC) coefficients [49] to obtain the residual signal. Next, a peak-picking algorithm presented in [20] is applied to the residual signal to obtain the timings of the peaks that are approximated as the glottal closure instants (GCI). After the determination of the GCIs, the signal is cut into GCI-centered, two pitch-period long frames that are Hanning windowed.

The footmark of the frames is condensed by the means of Resampling and Normalizing (RN) each frame. The residual frames are resampled to 20 samples, and then normalized in energy, making them amenable to clustering. K-means clustering around 100 centroids is applied to the RM-modified frames to obtain the RN codebook. A representative original frame is selected from each cluster of the RN codebook with the criterion that it is the longest frame within 10 closest frames to the cluster centroid. A codebook of original frames is constructed from such frames, which are linked to their RN codebook counterparts.

The main analysis phase is carried out for each frame as follows: First, the $F0$ and MGC coefficients are determined. Next, the residual signal is obtained with the analysis filter derived from the MGC coefficients, and the RN procedure is applied to the residual. The RN-modified frames can be used as such in pure analysis/synthesis, but they are heavily correlated and as such unusable in statistical modeling with HMMs. That is why Principal Component Analysis (PCA) without dimensionality reduction is used to linearly transform and decorrelate the coefficients. The HMMs are trained with these PCA-transformed versions of the residual signals.

Table 4.5: The analysis vector of the pitch-synchronous residual codebook vocoder, where p denotes the order number of the spectral analysis.

Excitation parameters	$1 \times F0$
	$20 \times$ PCA-transformed RN frame samples
Spectral parameters	$p \times$ MGC coefficients

The overall feature vector produced by the proposed method is shown in Table 4.5. The spectral properties are modeled with the mel-generalized cepstral coefficients, and the excitation is modeled using the $F0$ value as well as the 20 coefficients representing the PCA-transformed RN frame.

Synthesis

The synthesis phase for the Pitch-Synchronous Residual Codebook method (Figure 4.8) is carried as follows: First, the PCA-transformed Resampled and Normalized (RN) residual frames are converted back to their RN form. These frames are compared to the frames in the RN codebook, and the RN codebook frame that minimizes the euclidean distance (mean squared error) is selected as the synthesis frame index.

The frame corresponding to the synthesis frame index is selected from the original signal codebook, and its pitch and energy is modified to match the target frame.

The final synthetic signal is obtained by applying the Pitch-Synchronous Overlap-Add algorithm [56] to each excitation frame, and filtering them with the Mel-Generalized Log Spectral Approximation (MGLSA) filter derived from the MGC coefficients.

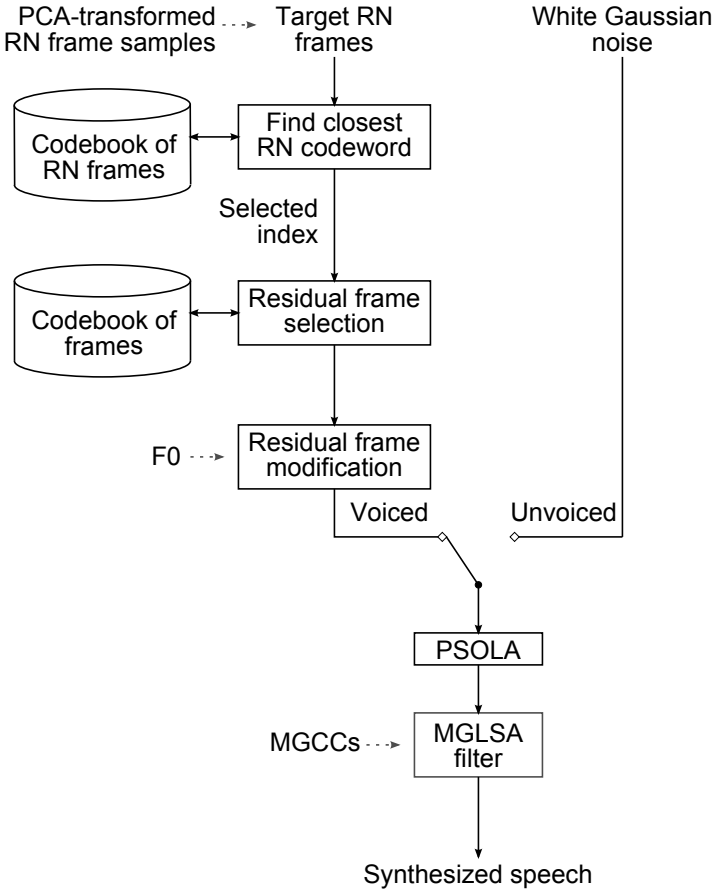


Figure 4.8: The block diagram of the CLT synthesis scheme.

The synthesis quality of the method was tested in two different ways: First, the analysis/synthesis quality of the method was tested by a Mean Opinion Score (MOS) test where the reference points were set by the original speech sample and the simple pulse excitation vocoder (see Section 4.1). The MOS scores were found to lay in between the reference points, with original samples around 4.5, PSRC samples around 3-3.5 and the pulse excitation samples around 1.5-2 [22].

The second test performed used the HMM synthesis framework, where the proposed PSRC method was compared to the simple pulse excitation vocoder in a pair comparison test. In the pair comparison test subjects select which sound they prefer (or claim no preference) out of two presented samples. For male speakers, the proposed method was preferred over 80% of the time, but for a female test voice the proposed method was not clearly preferred, with over 50% of the answers falling into the "No Preference" category [22].

4.3.3 Deterministic plus Stochastic Model

The Deterministic plus Stochastic Model (DSM), first described by Drugman et al. in [23] and more recently with changes in [21], is a refinement of the Pitch Synchronous Residual Codebook (PSRC) vocoder desrcibed in Section 4.3.2. More precisely, the DSM vocoder can be seen as a hybrid of the PSRC vocoder and the Harmonic plus Noise Model (HNM) described in Section 4.2.3. Because of this reason, the method has some traits of the Mixed Multi-band Excitation vocoders as well as the Residual Modeling vocoders. The decision to include the DSM vocoder in the residual modeling category was based on the fact how the excitation is modeled in a least squares (LS) sense, as can be seen in the *Analysis* section.

The speech production model used by the DSM vocoder is essentially the same that is used for the HNM/Two-Band Excitation method: Low-band excitation (residual) is modeled as a harmonic component that is spearated by a *maximum voiced frequency* F_m from the high-band excitation that is modeled as time- and frequency-domain modulated white Gaussian noise.

Instead of a low-pass filtered pulse train, the deterministic component is modeled for each speaker and style by the means of residual modeling from a database of Glottal Closure Instant (GCI) centered residual frames. With the pre-modeled parameters, the DSM vocoder uses only the $F0$ and mel-generalized cepstral (MGC) coefficients [49] for the statistical parametric representation of the speech signal.

Analysis

The main interest of the analysis part of the DSM vocoder is the determination of the speaker and style dependent deterministic and stochastic components $r_d(t)$ and $r_s(t)$, which is done utilizing a database of training data. The database of Glottal Closure Instant (GCI) centered residual frames is constructed as follows: First, the residual signal of each speech sample is obtained by inverse filtering the original signal with the MLSA filter obtained from the MGC coefficients. Next, the GCIs are detected from the residual signal by using a detection method described in [20]. The GCIs correspond to the point of the highest excitation of the glottal flow derivative in the speech signal, which correspond to the high energy peaks in the residual signal. After the GCIs are determined, the residual is Blackman windowed GCI-synchronously into GCI-centered, two-pitch-period long frames to form the database.

For the maximum voiced frequency F_m, a fixed value is selected depending on the speaking style of the dataset. It is argued in [21] that the use of a fixed F_m is justified, because without a great loss in quality, it circumvents the problem of estimating accurate F_m trajectories, and it alleviates the problem of too low estimates of F_m which add unpleasant noise to the synthesized voice. The fixed value is selected so that it is higher for loud speech (for example 4600 Hz) and lower for soft speech (for example 2460 Hz).

The modeling of the low-frequency deterministic component is done by decomposing the pitch synchronous residual database on an orthogonal basis obtained by Principal Component Analysis (PCA) [39]. Preliminarily to the PCA the residual frames in the database are normalized in pitch and energy, which ensures the coherence of the dataset. The normalized fundamental frequency $F0^*$ must meet the following criterion to ensure that there are no energy holes (band of the low-frequency component not reaching the maximum voiced frequency F_m) in the normalized database with respect to synthesis:

$$F0^* \leq \frac{F_N}{F_m} \cdot F0_{min}, \tag{4.31}$$

where F_N is the Nyqvist frequency, and $F0_{min}$ is the smallest $F0$ value in the database.

The PCA is applied to the dataset of N normalized m-length residual frames by calculating the eigenvectors and eigenvalues of the data covariance matrix (the $m \times m$ covariance matrix is estimated from the N residual frames). This computation leads to m eigenvalues λ_i and their corresponding eigenvectors μ_i (of length m). Because of the orthogonal linear transformation of the PCA, the eigenvectors point into the directions that maximize the data dispersion along the axes (or gives the best representation of the data in least squared sense). λ_i represent the amount of data dispersion along the axis μ_i, and thus efficient dimensionality reduction can be carried out by selecting only the k largest eigenvalues and their corresponding eigenvectors to the modeled representation. Moreover, it is shown in [21] that only the eigenvector, or *eigenresidual*, μ_1 corresponding to the largest eigenvalue λ_1 is needed for an accurate representation of the low-band residual signal. Thus, the first eigenresidual $\mu_1(n)$ is selected to model the deterministic component of the DSM.

The model used for the stochastic part $r_s(t)$ of the DSM vocoder consists of white Gaussian noise $n(t)$ convolved with an autoregressive model $h(t)$, with its time structure controlled by an energy envelope $e(t)$:

$$r_s(t) = e(t) \cdot [h(t) \circledast n(t)] \tag{4.32}$$

For the estimation of $h(t)$ and $e(t)$, the original residual signal database is modified as follows: First, the signals are normalized in energy, and then they are high-pass filtered with a cut-off frequency of F_m. From these signals, $h(t)$ is estimated as the linear predictive envelope of the average amplitude spectrum. As most of the spectral information is absent in the residual signal due to the inverse filtering, the estimated $h(t)$ in practice acts as a high-pass filter with a cut-off frequency of F_m.

The energy envelope $e(n)$ is determined as the average Hilbert envelope of the high-pass filtered dataset re-sampled to the normalized pitch value $F0^*$ [60].

After the deterministic and stochastic components for the target speaker are determined from the training database, sample sounds can be analyzed. The only coefficients needed to analyze at this point are only the $F0$ and mel-generalized

cepstral coefficients, which are also the only input streams for the HMM training. This property makes the proposed method very viable for applications with limited computational resources.

Table 4.6: The analysis vector of the DSM vocoder.

Excitation parameters	$1 \times F0$
Spectral parameters	$p \times$ MGC coefficients

Synthesis

The block diagram for the synthesis phase of the DSM vocoder can be seen in Figure 4.9. The pre-determined eigenresidual μ_1 is input as the basis of the voiced excitation, and it is re-sampled according to the $F0$ value of the frame, and low-pass filtered according to the maximum voiced frequency F_m. Unvoiced excitation is generated by frequency- and time modulating white Gaussian noise by the $h(n)$ and $e(n)$ envelopes respectively. The $h(n)$ envelope is obtained from the pre-determined LPC coefficients, and the $e(n)$ envelope is obtained by re-sampling the pre-determined envelope with respect to $F0$.

The voiced and unvoiced excitations are added together to form mixed excitation that is overlap-added [56] to form the final excitation signal. The excitation is input to the MLSA filter derived from the MGC coefficients.

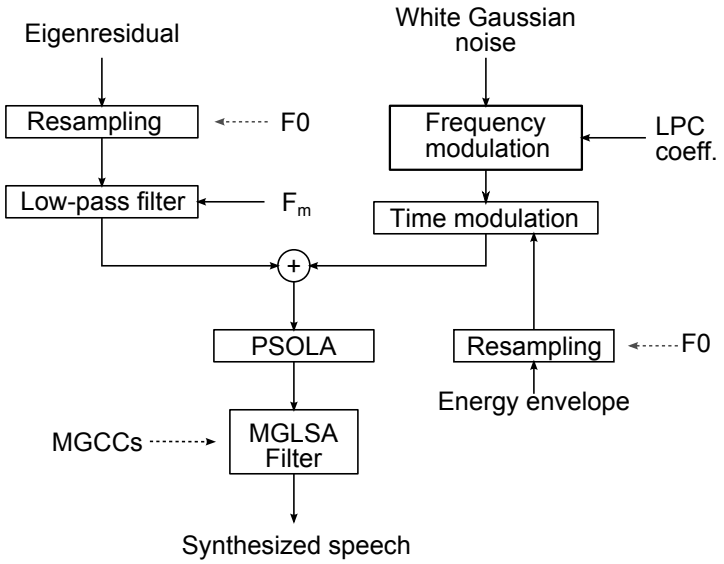


Figure 4.9: The synthesis block diagram of the DSM vocoder.

The quality of the DSM vocoder was tested against the STRAIGHT vocoder (see Section 4.2.2) and the pulse excitation vocoder (see Section 4.1) in a subjective Comparative Mean Opinion Score (CMOS) test [21]. The results gave nearly identical performance compared to STRAIGHT, and clearly better performance against

the pulse excitation. The results were found to be slightly better for male voices than female voices. The advantage of the DSM vocoder over STRAIGHT is its light computational footprint, but its main problem is that because of the reliance in non-parametric characteristics, the method is poorly flexible.

4.4 Glottal Source Modeling Vocoders

Glottal source modeling vocoders are motivated by the fact that they use estimated characteristics of the real glottal pulse in the determination of the excitation signal. These methods include the GlottHMM vocoder [64], Glottal Post-Filtering vocoder [12], and the Glottal Spectral Separation vocoder [11].

4.4.1 GlottHMM

The GlottHMM vocoder was first proposed by Raitio in [62], and later in refined form by Raitio et al. in [64]. The main idea behind the GlottHMM vocoder is that it estimates the real glottal pulse signal $G(z)$ and the real vocal tract filter $V(z)$ associated with it. This is done for example by utilizing a method called Iterative Adaptive Inverse Filtering (IAIF) [4]. As described in Section 2.2, conventional source-filter models estimate all of spectral envelope features into the filter part $F(z)$, and the rest of the signal into the source part $E(z)$:

$$S(z) = E(z)F(z) \tag{4.33}$$

The conventional spectral envelope includes the spectral properties of the glottal pulse, but with the approach of GlottHMM, the speech signal can be represented as:

$$S(z) = G(z)V(z)L(z), \tag{4.34}$$

where $L(z)$ is the lip radiation effect (see Section 2.1), and all parts are estimates of real physical properties. $G(z)$ can be written as

$$G(z) = E(z)F_G(z), \tag{4.35}$$

where $F_G(z)$ is a filter containing the spectral envelope of the glottal pulse. With this, the relation between $F(z)$ and $V(z)$ can be written as:

$$F(z) = V(z)F_G(z)L(z) \tag{4.36}$$

$$V(z) = \frac{F(z)}{F_G(z)L(z)} \tag{4.37}$$

The advantage of using the proposed method for the representation of speech is that real glottal pulses can be used as the excitation for the synthetic speech signal, which provides more natural synthesis quality compared to the pulse train excitation. Also, the glottal flow spectrum can easily be adapted and/or modified.

Analysis

The analysis phase of the GlottHMM vocoder works as follows: First, the speech signal is high-pass filtered (cut-off frequency of 70 Hz) and windowed into fixed length rectangular frames, from which the signal log energy is calculated as a feature parameter. Second, the Iterative Adaptive Inverse Filtering (IAIF) algorithm [4] is applied to each frame, which results in the LPC representation of the vocal tract spectrum and the waveform representation of the voice source. The LPC spectral envelope estimate of the voice source is calculated, and along with the LPC estimate of the vocal tract spectral envelope, is converted into a LSF representation (see Section 2.3.2). The glottal flow waveform is used also for the acquisition of the $F0$ value as well as the Harmonic-to-Noise Ratio (HNR) values for a predetermined amount of sub-bands of frequency.

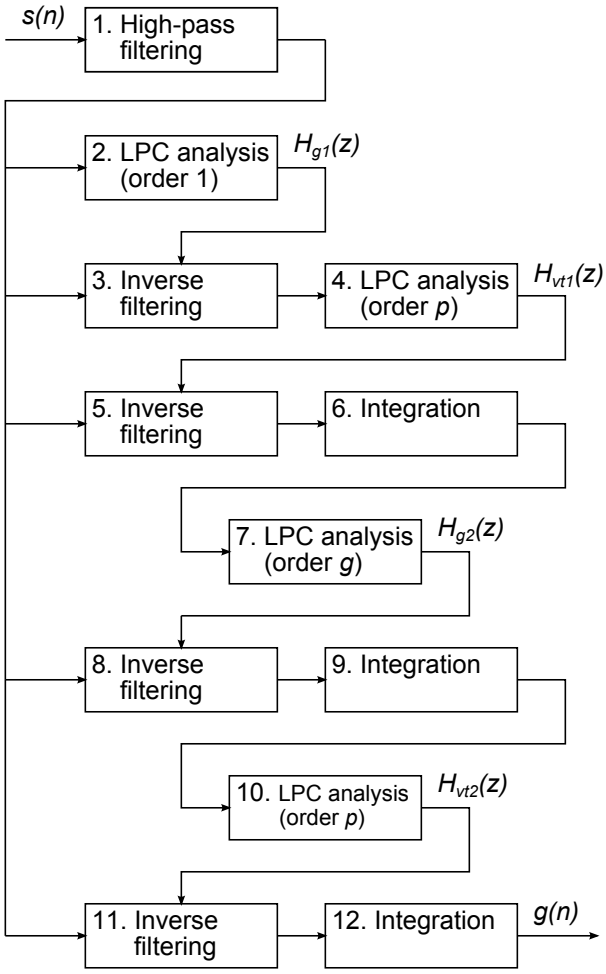


Figure 4.10: The block diagram of the IAIF algorithm.

The flow-chart of the IAIF algorithm is presented in Figure 4.10. For input, the algorithm needs only the high-pass filtered speech pressure signal $s(n)$, and for output it gives out the estimated vocal tract filter $H_{vt2}(z)$ and the glottal source pressure signal $g(n)$. The high-pass filtered signal frame $s(n)$ is LPC analyzed (with

order of 1) to obtain an initial estimate of the glottal source spectrum $H_{g1}(z)$. Signal frame $s(n)$ is inverse filtered with $H_{g1}(z)$, and the obtained signal is LPC analyzed (with order p) to obtain the initial estimate for the spectral envelope $H_{vt1}(z)$. Next, $s(n)$ is inverse filtered with $H_{vt1}(z)$ to cancel out the effects of the vocal tract and integrated (inverse filtered with $L(z)$) to cancel out the effects of the lip radiation effect. The resulting glottal flow signal estimate is LPC analyzed (order g) to obtain the refined spectral envelope estimate of the glottal flow $H_{g2}(z)$. Again, $s(n)$ is inverse filtered by $H_{g2}(z)$ and $L(z)$ to obtain the final signal, whose spectral envelope $H_{vt2}(z)$ is estimated with LPC analysis of order p. The final estimate of the glottal flow signal $g(n)$ is obtained by inverse filtering $s(n)$ with $H_{vt2}(z)$ and $L(z)$.

IAIF has been shown to produce satisfactory performance in estimating the glottal source parameters in [5].

The estimated glottal flow signal $g(n)$ is used to produce the rest of the analysis parameters. A voicing decision is made based on the amount of zero-crossings and on low-band (less than 1 kHz) energy. For voiced frames, the $F0$ value of the frame is estimated using the autocorrelation method [32]. The Harmonic-to-Noise Ratio (HNR) [57] is calculated from $g(n)$ as follows: The Fourier transform of the signal is calculated, from which the cepstrum of each frequency band is evaluated (see Section 2.4). For each frequency band, the degree of harmonicity is determined by the strength of the cepstral peak (defined by $F0$) in ratio to the averaged value of other quefrencies of the cepstrum. For unvoiced frames, the $F0$ and HNR values are set to zero.

Table 4.7: The analysis vector of the GlottHMM vocoder, where p denotes the order number of the vocal tract spectral analysis, m denotes the number of HNR sub-bands, and n denotes the order number of the source spectral analysis.

Excitation parameters	$1 \times F0$
	$1 \times$ log energy
	$m \times$ HNR
	$n \times$ glottal source LSF
Spectral parameters	$p \times$ vocal tract LSF

The final analysis vector of the GlottHMM consists of the parameters depicted in Table 4.7. Single parameters are used for the $\log F0$ and log energy, m parameters (typically around 5) are used for the HNR coefficient, n parameters (typically around 10-20) are used for the glottal source LSF parameters, and p parameters (typically around 20-30) are used for the vocal tract LSF parameters. Compared to the simple impulse excitation vocoder (see Section 4.1), the GlottHMM vocoder has significantly more parameters dedicated for the modeling of the excitation: The $F0$, HNR, and the source LSF coefficients all are used to model the excitation (or source) signal that is filtered by the vocal tract filter.

Synthesis

The synthesis block diagram for the GlottHMM vocoder is depicted in Figure 4.11. Unlike most of the state-of-the-art vocoders, the GlottHMM vocoder does not use a traditional mixed excitation model for the excitation generation. The method used for the excitation generation is based on the voiced/unvoiced decision.

For voiced frames, the heart of the synthesis procedure is a fixed library pulse that is obtained by glottal inverse filtering a sustained vowel signal. The library pulse is interpolated to match the target $F0$ value using cubic spline interpolation, and its energy is set to match the target gain obtained from the analysis vector.

Next, a HNR analysis is done to the library pulse in a similar way as in the analysis phase. For each sub-band, noise is added to the real and imaginary parts of the FFT vector according to the differences between the obtained and the target HNR values. This acts like the mixed excitation for voiced frames.

The spectrum of the library pulse is matched to the spectrum of the target glottal pulse obtained from the analysis vector. The spectral matching is done by performing LPC analysis (order m) to the library pulse, and then filtering the obtained residual with the target synthesis filter (order m). Finally, the lip radiation effect is added to the excitation by filtering it with a fixed differentiator.

For unvoiced frames, the excitation is generated as white Gaussian noise whose gain is set by the energy parameter of the analysis vector.

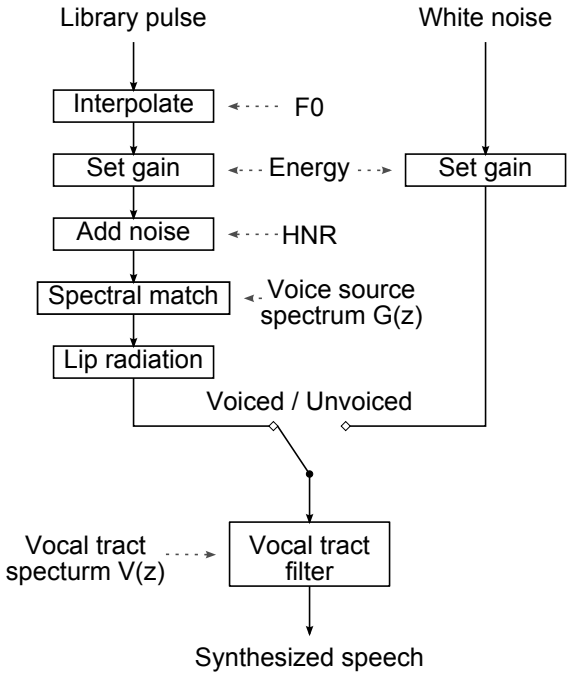


Figure 4.11: The synthesis block diagram of the GlottHMM vocoder.

The excitation is combined in the time domain by overlap-adding [56] target frames, and the final synthetic signal is generated by filtering the excitation with the vocal tract filter derived from the vocal tract LSFs obtained from the analysis

vector.

The synthesis quality of the GlottHMM vocoder has had encouraging results especially with a male voice, where it has clearly surpassed the STRAIGHT vocoder (see Section 4.2.2) in a subjective listening test [64]. However, female voices have been shown to be more problematic with the method [71]. Also, the GlottHMM vocoder uses comparably large amounts of excitation parameters, which makes its computational footprint large.

4.4.2 GlottHMM with Pulse Library Technique

Recently, GlottHMM has undergone some modifications from its first proposed form. In the newly proposed version [71], [63], a glottal pulse library is constructed from a speech database to ensure even more natural glottal pulse excitation to the synthesis. Also, the IAIF procedure has been simplified to yield more robust estimates at the expense of exactness.

The pulse library version of the GlottHMM vocoder works in a same way as the original version, but instead of a fixed single library pulse, it uses a *pulse library* constructed of various glottal inverse filtered pulses. A pulse that matches the selected attributes of the target pulse (with the target cost based on the root mean square, rms, of selected attributes) is selected as the one to be modified. The attributes used to determine the target cost include the spectral envelope, $F0$, HNR, spectral tilt, and energy of the glottal pulse.

Analysis

The analysis phase is carried out similarly to the original GlottHMM vocoder (see Section 4.4.1), including the $F0$ and HNR estimation. The source and vocal tract filter estimation is also done in a similar fashion, but a modified version of the IAIF algorithm is used. That is because the original IAIF algorithm yields accurate estimates of the voice source at its best, but in adverse conditions the estimates may vary greatly from frame to frame [71].

The block diagram for the modified IAIF algorithm is presented in Figure 4.12. It has just one iteration of the parameter estimation steps, down from two from the version in Section 4.4.1. The use of stabilized weighted linear prediction (SWLP) [51] in the estimation of the vocal tract filter mitigates the effect that the harmonic peaks have on the modeled formants.

The construction of the pulse library is performed preliminarily to the analysis phase by taking a segment of speech from the target speaker, and applying the (modified) IAIF algorithm to it so that the glottal excitation signal is obtained. From this signal, the glottal closure instants (GCIs) are detected, and GCI centered two-period long segments are extracted and windowed with Hann windowing. The obtained glottal pulses are normalized in energy, and saved in the pulse library with their voice source parameters (all parameters of the analysis vector (Table 4.8)).

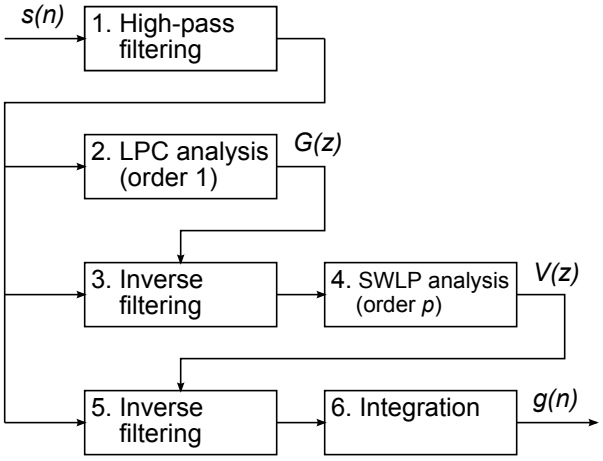


Figure 4.12: The synthesis block diagram of the modified IAIF algorithm.

Table 4.8: The analysis vector of the GlottHMM vocoder with pulse library, where p denotes the order number of the vocal tract spectral analysis, m denotes the number of HNR sub-bands, and n denotes the order number of the source spectral analysis.

Excitation parameters	$1 \times F0$
	$1 \times$ log energy
	$m \times$ HNR
	$n \times$ glottal source LSF
Spectral parameters	$p \times$ vocal tract LSF

Synthesis

The synthesis phase of the GlottHMM vocoder with the pulse library technique is more simple than with the normal GlottHMM vocoder. Instead of modifying a single fixed library pulse, the closest glottal pulse is selected from the pulse library, and it is used without further modifications.

The pulse is selected from the pulse library so that it minimizes the target and concatenation costs. The target cost is composed of the RMS error between the voice source parameters of the pulse and the parameters generated by the HMM (or analysis phase in analysis/synthesis). The individual weights for the voice source parameters are set according to subjective experiments. Minimizing the target cost ensures that the selected pulse has the desired voice source characteristics. The concatenation cost is computed as the RMS error between consecutive down-sampled pulse waveforms in each full voiced section. Minimizing the concatenation cost ensures that adjacent pulse waveforms do not differ substantially from each other, which could degrade the quality. The best matching pulses that minimize the joint target and concatenation costs are searched from the pulse library by using the Viterbi algorithm [6].

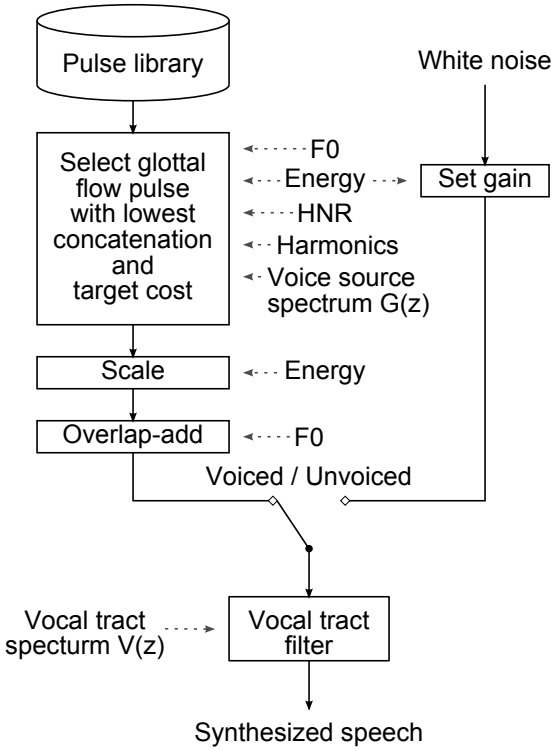


Figure 4.13: The synthesis block diagram of the GlottHMM vocoder with pulse library.

The block diagram of the synthesis phase is presented in Figure 4.13. Notice that the synthesis procedure is identical to the synthesis procedure of the original GlottHMM vocoder (Figure 4.11), with the exception that instead of using the single library pulse, the best matching pulses are searched from the pulse library and used without substantial modifications. The synthesis quality of the modified GlottHMM vocoder has been tested on a few sources. In the Blizzard Challenge 2011 [71], the goal was to construct a female voice. The results showed an improvement in the synthesis quality of the female voice compared to the old method, but compared to other state-of-the-art methods in the challenge, the quality was mediocre. In [63] the pulse library technique was compared to the single pulse version of GlottHMM for male voices, and the pulse library version was found to get slightly better results in CCR and pair comparison tests. The method is still in its early stages of development, and the authors are expecting better results with further experimentation on the pulse library and the vocal tract parametrization [71].

4.4.3 Glottal Post-Filtering

The Glottal Post-Filtering (GPF) vocoder proposed by Cabral et al. in 2007 [12], and later in more detail by Cabral in 2010 [10], proposes the use of the Liljencrants-Fant (LF) model [26] to model the voiced excitation instead of binary pulses in the framework of statistical parametric speech synthesis. The fixed, speaker specific pitch-adaptive excitation model is used alongside with the spectral model of the HTS

STRAIGHT vocoder [75] (see Section 4.2.2). The use of the LF-model is thought to have the advantage of a less harmonic structure at high-frequencies of the spectrum compared to the pulse train. Also, the use of the LF-model permits flexibility to transform voice quality by modifying the glottal parameters [12].

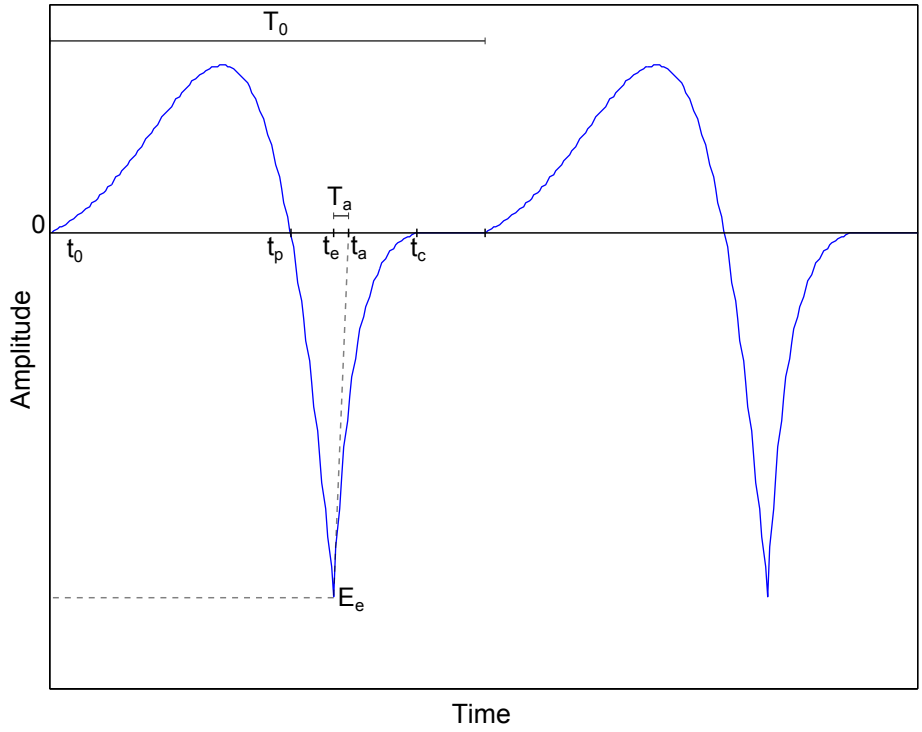


Figure 4.14: The Liljencrants-Fant model.

The LF-model approximates of the glottal-flow derivative waveform by using seven parameters (including the fundamental period T_0) that are estimated from the LPC residual signal by Cabral et al. (see Sections 2.2.2 and 2.3). The LF-model waveform is depicted in Figure 4.14 along with the LF-model parameters. The model is divided into three parts, expressed in mathematical form as:

$$e(t) = \begin{cases} E_0 e^{\alpha t} \sin(\omega_g t), & 0 \leq t \leq t_e \\ -\frac{E_e}{\epsilon t_\alpha}[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}], & t_e < t \leq t_c \\ 0, & t_c < t \leq T_0 \end{cases}, \tag{4.38}$$

where $\omega_g = \frac{\pi}{t_p}$, t_0 is the opening instant of the vocal folds, t_p is the instant of maximum airflow, t_e is the instant of maximum negative amplitude E_e, t_a is the duration from t_e to the point where a tangent to the exponential at $t = t_e$ hits the time axis (measuring the abruptness of the closure), t_c is the instant where the exponential part ends, and T_0 is the length of the fundamental period. The scaling parameters E_0, ϵ, and α can be calculated from Equation 4.38 by imposing $e(t_e) = E_e$ and the energy balance $\int_0^T e(t)dt = 0$ [10].

The spectral properties of the LF-model illustrated in Figure 4.15 can be characterized by three asymptotic lines with +6 dB/oct, -6 dB/oct, and -12 dB/oct

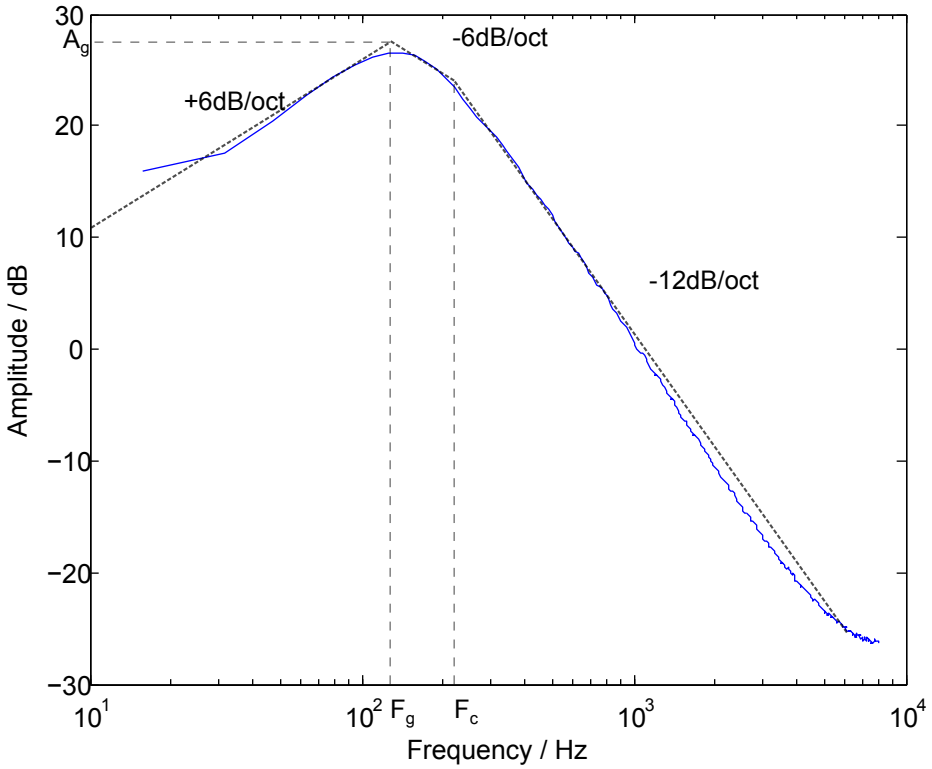


Figure 4.15: The spectrum of the LF-model waveform.

slopes [19]. The point of crossing of the first two lines denotes the *glottal spectral peak*, whose position is defined by the frequency F_g. The third line is formed due to the contribution of spectral tilt, which adds an additional -6 dB/oct above the frequency F_c. The frequency F_g can be estimated by:

$$F_g = \frac{1}{2\pi \frac{t_e + t_a}{T_0} T_0} \sqrt{\frac{e(\alpha_m)}{j(\alpha_m)}}, \tag{4.39}$$

where $j(\alpha_m)$ and $e(\alpha_m)$ are functions of the asymmetry coefficient $\alpha_m = \frac{t_p}{t_e - t_p} / (1 + \frac{t_p}{t_e - t_p})$ [16]. F_c can be estimated by the following expression [26]:

$$F_c = \frac{1}{t_a 2\pi} \tag{4.40}$$

Analysis

The main part of the analysis phase in the GPF vocoder consists of the determination of the speaker specific LF-model parameters to be used in synthesis, which is done preliminarily to the "real" analysis. The parameters are estimated for a number of utterances, after which the mean of their fundamental-period-normalized versions is computed to obtain the speaker-specific estimate. The LF-model parameters are estimated as follows:

The signals are preprocessed by low-pass filtering with a 4 kHz cut-off frequency, and they are 20 ms Hanning windowed (centered at glottal epochs). For each frame, the LPC residual signal is obtained by analysis filtering. The t_e and E_e can be estimated straight from the residual by selecting the location and amplitude of the residual peak respectively. The residual signal is high-pass filtered with a cut-off frequency of 80 Hz to reduce gross errors in the following integration. Next, an estimate of the glottal flow waveform is obtained by integrating the high-pass filtered residual signal. From the estimate, the location of the maximum glottal flow amplitude U_{max} gives the estimate of the parameter t_p, and the location of the minimum glottal flow amplitude U_{min} gives the estimate of the parameter t_c. t_0 can be estimated by

$$t_0 = \frac{2(U_{max} - U_{min})}{\pi E_{max}}, \tag{4.41}$$

where E_{max} is the maximum value of the residual in the period.

Finally, t_a is estimated by calculating the derivative of the residual signal, and finding its maximum value M, which is used to estimate t_a as follows:

$$t_a = \frac{E_e}{M F_s}, \tag{4.42}$$

where F_s is the sampling frequency of the signal.

With the exception of t_a and E_e, the LF-model parameters were found to increase linearly with the fundamental period T_0 [12], which is the reason that they are normalized by the pitch period. As the final steps in the estimation of the speaker-specific LF-parameters, the normalized estimates are median filtered, and their means are calculated to obtain the final estimate.

Table 4.9: The analysis vector of the GPF vocoder, where p denotes the order number of the vocal tract spectral analysis.

Excitation parameters	$1 \times F0$
	$5 \times$ aperiodicity measures
Spectral parameters	$p \times$ STRAIGHT MFCC

After the acquisition of the speaker-specific LF-parameters, the analysis phase is carried out identically to the STRAIGHT vocoder described in Section 4.2.2, with the F0, STRAIGHT mel-cepstrum, and aperiodicity coefficients making up the analysis feature vector (Table 4.9).

Synthesis

The synthesis phase of the GPF vocoder is illustrated in Figure 4.16. The voiced excitation is generated as a LF-model pulse according to Equation 4.38, where the pitch-normalized speaker-specific LF-model parameters are scaled according to the

$F0$ value of the frame. The LF-waveform can not be used as such for the excitation, because the STRAIGHT spectrum includes the spectral properties of the glottal pulse in its estimate. Because of this, the spectral envelope of the LF-waveform is flattened by the means of *post-filtering*. The post-filter is defined by three linear segments which are symmetric to the slopes of the LF-model spectrum (-6 dB/oct, +6 dB/oct, and +12 dB/oct). The cut-off frequencies F_g and F_c are calculated according to Equations 4.39 and 4.40 respectively.

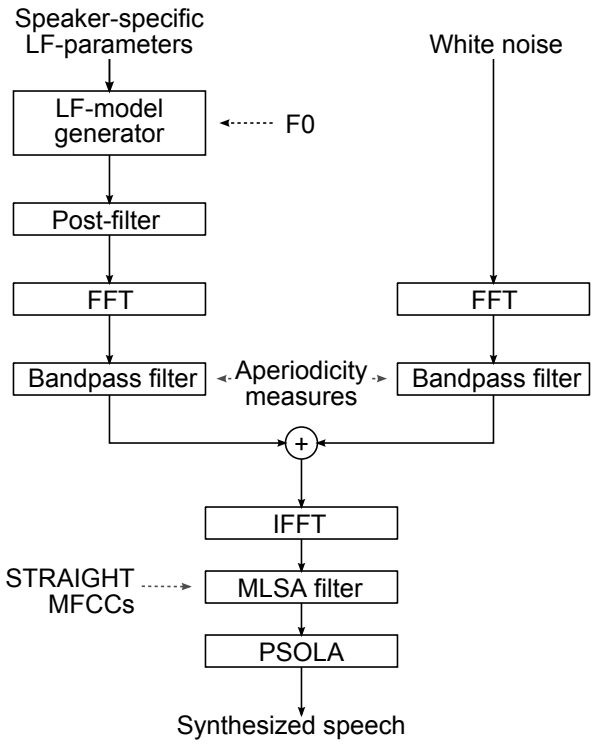


Figure 4.16: The synthesis block diagram of the GPF vocoder.

The rest of the synthesis phase is carried out as in the HMM-adapted HTS STRAIGHT vocoder (see Section 4.2.2, *Synthesis*): Voiced and unvoiced excitation are weighted in the frequency domain according to the aperiodicity measurements, phase manipulation is applied to the voiced excitation, and the combined mixed excitation is transformed back to the time-domain. The pitch-synchronous overlap-add algorithm [56] is applied to the generated frames to obtain the final excitation signal, which is filtered with the MLSA filter derived from the STRAIGHT MFCC coefficients to obtain the synthesized signal.

The synthesis quality of the GPF method has been shown to surpass the impulse excitation vocoder [12], but it has not been able to surpass the baseline HTS STRAIGHT system in quality [10]. The main problem of the GPF vocoder is the accuracy of the glottal post filtering, which is done using a simplified model of the glottal-source derivative spectrum [10].

4.4.4 Glottal Spectral Separation

The Glottal Spectral Separation (GSS) vocoder proposed by Cabral et al. in 2008 [11], and later is more detail by Cabral in 2010 [10], refines the basic idea of the GPF vocoder (see Section 4.4.3): The Liljencrants-Fant (LF) model of the glottal-flow derivative waveform is integrated into the HTS STRAIGHT system (see Section 4.2.2) to model the voiced excitation instead of a binary pulse. The refinements in the GSS vocoder include the replacing of the glottal post-filtering procedure with a *spectral separation* of the glottal-flow derivative spectrum from the STRAIGHT spectral envelope, as well as the statistical modeling of the LF-parameters.

Similarly to the GlottHMM vocoder (see Section 4.4.1), the GSS model for the source-filter separation is not the traditional model where all of the spectral envelope features are modeled by a filter $\hat{H}(\omega)$, and the spectrally flat residual component is modeled by $E(\omega)$. Instead, the GSS model separates the spectrum of the glottal source derivative $D(\omega)$ from the overall spectral envelope $\hat{H}(\omega)$ to obtain the a spectral representation of $V(\omega) = \frac{\hat{H}(\omega)}{D(\omega)}$. With these models, the speech signal $S(\omega)$ can be represented as:

$$S(\omega) = E(\omega)D(\omega)V(\omega) = E(\omega)D(\omega)\frac{\hat{H}(\omega)}{D(\omega)} = E(\omega)\hat{H}(\omega) \tag{4.43}$$

The LF-model used in the GSS vocoder is a slightly simplified version of the LF-model used in the GPF vocoder, where the zero part of the model is discarded:

$$e_{LF}(t) = \begin{cases} E_0 e^{\alpha t}\sin(\omega_g t), & 0 \leq t \leq t_e \\ -\frac{E_e}{\epsilon t_\alpha}[e^{-\epsilon(t-t_e)} - e^{-\epsilon(t_c-t_e)}], & t_e < t \leq T_0 \end{cases} \tag{4.44}$$

For more details of the LF-model, see Section 4.4.3.

Analysis

The analysis phase of the GSS vocoder consists of the following parts: The estimation of the LF-model parameters for each two-pitch-period long frame centered at the glottal closure instant, the construction of the LF-model waveform $e_{LF}(t)$ according to the LF-model parameters, the estimation of the STRAIGHT spectrum $\hat{H}(\omega)$, and the spectral separation of the LF-waveform's spectrum $D_{LF}(\omega)$ from $\hat{H}(\omega)$ to obtain the spectral estimate $V(\omega)$ used in the synthesis phase.

The estimation of the LF-model parameters is done as described in Section 4.4.3, *Analysis*, and the estimate of the glottal-flow derivative waveform $e_{LF}(t)$ is constructed according to Equation 4.44. The FFT algorithm is used to calculate the spectrum $D_{LF}(\omega)$ of the zero-padded $e_{LF}(t)$.

The estimation of the overall spectral envelope $\hat{H}(\omega)$ is done using the STRAIGHT vocoder, which is described in Section 4.2.2, *Analysis*. The GSS spectral envelope $V(\omega)$ is obtained by dividing $\hat{H}(\omega)$ with $D_{LF}(\omega)$, which is converted into MFC coefficients.

The LF-parameters, with the exception of E_e and t_a, are normalized by the pitch period as in the GPF vocoder. After the normalization, the logarithm of their

inverse value is calculated to be used for the statistical modeling of the LF-model parameters.

Table 4.10: The analysis vector of the GSS vocoder, where p denotes the order number of the vocal tract spectral analysis.

Excitation parameters	$1 \times F0$
	$5 \times$ aperiodicity measures
	$6 \times$ *log* inverted LF-model parameters
Spectral parameters	$p \times$ STRAIGHT/GSS MFCC

The analysis vector of the GSS vocoder can be seen in Table 4.10. The vector consists of the normal STRAIGHT parameters; $F0$, aperiodicity measures, and the MFCCs (with the exception that for voiced frames, the MFCCs model the GSS spectral envelope $V(\omega)$ instead of the STRAIGHT spectral envelope $\hat{H}(\omega)$). In addition of the STRAIGHT parameters, the logarithms of the inverted LF-model parameters are included in the analysis vector. [10]

Synthesis

The block diagram of the synthesis phase of the GSS vocoder can be seen in Figure 4.17. As with the GPF vocoder (see Section 4.4.3, *Synthesis*), the synthesis procedure is embedded around the framework of STRAIGHT mixed excitation synthesis.

For each frame, voiced excitation is generated by generating a LF-model waveform (Equation 4.44) with a length of two pitch-periods (centered at the peak value) according to the $F0$ and LF-model parameters obtained from the analysis vector. The obtained waveform $e_{LF}(t)$ is Fourier transformed into the spectral representation $D(\omega)$, and it is weighted according to the aperiodicity measures of the analysis vector. The noise part of the voiced excitation is generated by multiplying white Gaussian noise in the frequency domain with the spectral envelope $E_p(\omega)$ of $e_{LF}(t)$ to compensate for the missing spectral features of the glottal pulse in the vocal tract filter $V(\omega)$. Next, the unvoiced excitation is scaled in energy and weighted according to the aperiodicity measures of the analysis vector. For unvoiced frames, only white Gaussian noise is used for the excitation.

The mixed excitation signal is formed in the frequency domain by adding the voiced and unvoiced excitation signals together, and then the inverse Fourier transform (IFFT) is applied to it. Each frame is filtered with the MLSA filter $V(\omega)$ obtained from the MFCCs, and the pitch-synchronous overlap-add algorithm [56] is applied to the frames to get the synthesized speech signal.

The synthesis quality of the GSS vocoder has been shown to outperform the simple impulse response vocoder [11], but like the similar GPF vocoder, it was clearly outperformed by the HTS STRAIGHT system in voice quality [10]. Also like the GPF vocoder, the main interest in the use of the GSS vocoder is its flexibility in controlling the glottal source parameters for use in voice transformation [10].

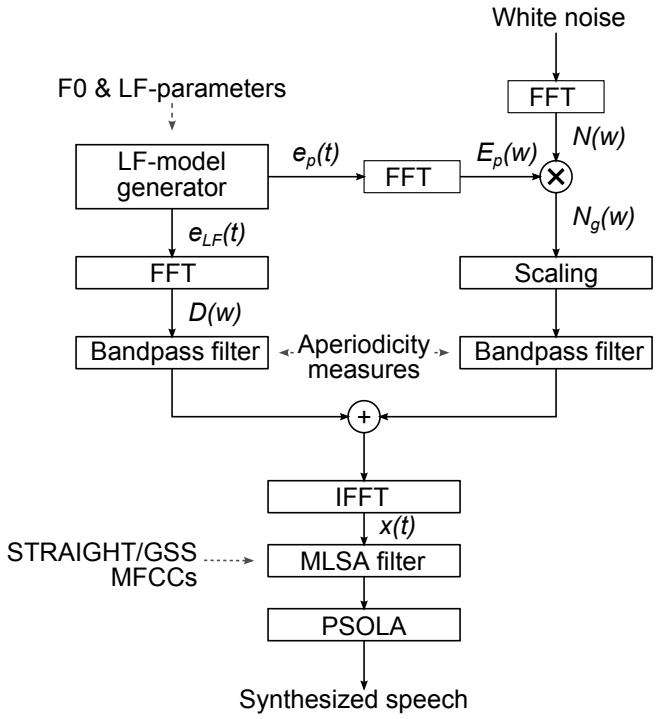


Figure 4.17: The synthesis block diagram of the GSS vocoder.

4.5 Vocoders Based on Sinusoidal Modeling

Sinusoidal modeling methods are defined by their use of the information of individual sinusoidal components in their analysis and synthesis phases. In the context of statistical parametric speech synthesis, sinusoidal modeling is used in the Multiband Excitation vocoder of Abdel-Hamind [1], and in the Harmonic/Stochastic Model of Banos [8]. The original Harmonic plus Noise Model (HNM) of Stylianou [69], [70] is also a sinusoidal modeling method, but the statistical parametric adaptations of the method have circumvented the sinusoidal modeling *per se* (see Sections 4.2.3 and 4.3.3), and thus in the context of this thesis are not considered as sinusoidal modeling methods.

4.5.1 Multiband Excitation

The statistical parametric synthesis adapted version of the Multiband Excitation (MBE) vocoder [28] was proposed by Abdel-Hamid et al. in 2006 [1]. The MBE vocoder is very similar to the Mixed Excitation vocoder (see Section 4.2.1) in that it estimates the degrees of voicing in multiple sub-bands of frequency, and in the synthesis phase mixed excitation is determined according to the voicing ratio of each band.

In the HMM-adapted version, the authors argue that the use of MFCCs for the representation of the spectral envelope may degrade the quality of the synthetic speech, because MFCCs are known to discriminate between speaker-specific characteristics in favor of phoneme-specific characteristics [1]. To overcome this problem,

the authors propose a sinusoidal modeling approach, where the spectral envelope is sampled at fixed positions. These samples can be interpolated to form a continuous spectral envelope, so they are used as the spectral modeling parameters in the HMMs.

Analysis

The analysis phase of the MBE vocoder for statistical parametric speech synthesis consists of three parts: $F0$ extraction, spectral envelope amplitude estimation, and sub-band voicing estimation. The $F0$ estimation must be done accurately, because the spectral envelope estimation is dependent on it.

The spectral envelope estimation is done by calculating the root mean squared (rms) values of harmonic component centered frequency bins from the STFT spectrum of each frame. The estimation of the rms value instead of the harmonic amplitude peaks makes the proposed method usable also for unvoiced frames. For unvoiced frames, an arbitrary non-zero value is selected as the $F0$ value for this purpose. In mathematical form the estimation of the harmonic bands is expressed as:

$$a_i = \sqrt{\sum_{k=(i-0.5)\cdot h+0.5}^{(i+0.5)\cdot h+0.5} s_k^2}, \tag{4.45}$$

where a_i is the amplitude of the ith harmonic band, s_k is the kth STFT sample, and h is the number of STFT samples per pitch period, given by:

$$h = 2F0N/s_r, \tag{4.46}$$

where N is the window length in samples and s_r is the sample rate.

After the estimation of the harmonic band amplitudes a_i, the amplitudes are interpolated to form an estimate of the whole spectral envelope. Next, the spectral envelope estimate is sampled at fixed positions to acquire the spectral amplitude values used for the analysis vector.

The voicing estimation for the selected m frequency bands is done by computing the ratio of energy around the hypothesized harmonic peaks, and the energy in the valleys around them. To compute this, the total amount of energy in each band is summed as b, and the energy for the frequencies around the harmonics in the band with a distance less than $F0/4$ is summed as v. The voicing value is the ratio v/b, which is expected to be around 0.5 for unvoiced bands, and near one for voiced bands.

The analysis vector of the HMM-adapted MBE vocoder can be seen in Table 4.11. It consists of the $\log F0$ value, voicing band strengths, and the spectral envelope samples. The values used for the amount of voicing bands and spectral envelope samples in [1] were 17 and 80, respectively.

Synthesis

The synthesis phase of the HMM-adapted Multiband Excitation vocoder is presented in Figure 4.18. First, voiced and unvoiced signals are fully generated according to the

Table 4.11: The analysis vector of the MBE vocoder, where p denotes the order number of the vocal tract spectral analysis and m denotes the number of voicing bands.

Excitation parameters	$1 \times F0$
	$m \times$ voicing band strengths
Spectral parameters	$p \times$ spectral envelope samples

analysis vector values, after which they are mixed according to the voicing strengths.

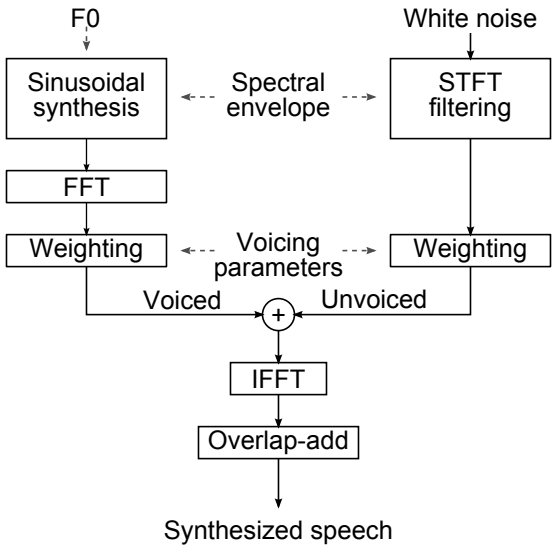


Figure 4.18: The synthesis block diagram of the MBE vocoder.

The voiced signal is based on a sinusoidal harmonic model: For each frame, it is modeled as a summation of sinusoids:

$$s_i(t) = \sum_{h=1}^{N_i} a_{i,h}(t) sin(\theta_{i,h}(t)), \tag{4.47}$$

where i is the frame index, N_i is the number of harmonic components in frame i, t is the time from the beginning of the frame, $a_{i,h}(t)$ is the amplitude track of the hth harmonic, and $\theta_{i,h}(t)$ is the phase track of the hth harmonic.

The amplitudes $a_{i,h}(0)$ at the beginning of each frame are taken from the spectral envelope, which is generated by interpolating the spectral envelope samples from the analysis vector. The amplitude track $a_{i,h}(t)$ is generated as a linear interpolation of the amplitude at the beginning and the end of the frame as:

$$a_{i,h}(t) = a_{i,h}(0) + \frac{t(a_{i+1,h}(0) - a_{i,h}(0))}{T}, \tag{4.48}$$

where T is the total frame length.

The phase tracks $\theta_{i,h}(t)$ are computed based on the $F0$ track and the phase $\theta_{i,h}(0)$ at the beginning of the frame as follows:

$$\theta_{i,h}(t) = \theta_{i,h}(0) + \int_0^t h\delta_i(\tau)d\tau, \tag{4.49}$$

where $\delta_i(t)$ is the fundamental frequency track estimated as a linear interpolation between the fundamental frequencies at the beginning and end of the frame:

$$\delta_i(t) = F0_i + \frac{t(F0_{i+1} - F0_i)}{T}, \tag{4.50}$$

where $F0_i$ is the fundamental frequency at the beginning of frame i. When the starting phase $\theta_{0,h}(0)$ is assumed to be zero, and the phase at the beginning of the frame is considered as the phase at the end of the previous frame ($\theta_{i,h}(0) = \theta_{i-1,h}(T)$), Equation 4.49 can be written in the following form:

$$\theta_{i,h}(t) = h\theta_{i-1,0}(T) + 2\pi h\left(F0_i t + \frac{t^2(F0_{i+1} - F0_i)}{2T}\right) \tag{4.51}$$

The unvoiced signal is generated by weighting white Gaussian noise in the frequency domain with the spectral envelope interpolated from the spectral envelope amplitudes.

The final part of the synthesis procedure is the mixing of the voiced and unvoiced signals according to the voicing ratios. This is done in the frequency domain, where each sub-band is weighted with its respective voicing ratio.

The quality of the synthetic voice signal of the HMM-adapted MBE vocoder was evaluated in a subjective Mean Opinion Score (MOS) test. Three vocoders were compared in the test: Traditional impulse response vocoder, the MBE vocoder with a MFCC representation of the spectral envelope, and the proposed MBE vocoder. The MOS scores were around 2.75, 3, and 3.6, respectively. The results of the MOS test are difficult to compare to other vocoders, but the results indicate that the use of the proposed spectral envelope representation clearly improves the quality compared to traditional MFCCs. [1]

4.5.2 Harmonic/Stochastic Model

The Harmonic/Stochastic Model (HSM) was first developed by Erro et al. [24], [25] as a pitch-asynchronous sinusoidal modeling method for concatenative speech synthesis. The adaptation of the HSM into the framework of statistical parametric speech synthesis was proposed by Banos et al. in 2008 [8]. The adaptation follows the basic principles of the original HSM method, but to make the parameters trainable in a HMM framework, an approximation has to be made in the process that degrades the quality of the statistical parametric implementation.

In HSM, the speech signal is modeled as the superposition of two components: a harmonic component and a stochastic (aperiodic) component. The harmonic component is constructed from a sum of sinusoids, and the stochastic component is the

result when the harmonic component is subtracted from the original signal, which is modeled by spectrally weighted white Gaussian noise:

$$s^{(k)}[n] = \sum_l A_l^2 \cos(2\pi l f_0^{(k)} \frac{n}{f_s} + \phi_l^{(k)}) + \sigma[n] \circledast h_{LPC}^{(k)}[n], \tag{4.52}$$

where k is the frame number, l is the harmonic number, A_l is the amplitude of harmonic l, ϕ_l is the phase of the lth harmonic, $f_0^{(k)}$ is the fundamental frequency in frame k, f_s is the sampling frequency, $\sigma[n]$ is white Gaussian noise, and $h_{LPC}[n]$ is the all-pole filter that contains the LPC spectral envelope of the stochastic part.

This structure, and the pitch-asynchronous design of the analysis phase, make the HSM a simple and effective tool for speech manipulation [25].

Analysis

The analysis phase of the HSM vocoder consists of the following parts: First, the harmonic component is determined by estimating the phases and amplitudes of the harmonic components in each frame, and then the harmonic waveform between adjacent frames is generated by interpolating the amplitude and phase trajectories. Second, the stochastic component is determined by subtracting the harmonic component from the corresponding original signal segment. Finally, the harmonic and stochastic components are parametrized to be used in the framework of HMMs.

The estimation of the amplitudes $A_l^{(k)}$ and phases $\phi_l^{(k)}$ of the harmonic components is done utilizing the algorithm proposed by Depalle et al. in [18]: An initial estimate of the amplitudes, phases, and their locations in each frame is obtained for example by peak picking the Fourier transform spectrum of the frame. The initial estimates of the phases and amplitudes, and the frequencies of the harmonics are refined in an iterative loop using least squares optimization.

After the frequencies, amplitudes and phases of each harmonic component in each frame are determined, the harmonic signal between adjacent frames is constructed by utilizing the interpolation algorithm proposed by McAulay et al. [54]: The harmonic peaks are matched frame-to-frame so that close peaks are connected, and peaks with no close counterpart in the other frame are phased off (or created). With the harmonic trajectories set, the amplitudes and instantaneous phases of the components are determined by interpolation. The amplitude trajectories are obtained by linear interpolation given by:

$$\tilde{A}_l(n) = A_l^{(k)} + \frac{(A_l^{(k+1)} - A_l^{(k)})}{S} n, \tag{4.53}$$

where $n = 0, 1, \ldots, S-1$ is the time sample into the kth frame, and S is the length of the harmonic component frame (given by the step-size of the analysis frame).

The phase trajectory can not be obtained by simple linear interpolation, because the phase values $\phi_l^{(k)}$ are obtained modulo 2π. The solution is to use cubic interpolation that unwraps the phase trajectory given by:

$$\tilde{\phi}(t) = \phi_l^{(k)} + \omega_l^{(k)} t + \alpha(M_l^*) t^2 + \beta(M_l^*) t^3, \tag{4.54}$$

where t is a time variable with $t = 0$ corresponding to frame k and $t = T$ corresponding to frame $k+1$, $\omega_l^{(k)} = 2\pi f_0^{(k)}$ is the fundamental angular velocity, $\alpha(\ )$ and $\beta(\ )$ are functions described in detail in [54], and M_l^* is the integer multiple of 2π that gives the optimally smooth trajectory given by:

$$M_l^* = \frac{1}{2\pi}\left[(\phi_l^{(k)} + \omega_l^{(k)}T - \phi_l^{(k+1)}) + (\omega_l^{(k+1)} - \omega_l^{(k)})\frac{T}{2}\right], \tag{4.55}$$

After the amplitude and phase trajectories of all of the components are determined, the harmonic signal $\tilde{s}_h(n)$ is constructed by

$$\tilde{s}_h(n) = \sum_l \tilde{A}_l(n)\cos(\tilde{\phi}_l(n)) \tag{4.56}$$

The stochastic signal $\tilde{s}_s(n)$ is obtained by subtracting the harmonic signal from the corresponding segment of the original signal $s(n)$:

$$\tilde{s}_s(n) = s(n) - \tilde{s}_h(n) \tag{4.57}$$

The estimated harmonic and stochastic signals are finally converted into a parametric representation. LPC representation was selected as the parametric representation for both parts in [8], discarding the phase information of the harmonic sinusoids which is difficult to model with HMMs. The LSF coefficients of $\tilde{s}_h(n)$ and $\tilde{s}_s(n)$ are calculated using the autocorrelation method, and along with the LPC gain coefficients and the $F0$ value, they form the analysis vector of the HSM vocoder (Table 4.12).

Table 4.12: The analysis vector of the HSM vocoder, where p denotes the order number of the vocal tract spectral analysis.

Excitation parameters	$1 \times F0$
Spectral parameters	$p \times$ harmonic envelope LSF
	$p \times$ stochastic envelope LSF

Synthesis

The synthesis phase of the HSM vocoder for statistical parametric speech synthesis (illustrated in Figure 4.19) is relatively straightforward compared to the analysis phase. Each 2N-length frame of the signal is constructed according to Equation 4.52, and the signals are pitch-asynchronously overlap-added according to:

$$s(kN + m) = (\frac{N - m}{N})s^{(k)}(m) + (\frac{m}{N})s^{(k+1)}(m - N), \tag{4.58}$$

where N is the hop size of the analysis frames, $m = 0, 1, \ldots, N-1$, and k is the frame number.

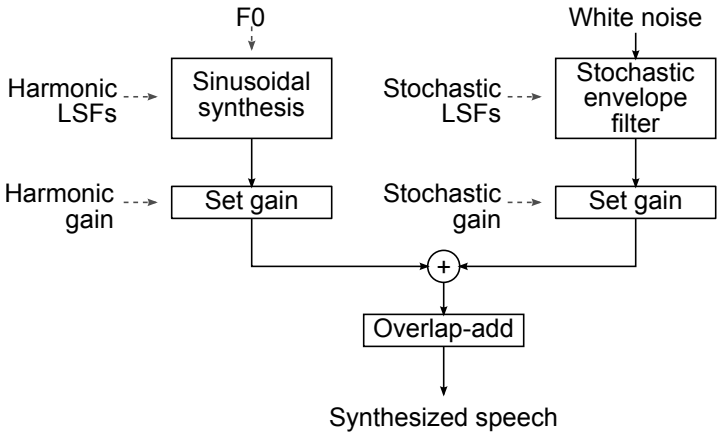


Figure 4.19: The synthesis block diagram of the HSM vocoder.

The amplitudes of the harmonic components of each frame are determined by sampling the amplitude spectrum of the harmonic LPC envelope $H(f)$ at the multiples of the fundamental frequency.

$$A_l^{(k)} = |H^{(k)}(lf_0^{(k)})| \tag{4.59}$$

The phase of each harmonic is also determined from the minimum phase response LPC envelope, but a linear phase term α has to be added to them in order to keep them coherent with those of the previous frame. Parameter α is obtained by assuming that the fundamental frequency varies linearly from frame $k-1$ to frame k.

$$\phi_l^{(k)} = l\alpha^{(k)} + \arg\{H^{(k)}(lf_0^{(k)})\}, \tag{4.60}$$

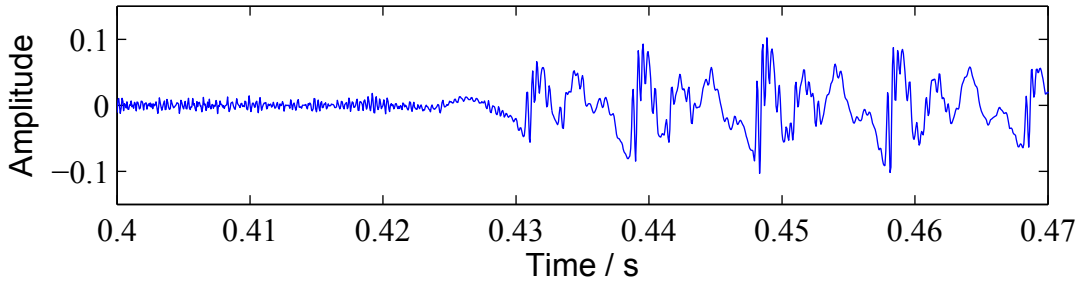
$$\alpha^{(k)} = \alpha^{(k-1)} + \pi\frac{N}{f_s}(f_0^{(k-1)} + f_0^{(k)}) \tag{4.61}$$

After the amplitudes and phases for each frame are determined from Equations 4.59 and 4.60, they are used in Equation 4.52 to construct the harmonic part of the frame.

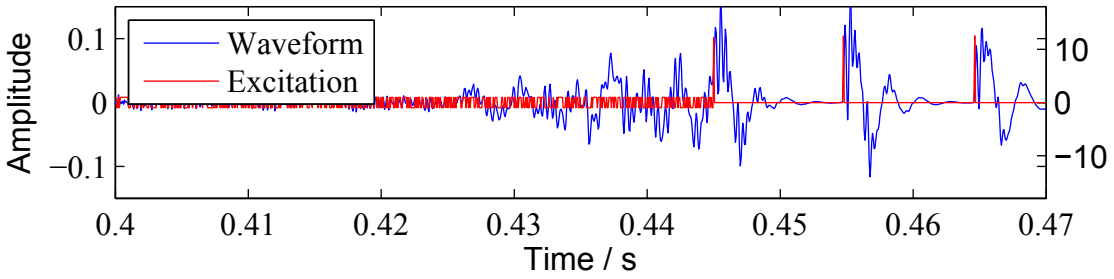
The stochastic part of each frame is constructed by filtering white Gaussian noise $\sigma(n)$ with the LPC synthesis filter $h_{LPC}(n)$ obtained from the LSF coefficients of the stochastic part. The stochastic part of each frame is added to the harmonic part as described in Equation 4.58. For unvoiced frames, only the stochastic part is modeled, and the stochastic spectral envelope is obtained by LPC analyzing the frame.

After the synthesis frames have been generated, they are overlap-added according to Equation 4.58 to obtain the synthetic signal. As mentioned before, the quality of the synthetic signal is degraded compared to the original HSM algorithm, because the phase information of the harmonic components is discarded in the analysis phase. However, Banos et al. argue in [8] that in their experience representing the harmonic component with an all-pole filter leads to synthetic speech with reasonable quality. The quality of the synthetic signal has been tested in [8] against the baseline HTS STRAIGHT system (see Section 4.2.2) in a perceptual pair comparison test. The

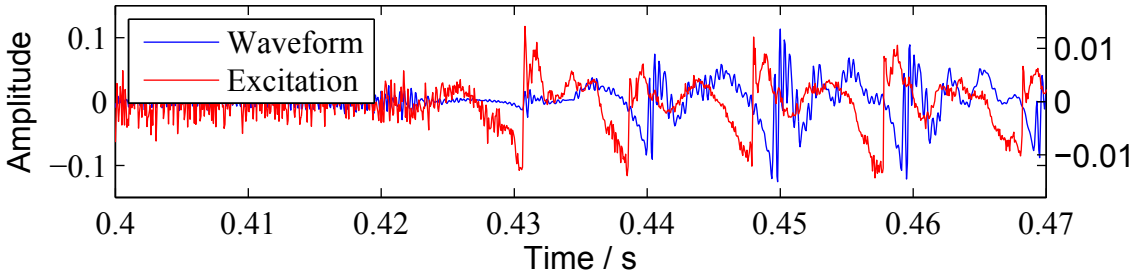
HSM vocoder was found to slightly outperform the STRAIGHT vocoder in similarity and quality tests. The main down-side of the HSM vocoder is that it needs double the amount of parameters for the spectral representation, since the harmonic and stochastic parts are modeled separately. To achieve high-quality spectral representations, the LPC orders used are usually in the range of 20 to 30 per spectrum. This amount of parameters used in the modeling makes the computational footprint of the HMMs large compared to other methods with similar spectral resolution.



(a) Part of the waveform of the original utterance.



(b) Waveform and excitation of the impulse excitation vocoded signal.



(c) Waveform and excitation of the GlottHMM vocoded signal.

Figure 4.20: The speech signal and excitation signal waveforms.

4.6 Representative Example

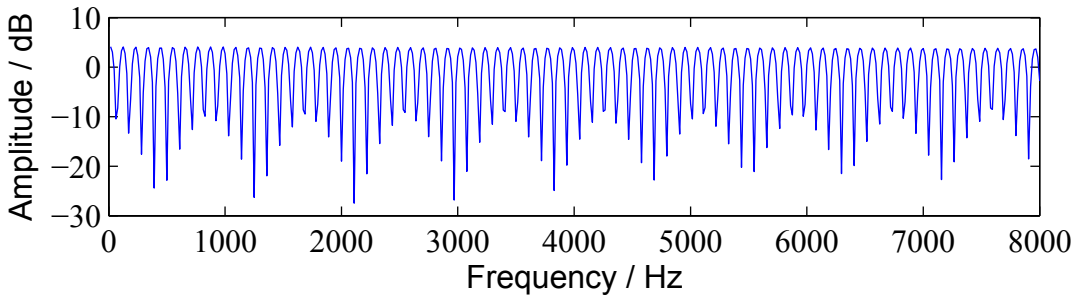
To illuminate the operations of vocoders, representative examples of the impulse excitation vocoder (Section 4.1) and the GlottHMM vocoder (Section 4.4.1) are presented in detail. An utterance is vocoded using both methods, and the waveform

of the resultant excitation signal and synthetic signal is studied, as well as the spectral characteristics of a selected frame.

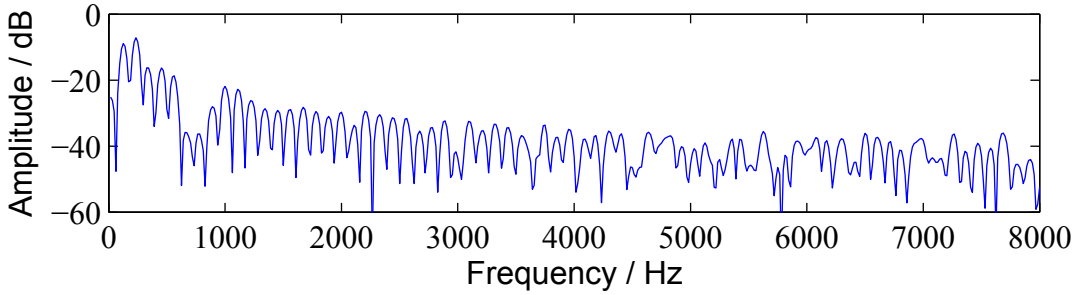
The original signal is presented in Figure 4.20(a), along with its impulse excitation and GlottHMM vocoded signals in Figures 4.20(b) and (c), respectively. The impulse excitation vocoding was done utilizing MFCCs of order 24, and the GlottHMM vocoding was done utilizing 30 LSF coefficients for the vocal tract, 10 LSF coefficients for the glottal source, and 5 coefficients for the HNR.

The original utterance is a transitional waveform from [s] to [i] in the Finnish word "yksi". The impulse vocoded version shows clearly that it cannot sufficiently model the transitional frames from voiced to unvoiced sounds, whereas the more sophisticated GlottHMM vocoder produces a similar waveform to the original signal. Also, the impulse vocoded signal has a minimum-phase time structure, which makes the energy of the voiced parts gravitate towards the excitation pulses.

The spectra of a voiced excitation signal frame of the impulse excitation vocoder and the GlottHMM vocoder are presented in Figure 4.21. The spectrum of the impulse excitation signal shows a perfect harmonic structure with a flat envelope, whereas the GlottHMM vocoded excitation has a relatively harmonic spectrum on low frequencies, but above 3 kHz frequencies the excitation begins to become aperiodic. Also, the GlottHMM excitation's spectral envelope is a declining slope towards higher frequencies.



(a) Impulse excitation spectrum.

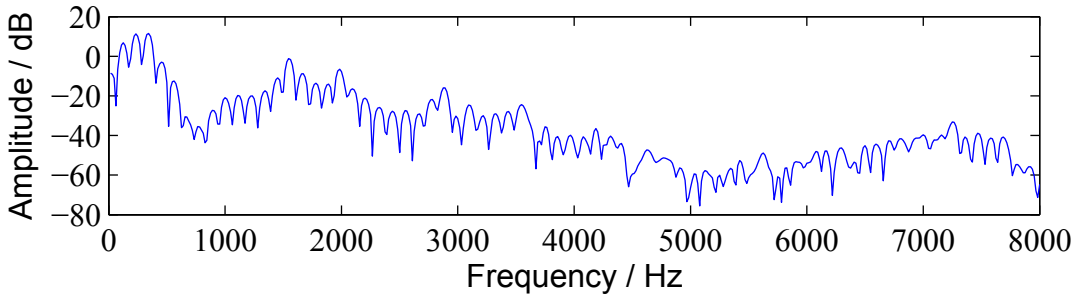


(b) GlottHMM excitation spectrum.

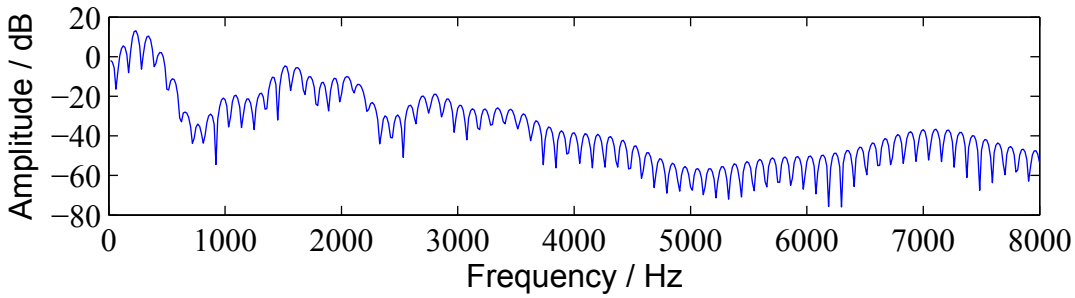
Figure 4.21: The excitation signal spectra.

The spectra of the original and synthesized signals are presented in Figure 4.22. Most notable difference in the spectra is the perfect harmonic structure of the impulse excitation vocoded signal. The harmonic structure of the GlottHMM vocoded

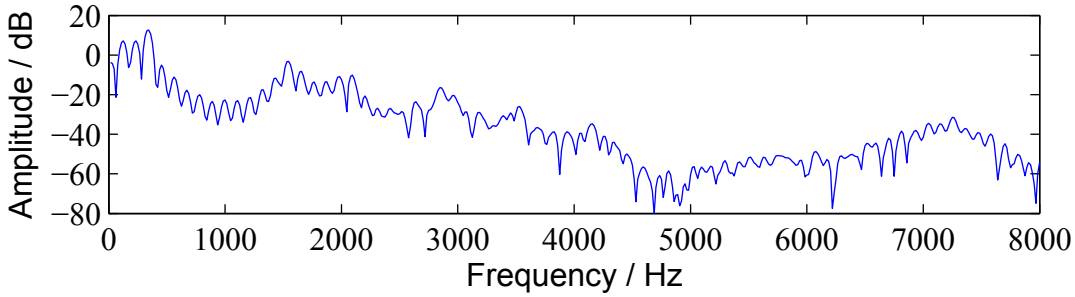
signal is seemingly very close to the harmonic structure of the original signal. The spectral envelope is however modeled more accurately by the MFCCs of the impulse excitation vocoder than the LSFs of the GlottHMM vocoder. This can be best seen in the spectral valley around 800 Hz, which is a lot deeper in reality than in the LPC based envelope.



(a) Spectrum of the original frame.



(b) Spectrum of the impulse excitation vocoded frame.



(c) Spectrum of the GlottHMM vocoded frame.

Figure 4.22: The original and vocoded speech signal spectra.

5 Statistical Analysis of Vocoder Parameters

The vocoders discussed in Section 4 have been sufficiently documented in the literature for their operation, but no detailed studies have been published on the statistical properties of the vocoder parameters. In HMM-based speech synthesis, the vocoder parameter distributions are modeled commonly by a multivariate Gaussian distribution that is fitted into the distribution of the data [76]. Because the HMM parameter generation algorithm is mainly interested in the mean value of the distribution to maximize the likelihood of the generated parameter track, the Gaussian distribution assumption is sufficient in most cases.

However, if the parameter distributions would have multiple peaks or high skewness, the Gaussian model would be highly inaccurate in terms of what the mean value represents: With multiple peaks, the mean value could be at a valley of the distribution, where the likelihood that a parameter with such value would be generated would be small. Still, the Gaussian model would generate parameters around this value. With high skewness in the original distribution, the mean value would be placed along the tail of the distribution, far from the main peak which could be arguably more optimal place for the mean of the Gaussian model. The knowledge of the vocoder parameter distributions is thus important in justifying the viability of the selected statistical modeling method.

Another interesting aspect of the vocoder parameters is that if statistical deviation would be found in the parameters based on the *style* or *emotion* of the speech, the information could be useful in the synthesis of emotional speech: Parameters with higher emotional dependence could hypothetically have better controllability and thus better quality in the synthesis of emotional speech. Also the knowledge of statistical variance in the vocoder parameters based on the speech properties could be used in applications such as emotion detection. For example, MFCCs are already widely used in speaker recognition [48].

In this thesis, the statistical properties of three vocoders were studied: GlottHMM (Section 4.4.1), STRAIGHT (Section 4.2.2) and Harmonic/Stochastic Model (HSM) (Section 4.5.2). The selection criteria for these vocoders were that they:

1. Represented different vocoding approaches (Glottal Source Modeling 4.4, Multi-Band Mixed Excitation 4.2, and Sinusoidal Modeling 4.5),

2. Have shown good performance in previous studies, and

3. Were available for the study.

The GlottHMM (version 1.0.5) and STRAIGHT (version 40_003) vocoders were obtained readily, and the HSM vocoder was implemented according to articles [8], [25], [18], and [54]. The CLT vocoder of Maia [52] was also obtained and considered to be representing the Residual modeling vocoders (Section 4.3), but as discussed in Section 4.3.1, the CLT vocoder needs HMM-labeled data to function, which was not possible for the analysis/synthesis experiments that were conducted.

5.1 Test Setup

The goal of the statistical analysis test was to observe the statistical distributions of vocoder parameters analyzed from a database of different forms of emotional speech, and draw conclusions regarding the distributions' Gaussianity and variability. Additionally, the hypothesis that the speakers' emotions affect the values of the vocoder parameters was tested with a statistical analysis to justify recommendations for possible future research topics.

The Speech Database

The speech database used for the study was the Finnish Emotion Study Recording Material database recorded at the University of Oulu in 2003 [2]. The database consists of four female and five male professional actors reading the same approximately one minute long text passage in five different emotions: "neutral" (neutraali), "sad" (surullinen), "joyful" (iloinen), "angry" (vihainen) and "affective" (hellä). The speech within each subset of the database was very broad in terms of the emotion: For example angry speech can be both "hot" and "cold", where the the extreme cases of hot anger are straight up yelling, and cold anger is a more subtle passive aggressive style.

For the study of the parameter distributions, only the three emotions "neutral", "angry", and "sad" were selected to reduce the scope of the produced data. The emotions were selected so that different excitation modes of speech would be represented: angry speech tends to be a "hot" emotion, whereas sad speech can be considered more as a "cold" emotion, with neutral speech being in the middle. All five emotions were used in the statistical testing of emotional effects in the vocoder parameters.

Vocoder Setup

The vocoders were set up by using their default number of parameters in all coefficients to obtain characteristic parameter distributions. An exception for this was the HSM vocoder, where the number of LSF parameters was tuned higher than the reported article length [8] to obtain more meaningful comparisons with the LSF parameters of the GlottHMM vocoder. The increase of the number of parameters did not affect negatively to the synthesis quality. The parameter lengths for each vocoder are reported in Table 5.1.

The fundamental frequency estimation for the HSM vocoder was implemented using the YIN algorithm [17], and the fundamental frequency estimation for the STRAIGHT vocoder was done using the SWIPE algorithm [13] implemented in the Speech Processing Toolkit (SPTK [68]), because the obtained STRAIGHT version did not have integrated $F0$ estimation.

The analysis was done utilizing a 5 ms frame shift for the GlottHMM and STRAIGHT vocoders, and a 10 ms frame shift for the HSM vocoder. Longer frame shift was used for the HSM vocoder because it produced better synthesis quality.

Frame length was 30 ms for the HSM vocoder, and the STRAIGHT and GlottHMM vocoders used pitch-adaptive frame lengths.

Table 5.1: The vector lengths of the selected vocoders for the statistical analysis.

Vocoder	GlottHMM	STRAIGHT	HSM
Parameters	$1 \times F0$	$1 \times F0$	$1 \times F0$
	$1 \times$ Energy	$5 \times$ Aperiodicity	$1 \times$ Stochastic Energy
	$5 \times$ HNR	$25 \times$ MFCC	$1 \times$ Harmonic Energy
	$20 \times$ Source LSF		$20 \times$ Stochastic LSF
	$30 \times$ Vocal Tract LSF		$30 \times$ Harmonic LSF
Sum	57	31	53

Depending on the vocoder parameter, the analysis was done either on both voiced and unvoiced frame coefficients, or only for voiced frame coefficients. For example, the HNR and source LSF coefficients of the GlottHMM vocoder, the aperiodicity coefficients of the STRAIGHT vocoder and the harmonic LSF coefficients of the HSM vocoder are justified to be used only in voiced frames. The vocoders compute their values for all frames, so a large bulk of the data used in the analysis would be meaningless if the unvoiced frames' values would be included. Additionally, the analysis of $F0$ parameters was omitted, because in theory their distributions are identical. Any differences found in distributions would be caused mainly by the different pitch estimation algorithms that were used, whose in depth study was outside the scope of this thesis work.

5.2 Analysis Methods

The distributions of the vocoder parameters were studied by computing their main statistics. These statistics were the mean, variance, skewness, kurtosis and negentropy of each distribution. The distributions were obtained from each subgroup based on gender and emotion. Whether the emotion type has an effect on the vocoder parameters obtained was tested with the Friedman's test. Moreover, if the effect was found to be significant, a simple post-hoc test comparing the the 95% confidence intervals of parameter means was applied to obtain information about the nature of the effect.

5.2.1 Statistical Measures

The statistical measures selected were in effect the first moment, second central moment, and third and fourth normalized moments of the parameter distribution, in addition to *negentropy* [36], which is a robust measure of Gaussianity. The four moments give good basic characteristics of the distributions, and when combined

with the negentropy value, it is possible to evaluate the accuracy of the Gaussian estimation of the distributions.

Mean

The *mean* μ_1 of a parameter distribution is ideally the *expected value*, or *expectation* of a random variable pulled from the distribution. Expectation in turn is the weighted average of all possible values that the random variable can have. For a continuous distribution with a probability density function $f(x)$, the expectation is defined as:

$$E[X] = \int_{-\infty}^{\infty} x f(x) dx \tag{5.1}$$

However, with a finite number of data points drawn from the distribution, the accurate estimation of the probability density function is difficult, and the expectation is approximated as the arithmetical mean of the samples:

$$E[X] = \mu_1 \approx \bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i, \tag{5.2}$$

where N is the total number of samples.

The mean of a distribution is the point which divides 50% of the probability mass on the probability distribution function. Other mean-like statistical measures are the *median*, which tells separates the upper half of the sample distribution from the lower half, and the *mode*, which tells the most probable value in the parameter distribution. For a Gaussian distribution, the mean, median and mode have the same value.

Variance

Variance is the second *central moment* of a statistical distribution, meaning that it is computed relative to the mean value of the distribution:

$$\mathrm{Var}(X) = \sigma^2 = E[(X - E[X])^2] = \bar{X^2} - \bar{X}^2 = \bar{X^2} - \mu_1^2, \tag{5.3}$$

where σ is the standard deviation of the distribution, defined as the square root of the variance.

Variance gives the measure of how far the distribution's values have spread out from the mean value: A small variance means that the distribution is centered tightly around the mean value, whereas a large variance value means that the distribution is scattered loosely around the mean value. Even though they represent essentially the same information, the square root of variance, the standard deviation σ, is generally a more intuitive type of representation, because it is reported in the same units as the measurement data. The 95% confidence intervals for a given measurement can be obtained by adding 2σ to represent the upper and lower bounds of the interval.

For a Gaussian distribution, the variance and mean values are the only parameters defining the distribution completely. However, multivariate distributions require

information also about the *covariance* of different parameters. Covariance can be thought to represent how much two random variables change together: If variable X is big when variable Y is big and vice versa, the covariance between the two variables is a positive number. If variable X is small when variable Y is big and vice versa, the covariance is a negative number. If no such patterns are present, the covariance is zero, which is a requirement for two parameters being statistically independent. Covariance between two parameters is computed as:

$$\mathrm{Cov}(X,Y) = E[(X - E[X])(Y - E[Y])] = \bar{XY} - \bar{X}\bar{Y} \tag{5.4}$$

When the covariance between all of the parameters is calculated, the information can be represented as a *covariance matrix*, where the diagonal consists of the variance values of each parameter, and the off-diagonal cells are composed of the respective covariance values.

In the context of HMM modeling, the vocoder parameters are assumed to be statistically independent, meaning that the covariance between the parameters is assumed to be zero, and the covariance matrix is diagonal.

Skewness

The skewness of a statistical distribution is its *third standardized moment*, defined as

$$\mathrm{Skew}(X) = E\left[\left(\frac{X-\mu_1}{\sigma}\right)^3\right] \approx \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^3}{(\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2)^{3/2}}, \tag{5.5}$$

where μ_1 is the mean value and σ is the standard deviation of the distribution. The value is normalized in relation to the standard deviation so that the skewness values between different distributions can be directly compared.

Skewness tells the measure of asymmetry of the distribution relative to the mean value. A distribution has positive skewness, if the left-side tail of its probability distribution function is shorter than the right-side tail. In other words this means that the bulk of the observed values lie to the left of the mean value. A distribution has negative skewness in the opposite case. A perfectly symmetrical distribution such as the Gaussian distribution has zero skewness.

For a highly skewed distribution, the modeling using a Gaussian distribution might be problematic if the mean value of the distribution is used as the mean value of the Gaussian distribution. That is because the bulk of the mass of the probability density function lie far away from the mean value. The mode or median of such a distribution might be more useful in representing the mean value of the Gaussian model.

Kurtosis

The kurtosis of a statistical distribution is its *fourth standardized moment* subtracted by three to obtain zero kurtosis for the Gaussian distribution:

$$\mathrm{Kurt}(X) = E\left[\left(\frac{X-\mu_1}{\sigma}\right)^4\right] - 3 \approx \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^4}{(\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2)^2} - 3 \tag{5.6}$$

Kurtosis can be considered as the measure of "peakedness" of the statistical distribution: Given the same variance, a distribution with high kurtosis is centered around the mean value more than a distribution with low kurtosis. Also, a distribution with high kurtosis has short tails, whereas a low kurtosis distribution has long tails. As the definition of kurtosis suggests, the Gaussian distribution has zero kurtosis, whereas for example the uniform distribution has a kurtosis of -1.2, and the Laplace distribution has a kurtosis of 3. Distributions with above-zero kurtosis are called *super-Gaussian*, and distributions with below-zero kurtosis are called *sub-Gaussian*. This property makes kurtosis also a satisfactory measure of Gaussianity in a distribution, and it is used in applications such as Independent Component Analysis (ICA [36]).

Unfortunately, kurtosis can not be considered as an accurate measure of peakedness for asymmetric (non-zero skewness) distributions [7], and as such it is not a robust measure for comparative uses [36]. The use of kurtosis in this thesis was selected because it nevertheless gives an approximation of the scale of the peakedness in distributions that at the moment are modeled as Gaussian distributions.

Negentropy

Negentropy is an information theoretical concept, which measures the Gaussianity of a distribution, or more precisely, the difference in entropy between the observed distribution and the Gaussian distribution. A fundamental result of information theory is that a Gaussian variable has the largest entropy among all random variables of equal variance [36].

The information theoretical concept of entropy for a random variable y with probability density function $f(y)$ is defined for a continuous distribution as

$$H(y) = -\int f(y) \log f(y) \mathrm{d}y, \tag{5.7}$$

and it represents the degree of information that the observation of the variable gives. The more unpredictable (random) the variable, the higher its entropy becomes. Using the definition of entropy, the negentropy J is defined as

$$J(y) = H(y_{Gauss}) - H(y), \tag{5.8}$$

where y_{Gauss} is a Gaussian random variable with the same variance as y. Due to the above-mentioned properties, the negentropy is always non-negative, and zero only if the variable y has a Gaussian distribution.

The problem with using the negentropy based on its definition is that the computation of the differential entropy requires knowledge of the probability density function of the distribution, which is a highly theoretical concept in practical applications. Hyvärinen has developed an approximation for negentropy based on the maximum entropy principle [35], acquiring a general form for the approximation as:

$$J(y) \approx \sum_{i=1}^{p} k_i (E[G_i(y)] - E[G_i(v)])^2, \tag{5.9}$$

where k_i are positive constants, v is a Gaussian variable with zero mean and unit variance, variable y has zero mean and unit variance, and functions G_i are some non-quadratic functions. Robust choices for the functions have been found to be:

$$G_1(u) = \frac{1}{a_1} \log \cosh a_1 u \tag{5.10}$$

$$G_2(u) = -\exp(-\frac{u^2}{2}) \tag{5.11}$$

where $1 \leq a_1 \leq 2$ is some suitable constant [36]. The numerical parameters for this thesis were chosen to be: $p = 2$, $k_1 = k_2 = 100$, and $a_1 = 1$. Negentropy values calculated from known distributions using these parameter values are presented in Table 5.2.

Table 5.2: Negentropy values for known distributions.

Distribution	Negentropy
Gaussian (generated)	0.0003
Uniform	0.26
Gamma (k=2, θ=1)	0.14
Gamma (k=1, θ=1)	0.48
Exponential (μ =1)	0.48
Laplace (from speech data)	2.31

Covariance Diagonality

As described in the Variance section, a multivariate Gaussian distribution is defined by its mean vector and covariance matrix where the diagonal consists of the variance of each parameter, and the other cells consist of the covariance between the respective parameters. In mathematical form, the covariance matrix Σ is expressed as:

$$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \sigma_{13}^2 & \cdots & \sigma_{1N}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \sigma_{23}^2 & \cdots & \sigma_{2N}^2 \\ \sigma_{31}^2 & \sigma_{32}^2 & \sigma_{33}^2 & \cdots & \sigma_{3N}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sigma_{N1}^2 & \sigma_{N2}^2 & \sigma_{N3}^2 & \cdots & \sigma_{NN}^2 \end{bmatrix} \tag{5.12}$$

where σ_{ii}^2 is the variance of the ith parameter, and σ_{ij}^2 is the covariance between the ith and the jth parameter.

For statistically independently distributed parameters, the covariance matrix is diagonal, meaning that the covariance between all parameters is zero. Statistical

independence between parameters is a common assumption in mathematical modeling, and it is also done in HMM modeling when each parameter is modeled only by a independent Gaussian distribution. Thus it would be interesting to know how valid this assumption is for the studied vocoder parameters, or in other words, it would be interesting to measure the degree of *diagonality* of the covariance matrices of the parameters. A simple measure for the diagonality is to look at the mass ratio between the variances at the diagonal to the whole mass of the matrix, and thus the following simple measure was developed:

$$\text{Diag}(\Sigma) = \frac{\sum_{i=1}^{N} |\sigma_{ii}^2|}{\sum_{i=1}^{N} \sum_{j=1}^{N} |\sigma_{ij}^2|}, \tag{5.13}$$

where σ_{ij}^2 are the cell values of the $N \times N$ covariance matrix Σ. The absolute value is taken from the cell values to acquire the total covariance mass of the cells: without the absolute value the negative and positive covariances could cancel out each other, which is not desirable for the measurement; it is desirable to obtain the total covariance in the terms of accumulated "mass", be it negative or positive. By using Equation 5.13, the diagonality for a diagonal covariance matrix is 1, and the diagonality for a constant valued $N \times N$ matrix would be $1/N$, meaning that 100 % and 100/N % of the value mass of the covariance matrix is located in the diagonal, respectively.

5.2.2 Statistical Testing

The statistical testing of the effects of speaker emotion on the vocoder parameter distribution was done by utilizing the Friedman test to determine whether the different emotion groups had an effect on the overall values of the vocoder parameters. More detailed analysis was done by post-hoc testing based on the results of the Friedman's test. The Friedman's test and the post-hoc test are presented in the following sections.

Friedman Test

The Friedman test is a non-parametric statistical test used to find differences in comparative parameters across multiple cases. The non-parametricality means that the test does not make any assumptions about the statistical distributions of the studied parameters, which is more preferred in this case than the similar parametric ANOVA (Analysis of Variance) test. The ANOVA test assumes that the parameter distributions have a Gaussian distribution, which might not be the case for all of the tested parameters.

The test is carried by putting the tested parameter vectors into a matrix where each column contains the obtained parameters of each frame for a certain emotion. Next, each row is *ranked* from lowest to highest (i.e. 1 for lowest, 5 for highest for five columns) value, and the average rank M_g for each column is computed, in addition to the average rank M_{all} of all columns (which is always equal to $(k+1)/2$, where k is the number of columns). The emotion database used contained five different

emotions which were all included, so $k = 5$ for the experiment. Next, the sum of squared deviates SS_{bg} is computed as

$$SS_{bg} = \sum_{g=1}^{k} \left[n_g (M_g - M_{all})^2 \right], \tag{5.14}$$

where n_g is the number of samples in the gth column. The sum of squared deviates can be converted into the test statistic χ^2 by

$$\chi^2 = \frac{SS_{bg}}{k(k+1)/12}, \tag{5.15}$$

which is the point at the χ^2-distribution with $k - 1$ degrees of freedom from which the confidence interval can be calculated.

The Friedman test indicates whether the parameter distributions change as a function of at least one of the emotions. It does not tell which emotions have an effect and how much, so some kind of post-hoc testing is required after the initial test to find out the details.

Post-hoc Testing

The post-hoc test is used to gain insight about the nature of the effect(s) that the initial statistical test found to be significant. In the case of the performed test, this means that the post-hoc analysis is used to find the patterns of how different emotions affect different vocoder parameters, given that the Friedman test has concluded statistical significance in the emotion dependencies.

The post-hoc testing was done by calculating and comparing the 95% confidence intervals of the mean values of each subset distribution. If the confidence intervals did not overlap between two emotions, the difference was considered statistically significant. This method is considered statistically significant in more established post-hoc tests, such as the Fisher's Least Significant Difference (LSD) test.

Each emotion was studied versus other emotions for statistically significant differences for all vocoder parameters, and the number of the emotions that discriminated significantly against all other emotions was counted for each parameter. With the number of unambiguously discriminating emotions for each vocoder parameter obtained, an approximation of each *parameter group's* (for example spectral envelope MFCCs or the HNR coefficients) emotional sensitivity was computed by calculating the mean value of each group's numbers. An assumption is that this measure will give insight on the overall sensitivity for different emotions for the selected vocoder parameter type. However, because this approach is highly simplifying, and because the material in the speech database did not consist of uniform emotional output, the results should be considered mostly as trendsetting for possible future research.

5.3 Analysis Results

The results of the conducted analysis are presented in two parts. In the first part, the data for the vocoder parameter distributions are presented, and in the second

part, the results for the effect of speaker emotion in parameter distributions testing are presented.

5.3.1 Parameter Distributions

The parameter distributions for each vocoder were obtained for both genders for the three emotions "neutral", "sad" and "angry". The statistical parameters described in Section 5.2.1 were computed for each distribution, and they are presented as whole in Appendix A. However, it is very tedious to draw any meaningful information about the overall properties of the distributions from such a large array of data.

Instead, the data was simplified in the following ways to acquire the form presented in Tables 5.3, 5.4 and 5.5: First of all, every coefficient of the same type was clumped up into the same category: For example, the HNR coefficients 1 to 5 were thought of just as representations of the HNR coefficient group. Second, the data for each emotion and gender was put together, and this simplified pool was searched for the minimum, median and maximum values of all of the statistical parameters. This representation has the following properties:

- Because the statistical measures from which the values are searched for are computed from a varying pool of emotions and genders, the parameter distributions will likely have variance which will bring out the characteristic range of the vocoder parameters. The minimum and maximum values reflect this.

- Even though each parameter type is treated as one parameter group instead of multiple individual parameters with their own distributions (which they are in reality), the obtained values give meaningful information about the characteristics of the parameters: When compared to the minimum and maximum values, the median value reflects how the average number of values is scattered in the parameters.

- The values in the table are real values from some parameter in the pool: they are not averaged, and so do not have any bias that might be caused by averaging a relatively small number of values, where some values might have a large noise component. This is the main reason why the median value was chosen over the mean value to represent the "average" values.

- If some parameters are deviating vastly from the average values, it is reflected in the obtained minimum/maximum values even though only a few parameters out of many show this behavior. This is desirable in the sense that it tells about the robustness of different parameters to highly varying forms of excitation. Outlier values might also be caused by an implementation bug in the vocoder (for example, an unvoiced frame is classified as voiced, and this distorts the computation of some parameter), and not solely a fundamental property of the parameters, which is also useful information.

Table 5.3: The compressed statistical measures of the GlottHMM parameter distributions taken from the "Neutral", "Sad", and "Angry" emotion databases for both genders.

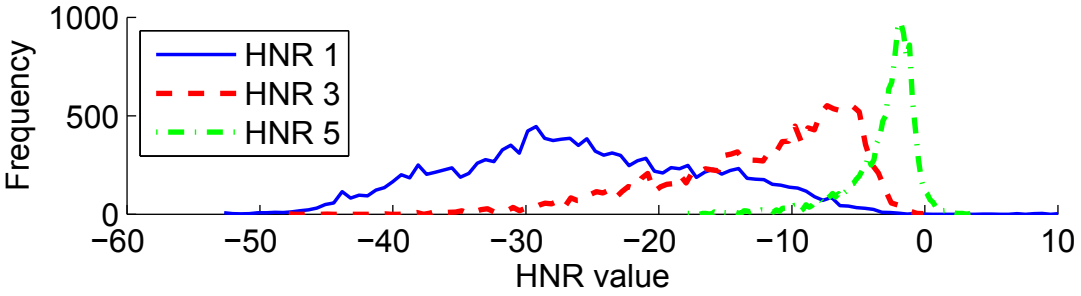
		LSF VT (V)	LSF VT (UV)	LSF Source	HNR
Mean	min	0.14	0.06	0.07	-25.97
	med	1.55	1.54	1.45	-12.24
	max	3.04	3.03	2.98	-1.99
Deviation	min	0.02	0.02	0.01	1.41
	med	0.05	0.05	0.02	7.09
	max	0.08	0.07	0.04	11.51
Skewness	min	-1.50	-1.46	-1.39	-2.69
	med	0.06	-0.24	0.02	-0.92
	max	0.66	1.47	2.46	0.20
Kurtosis	min	-0.98	-0.62	0.46	-1.03
	med	0.01	1.73	2.50	0.60
	max	4.55	5.44	14.74	13.97
Negentropy	min	0.00	0.00	0.00	0.00
	med	0.01	0.22	0.18	0.11
	max	0.82	0.94	2.21	1.32
Cov diagonality	min	0.51	0.38	0.69	0.65
	med	0.54	0.42	0.83	0.69
	max	0.56	0.47	0.87	0.70

GlottHMM Parameters

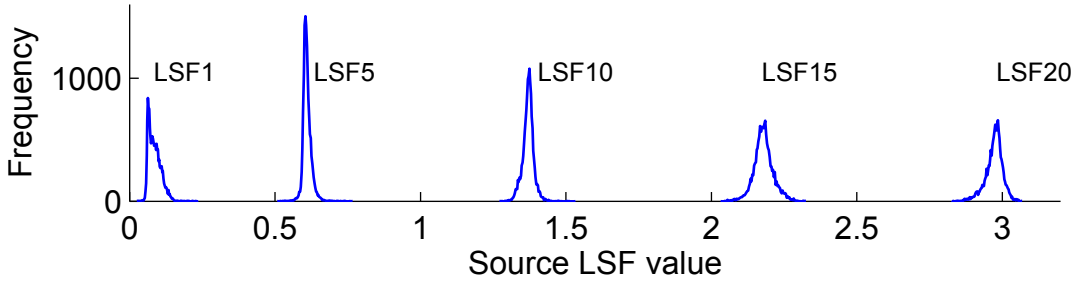
The compressed statistical measure table of the GlottHMM vocoder is presented in Table 5.3, and representative histograms from the "angry male" database are presented in Figure 5.1. The Vocal Tract (VT) LSF parameters are divided into the "Voiced" (V) and "Unvoiced" (UV) categories, because the estimation method used by the vocoder differs for these cases. Also, the unvoiced category contains a substantial amount of values that are computed from silent frames where no speech is present, so the voiced category can be thought of as a more accurate representation of parameter distributions for phonemes.

The statistical measures for all LSF parameters reflect a very similar overall behavior: The LSF parameters' mean values rise very linearly from slightly over 0 to slightly under π, which is expected, given that the LSF parameters represent a frequency from 0 to π in ascending order (see Section 2.3.2). A notable difference in the mean values is present in the voiced LSF VT coefficients: the minimum value is substantially higher than for the other LSF coefficients, which is because the spectral tilt of the glottal excitation is removed from the spectrum of the voiced LSF VT coefficients, making the LPC analysis allocate more emphasis on higher frequencies.

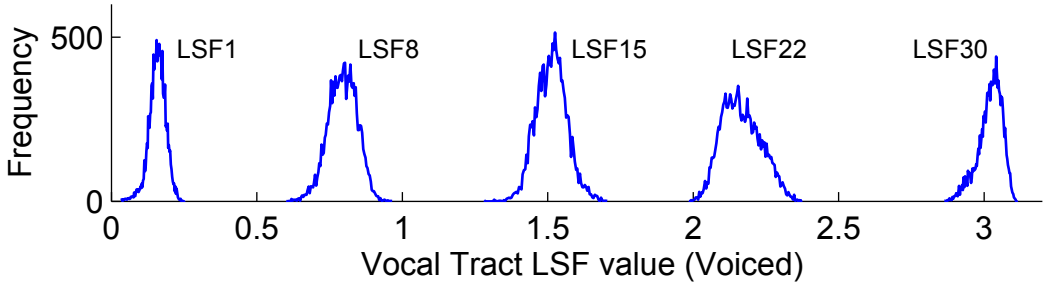
The standard deviations for the LSF distributions are lower for the source parameters than for the vocal tract parameters, which is explained by the more similar



(a) Histograms for the selected HNR parameter distributions.



(b) Histograms for the selected Source LSF parameter distributions.



(c) Histograms for the selected Vocal Tract LSF parameter distributions.

Figure 5.1: Histograms of the GlottHMM parameter distributions for the "Angry male" database.

overall shape of the glottal pulse spectrum compared to the overall shape of the vocal tract envelope.

Another similarity between the LSF parameters is the skewness of the distributions: The skewness changes gradually as a function of the LSF coefficient number from significantly negative values to significantly positive values, with nearly non-skewed values in the middle. This observation is in line with previous studies [66], [67]. The kurtosis values indicate that with the exception of the voiced LSF VT parameters, most of the LSF distributions are significantly super-Gaussian (median greater than 1 and large maximum values). The maximum kurtosis value for the source parameters is significantly larger than the median value, which might be indicative of erroneous parameter estimation caused by the unrobustness of the IAIF algorithm in some cases (for example in the case of a voiced/unvoiced decision error).

Similar to the kurtosis values, the negentropy values reflect that the median val-

ues for the voiced vocal tract LSF parameters are fairly Gaussian, but the other LSF parameters are more in the non-Gaussian territory (with little lower negentropy than the uniform distribution). The maximum negentropy values for all LSF parameters are clearly in the non-Gaussian territory.

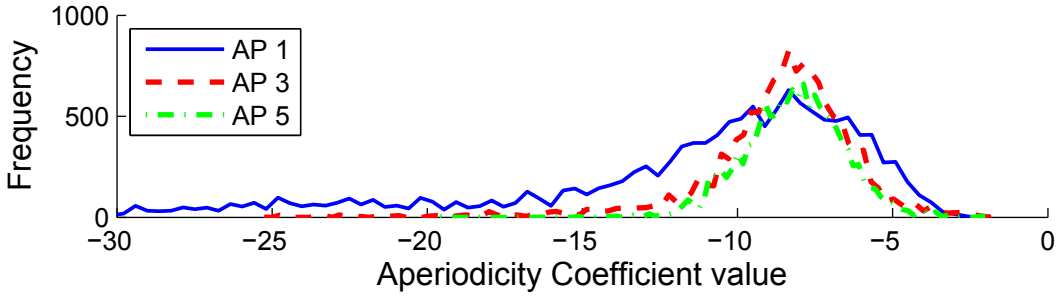
The distributions for the HNR coefficients have greatly varying mean and deviation values. They are mostly negatively skewed (median = -0.92), and the kurtosis values lie quite evenly on both sides of zero. The exception is the maximum kurtosis value, which is most probably caused by an error similar to the LSF source coefficients. Further investigation of the data in Appendix A supports this hypothesis (the high value is only present in one of the distributions). The median negentropy value of the HNR coefficients can be considered fairly Gaussian.

The GlottHMM parameters have significant amounts of covariance between each other, as illustrated by the covariance diagonality measures. Especially the vocal tract LSF coefficients have a low diagonal-to-all ratio. This might be caused by the property of the LSF coefficients, where the higher order coefficients get higher values than the preceding coefficients, thus creating covariance.

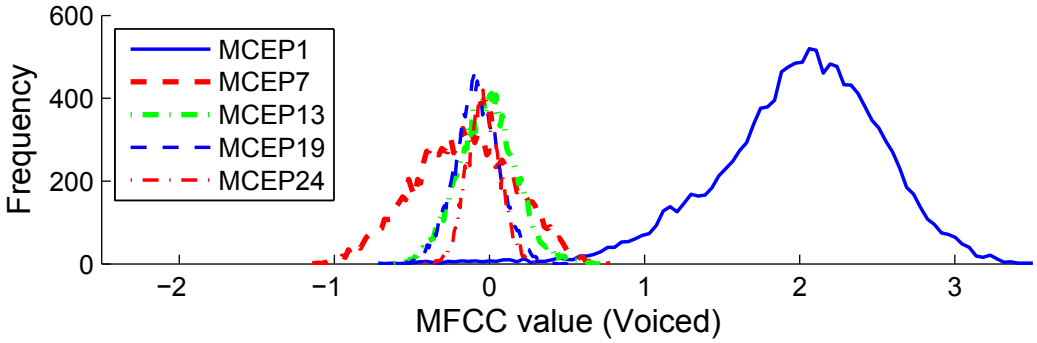
STRAIGHT Parameters

Table 5.4: The compressed statistical measures of the STRAIGHT parameter distributions taken from the "Neutral", "Sad", and "Angry" emotion databases for both genders.

		MFCC (V)	MFCC (UV)	AP
Mean	min	-0.34	-0.07	-13.15
	med	-0.04	-0.01	-8.71
	max	2.29	1.21	-8.13
Deviation	min	0.09	0.08	1.49
	med	0.18	0.14	2.67
	max	0.67	0.85	7.79
Skewness	min	-1.27	-1.00	-2.46
	med	-0.03	-0.10	-1.30
	max	0.45	1.28	-0.02
Kurtosis	min	-0.62	0.08	0.28
	med	0.04	0.63	2.27
	max	2.25	4.89	9.85
Negentropy	min	0.00	0.00	0.00
	med	0.00	0.02	0.20
	max	0.27	1.44	1.84
Cov diagonality	min	0.85	0.93	0.79
	med	0.87	0.95	0.83
	max	0.94	0.97	0.92



(a) Histograms for the selected AP parameter distributions.



(b) Histograms for the selected MFCC parameter distributions.

Figure 5.2: Histograms of the STRAIGHT parameter distributions for the "Angry male" database.

The compressed parameter distribution table for the STRAIGHT vocoder is presented in Table 5.4, and the representative examples for the STRAIGHT parameter distributions are presented as histograms in Figure 5.2. As expected, The distributions for the mel-cepstral coefficients are remarkably Gaussian: Even the maximum negentropy values are moderately small for the voiced MFCC distributions. The MFCC distributions' mean and variance deviate at lower coefficient values, but for higher coefficient values, the distributions are close to zero-mean Gaussian distributions with a standard deviation of about 0.18. The skewness of the MFCCs is distributed fairly evenly on both sides, with a near-zero median value. The diagonality of the MFCC covariance matrix is also very high, with nearly 90% of the mass of the matrix laying on the diagonal.

The aperiodicity coefficients have a similar form that is in common with each coefficient: Their mode values are placed around -7 and -8, and the distributions have a lengthy tail to the left side, which makes them negatively skewed. Figure 5.2 (a) reveals that apart from the tails, the distributions have a very Gaussian shape. The tails of the distributions increases also their kurtosis and negentropy values to a fairly non-Gaussian territory. These observations suggest that the aperiodicity coefficients would be better suited for the single Gaussian modeling, if the tails of the distributions would be truncated. The diagonality of the aperiodicity coefficients' covariance matrix is excellent, being nearly as high as for the MFCCs.

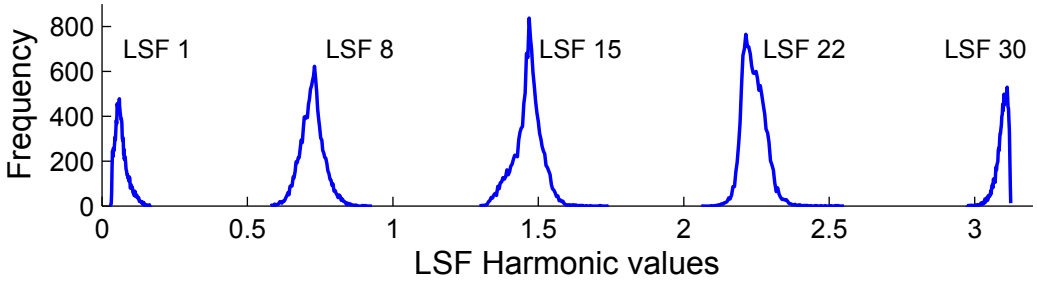
HSM Parameters

Table 5.5: The compressed statistical measures of the HSM parameter distributions taken from the "Neutral", "Sad", and "Angry" emotion databases for both genders.

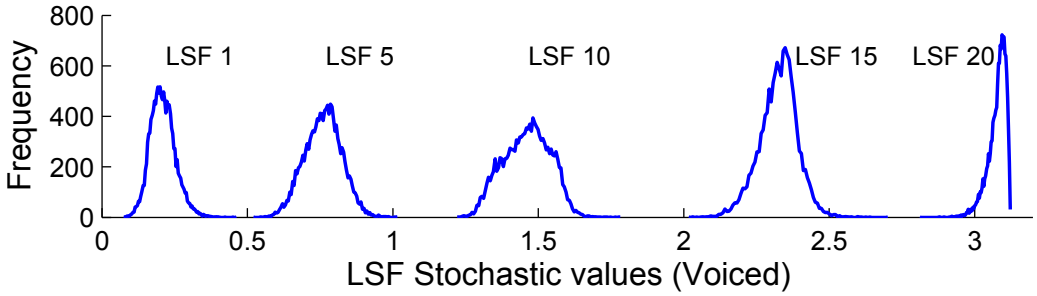
	LSF S (V)	LSF S (UV)	LSF H
min	0.21	0.10	0.07
Mean — med	1.55	1.60	1.53
max	3.08	3.09	3.10
min	0.03	0.02	0.02
Deviation — med	0.07	0.05	0.04
max	0.10	0.07	0.05
min	-1.36	-1.76	-1.38
Skewness — med	-0.20	-0.43	0.14
max	0.51	1.84	1.13
min	-0.77	0.61	-0.37
Kurtosis — med	0.02	2.58	0.50
max	2.71	6.05	3.15
min	0.00	0.01	0.00
Negentropy — med	0.01	0.32	0.02
max	0.16	0.97	0.24
min	0.55	0.46	0.42
Cov diagonality — med	0.59	0.51	0.45
max	0.62	0.59	0.50

The compressed statistical measure table of the HSM vocoder is presented in Table 5.5, and representative histograms from the "angry male" database are presented in Figure 5.3. The overall behavior of the LSF parameters is the same as for the GlottHMM LSF parameters: The parameters have increasing mean values in proportion to the coefficient number from 0 to π, and the skewness of the distributions shifts from mostly positively skewed for low coefficient numbers to mostly negatively skewed for high coefficient numbers.

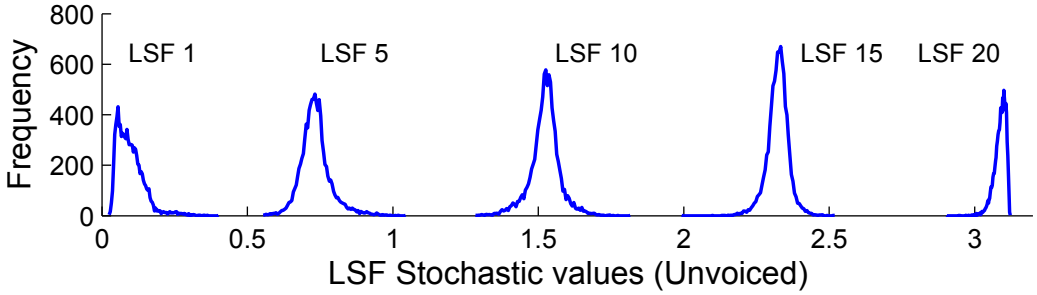
The distributions for the stochastic LSF coefficients differ greatly from voiced to unvoiced parameters: the voiced coefficients (computed from the residual signal after the harmonic part has been subtracted) have higher minimum mean values, which is caused by the removal of the spectral tilt (which is included in the harmonic part), which puts more emphasis on higher frequencies. The voiced coefficients have also higher deviation, as the median value of the voiced coefficients is the same as the maximum value of unvoiced coefficients. The skewness values for the voiced coefficients are closer to zero, and especially the maximum skewness value is significantly smaller than for the unvoiced coefficients. The kurtosis values for the voiced coefficients are especially different from the other studied LSF parameters: The median kurtosis value is very close to zero, and the maximum values are also comparatively small. The similar trend is also present in the negentropy values: the



(a) Histograms for the selected Harmonic LSF parameter distributions.



(b) Histograms for the selected voiced Stochastic LSF parameter distributions.



(c) Histograms for the selected unvoiced Stochastic LSF parameter distributions.

Figure 5.3: Histograms of the HSM vocoder parameter distributions for the "Angry male" database.

voiced stochastic LSFs have very low negentropy values, actually getting a lower maximum value than the STRAIGHT vocoder's MFCCs. The unvoiced coefficients have significantly higher kurtosis and negentropy values, and they can be considered as super-Gaussian distributions. The unvoiced stochastic LSF coefficients' high kurtosis is likely due to the fact that the silent unvoiced frames that make up a large portion of the database used produce LSF coefficients similar to each other, which makes the histogram peaks relatively high. Because of this, the unvoiced frames' coefficients are not completely suitable for the estimation of the unvoiced stochastic LSFs' behavior in the case of unvoiced phonemes. The diagonality of the covariance matrices is on the weaker side of all of the studied vocoder parameter types: they vary roughly from 45% to 60%.

The harmonic LSF coefficients' distributions are located in between the voiced and unvoiced stochastic LSF coefficients' distributions in terms of their properties:

The mean and deviation values for the harmonic LSFs are close to the unvoiced stochastic LSFs, the skewness values are in between, and the kurtosis and negentropy values are close to the voiced stochastic LSFs. This means that also the harmonic LSF coefficients have near-Gaussian distributions. However, the covariance diagonality is even weaker than for the stochastic LSF coefficients, being around 42% and 50%.

Comparative Review of the Statistics

The statistics of the vocoder parameter distributions can be compared in terms of their form and Gaussianity, but it is important to remember that the studied distributions are obtained from context-independent data. In actual HMM-modeling, the data is divided into contexts (e.g. monophones) whose distributions are modeled in the HMM framework. The context dependent distributions differ from the overall distribution that is in reality the sum of every context-dependent distribution. However, it is reasonable to assume that the context-dependent distributions do not substantially differ from the form of the overall distribution.

Compared to the LSF coefficients of the GlottHMM and HSM vocoders, the MFCC coefficients are from clearly to slightly more Gaussian, and are very clearly more statistically independent, which means that the current single Gaussian modeling of the parameters for HMM synthesis is the most accurate for MFCCs. However, this does not directly imply that MFCCs would necessarily be the best choice for the spectral envelope model, because LSFs have been documented to have better interpolation properties [59], and in some cases outperforming the MFCCs in synthesis quality [47]. Nevertheless, the great suitability of the MFCCs for statistical parametric modeling is not a thing to be overlooked.

The LSF coefficients of the HSM vocoder (median negentropy values of 0.01, 0.32, and 0.02) are more Gaussian than the LSF coefficients of the GlottHMM vocoder (median negentropy values of 0.01, 0.22, and 0.18), which is rather surprising.

Interestingly, the HNR coefficients of the GlottHMM vocoder were much more Gaussian than the Aperiodicity coefficients of the STRAIGHT vocoder (negentropy mean values of 0.11 and 0.20, respectively), even though they convey information about essentially the same properties (the amount of noise in each sub-band in voiced frames) in slightly different forms. The covariance matrix diagonality of the aperiodicity coefficients is however slightly higher than the HNR coefficients' diagnoality.

A remarkable notion about the parameter distributions is that no dual-peakedness was found in any parameter's distribution with the used restrictions (separate statistics for voiced and unvoiced frames where justifiable).

5.3.2 Effect of Speaker Emotion in Parameter Distributions

The study for the effect of speaker emotion in parameter distributions was done as described in Section 5.2.2. Friedman's test with the significance level of 5% (with $p \ll 0.001$) confirmed that the emotion has an significant effect on the parameter

value. This is expected, given the full spectrum of speech styles contained in the database.

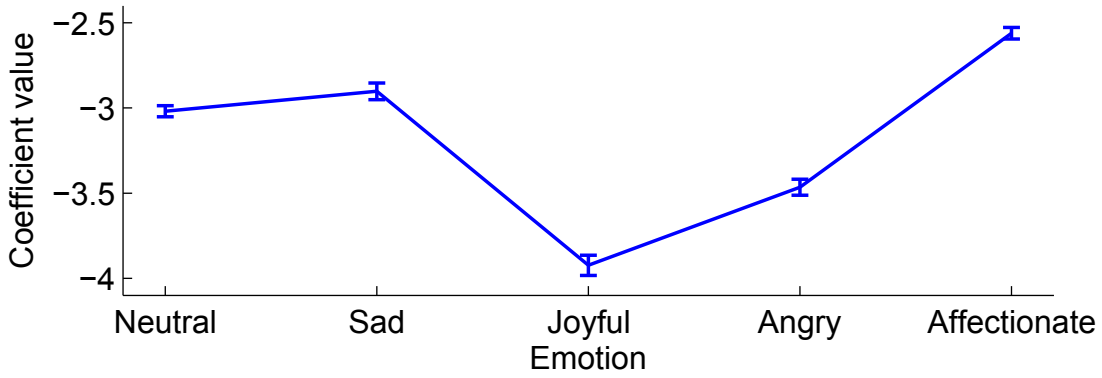


Figure 5.4: The mean values of the GlottHMM vocoder's HNR5 coefficient with their 95% confidence intervals as the function of different emotions.

With the validation of the Friedman test, the post-hoc testing was applied to all of the studied parameters of the selected vocoders. An example of the post-hoc analysis of the GlottHMM vocoder's HNR5 coefficient is presented in Figure 5.4. The post-hoc test confirmed that all of the parameter values of the different emotions differ significantly from each other. The number of unambiguously discriminated emotions for each vocoder parameter was obtained from the post-hoc graphs, and the average number of emotions for each parameter type of each vocoder is presented in Table 5.6:

Table 5.6: The average number of unambiguously discriminated emotions (out of five) for each vocoder parameter type.

	Parameter type	Avg	Avg (%)
GlottHMM —	LSF Source	3.35	67
	LSF Vocal Tract	3.83	77
	HNR	4	80
STRAIGHT —	MFCC	3.28	66
	AP	2.2	44
HSM —	LSF Stochastic	2.85	57
	LSF Harmonic	3.43	69

The results show that the vocoder parameters indeed have some descriptive value to the emotion of the speech, as most parameters could on average discriminate at least 3 emotions out of 5. For the tested vocoders, the GlottHMM parameter values seem to have the best discriminating power for the emotions both in terms of the maximum average value and the overall values. These results reinforce the findings of Lorenzo-Trueba et al. [50] that the GlottHMM parameters are powerful in expressive speech characterization.

It is notable that due to the non-uniform material (see Section 5.1, *The speech database*) of the used speech database, the results of this test must be interpreted as mostly illustrative. However, it is encouraging to find that the studied vocoder parameters could possibly be used to discriminate between very broad terms of emotions, which would have lots of real-world application potential.

6 Subjective Evaluation of Vocoder Quality

A subjective listening test was performed to find out the comparative analysis/synthesis *quality* of the three vocoders (GlottHMM, STRAIGHT and HSM). The testing of the analysis/synthesis quality does not give a direct rating on the quality of the vocoders in HMM-based speech synthesis, but rather it gives insight to the comparative vocoder qualities given (near) optimal parameter trajectories. This section describes the details and results of the conducted subjective listening test

6.1 Test Setup

The three vocoders used were the GlottHMM (version 1.0.5), STRAIGHT (version 40_003), and HSM (thesis author's own implementation), as described in Section 5.1. The lengths of the feature vectors were selected so that each method would have similar amounts of coefficients for spectral envelope information, and excitation information. Also, previously reported vector lengths by the methods' authors were taken into account. The selected vector lengths for the vocoders is presented in Table 6.1:

Table 6.1: The vector lengths of the selected vocoders for the subjective listening test.

Vocoder	GlottHMM	STRAIGHT	HSM
Parameters	1 × $F0$	1 × $F0$	1 × $F0$
	1 × Energy	5 × Aperiodicity	1 × Stochastic Energy
	5 × HNR	40 × MFCC	1 × Harmonic Energy
	10 × Source LSF		20 × Stochastic LSF
	30 × Vocal Tract LSF		20 × Harmonic LSF
Sum	57	46	43

The vocoder's control parameters were set mainly on default values, after some experimentation to ensure that they produce a quality that is well representative of the vocoder. A potential problem was identified with the vocoder's usage of different pitch detection algorithms (PDAs), which could produce uneven $F0$ estimation errors in the test samples, and thus affect the scores that the vocoders receive. As the goal of the subjective listening test was not to assess the quality of the PDAs or the quality of the vocoders enduring $F0$ errors, it was decided to select samples to the test that did not suffer from gross $F0$ estimation errors.

The selected speech samples for the listening test followed the principles of the speech material selected for the statistical analysis: Male and female samples using various emotions and quality were used to obtain a comprehensive take of different excitations for the vocoders. The selected emotions for the listening test were

"neutral", "sad" (representing "cold" emotions), and "joy" (representing "hot" emotions). Four samples were selected for each gender with each emotion, making a total of 24 speech samples ($4 \times 2 \times 3 = 24$). The "sad" and "joy" emotion samples were taken from the Finnish emotional speech database described in Section 5.1, and the "neutral" samples were taken from high-quality, HMM-training suitable databases. The samples were approximately 2 seconds long.

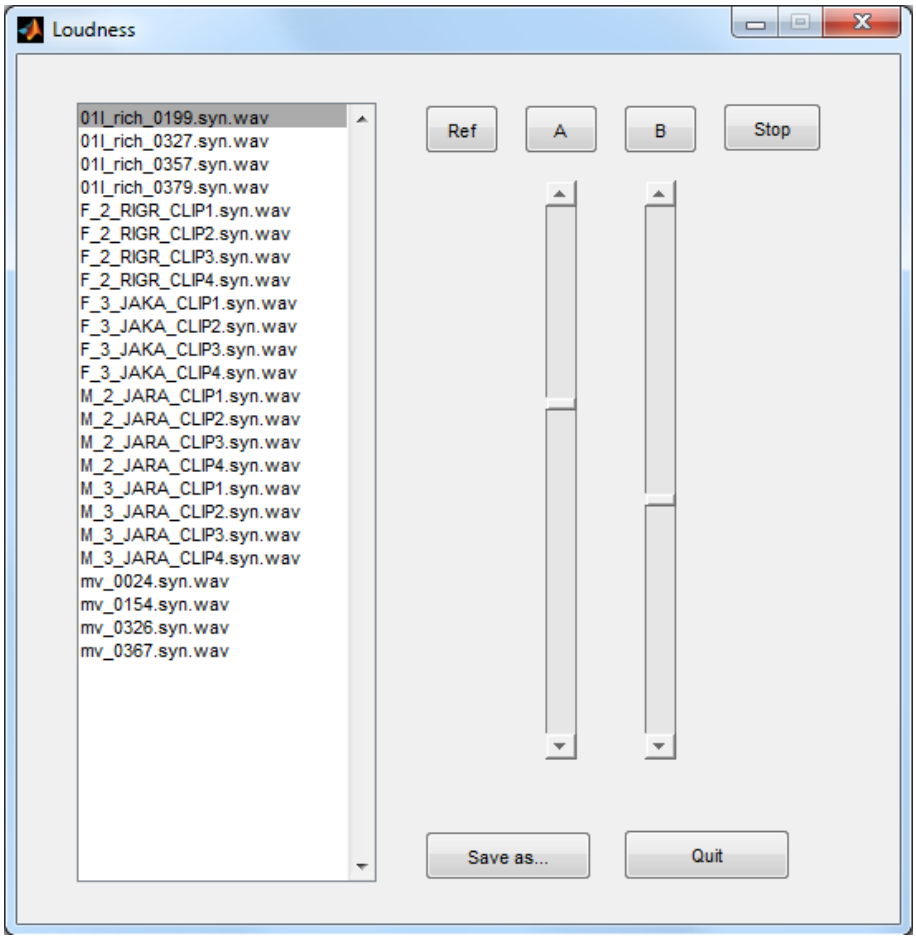


Figure 6.1: The user interface for the loudness normalization test.

The vocoded samples were found to be different in terms of *loudness* (the perceived volume), even though they were normalized in terms of signal energy. Thus a comparative listening test utilizing two expert listeners who rated the loudness of each sample was conducted to normalize the loudness differences. The test setup is presented in Figure 6.1: One vocoder's sample was held as the reference sample, and the volume of the other two vocoders was controllable using the sliders. The ratings for each vocoder were averaged, and the final samples were normalized according to the obtained average loudness of the vocoders.

The test was conducted using headphones in a soundproof listening booth designed for the conduction of subjective listening tests. The number of test subjects was 13 (9 male, 4 female), and they all were native Finnish speakers without docu-

mented hearing degradations, aged between 21 and 27.

6.2 CCR Test

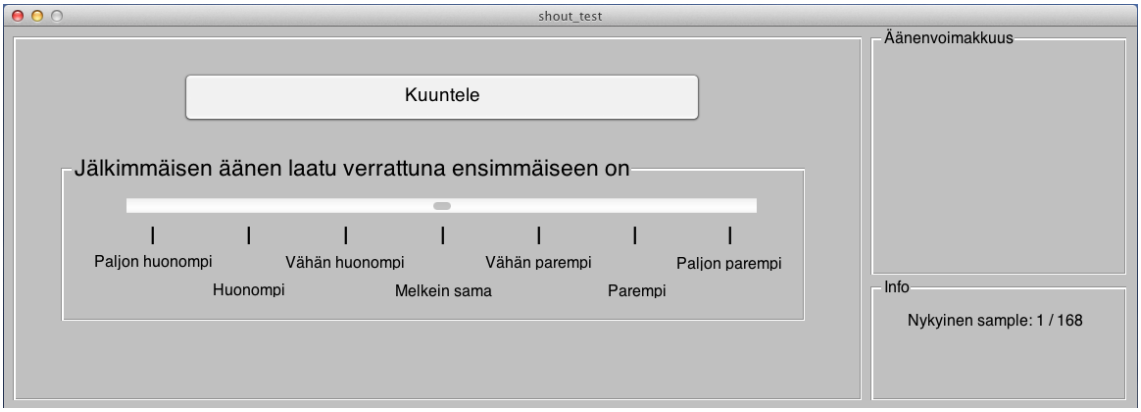


Figure 6.2: The user interface for the CCR test (in Finnish).

The test type used for the subjective listening test was the Comparison Category Rating (CCR) Test, which is an ITU-T P-800 standard [38]. On each trial of the CCR test, the listener is presented with a pair of speech samples using different vocodings. The listener is asked to evaluate the *quality* of the second sample compared to the quality of the first sample on a scale of -3 to 3, where the ratings correspond to:

3 : Much better

2 : Better

1 : Slightly better

0 : About the same

-1 : Slightly worse

-2 : Worse

-3 : Much worse

The user interface used in the conducted test is presented in Figure 6.2.

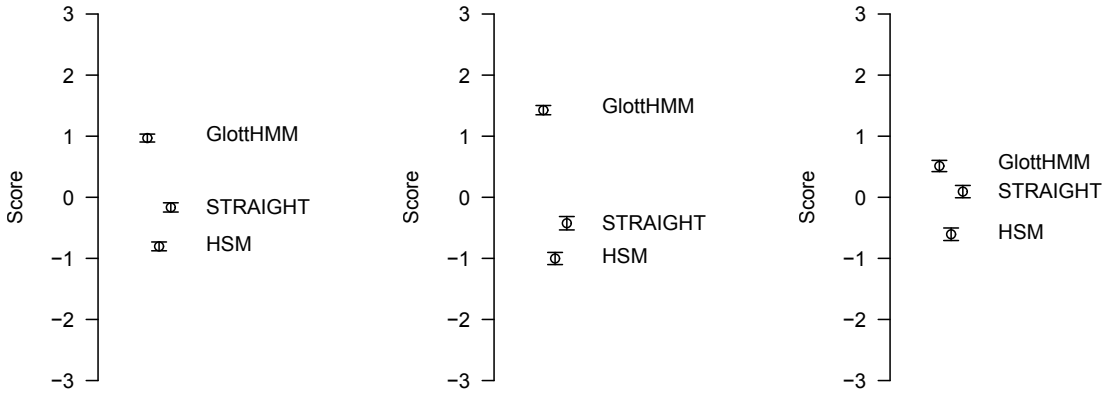
Sample pairs are formed by selecting every combination of different vocoders for each sample (including the null pairs where both samples are the same). Each sample pair (except for the null pairs) is included twice in the test so that they can be played in different orders. This eliminates any bias that the presentation order could introduce to the results, and it gives a method to evaluate the reliability of the listeners by comparing how consistently they rated the same sample pairs.

The total number of trials in the conducted CCR test thus became [number of combinations] $\times$ [number of repeats] $\times$ [number of samples] + [number of samples

(for null pairs)] $= 3 \times 2 \times 24 + 24 = 168$. The test was estimated to run for less than an hour, so a 5 minute break was added to the halfway point of the test to prevent listener exhaustion.

The data analysis of the CCR test is conducted by calculating the average of each vocoder's ratings (excluding the null pairs). For example, if a trial contains vocoder A first and vocoder B second, and the listener rates the pair as "-2", the rating for vocoder A is counted as "+2", and the rating for vocoder B is "-2". The resultant average rating for each vocoder is called its Comparison Category Rating (CCR). Even though the CCR scale is numerically in the same -3 to 3 scale as the listening test, the same interpretation of the numbers is no longer valid. Instead, the resulting ratings can be thought of as a *distance metric* that tells the relative differences in each vocoder's quality. This is also the reason why the original speech samples were not taken into the test: presumably the original sound quality would be greatly better than the vocoded samples, leading to a great gap between the original and vocoded ratings, but the differences between the vocoders would be smaller and more difficult to distinguish.

6.3 Listening Test Results



(a) CCR test results for all voice samples.

(b) CCR test results for male voice samples.

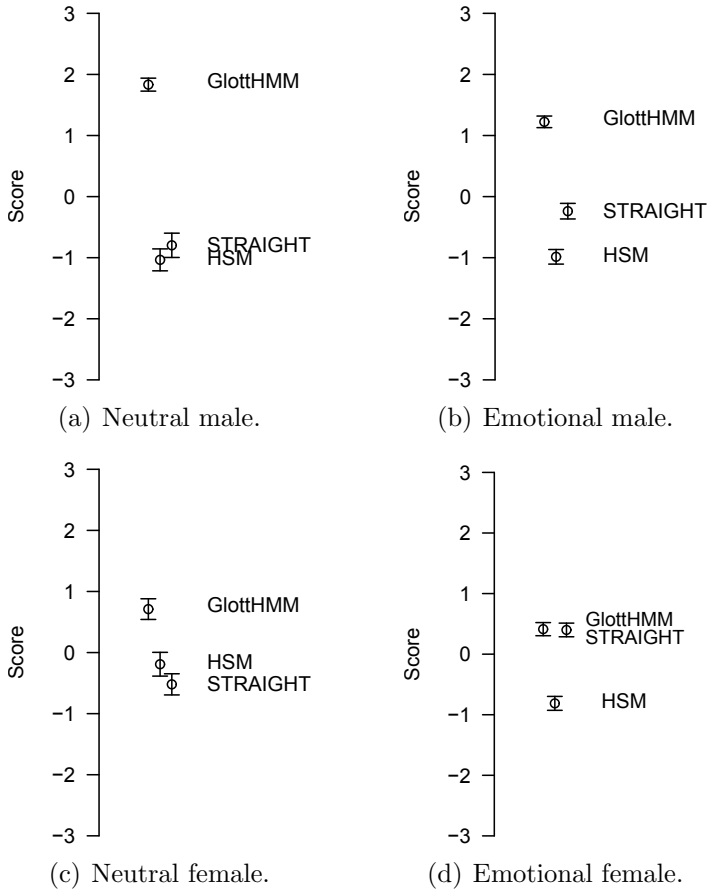
(c) CCR test results for female voice samples.

Figure 6.3: The CCR test results with their 95% confidence intervals.

The overall results of the subjective listening test are presented in Figures 6.3 (a)-(c). Figure 6.3 (a) presents the CCR ratings for the tested vocoders, when all test samples are evaluated. The GlottHMM vocoder has the best rating with a clear margin over the STRAIGHT vocoder, which in turn has a slightly smaller (but significant) margin compared to the HSM vocoder.

When the results are analyzed for only male or female speakers (Figures 6.3 (b) and (c)), a clear difference can be seen for the vocoder qualities by the gender: GlottHMM is clearly the best for male voices, but the differences get very small with female voices. The inconsistency of the GlottHMM vocoder quality regarding the gender is already a known issue [71]. It is likely the result of the unrobustness

of the inverse filtering procedure, which are known to be problematic for female voices. The STRAIGHT and HSM vocoders show consistency in their results, as their relative distance metric is quite fixed regardless of gender. This tells that the methods used in these vocoders are robust in quality regardless to the gender of the input speech, which makes the use of the vocoders more predictable compared to the GlottHMM vocoder.



(a) Neutral male. (b) Emotional male.

(c) Neutral female. (d) Emotional female.

Figure 6.4: The gender and emotion separated CCR test results with their 95% confidence intervals.

Figures 6.4 (a)-(d) present the CCR test results, when the type of speech and genders are separated. The speech types are sperated into the "neutral" and "emotional" categories, representing the different databases where the samples were taken from: "Neutral" samples are from HMM-training suitable high-quality databases, and the "emotional" samples are from the emotion database, which had sub-optimal SNR.

These results still show the overall trend of Figures 6.3 (a)-(c), but more subtle details are revealed. The HSM vocoder actually rises statistically on par to the STRAIGHT vocoder in the "neutral" samples, but is clearly weaker in the "emotional" samples. This tells that the HSM algorithm (or at least the used implementation) is unrobust in regards to noise and/or emotional speech input, which

was subjectively easy to verify. The difference in quality between the GlottHMM vocoder and the STRAIGHT vocoder shows also a trend beyond the gender separation: The samples from the HMM-training suitable "neutral" group received clearly higher scores for the GlottHMM vocoder for both genders than for the "emotional" group. This observation is consistent for all cases presented in Figure 6.4. Furthermore, this observation brings more backing to the claim that the quality of the GlottHMM vocoder is heavily influenced by the quality of the inverse filtering algorithm: In ideal conditions the vocoder quality is exquisite, but it degrades heavily with the introduction of sub-optimal voice and/or quality.

With the additional observations, the STRAIGHT vocoder can be seen as the most stable vocoder in terms of quality, as the changing trends in the CCR test results for the GlottHMM and the HSM vocoders can be attributed to observed problems in these methods.

7 Discussion and Conclusion

This section concludes the thesis by presenting a summary of the most important findings of the conducted analyses and tests. In the *Discussion*, a final look is given into the vocoders' test results along with ideas for future work. In the *Conclusion*, the thesis and its key results are presented in a brief and clear manner.

7.1 Discussion

The conducted tests, the statistical distribution analysis, the effect of emotion analysis, and the subjective evaluation of vocoder quality, provide a wide yet opaque picture of the studied vocoders. The wideness comes from the range of the conducted tests, which measure essential parts of a well-performing vocoder for HMM-based speech synthesis. The opaqueness comes from the sub-optimal suitability (non-uniformity of the emotional output, and the estimation of the distributions from context-independent data) of the used emotional speech database for the conducted tests; For more accurate results, the speech databases should consist of a consistent style of speech instead of mutually very different styles under a certain umbrella term. The discussion for each vocoder's observed properties is presented in the next sections.

The GlottHMM Vocoder

The GlottHMM vocoder had on overall the least Gaussian parameter distributions, along with low covariance matrix diagonalities for its parameter types. Specifically, many LSF parameters had high skewness and/or kurtosis values, which translates into mediocre accuracy in the single Gaussian modeled distribution used in HMM-based speech synthesis.

GlottHMM had the best discriminative power in the emotional effect test, with each parameter type performing above the average of the tested vocoders' parameter types. This suggests that the GlottHMM vocoder has good potential in the fields of emotion detection and emotional speech synthesis, which could be a topic for future research.

The subjective listening test ranked the GlottHMM vocoder as the best vocoder in terms of quality, but with a number of notable remarks: The quality of the vocoder output depends heavily on both, the gender of the speaker, and the quality of the input speech. This dependence was linked to the IAIF inverse filtering algorithm, which is known to have some problems in these cases. It is important to note that even if the GlottHMM vocoder outperformed the STRAIGHT and HSM vocoders in the analysis/synthesis test, its context-independent parameters' seemingly inferior suitability for a Gaussian model compared to the STRAIGHT vocoder will probably translate also into the context-dependent parameters. This would presumably lessen the quality compared to STRAIGHT in actual statistical TTS synthesis.

From the obtained results, the following suggestions for future research considering the GlottHMM vocoder can be made:

- The substitution of the LSF parameters used in spectral envelope modeling to some other parameter type (for example MFCCs).
- The suitability of the current GlottHMM vocoder parameters for emotion detection and emotional speech synthesis.
- The improvement of the inverse filtering algorithm, whose robustness is vital to the quality of the vocoder.

In the light of these observations, GlottHMM can be seen as an already powerful, but still unpolished vocoder with yet to be unravelled potential.

The STRAIGHT Vocoder

The STRAIGHT vocoder's MFCC coefficients had the most Gaussian distributions with the highest covariance matrix diagonality. Only some of the lower-order coefficients showed non-Gaussianity, but the distributions became rapidly very neatly Gaussian as the coefficient number increased. The aperiodicity (AP) coefficients had lower Gaussianity, which was attributed to the coefficient distributions' long asymmetrical tails. It was speculated that the single Gaussian modeling for the AP coefficients would be better if the asymmetric tail was truncated from the data from which the mean and variance are calculated for the Gaussian model.

The MFCCs had average discrimination power in the emotional effect test, but the AP coefficients had the weakest score out of all tested coefficient types. These results do not suggest that the STRAIGHT vocoder parameters would be well suited for emotion detection or manipulation.

STRAIGHT had the most stable performance in the subjective listening test, where it was placed consistently in the second place for the "all" and the gender differentiated sample groups. When the sample groups were differentiated by both emotion and gender, the STRAIGHT vocoder quality was on par to the HSM vocoder (no statistically significant difference) for the high-quality "neutral" samples, and on par with the GlottHMM vocoder for "emotional" female samples. The analysis of the subjective listening test proposed that the differences in the vocoders' positions were mostly due to problems in the GlottHMM and HSM vocoders in different sub-categories, which implies that the STRAIGHT vocoder has a stable output for all kinds of speech.

The obtained results suggest that the STRAIGHT vocoder is a very robust and stable vocoder that has well reached its maturity. This further reinforces the convention that the STRAIGHT vocoder is used as the benchmark vocoder for other state-of-the-art vocoders. Its coefficient types are well suited for single Gaussian modeling, and its output quality is stable across the board. The results do not suggest any major future research topics regarding the STRAIGHT vocoder.

The HSM Vocoder

The HSM vocoder's Harmonic LSF and voiced Stochastic LSF parameters had surprisingly high Gaussianity, which were almost on par with the STRAIGHT vocoder's

MFCCs. The unvoiced stochastic HSM parameters had high kurtosis values, which was the main source of non-Gaussianity in those parameters. However, this was mostly attributed to the amount of silent frames that the "unvoiced" frame category contained, which makes the obtained data hardly applicable for statistical models for any phonemes. The HSM parameters' covariance diagonalities were the among weakest of all of the parameter types that were studied, which can be attributed as a property of the LSF coefficients.

In the emotional effect test, both HSM parameter types had average performance, with both scoring near 60%. As the HSM vocoder was first developed as a tool for speech manipulation [25], the emotion sensitivity of the vocoder parameters might make the vocoder suitable for future research concerning emotional speech synthesis.

The HSM vocoder scored the lowest scores in the subjective listening test. Detailed analysis of the results showed that the HSM vocoder had difficulties mostly in the "emotional" speech categories, where the samples had relatively low SNR and/or HNR. On the "neutral" category, HSM performed on par with the STRAIGHT vocoder. This suggests that the HSM vocoder is sensitive to noise. Part of this can be attributed to the implementation, where very few things were done to improve noise robustness, but considering the operating principles of the HSM vocoder, the noise sensitivity is most probably the sum of these two things. The noise sensitivity is not a problem in the context of HMM-training, because the databases use high-quality speech samples. The noise sensitivity of the vocoder can be attributed to be the most significant factor affecting its rating in the conducted listening test. Because of this, the effect of the emotions to the vocoder quality can not be clearly distinguished.

The obtained results show that the HSM vocoder has at best a decent quality that is comparable to the STRAIGHT vocoder, but the quality can get easily a lot worse because of the noise sensitivity issues. The vocoder's main feature compared to the other tested vocoders is the simpleness and versatility of its synthesis procedure. Given these observations about the HSM vocoder, the following future research suggestions can be made:

- Improvement of the noise robustness of the HSM vocoder.
- The suitability of the HSM vocoder for emotional speech synthesis.

7.2 Conclusion

This thesis presented a literature study followed by an experimental part of the state-of-the-art vocoders utilized in statistical parametric speech synthesis. Based on the literature study, the GlottHMM, STRAIGHT and HSM vocoders were selected for the experimental part. The experiments conducted were the analysis of vocoder parameter distributions, the statistical effect of emotions to the parameter values, and a subjective listening test.

The parameter distribution analysis indicated that most vocoder parameters' distributions are sufficiently Gaussian to justify their single Gaussian distribution

modeling used in statistical parametric speech synthesis. However, clear differences were found between different vocoders. The MFCC coefficients used by the STRAIGHT vocoder were found to be the most Gaussian parameter type, clearly surpassing in Gaussianity the LSF coefficients used in the GlottHMM and HSM vocoders.

The emotion of the used speech was found to have an effect on the vocoder parameter values for all vocoders, with the GlottHMM vocoder parameters making the most number of unambiguous distinctions between the tested emotions.

In the subjective listening test, the relative analysis/synthesis quality of the vocoders was tested, and the GlottHMM vocoder overwhelmingly outperformed the other vocoders for the male voices, and slightly for the female voices. The STRAIGHT vocoder was found to be the most stable in quality and robust for various styles and qualities of speech.

References

[1] Abdel-Hamid, O., Abdou, S. M., and Rashwan, M. Improving Arabic HMM based speech synthesis quality. In *Interspeech* (2006), ISCA, pp. 1332–1335.

[2] Airas, M., and Alku, P. Emotions in vowel segments of continuous speech: Analysis of the glottal flow using the normalized amplitude quotient. *Phonetice 63*, 1 (2006), pp. 26–46.

[3] Akamine, M., and Kagoshima, T. Analytic generation of synthesis units by closed loop training for totally speaker driven text to speech system (TOS drive TTS). In *ICSLP* (1998), ISCA.

[4] Alku, P. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication 11*, 2-3 (1992), pp. 109–118.

[5] Alku, P., Story, B., and Airas, M. Estimation of the voice source from speech pressure signals: Evaluation of an inverse filtering technique using physical modelling of voice production. *Folia Phoniatrica et Logopaedica 58*, 2 (2006), pp. 102–113.

[6] Alpaydin, E. *Introduction to Machine Learning*. MIT Press, 2004.

[7] Balanda, K., and MacGillivray, H. Kurtosis: A critical review. *The American Statistician 42*, 2 (1988), pp. 111–119.

[8] Banos, E., Erro, D., Bonafonte, A., and Moreno, A. Flexible harmonic/stochastic modeling for HMM-based speech synthesis. *V Jornadas en Tecnología del Habla* (2008).

[9] Boashash, B. Estimating and interpreting the instantaneous frequency of a signal. I. Fundamentals. *Proceedings of the IEEE 80*, 4 (Apr. 1992), pp. 520–538.

[10] Cabral, J. *HMM-based Speech Synthesis Using an Acoustic Glottal Source Model*. PhD thesis, University of Edinburgh, 2010.

[11] Cabral, J., Renals, S., Richmond, K., and Yamagishi, J. Glottal spectral separation for parametric speech synthesis. pp. 1829–1832.

[12] Cabral, J. P., Renals, S., Richmond, K., and Yamagishi, J. Towards an improved modeling of the glottal source in statistical parametric speech synthesis. In *Proc. of the 6th ISCA Workshop on Speech Synthesis, Bonn, Germany* (2007), ISCA.

[13] Camacho, A. *SWIPE: A Sawtooth Waveform Iinspired Pitch Estimator for Speech and Music*. PhD thesis, University of Florida, 2007.

[14] Cotescu, M., and Gavat, I. Sources of increased variability in HMM synthetic voices. In *Speech Technology and Human-Computer Dialogue (SpeD), 2011 6th Conference on* (2011), pp. 1–6.

[15] Creer, S., Green, P., Cunningham, S., and Yamagishi, J. Building personalised synthesised voices for individuals with dysarthria using the HTS toolkit. In *Computer Synthesized Speech Technologies: Tools for Aiding Impairment*, J. W. Mullennix and S. E. Stern, Eds., 1st ed. IGI Global, 2009.

[16] d'Alessandro, C., and Doval, B. Voice quality modification for emotional speech synthesis,. In *Eurospeech* (2003), pp. 1653–1656.

[17] de Cheveigné, A., and Kawahara, H. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America 111*, 4 (2002), pp. 1917–1930.

[18] Depalle, P., and Hélie, T. Extraction of spectral peak parameters using a short-time fourier transform modeling and no sidelobe windows. In *IEEE 1997 Workshop on Applications of Signal Processing to Audio and Acoustics* (1997).

[19] Doval, B., d'Alessandro, C., and Henrich, N. The spectrum of glottal flow models. *Acta Acustica United With Acustica 92*, 6 (2006), pp. 1026–1046.

[20] Drugman, T., and Dutoit, T. Glottal closure and opening instant detection from speech signals. In *INTERSPEECH* (2009), ISCA, pp. 2891–2894.

[21] Drugman, T., and Dutoit, T. The deterministic plus stochastic model of the residual signal and its applications. *Audio, Speech, and Language Processing, IEEE Transactions on 20*, 3 (mar. 2012), pp. 968–981.

[22] Drugman, T., Moinet, A., Dutoit, T., and Wilfart, G. Using a pitch-synchronous residual codebook for hybrid HMM/frame selection speech synthesis. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on* (april 2009), pp. 3793–3796.

[23] Drugman, T., Wilfart, G., and Dutoit, T. A deterministic plus stochastic model of the residual signal for improved parametric speech synthesis. In *INTERSPEECH* (2009), ISCA, pp. 1779–1782.

[24] Erro, D., and Moreno, A. A pitch-asynchronous simple method for speech synthesis by diphone concatenation using the deterministic plus stochastic model. In *Proc. 10th Int. Conf. on Speech and Computer* (2005), pp. 321–324.

[25] Erro, D., Moreno, A., and Bonafonte, A. Flexible harmonic/stochastic speech synthesis. In *6th ISCA Workshop on Speech Synthesis* (2007), ISCA.

[26] Fant, G. The voice source in connected speech. *Speech Communication 22*, 2-3 (1997), pp. 125–139.

[27] Fukada, T., Tokuda, K., Kobayashi, T., and Imai, S. An adaptive algorithm for mel-cepstral analysis of speech. *Acoustics, Speech, and Signal Processing, IEEE International Conference on 1* (1992), pp. 137–140.

[28] Griffin, D., and Lim, J. Multiband excitation vocoder. *Acoustics, Speech and Signal Processing, IEEE Transactions on 36*, 8 (1988), pp. 1223–1235.

[29] Guerchi, D., and Mermelstein, P. Low-rate quantization of spectral information in a 4 kb/s pitch-synchronous CELP coder. In *IEEE Workshop on Speech Coding. Proceedings.* (2000), pp. 111–113.

[30] Han, S., Jeong, S., and Hahn, M. Optimum MVF estimation-based two-band excitation for HMM-based speech synthesis. *ETRI Journal 31*, 4 (Aug´. 2009), pp. 457–459.

[31] Hemptinne, C. Integration of the Harmonic plus Noise Model (HNM) into the Hidden Markov Model-Based Speech Synthesis System (HTS). Master's thesis, Idiap Research Institute, 2006.

[32] Hess, W. *Pitch Determination of Speech Signals: Algorithms and Devices.* Springer-Verlag, Berlin, 1983.

[33] HTS. HMM-based speech synthesis system, http://hts.sp.nitech.ac.jp/, referenced 24 September 2012.

[34] Huang, X., Acero, A., Acero, A., and Hon, H. *Spoken language processing: a guide to theory, algorithm, and system development.* Prentice Hall PTR, 2001.

[35] Hyvärinen, A. New approximations of differential entropy for independent component analysis and projection pursuit. In *Advances in Neural Information Processing Systems* (1998), vol. 10, MIT Press, pp. 273–279.

[36] Hyvärinen, A., and Oja, E. Independent component analysis: algorithms and applications. *Neural Networks 13*, 4-5 (2000), pp. 411–430.

[37] Imai, S. Cepstral analysis synthesis on the mel frequency scale. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '83.* (1983), vol. 8, pp. 93–96.

[38] ITU. *Methods for subjective determination of transmission quality, International Telecommunication Union, Recommendation ITU-T P.800*, 1996.

[39] Jolliffe, I. *Principal Component Analysis*, 2nd ed. Springer, New York, 2002.

[40] Karjalainen, M. Kommunikaatioakustiikka. Tech. rep., Helsinki University of Technology, 2000.

[41] Kawahara, H. Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited. In *IEEE ICASSP-97* (1997), vol. 2, pp. 1303–1306.

[42] Kawahara, H. STRAIGHT, exploitation of the other aspect of VOCODER: Perceptually isomorphic decomposition of speech sounds. *Acoustical Science and Technology 27*, 6 (2006), pp. 349–353.

[43] Kawahara, H., Estill, J., and Fujimura, O. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system straight. In *2nd MAVEBA* (2001).

[44] Kawahara, H., Katayose, H., de Cheveigné, A., and Patterson, R. D. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity. In *Eurospeech* (1999), ISCA.

[45] Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds. *Speech Communication 27*, 3-4 (Apr. 1999), pp. 187–207.

[46] Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., and Banno, H. Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on* (april 2008), pp. 3933–3936.

[47] Kim, S.-J., Kim, J.-J., and Hahn, M. HMM-based Korean speech synthesis system for hand-held devices. *Consumer Electronics, IEEE Transactions on 52*, 4 (nov. 2006), pp. 1384–1390.

[48] Kinnunen, T., and Li, H. An overview of text-independent speaker recognition: From features to supervectors. *Speech Communication 52*, 1 (2010), pp. 12–40.

[49] Koishida, K., Tokuda, K., Kobayashi, T., and Imai, S. Spectral representation of speech based on mel-generalized cepstral coefficients and its properties. *Electronics and Communications in Japan (Part III: Fundamental Electronic Science) 83*, 3 (2000), pp. 50–59.

[50] Lorenzo-Trueba, J., Barra-Chicote, R., Raitio, T., Obin, N., Alku, P., Yamagishi, J., and Montero, J. Towards glottal source controllability in expressive speech synthesis. In *Proc. of Interspeech* (2012).

[51] Magi, C., Pohjalainen, J., Bäckström, T., and Alku, P. Stabilised weighted linear prediction. *Speech Communication 51*, 5 (2009), pp. 401–411.

[52] Maia, R., Toda, T., Zen, H., Nankaku, Y., and Tokuda, K. *An excitation model for HMM-based speech synthesis based on residual modeling*, vol. 2. 2007, pp. 131–136.

[53] Makhoul, J., Viswanathan, R., Schwartz, R., and Huggins, A. A mixed-source model for speech compression and synthesis. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '78.* (apr 1978), vol. 3, pp. 163–166.

[54] McAulay, R., and Quatieri, T. Speech analysis/synthesis based on a sinusoidal representation. *Acoustics, Speech and Signal Processing, IEEE Transactions on 34*, 4 (aug 1986), pp. 744–754.

[55] McCree, A., and Barnwell, T.P., I. A mixed excitation lpc vocoder model for low bit rate speech coding. *Speech and Audio Processing, IEEE Transactions on 3*, 4 (jul 1995), pp. 242–250.

[56] Moulines, E., and Charpentier, F. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication 9*, 5-6 (1990), pp. 453–467.

[57] Murphy, P. J. Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis. *The Journal of the Acoustical Society of America 105*, 5 (1999), pp. 2866–2881.

[58] O'Shaughnessy, D. *Speech communications: human and machine*. IEEE.

[59] Paliwal, K., and Kleijn, W. Quantization of LPC parameters. *Speech Coding and Synthesis* (1995), pp. 433–466.

[60] Pantazis, Y., and Stylianou, Y. Improving the modeling of the noise part in the harmonic plus noise model of speech. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on* (31 2008-april 4 2008), pp. 4609 –4612.

[61] Rabiner, L., and Schafer, R. *Digital Processing of Speech Signals*. Prentice Hall, 1978.

[62] Raitio, T. Master thesis: Hidden markov model based finnish text-to-speech system utilizing glottal inverse filtering. Master's thesis, Helsinki University of Technology, 2008.

[63] Raitio, T., Suni, A., Pulakka, H., Vainio, M., and Alku, P. Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis. In *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on* (may 2011), pp. 4564–4567.

[64] Raitio, T., Suni, A., Yamagishi, J., Pulakka, H., Nurminen, J., Vainio, M., and Alku, P. HMM-based speech synthesis utilizing glottal inverse filtering. *Audio, Speech, and Language Processing, IEEE Transactions on 19*, 1 (jan. 2011), pp. 153–165.

[65] Schroeder, M., and Atal, B. Code-excited linear prediction (CELP): High-quality speech at very low bit rates. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85.* (apr 1985), vol. 10, pp. 937–940.

[66] Soong, F., and Juang, B. Line spectrum pair (LSP) and speech data compression. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '84.* (mar 1984), vol. 9, pp. 37–40.

[67] Soong, F., and Juang, B. Optimal quantization of LSP parameters. *Speech and Audio Processing, IEEE Transactions on 1*, 1 (jan 1993), pp. 15–24.

[68] SPTK. Speech signal processing toolkit, http://sp-tk.sourceforge.net/, referenced 24 September 2012.

[69] Stylianou, Y. *Harmonic plus Noise models for Speech, combined with Statistical Methods, for Speech and Speaker Modification.* PhD thesis, École nationale supérieure des télécommunication, 1996.

[70] Stylianou, Y. Applying the harmonic plus noise model in concatenative speech synthesis. *Speech and Audio Processing, IEEE Transactions on 9*, 1 (jan 2001), pp. 21–29.

[71] Suni, A., Raitio, T., Vainio, M., and Alku, P. The GlottHMM entry for Blizzard Challenge 2011: Utilizing source unit selection in HMM-based speech synthesis for improved excitation generation. In *Proc. of the ISCA Blizzard Challenge 2011 Workshop* (2011), ISCA.

[72] Yamagishi, J., Nose, T., Zen, H., Ling, Z., Toda, T., Tokuda, K., King, S., and Renals, S. Robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Transactions on Audio, Speech and Language Processing 17*, 6 (2009), pp. 1208–1230.

[73] Yamagishi, J., Usabaev, B., King, S., Watts, O., Dines, J., Tian, J., Hu, R., Guan, Y., Oura, K., Tokuda, K., Karhila, R., and Kurimo, M. Thousands of voices for HMM-based speech synthesis – analysis and application of TTS systems built on various ASR corpora. *IEEE Transactions on Audio, Speech and Language Processing 18*, 5 (July 2010), pp. 984–1004.

[74] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. Mixed excitation for HMM-based speech synthesis. In *Proc. Eurospeech 2001* (2001), ISCA, pp. 2263–2266.

[75] Zen, H., Toda, T., Nakamura, M., and Tokuda, K. Details of the Nitech HMM-Based Speech Synthesis System for the Blizzard Challenge 2005. *IEICE - Trans. Inf. Syst. E90-D*, 1 (Jan. 2007), pp. 325–333.

[76] Zen, H., Tokuda, K., and Black, A. W. Statistical parametric speech synthesis. *Speech Communication 51*, 11 (2009), pp. 1039–1064.

A Statistical Property Tables of Analyzed Vocoders

The complete statistical property tables for the tested vocoders (GlottHMM, STRAIGHT, and HSM) are presented in the following Tables A1-A9. The properties are computed from the Finnish emotional database individually for both genders and for the "neutral", "angry", and "sad" emotions, resulting in a total number of six tables per vocoder (= 18 tables).

Table A1: The statistical properties of the GlottHMM vocoder parameters for the "Neutral" databases.

GlottHMM	Male Mean	Neutral Var	Skew	Kurt	Neg
F0	101.89	517.5611	2.70	27.37	0.45
HNR				**Cov.d**	**0.65**
HNR1	-23.36	86.4447	0.04	-0.56	0.04
HNR2	-15.09	54.9878	-0.57	-0.06	0.02
HNR3	-10.96	33.4606	-0.92	0.79	0.00
HNR4	-7.44	20.2939	-1.37	2.53	0.13
HNR5	-2.49	5.0171	-2.69	13.97	1.32
Mean	-11.87	40.0408	-1.10	3.34	0.30
Min	-23.36	5.0171	-2.69	-0.56	0.00
Max	-2.49	86.4447	0.04	13.97	1.32
LSF VT V				**Cov.d**	**0.54**
LSF1	0.15	0.0007	-0.29	0.94	0.05
LSF2	0.22	0.0009	-0.04	0.58	0.01
LSF3	0.32	0.0018	-0.04	-0.39	0.03
LSF4	0.40	0.0024	0.00	-0.51	0.04
LSF5	0.47	0.0025	0.11	-0.67	0.03
LSF6	0.54	0.0020	0.13	-0.35	0.04
LSF7	0.64	0.0012	-0.02	0.85	0.04
LSF8	0.78	0.0024	-0.01	-0.49	0.07
LSF9	0.91	0.0032	-0.72	0.30	0.00
LSF10	1.03	0.0045	-0.91	0.43	0.01
LSF11	1.12	0.0052	-0.44	-0.71	0.10
LSF12	1.20	0.0052	0.02	-0.93	0.17
LSF13	1.29	0.0047	0.51	-0.30	0.00
LSF14	1.38	0.0033	0.48	0.21	0.00
LSF15	1.50	0.0024	0.10	0.35	0.01
LSF16	1.61	0.0020	0.08	0.54	0.04
LSF17	1.71	0.0023	0.16	0.20	0.01
LSF18	1.79	0.0022	0.03	0.17	0.00
LSF19	1.89	0.0021	0.30	0.56	0.03
LSF20	1.99	0.0022	0.04	0.55	0.01
LSF21	2.08	0.0021	0.18	0.73	0.05
LSF22	2.17	0.0031	0.66	0.06	0.00
LSF23	2.25	0.0044	0.22	-0.83	0.11
LSF24	2.33	0.0049	0.03	-0.64	0.07
LSF25	2.45	0.0065	-0.09	-0.32	0.01
LSF26	2.60	0.0053	-0.49	0.19	0.00
LSF27	2.73	0.0031	-0.59	0.55	0.00
LSF28	2.83	0.0017	-0.56	1.09	0.01
LSF29	2.92	0.0016	-0.25	0.21	0.01
LSF30	3.01	0.0016	-0.42	0.04	0.00
Mean	1.54	0.0029	-0.06	0.08	0.03
Min	0.15	0.0007	-0.91	-0.93	0.00
Max	3.01	0.0065	0.66	1.09	0.17
LSF VT UV				**Cov.d**	**0.38**
LSF1	0.06	0.0006	0.54	0.42	0.02
LSF2	0.12	0.0013	0.96	1.21	0.06
LSF3	0.21	0.0028	0.21	-0.33	0.01
LSF4	0.32	0.0027	-0.19	-0.23	0.01
LSF5	0.42	0.0027	-0.40	-0.07	0.00
LSF6	0.52	0.0039	-0.37	-0.51	0.04
LSF7	0.62	0.0051	-0.55	-0.41	0.02
LSF8	0.74	0.0043	-0.35	-0.09	0.00
LSF9	0.86	0.0037	-0.10	0.32	0.02
LSF10	0.97	0.0032	-0.01	1.47	0.14
LSF11	1.08	0.0035	-0.02	2.12	0.36
LSF12	1.18	0.0038	-0.47	1.78	0.33
LSF13	1.28	0.0039	-0.54	0.95	0.12
LSF14	1.38	0.0044	-0.30	0.64	0.07
LSF15	1.48	0.0044	0.08	1.28	0.16
LSF16	1.59	0.0035	0.60	2.30	0.37
LSF17	1.69	0.0030	0.82	2.71	0.43
LSF18	1.80	0.0034	0.56	2.31	0.31
LSF19	1.90	0.0039	0.33	1.46	0.25
LSF20	2.00	0.0037	0.07	1.41	0.18
LSF21	2.11	0.0028	0.00	1.95	0.34
LSF22	2.21	0.0023	0.03	1.39	0.16
LSF23	2.31	0.0025	-0.20	0.90	0.07
LSF24	2.40	0.0026	-0.55	1.05	0.12
LSF25	2.51	0.0024	-0.80	1.61	0.13
LSF26	2.61	0.0022	-0.89	1.91	0.21

GlottHMM	Female Mean	Neutral Var	Skew	Kurt	Neg
F0	154.82	1813.9903	1.44	6.95	1.08
HNR				**Cov.d**	**0.69**
HNR1	-25.97	132.5162	0.20	-1.03	0.21
HNR2	-17.81	72.3491	-0.14	-0.84	0.10
HNR3	-12.45	40.9932	-0.45	-0.53	0.06
HNR4	-7.40	16.6820	-1.03	1.00	0.05
HNR5	-3.04	3.9893	-1.55	3.86	0.47
Mean	-13.33	53.3060	-0.59	0.49	0.18
Min	-25.97	3.9893	-1.55	-1.03	0.05
Max	-3.04	132.5162	0.20	3.86	0.47
LSF VT V				**Cov.d**	**0.53**
LSF1	0.15	0.0005	-1.50	4.55	0.82
LSF2	0.22	0.0006	-0.38	1.12	0.03
LSF3	0.31	0.0013	0.05	-0.07	0.01
LSF4	0.38	0.0021	0.17	-0.47	0.02
LSF5	0.47	0.0027	0.13	-0.68	0.07
LSF6	0.53	0.0030	0.21	-0.77	0.14
LSF7	0.62	0.0030	0.14	-0.51	0.07
LSF8	0.74	0.0027	-0.19	-0.07	0.00
LSF9	0.87	0.0019	-0.42	0.40	0.00
LSF10	0.98	0.0020	-0.49	0.95	0.11
LSF11	1.09	0.0033	-0.44	0.07	0.00
LSF12	1.19	0.0046	-0.52	-0.34	0.04
LSF13	1.29	0.0056	-0.29	-0.71	0.10
LSF14	1.38	0.0068	0.16	-0.76	0.10
LSF15	1.48	0.0067	0.43	-0.51	0.04
LSF16	1.58	0.0052	0.42	0.16	0.00
LSF17	1.70	0.0034	0.32	0.28	0.01
LSF18	1.82	0.0025	0.00	-0.12	0.01
LSF19	1.91	0.0022	0.07	0.18	0.00
LSF20	1.99	0.0023	0.29	0.13	0.00
LSF21	2.08	0.0021	0.30	0.47	0.00
LSF22	2.19	0.0027	0.11	0.05	0.00
LSF23	2.31	0.0025	-0.35	0.33	0.01
LSF24	2.40	0.0018	0.02	0.23	0.02
LSF25	2.49	0.0022	0.35	0.17	0.00
LSF26	2.59	0.0027	0.21	-0.08	0.00
LSF27	2.69	0.0043	0.45	0.20	0.01
LSF28	2.81	0.0052	-0.05	-0.97	0.21
LSF29	2.94	0.0036	-0.86	0.25	0.02
LSF30	3.04	0.0014	-1.42	3.35	0.32
Mean	1.54	0.0030	-0.10	0.23	0.07
Min	0.15	0.0005	-1.50	-0.97	0.00
Max	3.04	0.0068	0.45	4.55	0.82
LSF VT UV				**Cov.d**	**0.40**
LSF1	0.06	0.0011	0.96	-0.04	0.03
LSF2	0.12	0.0019	0.64	0.21	0.01
LSF3	0.22	0.0026	0.41	0.16	0.01
LSF4	0.33	0.0023	0.03	0.03	0.00
LSF5	0.44	0.0025	-0.24	0.09	0.01
LSF6	0.54	0.0033	-0.71	0.25	0.00
LSF7	0.64	0.0041	-0.86	0.54	0.07
LSF8	0.76	0.0045	-0.92	1.02	0.17
LSF9	0.87	0.0035	-0.78	1.04	0.09
LSF10	0.97	0.0027	-0.39	1.37	0.21
LSF11	1.07	0.0027	0.16	1.27	0.14
LSF12	1.19	0.0031	-0.16	1.46	0.24
LSF13	1.29	0.0031	0.02	2.72	0.63
LSF14	1.39	0.0033	0.03	2.54	0.51
LSF15	1.49	0.0031	-0.09	2.08	0.41
LSF16	1.59	0.0032	0.10	2.46	0.44
LSF17	1.70	0.0031	0.74	3.41	0.73
LSF18	1.81	0.0028	0.93	3.30	0.66
LSF19	1.92	0.0022	1.06	4.18	0.94
LSF20	2.01	0.0023	1.00	4.26	0.80
LSF21	2.12	0.0026	0.65	3.59	0.82
LSF22	2.22	0.0026	0.39	3.19	0.77
LSF23	2.32	0.0018	0.13	3.01	0.73
LSF24	2.42	0.0014	0.01	2.75	0.47
LSF25	2.52	0.0017	-0.42	1.65	0.23
LSF26	2.62	0.0015	-0.88	2.02	0.38

Table A1: The statistical properties of the GlottHMM vocoder parameters for the "Neutral" databases.

GlottHMM	Male Neutral					GlottHMM	Female Neutral				
	Mean	Var	Skew	Kurt	Neg		Mean	Var	Skew	Kurt	Neg
LSF27	2.72	0.0015	-0.61	2.42	0.31	LSF27	2.72	0.0011	-0.86	3.27	0.42
LSF28	2.82	0.0009	-0.81	2.84	0.28	LSF28	2.82	0.0011	-0.82	2.36	0.49
LSF29	2.92	0.0006	-0.50	2.03	0.16	LSF29	2.93	0.0007	-1.04	4.34	0.74
LSF30	3.02	0.0006	-0.88	1.89	0.22	LSF30	3.02	0.0004	-0.30	2.19	0.14
Mean	1.53	0.0029	-0.14	1.22	0.16	Mean	1.54	0.0024	-0.04	2.02	0.38
Min	0.06	0.0006	-0.89	-0.51	0.00	Min	0.06	0.0004	-1.04	-0.04	0.00
Max	3.02	0.0051	0.96	2.84	0.43	Max	3.02	0.0045	1.06	4.34	0.94
LSF S				**Cov.d**	**0.86**	**LSF S**				**Cov.d**	**0.87**
LSF1	0.07	0.0002	1.31	2.35	0.07	LSF1	0.09	0.0002	1.45	4.69	0.41
LSF2	0.16	0.0003	1.82	7.10	0.65	LSF2	0.17	0.0004	2.35	9.29	1.62
LSF3	0.31	0.0002	1.85	8.26	0.97	LSF3	0.32	0.0004	2.06	12.29	2.17
LSF4	0.45	0.0002	1.29	8.90	0.69	LSF4	0.46	0.0002	2.03	13.92	1.65
LSF5	0.60	0.0001	0.94	14.74	1.47	LSF5	0.61	0.0002	1.06	12.76	1.74
LSF6	0.75	0.0001	0.15	13.43	1.45	LSF6	0.75	0.0002	0.23	10.96	1.40
LSF7	0.91	0.0001	-0.67	10.67	1.62	LSF7	0.91	0.0002	-0.42	10.28	1.12
LSF8	1.06	0.0001	-0.63	8.93	1.31	LSF8	1.06	0.0002	-0.23	6.87	0.80
LSF9	1.21	0.0002	-0.87	5.60	0.55	LSF9	1.22	0.0002	-0.11	6.16	0.52
LSF10	1.37	0.0002	-0.26	3.36	0.31	LSF10	1.37	0.0003	-0.12	3.35	0.39
LSF11	1.53	0.0004	-0.23	0.95	0.05	LSF11	1.54	0.0003	-0.34	2.50	0.25
LSF12	1.68	0.0007	-0.54	1.06	0.07	LSF12	1.69	0.0003	0.00	2.52	0.33
LSF13	1.85	0.0010	0.00	0.82	0.06	LSF13	1.85	0.0004	-0.15	1.83	0.13
LSF14	2.01	0.0010	-0.24	2.07	0.17	LSF14	2.01	0.0005	-0.36	1.20	0.06
LSF15	2.18	0.0010	0.22	1.63	0.18	LSF15	2.17	0.0007	-0.35	0.93	0.04
LSF16	2.34	0.0007	0.24	1.21	0.12	LSF16	2.32	0.0010	-0.06	0.72	0.06
LSF17	2.50	0.0006	-0.12	1.16	0.10	LSF17	2.50	0.0012	-0.19	0.86	0.06
LSF18	2.65	0.0007	-0.50	1.58	0.11	LSF18	2.65	0.0009	-0.19	1.22	0.17
LSF19	2.83	0.0007	0.01	1.69	0.14	LSF19	2.83	0.0006	0.29	1.26	0.18
LSF20	2.97	0.0009	-0.45	0.90	0.09	LSF20	2.98	0.0005	-0.32	1.60	0.13
Mean	1.47	0.0005	0.17	4.82	0.51	Mean	1.47	0.0005	0.33	5.26	0.66
Min	0.07	0.0001	-0.87	0.82	0.05	Min	0.09	0.0002	-0.42	0.72	0.04
Max	2.97	0.0010	1.85	14.74	1.62	Max	2.98	0.0012	2.35	13.92	2.17

Table A2: The statistical properties of the GlottHMM vocoder parameters for the "Sad" databases.

GlottHMM	Male Sad					GlottHMM	Female Sad				
	Mean	Var	Skew	Kurt	Neg		Mean	Var	Skew	Kurt	Neg
F0	97.22	637.5354	0.88	11.12	0.39	F0	181.35	6738.6537	0.55	-0.43	0.06
HNR				**Cov. d**	**0.70**	**HNR**				**Cov. d**	**0.68**
HNR1	-21.10	106.4190	0.11	-0.91	0.17	HNR1	-22.06	124.6732	-0.03	-0.94	0.17
HNR2	-13.04	51.1631	-0.46	-0.43	0.03	HNR2	-16.17	75.8539	-0.28	-0.60	0.10
HNR3	-8.90	25.0796	-0.91	0.81	0.01	HNR3	-12.23	64.2585	-1.03	0.71	0.03
HNR4	-5.88	12.1001	-1.27	2.29	0.12	HNR4	-8.94	53.5580	-1.61	2.21	0.31
HNR5	-1.99	1.9882	-1.95	8.20	0.71	HNR5	-3.80	13.4018	-2.37	6.69	1.32
Mean	-10.18	39.3500	-0.90	1.99	0.20	Mean	-12.64	66.3491	-1.07	1.62	0.39
Min	-21.10	1.9882	-1.95	-0.91	0.01	Min	-22.06	13.4018	-2.37	-0.94	0.03
Max	-1.99	106.4190	0.11	8.20	0.71	Max	-3.80	124.6732	-0.03	6.69	1.32
LSF VT V				**Cov. d**	**0.56**	**LSF VT V**				**Cov. d**	**0.54**
LSF1	0.14	0.0008	-0.65	1.38	0.13	LSF1	0.16	0.0013	-0.42	1.69	0.37
LSF2	0.22	0.0011	-0.15	1.56	0.05	LSF2	0.23	0.0008	0.06	0.72	0.00
LSF3	0.32	0.0016	0.12	-0.33	0.02	LSF3	0.31	0.0015	0.04	-0.47	0.05
LSF4	0.39	0.0024	0.06	-0.54	0.03	LSF4	0.39	0.0025	0.06	-0.44	0.01
LSF5	0.46	0.0023	0.18	-0.52	0.01	LSF5	0.47	0.0028	0.08	-0.48	0.04
LSF6	0.53	0.0019	0.32	-0.15	0.01	LSF6	0.54	0.0030	0.15	-0.50	0.02
LSF7	0.64	0.0013	0.12	0.56	0.00	LSF7	0.64	0.0027	0.05	-0.01	0.01
LSF8	0.77	0.0020	0.10	0.16	0.01	LSF8	0.75	0.0025	-0.25	0.28	0.01
LSF9	0.91	0.0033	-0.72	0.38	0.01	LSF9	0.88	0.0024	-0.27	0.51	0.02
LSF10	1.02	0.0051	-0.78	-0.02	0.00	LSF10	0.99	0.0030	-0.26	-0.09	0.00
LSF11	1.12	0.0062	-0.50	-0.62	0.09	LSF11	1.10	0.0042	-0.38	-0.45	0.01
LSF12	1.21	0.0056	-0.01	-0.83	0.10	LSF12	1.20	0.0057	-0.29	-0.79	0.10
LSF13	1.30	0.0044	0.48	-0.22	0.00	LSF13	1.30	0.0063	-0.03	-0.98	0.12
LSF14	1.39	0.0032	0.55	0.39	0.02	LSF14	1.39	0.0062	0.30	-0.58	0.04
LSF15	1.51	0.0022	0.24	0.04	0.00	LSF15	1.49	0.0048	0.52	-0.08	0.00
LSF16	1.62	0.0019	0.27	0.77	0.04	LSF16	1.60	0.0035	0.09	-0.13	0.00
LSF17	1.72	0.0025	0.35	0.36	0.01	LSF17	1.72	0.0022	-0.07	0.52	0.01
LSF18	1.81	0.0024	0.12	-0.06	0.00	LSF18	1.82	0.0019	0.15	0.32	0.00
LSF19	1.90	0.0022	0.18	0.71	0.05	LSF19	1.91	0.0020	0.19	0.38	0.00
LSF20	2.00	0.0023	0.09	0.14	0.01	LSF20	1.99	0.0023	0.20	0.24	0.00
LSF21	2.09	0.0023	0.26	0.58	0.01	LSF21	2.09	0.0022	0.13	0.21	0.00
LSF22	2.18	0.0036	0.66	0.08	0.01	LSF22	2.20	0.0023	0.06	0.00	0.00
LSF23	2.26	0.0046	0.32	-0.71	0.11	LSF23	2.30	0.0024	0.10	-0.14	0.00
LSF24	2.36	0.0048	-0.06	-0.52	0.02	LSF24	2.38	0.0026	0.17	-0.48	0.03
LSF25	2.48	0.0051	-0.26	-0.22	0.00	LSF25	2.46	0.0034	0.19	-0.57	0.03
LSF26	2.60	0.0045	-0.24	-0.22	0.01	LSF26	2.55	0.0047	0.02	-0.78	0.08
LSF27	2.71	0.0040	-0.35	-0.51	0.04	LSF27	2.67	0.0055	-0.18	-0.36	0.01
LSF28	2.83	0.0029	-0.88	1.14	0.10	LSF28	2.82	0.0041	-0.54	0.04	0.01
LSF29	2.94	0.0017	-0.58	0.66	0.01	LSF29	2.94	0.0022	-0.72	1.03	0.04
LSF30	3.03	0.0011	-0.86	1.44	0.04	LSF30	3.04	0.0011	-0.92	1.44	0.03
Mean	1.55	0.0030	-0.05	0.16	0.03	Mean	1.54	0.0031	-0.06	0.00	0.03
Min	0.14	0.0008	-0.88	-0.83	0.00	Min	0.16	0.0008	-0.92	-0.98	0.00
Max	3.03	0.0062	0.66	1.56	0.13	Max	3.04	0.0063	0.52	1.69	0.37
LSF VT UV				**Cov. d**	**0.45**	**LSF VT UV**				**Cov. d**	**0.47**
LSF1	0.06	0.0006	0.92	0.94	0.00	LSF1	0.06	0.0012	1.47	1.75	0.07
LSF2	0.12	0.0017	1.07	1.07	0.07	LSF2	0.11	0.0021	0.84	0.10	0.01
LSF3	0.22	0.0027	0.23	-0.13	0.00	LSF3	0.22	0.0027	0.22	-0.04	0.00
LSF4	0.33	0.0023	-0.24	0.32	0.01	LSF4	0.34	0.0019	0.14	0.25	0.01
LSF5	0.44	0.0024	-0.59	0.60	0.04	LSF5	0.44	0.0022	-0.38	0.57	0.02

Table A2: The statistical properties of the GlottHMM vocoder parameters for the "Sad" databases.

GlottHMM	Male Sad					GlottHMM	Female Sad				
	Mean	Var	Skew	Kurt	Neg		Mean	Var	Skew	Kurt	Neg
LSF6	0.54	0.0032	-0.73	0.45	0.02	LSF6	0.55	0.0028	-0.80	0.80	0.06
LSF7	0.65	0.0037	-0.84	0.62	0.05	LSF7	0.66	0.0033	-1.04	1.28	0.16
LSF8	0.75	0.0032	-0.61	1.13	0.14	LSF8	0.77	0.0030	-0.99	2.04	0.32
LSF9	0.87	0.0027	-0.23	1.09	0.09	LSF9	0.88	0.0022	-0.64	1.61	0.16
LSF10	0.98	0.0023	-0.04	2.78	0.49	LSF10	0.98	0.0019	-0.34	2.40	0.38
LSF11	1.08	0.0027	0.17	2.52	0.37	LSF11	1.08	0.0023	0.20	2.18	0.27
LSF12	1.19	0.0028	-0.39	2.55	0.44	LSF12	1.19	0.0026	-0.49	2.44	0.43
LSF13	1.29	0.0027	-0.55	1.72	0.31	LSF13	1.29	0.0025	-0.74	3.24	0.71
LSF14	1.38	0.0030	-0.29	1.41	0.17	LSF14	1.39	0.0026	-0.52	2.29	0.42
LSF15	1.49	0.0031	0.05	1.70	0.17	LSF15	1.49	0.0027	-0.46	1.58	0.30
LSF16	1.60	0.0024	0.51	3.09	0.49	LSF16	1.59	0.0025	-0.31	2.02	0.36
LSF17	1.70	0.0020	0.79	3.40	0.41	LSF17	1.70	0.0019	0.44	3.62	0.51
LSF18	1.80	0.0024	0.55	3.04	0.41	LSF18	1.81	0.0017	0.48	2.82	0.36
LSF19	1.90	0.0029	0.15	2.29	0.35	LSF19	1.91	0.0015	0.39	3.98	0.56
LSF20	2.01	0.0025	-0.07	2.24	0.35	LSF20	2.01	0.0018	0.25	3.36	0.51
LSF21	2.11	0.0020	-0.14	2.69	0.46	LSF21	2.11	0.0021	-0.13	2.21	0.32
LSF22	2.22	0.0018	-0.31	2.16	0.38	LSF22	2.21	0.0018	-0.33	2.70	0.42
LSF23	2.32	0.0017	-0.68	2.27	0.30	LSF23	2.32	0.0013	-0.53	3.47	0.46
LSF24	2.42	0.0016	-0.78	2.71	0.31	LSF24	2.42	0.0012	-0.78	2.52	0.30
LSF25	2.52	0.0016	-0.73	1.26	0.10	LSF25	2.52	0.0013	-1.00	2.32	0.22
LSF26	2.62	0.0016	-1.16	2.01	0.30	LSF26	2.62	0.0013	-1.42	4.26	0.56
LSF27	2.72	0.0013	-1.19	3.34	0.53	LSF27	2.72	0.0011	-1.46	5.44	0.68
LSF28	2.82	0.0009	-1.01	3.29	0.46	LSF28	2.83	0.0008	-0.89	3.95	0.61
LSF29	2.93	0.0006	-0.74	4.39	0.38	LSF29	2.93	0.0005	-0.45	3.71	0.43
LSF30	3.02	0.0005	-1.04	3.15	0.30	LSF30	3.03	0.0004	-0.32	2.59	0.24
Mean	1.54	0.0022	-0.26	2.00	0.26	Mean	1.54	0.0019	-0.32	2.38	0.33
Min	0.06	0.0005	-1.19	-0.13	0.00	Min	0.06	0.0004	-1.46	-0.04	0.00
Max	3.02	0.0037	1.07	4.39	0.53	Max	3.03	0.0033	1.47	5.44	0.71
LSF S				**Cov. d**	**0.85**	**LSF S**				**Cov. d**	**0.69**
LSF1	0.07	0.0001	1.45	5.81	0.14	LSF1	0.10	0.0008	0.82	0.62	0.03
LSF2	0.16	0.0002	1.54	7.75	0.65	LSF2	0.19	0.0013	1.52	2.66	0.44
LSF3	0.31	0.0001	0.93	9.83	0.85	LSF3	0.34	0.0013	1.92	4.53	0.78
LSF4	0.45	0.0001	0.42	9.30	0.90	LSF4	0.47	0.0009	2.26	8.06	1.43
LSF5	0.60	0.0001	-0.61	7.49	1.12	LSF5	0.62	0.0009	2.46	9.63	2.21
LSF6	0.75	0.0002	-0.89	7.95	1.26	LSF6	0.76	0.0006	1.70	7.12	1.50
LSF7	0.90	0.0002	-1.39	6.74	1.15	LSF7	0.91	0.0004	0.78	4.06	0.78
LSF8	1.06	0.0002	-1.22	5.14	0.88	LSF8	1.06	0.0003	-0.09	3.97	0.56
LSF9	1.21	0.0003	-1.33	3.18	0.51	LSF9	1.22	0.0003	-0.27	4.59	0.40
LSF10	1.36	0.0004	-0.41	2.74	0.38	LSF10	1.37	0.0003	-0.09	2.82	0.29
LSF11	1.53	0.0006	-0.36	0.90	0.08	LSF11	1.53	0.0003	-0.41	3.34	0.18
LSF12	1.68	0.0007	-0.18	0.46	0.01	LSF12	1.68	0.0004	-0.12	2.35	0.06
LSF13	1.85	0.0009	-0.18	0.51	0.04	LSF13	1.84	0.0007	-0.23	1.24	0.04
LSF14	2.01	0.0010	-0.08	1.21	0.11	LSF14	2.00	0.0010	-0.18	1.11	0.07
LSF15	2.16	0.0016	-0.42	0.66	0.04	LSF15	2.17	0.0013	0.05	0.53	0.03
LSF16	2.34	0.0016	-0.10	1.47	0.26	LSF16	2.33	0.0011	0.04	0.55	0.02
LSF17	2.50	0.0011	0.68	1.40	0.09	LSF17	2.50	0.0010	0.07	0.73	0.03
LSF18	2.65	0.0008	-0.09	0.81	0.04	LSF18	2.65	0.0008	0.03	0.96	0.08
LSF19	2.83	0.0006	0.24	1.19	0.09	LSF19	2.83	0.0006	0.14	0.72	0.04
LSF20	2.98	0.0006	-0.36	1.13	0.05	LSF20	2.98	0.0005	-0.38	2.02	0.13
Mean	1.47	0.0006	-0.12	3.78	0.43	Mean	1.48	0.0007	0.50	3.08	0.45
Min	0.07	0.0001	-1.39	0.46	0.01	Min	0.10	0.0003	-0.41	0.53	0.02
Max	2.98	0.0016	1.54	9.83	1.26	Max	2.98	0.0013	2.46	9.63	2.21

Table A3: The statistical properties of the GlottHMM vocoder parameters for the "Angry" databases.

GlottHMM	Male Angry					GlottHMM	Female Angry				
	Mean	Var	Skew	Kurt	Neg		Mean	Var	Skew	Kurt	Neg
F0	132.29	1595.6273	1.23	2.29	0.02	F0	176.22	3662.3354	1.24	2.54	0.30
HNR				**Cov. d**	**0.65**	**HNR**				**Cov. d**	**0.69**
HNR1	-25.96	91.6193	0.12	-0.49	0.03	HNR1	-22.85	116.5609	-0.15	-0.93	0.17
HNR2	-17.16	66.1527	-0.47	-0.46	0.03	HNR2	-15.90	62.3832	-0.48	-0.50	0.06
HNR3	-12.42	49.2974	-0.95	0.49	0.00	HNR3	-12.25	51.7807	-0.92	0.35	0.00
HNR4	-8.24	30.3791	-1.45	2.09	0.11	HNR4	-8.27	32.9767	-1.60	2.76	0.30
HNR5	-3.05	6.9499	-1.88	4.76	0.87	HNR5	-3.59	8.6651	-2.22	6.94	1.03
Mean	-13.37	48.8797	-0.93	1.28	0.21	Mean	-12.57	54.4733	-1.07	1.73	0.31
Min	-25.96	6.9499	-1.88	-0.49	0.00	Min	-22.85	8.6651	-2.22	-0.93	0.00
Max	-3.05	91.6193	0.12	4.76	0.87	Max	-3.59	116.5609	-0.15	6.94	1.03
LSF VT V				**Cov. d**	**0.54**	**LSF VT V**				**Cov. d**	**0.51**
LSF1	0.16	0.0008	-0.49	1.00	0.06	LSF1	0.16	0.0011	-0.66	2.87	0.41
LSF2	0.23	0.0013	0.17	0.16	0.00	LSF2	0.24	0.0013	0.33	2.12	0.08
LSF3	0.32	0.0021	0.02	-0.51	0.05	LSF3	0.33	0.0021	0.16	-0.48	0.06
LSF4	0.40	0.0030	0.10	-0.76	0.07	LSF4	0.41	0.0032	0.14	-0.74	0.13
LSF5	0.48	0.0030	0.24	-0.64	0.10	LSF5	0.49	0.0038	0.22	-0.68	0.10
LSF6	0.55	0.0027	0.28	-0.51	0.05	LSF6	0.56	0.0043	0.40	-0.32	0.04
LSF7	0.66	0.0020	0.17	0.16	0.00	LSF7	0.65	0.0041	0.34	-0.22	0.01
LSF8	0.79	0.0025	-0.20	-0.04	0.00	LSF8	0.78	0.0030	-0.32	0.37	0.03
LSF9	0.93	0.0036	-0.85	0.47	0.01	LSF9	0.90	0.0026	-0.60	1.29	0.08
LSF10	1.03	0.0048	-0.83	0.07	0.00	LSF10	1.01	0.0031	-0.57	0.62	0.02
LSF11	1.13	0.0057	-0.48	-0.68	0.07	LSF11	1.11	0.0041	-0.56	-0.18	0.00
LSF12	1.21	0.0058	-0.07	-0.88	0.13	LSF12	1.21	0.0052	-0.45	-0.56	0.04
LSF13	1.30	0.0053	0.29	-0.61	0.05	LSF13	1.30	0.0061	-0.21	-0.87	0.10
LSF14	1.40	0.0041	0.34	-0.03	0.01	LSF14	1.40	0.0067	0.13	-0.75	0.09
LSF15	1.51	0.0028	0.06	0.17	0.00	LSF15	1.50	0.0059	0.46	-0.25	0.00
LSF16	1.62	0.0021	0.35	0.71	0.07	LSF16	1.61	0.0047	0.40	0.01	0.00
LSF17	1.72	0.0022	0.43	0.17	0.00	LSF17	1.73	0.0034	0.14	0.10	0.00
LSF18	1.80	0.0022	0.34	-0.01	0.00	LSF18	1.83	0.0028	0.18	0.08	0.00

Table A3: The statistical properties of the GlottHMM vocoder parameters for the "Angry" databases.

GlottHMM	Male Angry					GlottHMM	Female Angry				
	Mean	**Var**	**Skew**	**Kurt**	**Neg**		**Mean**	**Var**	**Skew**	**Kurt**	**Neg**
LSF19	1.88	0.0023	0.36	0.64	0.00	LSF19	1.92	0.0027	0.17	0.11	0.00
LSF20	1.98	0.0027	0.45	0.42	0.01	LSF20	2.00	0.0026	0.20	-0.02	0.01
LSF21	2.07	0.0036	0.38	-0.22	0.00	LSF21	2.09	0.0029	0.22	-0.02	0.00
LSF22	2.17	0.0045	0.28	-0.50	0.04	LSF22	2.19	0.0031	-0.02	-0.20	0.00
LSF23	2.26	0.0050	0.05	-0.85	0.13	LSF23	2.29	0.0025	-0.03	0.62	0.01
LSF24	2.36	0.0041	-0.15	-0.45	0.01	LSF24	2.38	0.0025	0.29	0.33	0.01
LSF25	2.48	0.0046	0.02	-0.06	0.00	LSF25	2.47	0.0035	0.48	-0.09	0.01
LSF26	2.61	0.0047	-0.36	-0.37	0.02	LSF26	2.55	0.0050	0.28	-0.65	0.15
LSF27	2.74	0.0027	-0.48	0.53	0.03	LSF27	2.66	0.0069	0.09	-0.59	0.03
LSF28	2.83	0.0019	-0.23	0.02	0.01	LSF28	2.81	0.0057	-0.49	-0.32	0.00
LSF29	2.93	0.0018	-0.18	-0.39	0.01	LSF29	2.94	0.0025	-0.94	1.48	0.12
LSF30	3.02	0.0017	-0.73	0.34	0.01	LSF30	3.04	0.0010	-0.64	0.55	0.01
Mean	1.55	0.0032	-0.02	-0.09	0.03	Mean	1.55	0.0036	-0.03	0.12	0.05
Min	0.16	0.0008	-0.85	-0.88	0.00	Min	0.16	0.0010	-0.94	-0.87	0.00
Max	3.02	0.0058	0.45	1.00	0.13	Max	3.04	0.0069	0.48	2.87	0.41
LSF VT UV				**Cov. d**	**0.41**	**LSF VT UV**				**Cov. d**	**0.42**
LSF1	0.08	0.0016	0.74	0.15	0.00	LSF1	0.07	0.0016	1.24	1.40	0.01
LSF2	0.14	0.0025	0.39	-0.62	0.06	LSF2	0.13	0.0026	0.49	-0.24	0.04
LSF3	0.24	0.0031	0.00	-0.33	0.01	LSF3	0.24	0.0031	0.06	-0.21	0.00
LSF4	0.34	0.0028	-0.20	0.02	0.00	LSF4	0.34	0.0025	0.22	0.24	0.01
LSF5	0.44	0.0027	-0.36	0.22	0.01	LSF5	0.44	0.0025	0.09	0.39	0.02
LSF6	0.53	0.0033	-0.40	-0.17	0.00	LSF6	0.54	0.0031	-0.24	0.38	0.02
LSF7	0.64	0.0042	-0.48	-0.17	0.00	LSF7	0.65	0.0041	-0.42	0.58	0.08
LSF8	0.75	0.0038	-0.36	0.26	0.01	LSF8	0.76	0.0043	-0.54	0.79	0.09
LSF9	0.87	0.0031	-0.25	0.75	0.05	LSF9	0.88	0.0035	-0.45	0.90	0.12
LSF10	0.98	0.0032	-0.39	1.51	0.18	LSF10	0.98	0.0030	-0.11	1.15	0.11
LSF11	1.09	0.0035	-0.48	1.90	0.21	LSF11	1.08	0.0032	0.26	0.96	0.07
LSF12	1.19	0.0039	-0.57	1.72	0.25	LSF12	1.19	0.0034	0.05	1.27	0.15
LSF13	1.29	0.0040	-0.56	0.94	0.12	LSF13	1.29	0.0035	0.13	2.09	0.38
LSF14	1.38	0.0042	-0.30	0.60	0.06	LSF14	1.39	0.0037	0.09	1.94	0.39
LSF15	1.49	0.0042	0.05	0.70	0.03	LSF15	1.49	0.0038	0.01	1.36	0.30
LSF16	1.60	0.0033	0.33	1.69	0.18	LSF16	1.60	0.0035	-0.02	1.72	0.31
LSF17	1.71	0.0026	0.79	2.01	0.18	LSF17	1.70	0.0030	0.49	2.64	0.40
LSF18	1.80	0.0029	0.60	1.95	0.16	LSF18	1.81	0.0028	0.70	2.28	0.30
LSF19	1.90	0.0035	0.35	1.61	0.17	LSF19	1.91	0.0027	0.61	2.00	0.36
LSF20	2.00	0.0038	0.09	1.28	0.16	LSF20	2.01	0.0028	0.52	2.24	0.38
LSF21	2.11	0.0033	-0.07	1.34	0.15	LSF21	2.11	0.0029	0.26	2.25	0.41
LSF22	2.21	0.0033	-0.47	1.11	0.17	LSF22	2.21	0.0028	0.03	1.90	0.35
LSF23	2.31	0.0035	-0.85	1.42	0.18	LSF23	2.32	0.0023	-0.24	2.10	0.38
LSF24	2.40	0.0038	-1.17	2.13	0.30	LSF24	2.42	0.0019	-0.30	1.87	0.22
LSF25	2.51	0.0031	-1.30	3.24	0.31	LSF25	2.51	0.0022	-0.46	1.65	0.23
LSF26	2.62	0.0021	-1.09	2.29	0.24	LSF26	2.62	0.0021	-0.98	2.82	0.51
LSF27	2.73	0.0012	-0.77	2.98	0.43	LSF27	2.72	0.0017	-1.33	4.84	0.70
LSF28	2.83	0.0009	-0.62	1.78	0.16	LSF28	2.82	0.0015	-0.84	3.67	0.51
LSF29	2.93	0.0007	-0.62	1.70	0.12	LSF29	2.93	0.0008	-0.81	4.38	0.48
LSF30	3.02	0.0006	-1.03	2.78	0.27	LSF30	3.02	0.0005	-0.07	1.33	0.14
Mean	1.54	0.0030	-0.30	1.23	0.14	Mean	1.54	0.0027	-0.05	1.69	0.25
Min	0.08	0.0006	-1.30	-0.62	0.00	Min	0.07	0.0005	-1.33	-0.24	0.00
Max	3.02	0.0042	0.79	3.24	0.43	Max	3.02	0.0043	1.24	4.84	0.70
LSF S				**Cov. d**	**0.80**	**LSF S**				**Cov. d**	**0.71**
LSF1	0.09	0.0005	0.97	1.67	0.00	LSF1	0.10	0.0007	1.14	2.16	0.08
LSF2	0.18	0.0006	1.33	2.96	0.09	LSF2	0.19	0.0010	1.27	2.50	0.11
LSF3	0.32	0.0004	1.58	4.43	0.36	LSF3	0.33	0.0008	1.34	3.95	0.38
LSF4	0.46	0.0003	1.42	5.95	0.35	LSF4	0.47	0.0006	1.60	7.56	0.86
LSF5	0.61	0.0002	1.38	10.74	0.76	LSF5	0.62	0.0005	1.86	9.11	1.12
LSF6	0.76	0.0002	1.17	13.71	1.09	LSF6	0.76	0.0004	1.27	7.76	1.19
LSF7	0.91	0.0002	0.93	11.41	0.98	LSF7	0.92	0.0004	0.68	8.26	1.03
LSF8	1.06	0.0002	0.73	10.07	0.60	LSF8	1.06	0.0003	0.20	5.22	0.83
LSF9	1.22	0.0003	-0.12	6.44	0.57	LSF9	1.22	0.0003	0.59	4.84	0.63
LSF10	1.37	0.0004	0.09	2.83	0.24	LSF10	1.38	0.0003	0.48	4.42	0.53
LSF11	1.53	0.0005	0.37	1.67	0.21	LSF11	1.54	0.0004	0.23	3.87	0.43
LSF12	1.69	0.0006	-0.14	1.59	0.12	LSF12	1.69	0.0004	0.24	3.39	0.27
LSF13	1.85	0.0009	-0.24	1.21	0.07	LSF13	1.85	0.0006	-0.03	1.90	0.13
LSF14	2.01	0.0011	-0.01	0.88	0.06	LSF14	2.00	0.0009	-0.34	1.01	0.05
LSF15	2.18	0.0011	0.08	1.10	0.09	LSF15	2.17	0.0012	-0.30	1.19	0.11
LSF16	2.34	0.0008	0.37	1.22	0.07	LSF16	2.33	0.0014	0.10	0.58	0.06
LSF17	2.50	0.0006	-0.08	1.66	0.12	LSF17	2.51	0.0011	0.11	1.29	0.13
LSF18	2.65	0.0007	-0.25	0.88	0.06	LSF18	2.66	0.0009	-0.11	1.13	0.12
LSF19	2.83	0.0007	-0.04	1.65	0.16	LSF19	2.83	0.0007	0.15	1.12	0.11
LSF20	2.97	0.0009	-0.72	1.42	0.10	LSF20	2.98	0.0005	-0.43	1.52	0.12
Mean	1.48	0.0006	0.44	4.17	0.30	Mean	1.48	0.0007	0.50	3.64	0.41
Min	0.09	0.0002	-0.72	0.88	0.00	Min	0.10	0.0003	-0.43	0.58	0.05
Max	2.97	0.0011	1.58	13.71	1.09	Max	2.98	0.0014	1.86	9.11	1.19

Table A4: The statistical properties of the STRAIGHT vocoder parameters for the "Neutral" databases.

STRAIGHT	Male Neutral					STRAIGHT	Female Neutral				
	Mean	**Var**	**Skew**	**Kurt**	**Neg**		**Mean**	**Var**	**Skew**	**Kurt**	**Neg**
F0	99.86	573.4645	2.81	16.69	1.08	F0	159.16	975.7279	1.41	2.88	0.24
AP				**Cov. d**	**0.81**	**AP**				**Cov. d**	**0.84**
AP1	-12.10	42.5967	-1.25	1.02	0.09	AP1	-13.15	60.6909	-1.15	0.28	0.01
AP2	-9.27	17.0961	-1.75	4.24	0.79	AP2	-10.07	20.1502	-1.63	2.82	0.53
AP3	-8.51	7.8604	-1.83	6.85	0.89	AP3	-8.91	6.2011	-1.59	4.68	0.58
AP4	-8.17	2.8425	-0.48	1.93	0.08	AP4	-8.44	3.2927	-0.48	1.86	0.11
AP5	-8.13	2.7130	-0.28	1.49	0.02	AP5	-8.25	2.6388	-0.07	0.84	0.03
Mean	-9.24	14.6217	-1.12	3.10	0.37	Mean	-9.76	18.5947	-0.98	2.10	0.25

Table A4: The statistical properties of the STRAIGHT vocoder parameters for the "Neutral" databases.

STRAIGHT	**Male**	**Neutral**				**STRAIGHT**	**Female**	**Neutral**			
	Mean	**Var**	**Skew**	**Kurt**	**Neg**		**Mean**	**Var**	**Skew**	**Kurt**	**Neg**
Min	-12.10	2.7130	-1.83	1.02	0.02	Min	-13.15	2.6388	-1.63	0.28	0.01
Max	-8.13	42.5967	-0.28	6.85	0.89	Max	-8.25	60.6909	-0.07	4.68	0.58
MFCC V				**Cov. d**	**0.87**	**MFCC V**				**Cov. d**	**0.88**
MFCC0	3.95	1.6540	-0.24	0.04	0.02	MFCC0	4.14	1.0938	-0.29	-0.12	0.00
MFCC1	2.19	0.3366	-1.27	2.25	0.27	MFCC1	2.29	0.3374	-0.91	1.78	0.05
MFCC2	0.13	0.2830	-0.31	-0.33	0.01	MFCC2	0.37	0.2253	-0.08	-0.51	0.02
MFCC3	0.61	0.1270	0.44	0.26	0.00	MFCC3	0.61	0.2108	0.11	-0.29	0.00
MFCC4	0.30	0.1565	-0.26	-0.47	0.02	MFCC4	0.08	0.1452	-0.37	-0.19	0.01
MFCC5	0.03	0.1031	-0.56	0.04	0.00	MFCC5	0.02	0.0768	-0.19	-0.09	0.00
MFCC6	0.10	0.0615	-0.16	0.04	0.00	MFCC6	-0.05	0.0700	-0.07	-0.25	0.01
MFCC7	-0.19	0.0973	-0.19	-0.36	0.01	MFCC7	-0.22	0.0615	-0.25	-0.07	0.00
MFCC8	-0.17	0.0541	-0.23	0.06	0.00	MFCC8	-0.17	0.0509	0.07	0.03	0.00
MFCC9	-0.03	0.0465	-0.22	-0.18	0.00	MFCC9	-0.17	0.0529	0.08	-0.17	0.00
MFCC10	-0.08	0.0535	-0.14	-0.12	0.00	MFCC10	0.01	0.0343	0.23	0.09	0.00
MFCC11	0.12	0.0373	0.14	0.07	0.00	MFCC11	-0.18	0.0320	0.03	-0.23	0.01
MFCC12	-0.11	0.0325	0.31	0.15	0.00	MFCC12	0.01	0.0277	-0.06	-0.05	0.00
MFCC13	0.00	0.0269	-0.34	0.42	0.05	MFCC13	-0.01	0.0276	0.09	-0.10	0.00
MFCC14	-0.01	0.0269	-0.17	-0.16	0.00	MFCC14	-0.10	0.0238	-0.08	-0.02	0.00
MFCC15	-0.02	0.0252	-0.29	-0.10	0.00	MFCC15	-0.08	0.0248	0.00	-0.12	0.00
MFCC16	-0.06	0.0265	0.15	0.04	0.00	MFCC16	-0.10	0.0185	0.04	0.24	0.01
MFCC17	-0.04	0.0185	-0.11	0.05	0.00	MFCC17	-0.12	0.0184	0.00	0.23	0.01
MFCC18	-0.06	0.0168	0.07	0.03	0.00	MFCC18	-0.06	0.0173	-0.09	0.10	0.00
MFCC19	-0.08	0.0139	0.03	0.10	0.00	MFCC19	-0.05	0.0167	-0.15	0.06	0.00
MFCC20	-0.05	0.0135	0.11	0.12	0.01	MFCC20	-0.08	0.0133	-0.04	0.08	0.00
MFCC21	-0.06	0.0111	0.12	0.24	0.01	MFCC21	-0.03	0.0114	0.00	0.10	0.00
MFCC22	-0.06	0.0129	-0.16	0.29	0.01	MFCC22	-0.06	0.0113	0.02	0.02	0.00
MFCC23	-0.05	0.0104	-0.11	0.15	0.03	MFCC23	-0.02	0.0102	-0.03	-0.02	0.00
MFCC24	-0.03	0.0101	0.13	0.12	0.00	MFCC24	-0.05	0.0091	-0.04	0.09	0.00
Mean	0.25	0.1302	-0.13	0.11	0.02	Mean	0.24	0.1048	-0.08	0.02	0.01
Min	-0.19	0.0101	-1.27	-0.47	0.00	Min	-0.22	0.0091	-0.91	-0.51	0.00
Max	3.95	1.6540	0.44	2.25	0.27	Max	4.14	1.0938	0.23	1.78	0.05
MFCC UV				**Cov. d**	**0.93**	**MFCC UV**				**Cov. d**	**0.96**
MFCC0	1.98	2.1235	0.89	0.06	0.01	MFCC0	1.86	1.5233	1.29	0.74	0.02
MFCC1	1.21	0.7203	-0.16	0.47	0.01	MFCC1	0.94	0.5436	0.06	2.28	0.66
MFCC2	0.18	0.1721	-0.22	0.81	0.08	MFCC2	0.17	0.1370	-0.03	1.04	0.20
MFCC3	0.38	0.1060	0.85	1.28	0.05	MFCC3	0.26	0.0859	1.10	3.64	0.49
MFCC4	0.18	0.0901	-0.07	0.98	0.09	MFCC4	0.08	0.0512	-0.80	2.38	0.31
MFCC5	0.14	0.0593	-0.55	1.49	0.12	MFCC5	0.13	0.0344	-0.43	1.56	0.13
MFCC6	0.07	0.0424	-0.10	0.32	0.02	MFCC6	0.01	0.0402	-0.47	0.90	0.05
MFCC7	-0.03	0.0614	-0.65	0.34	0.00	MFCC7	0.01	0.0367	-0.80	1.09	0.05
MFCC8	-0.03	0.0361	-0.46	0.67	0.03	MFCC8	-0.01	0.0282	-0.49	0.60	0.02
MFCC9	0.04	0.0302	-0.20	0.59	0.05	MFCC9	0.00	0.0259	-0.42	1.04	0.08
MFCC10	0.01	0.0310	-0.14	0.95	0.07	MFCC10	-0.02	0.0225	0.43	1.09	0.04
MFCC11	0.08	0.0305	0.52	1.06	0.06	MFCC11	-0.02	0.0203	-0.38	0.70	0.03
MFCC12	-0.01	0.0233	-0.34	0.34	0.01	MFCC12	0.02	0.0178	0.09	0.71	0.03
MFCC13	0.04	0.0209	-0.09	0.76	0.04	MFCC13	0.03	0.0163	-0.05	0.39	0.02
MFCC14	0.00	0.0215	0.01	0.56	0.03	MFCC14	-0.04	0.0167	-0.20	0.40	0.01
MFCC15	0.00	0.0180	-0.08	0.59	0.02	MFCC15	-0.03	0.0171	-0.09	0.37	0.01
MFCC16	-0.04	0.0184	-0.06	0.77	0.05	MFCC16	-0.06	0.0146	-0.11	0.20	0.00
MFCC17	0.00	0.0153	0.04	0.33	0.00	MFCC17	-0.05	0.0153	-0.21	0.33	0.01
MFCC18	-0.03	0.0143	-0.03	0.27	0.00	MFCC18	-0.02	0.0130	-0.05	0.20	0.00
MFCC19	-0.02	0.0129	-0.14	0.32	0.00	MFCC19	0.00	0.0125	-0.10	0.18	0.00
MFCC20	-0.03	0.0115	0.14	0.30	0.01	MFCC20	-0.04	0.0110	0.02	0.19	0.00
MFCC21	-0.04	0.0104	0.02	0.20	0.00	MFCC21	-0.04	0.0104	0.10	0.16	0.00
MFCC22	-0.04	0.0100	0.06	0.26	0.00	MFCC22	-0.05	0.0087	0.01	0.15	0.00
MFCC23	-0.01	0.0084	-0.02	0.17	0.00	MFCC23	-0.01	0.0079	0.10	0.17	0.00
MFCC24	-0.01	0.0078	0.00	0.17	0.00	MFCC24	-0.02	0.0073	-0.13	0.14	0.00
Mean	0.16	0.1478	-0.03	0.56	0.03	Mean	0.12	0.1087	-0.06	0.83	0.09
Min	-0.04	0.0078	-0.65	0.06	0.00	Min	-0.06	0.0073	-0.80	0.14	0.00
Max	1.98	2.1235	0.89	1.49	0.12	Max	1.86	1.5233	1.29	3.64	0.66

Table A5: The statistical properties of the STRAIGHT vocoder parameters for the "Sad" databases.

STRAIGHT	**Male**	**Sad**				**STRAIGHT**	**Female**	**Sad**			
	Mean	**Var**	**Skew**	**Kurt**	**Neg**		**Mean**	**Var**	**Skew**	**Kurt**	**Neg**
F0	104.85	721.6326	4.13	29.18	2.72	F0	191.27	4098.2530	0.64	-0.78	0.22
AP				**Cov. d**	**0.92**	**AP**				**Cov. d**	**0.83**
AP1	-10.26	23.8477	-1.51	2.20	0.32	AP1	-12.01	56.3349	-1.69	2.34	0.41
AP2	-8.78	7.8108	-1.08	2.75	0.11	AP2	-9.73	18.3863	-2.22	6.21	1.29
AP3	-8.29	3.3852	-0.47	2.77	0.17	AP3	-8.91	8.8732	-2.46	9.85	1.84
AP4	-8.17	2.2224	-0.05	0.87	0.02	AP4	-8.52	4.3578	-1.71	7.87	0.55
AP5	-8.15	2.2872	-0.05	0.52	0.00	AP5	-8.33	2.7623	-0.51	1.65	0.03
Mean	-8.73	7.9107	-0.63	1.82	0.13	Mean	-9.50	18.1429	-1.72	5.58	0.82
Min	-10.26	2.2224	-1.51	0.52	0.00	Min	-12.01	2.7623	-2.46	1.65	0.03
Max	-8.15	23.8477	-0.05	2.77	0.32	Max	-8.33	56.3349	-0.51	9.85	1.84
MFCC V				**Cov. d**	**0.85**	**MFCC V**				**Cov. d**	**0.85**
MFCC0	3.98	1.3873	-0.42	-0.05	0.00	MFCC0	4.09	1.5442	0.22	-0.38	0.03
MFCC1	2.07	0.3183	-0.99	1.99	0.13	MFCC1	2.10	0.3079	-0.58	0.78	0.01
MFCC2	0.22	0.2514	-0.23	-0.37	0.01	MFCC2	0.22	0.2686	-0.47	-0.01	0.00
MFCC3	0.64	0.1186	0.38	0.10	0.00	MFCC3	0.62	0.1908	0.45	0.08	0.00
MFCC4	0.31	0.1422	-0.23	-0.32	0.02	MFCC4	0.08	0.1430	-0.37	-0.38	0.02
MFCC5	0.05	0.1092	-0.47	-0.18	0.01	MFCC5	-0.01	0.0913	-0.12	-0.35	0.02
MFCC6	0.06	0.0659	-0.46	0.85	0.00	MFCC6	0.00	0.0834	-0.20	-0.22	0.01
MFCC7	-0.17	0.1172	-0.19	-0.39	0.02	MFCC7	-0.25	0.0994	-0.30	-0.42	0.04
MFCC8	-0.13	0.0495	-0.25	-0.14	0.00	MFCC8	-0.11	0.0539	-0.03	-0.16	0.00
MFCC9	-0.06	0.0464	0.00	0.02	0.00	MFCC9	-0.20	0.0541	-0.01	0.01	0.00

Table A5: The statistical properties of the STRAIGHT vocoder parameters for the "Sad" databases.

STRAIGHT	**Male**	**Sad**				**STRAIGHT**	**Female**	**Sad**			
	Mean	**Var**	**Skew**	**Kurt**	**Neg**		**Mean**	**Var**	**Skew**	**Kurt**	**Neg**
MFCC10	0.01	0.0595	0.01	-0.23	0.02	MFCC10	-0.02	0.0393	0.20	0.24	0.03
MFCC11	0.02	0.0293	0.18	0.26	0.00	MFCC11	-0.15	0.0357	0.23	0.01	0.00
MFCC12	-0.08	0.0337	0.17	-0.01	0.00	MFCC12	-0.08	0.0274	0.00	0.21	0.00
MFCC13	-0.02	0.0279	0.00	0.37	0.00	MFCC13	0.03	0.0267	0.12	-0.10	0.00
MFCC14	-0.01	0.0247	-0.03	0.00	0.00	MFCC14	-0.10	0.0283	0.35	0.31	0.00
MFCC15	-0.01	0.0230	-0.16	-0.03	0.00	MFCC15	-0.07	0.0220	0.01	0.10	0.01
MFCC16	-0.08	0.0279	0.04	-0.07	0.01	MFCC16	-0.08	0.0189	-0.11	0.13	0.00
MFCC17	-0.04	0.0237	-0.12	-0.03	0.00	MFCC17	-0.10	0.0179	-0.04	0.32	0.01
MFCC18	-0.05	0.0192	-0.18	0.07	0.00	MFCC18	-0.05	0.0184	-0.11	0.09	0.00
MFCC19	-0.06	0.0206	0.04	0.17	0.00	MFCC19	-0.07	0.0137	0.06	0.47	0.02
MFCC20	-0.06	0.0158	0.13	0.12	0.00	MFCC20	-0.06	0.0141	-0.30	0.23	0.00
MFCC21	-0.04	0.0157	-0.03	-0.02	0.01	MFCC21	-0.03	0.0116	-0.19	0.38	0.01
MFCC22	-0.07	0.0144	-0.01	0.11	0.00	MFCC22	-0.04	0.0100	-0.02	0.32	0.00
MFCC23	-0.03	0.0144	0.13	0.08	0.01	MFCC23	-0.02	0.0094	-0.01	0.43	0.03
MFCC24	-0.04	0.0114	-0.02	0.02	0.00	MFCC24	-0.02	0.0088	-0.11	0.30	0.02
Mean	0.26	0.1187	-0.11	0.09	0.01	Mean	0.23	0.1255	-0.05	0.10	0.01
Min	-0.17	0.0114	-0.99	-0.39	0.00	Min	-0.25	0.0088	-0.58	-0.42	0.00
Max	3.98	1.3873	0.38	1.99	0.13	Max	4.09	1.5442	0.45	0.78	0.04
MFCC UV				**Cov. d**	**0.93**	**MFCC UV**				**Cov. d**	**0.95**
MFCC0	1.71	1.6603	1.13	0.39	0.00	MFCC0	1.65	1.0724	1.69	2.23	0.28
MFCC1	1.02	0.4766	0.13	1.52	0.22	MFCC1	0.83	0.2858	0.48	4.29	1.44
MFCC2	0.15	0.1181	-0.13	1.43	0.31	MFCC2	0.09	0.0885	-0.26	2.69	0.59
MFCC3	0.30	0.0944	1.18	2.09	0.24	MFCC3	0.19	0.0644	1.28	4.89	0.71
MFCC4	0.14	0.0676	0.32	1.81	0.30	MFCC4	0.07	0.0385	-0.50	3.13	0.39
MFCC5	0.14	0.0442	-0.52	2.42	0.24	MFCC5	0.13	0.0270	-0.37	1.58	0.11
MFCC6	0.04	0.0400	-0.31	1.21	0.09	MFCC6	0.03	0.0305	-0.49	1.44	0.12
MFCC7	-0.01	0.0563	-0.75	0.68	0.02	MFCC7	0.03	0.0346	-1.00	1.78	0.09
MFCC8	-0.01	0.0299	-0.47	0.81	0.04	MFCC8	0.00	0.0237	-0.56	1.03	0.06
MFCC9	0.02	0.0256	-0.31	0.83	0.07	MFCC9	0.02	0.0236	-0.62	1.67	0.12
MFCC10	0.02	0.0276	0.42	1.25	0.11	MFCC10	-0.01	0.0195	0.36	1.20	0.05
MFCC11	0.04	0.0210	0.30	1.05	0.08	MFCC11	-0.01	0.0198	-0.39	0.82	0.05
MFCC12	-0.01	0.0221	-0.19	0.87	0.03	MFCC12	0.00	0.0177	-0.10	0.68	0.03
MFCC13	0.03	0.0194	0.01	0.89	0.04	MFCC13	0.04	0.0164	0.07	0.58	0.02
MFCC14	0.00	0.0175	0.09	0.78	0.03	MFCC14	-0.03	0.0159	-0.14	0.73	0.03
MFCC15	0.00	0.0157	0.06	0.57	0.02	MFCC15	-0.03	0.0152	-0.10	0.53	0.01
MFCC16	-0.05	0.0168	-0.25	0.55	0.01	MFCC16	-0.05	0.0134	-0.04	0.33	0.00
MFCC17	0.00	0.0150	-0.13	0.70	0.02	MFCC17	-0.03	0.0153	-0.31	0.49	0.01
MFCC18	-0.01	0.0146	-0.28	0.67	0.01	MFCC18	-0.01	0.0132	-0.14	0.27	0.01
MFCC19	-0.01	0.0137	-0.15	0.47	0.01	MFCC19	0.01	0.0119	-0.15	0.22	0.00
MFCC20	-0.05	0.0109	0.03	0.40	0.01	MFCC20	-0.05	0.0111	-0.02	0.37	0.01
MFCC21	-0.03	0.0117	0.12	0.27	0.00	MFCC21	-0.05	0.0106	0.15	0.08	0.00
MFCC22	-0.04	0.0103	0.05	0.25	0.01	MFCC22	-0.03	0.0086	0.10	0.17	0.00
MFCC23	0.00	0.0087	0.11	0.37	0.01	MFCC23	0.00	0.0078	0.04	0.14	0.00
MFCC24	-0.01	0.0077	-0.09	0.24	0.01	MFCC24	-0.01	0.0072	-0.06	0.13	0.00
Mean	0.14	0.1138	0.01	0.90	0.08	Mean	0.11	0.0757	-0.04	1.26	0.17
Min	-0.05	0.0077	-0.75	0.24	0.00	Min	-0.05	0.0072	-1.00	0.08	0.00
Max	1.71	1.6603	1.18	2.42	0.31	Max	1.65	1.0724	1.69	4.89	1.44

Table A6: The statistical properties of the STRAIGHT vocoder parameters for the "Angry" databases.

STRAIGHT	**Male**	**Angry**				**STRAIGHT**	**Female**	**Angry**			
	Mean	**Var**	**Skew**	**Kurt**	**Neg**		**Mean**	**Var**	**Skew**	**Kurt**	**Neg**
F0	138.90	2083.2241	0.95	0.59	0.00	F0	184.71	3023.6234	1.30	1.52	0.10
AP				**Cov. d**	**0.83**	**AP**				**Cov. d**	**0.79**
AP1	-11.78	41.3843	-1.46	1.65	0.25	AP1	-12.72	57.2323	-1.35	0.99	0.08
AP2	-9.27	15.0547	-1.75	4.70	0.72	AP2	-10.15	24.7783	-1.99	4.52	0.93
AP3	-8.64	6.4201	-1.79	7.00	0.82	AP3	-8.98	9.0298	-2.13	7.36	1.19
AP4	-8.28	2.8530	-0.39	1.40	0.07	AP4	-8.41	4.0194	-0.93	3.91	0.23
AP5	-8.20	2.6239	-0.26	1.42	0.04	AP5	-8.23	2.6866	-0.02	1.21	0.07
Mean	-9.23	13.6672	-1.13	3.24	0.38	Mean	-9.69	19.5493	-1.29	3.60	0.50
Min	-11.78	2.6239	-1.79	1.40	0.04	Min	-12.72	2.6866	-2.13	0.99	0.07
Max	-8.20	41.3843	-0.26	7.00	0.82	Max	-8.23	57.2323	-0.02	7.36	1.19
MFCC V				**Cov. d**	**0.94**	**MFCC V**				**Cov. d**	**0.87**
MFCC0	4.34	2.7677	-0.50	-0.39	0.03	MFCC0	5.09	2.0385	-0.17	0.08	0.00
MFCC1	1.98	0.2925	-0.69	1.37	0.06	MFCC1	2.04	0.4190	-0.43	0.24	0.00
MFCC2	-0.03	0.3441	-0.03	-0.59	0.04	MFCC2	0.02	0.4438	-0.11	-0.62	0.05
MFCC3	0.54	0.1595	0.31	-0.01	0.00	MFCC3	0.65	0.2569	0.23	-0.39	0.03
MFCC4	0.31	0.2084	-0.29	-0.45	0.01	MFCC4	-0.11	0.1875	-0.41	-0.21	0.02
MFCC5	-0.18	0.1573	-0.41	-0.34	0.03	MFCC5	-0.08	0.0675	-0.05	0.00	0.00
MFCC6	0.05	0.0613	0.00	-0.01	0.00	MFCC6	-0.02	0.0778	-0.06	-0.14	0.00
MFCC7	-0.17	0.0999	-0.07	-0.40	0.03	MFCC7	-0.34	0.0846	-0.16	-0.15	0.00
MFCC8	-0.27	0.0597	-0.18	-0.22	0.00	MFCC8	-0.12	0.0640	0.10	0.16	0.01
MFCC9	-0.10	0.0528	-0.09	0.01	0.00	MFCC9	-0.26	0.0575	-0.04	-0.20	0.00
MFCC10	-0.08	0.0643	0.00	-0.33	0.01	MFCC10	0.02	0.0488	-0.05	-0.04	0.01
MFCC11	0.01	0.0429	0.28	0.43	0.00	MFCC11	-0.15	0.0371	-0.09	-0.15	0.00
MFCC12	-0.11	0.0362	0.00	0.04	0.00	MFCC12	-0.03	0.0334	0.04	-0.15	0.01
MFCC13	-0.01	0.0351	0.12	0.18	0.01	MFCC13	0.01	0.0324	0.17	0.22	0.02
MFCC14	-0.04	0.0308	-0.01	0.07	0.00	MFCC14	-0.11	0.0323	0.05	0.02	0.00
MFCC15	-0.04	0.0266	0.02	0.19	0.00	MFCC15	-0.05	0.0265	0.12	0.12	0.00
MFCC16	-0.08	0.0281	0.06	0.25	0.00	MFCC16	-0.08	0.0215	0.11	0.84	0.01
MFCC17	-0.07	0.0203	-0.09	0.07	0.00	MFCC17	-0.09	0.0194	0.08	0.26	0.00
MFCC18	-0.05	0.0175	0.07	0.22	0.00	MFCC18	-0.05	0.0203	-0.04	-0.17	0.00
MFCC19	-0.08	0.0210	0.07	0.11	0.00	MFCC19	-0.05	0.0164	-0.17	0.31	0.00
MFCC20	-0.05	0.0177	0.02	0.06	0.00	MFCC20	-0.06	0.0129	-0.09	0.18	0.00
MFCC21	-0.04	0.0140	0.13	0.09	0.00	MFCC21	-0.02	0.0124	-0.10	0.23	0.01
MFCC22	-0.05	0.0144	-0.09	0.09	0.00	MFCC22	-0.03	0.0111	0.05	0.39	0.02

Table A6: The statistical properties of the STRAIGHT vocoder parameters for the "Angry" databases.

STRAIGHT	**Male**	**Angry**				**STRAIGHT**	**Female**	**Angry**			
	Mean	**Var**	**Skew**	**Kurt**	**Neg**		**Mean**	**Var**	**Skew**	**Kurt**	**Neg**
MFCC23	-0.02	0.0110	-0.02	0.02	0.00	MFCC23	-0.03	0.0104	-0.07	0.47	0.01
MFCC24	-0.04	0.0107	0.02	-0.01	0.01	MFCC24	-0.02	0.0091	0.01	0.38	0.01
Mean	0.23	0.1838	-0.06	0.02	0.01	Mean	0.24	0.1616	-0.04	0.07	0.01
Min	-0.27	0.0107	-0.69	-0.59	0.00	Min	-0.34	0.0091	-0.43	-0.62	0.00
Max	4.34	2.7677	0.31	1.37	0.06	Max	5.09	2.0385	0.23	0.84	0.05
MFCC UV				**Cov. d**	**0.96**	**MFCC UV**				**Cov. d**	**0.97**
MFCC0	1.71	3.4505	0.75	-0.22	0.02	MFCC0	2.47	3.0434	0.76	-0.24	0.03
MFCC1	1.05	0.6431	-0.10	0.66	0.03	MFCC1	1.01	0.5696	-0.28	1.88	0.33
MFCC2	0.12	0.1901	-0.38	1.01	0.17	MFCC2	0.06	0.2131	-0.52	1.00	0.13
MFCC3	0.30	0.0987	1.02	1.80	0.14	MFCC3	0.24	0.1078	0.54	1.80	0.27
MFCC4	0.15	0.0784	-0.06	1.51	0.26	MFCC4	0.03	0.0738	-0.98	2.49	0.30
MFCC5	0.08	0.0669	-0.93	2.60	0.29	MFCC5	0.08	0.0514	-0.71	1.23	0.11
MFCC6	0.02	0.0422	0.05	0.56	0.01	MFCC6	0.01	0.0443	-0.09	0.39	0.01
MFCC7	-0.05	0.0525	-0.53	0.62	0.02	MFCC7	-0.03	0.0541	-0.77	0.98	0.04
MFCC8	-0.06	0.0352	-0.33	0.45	0.01	MFCC8	-0.05	0.0353	-0.16	0.09	0.00
MFCC9	0.00	0.0301	-0.12	0.70	0.04	MFCC9	-0.04	0.0352	-0.25	0.42	0.01
MFCC10	0.01	0.0291	0.26	0.91	0.07	MFCC10	-0.01	0.0296	0.39	0.94	0.06
MFCC11	0.03	0.0280	0.64	1.36	0.05	MFCC11	-0.02	0.0245	-0.30	0.52	0.02
MFCC12	-0.05	0.0238	-0.40	0.78	0.04	MFCC12	-0.01	0.0224	0.03	0.66	0.02
MFCC13	0.03	0.0228	0.35	1.07	0.04	MFCC13	0.02	0.0193	-0.10	0.48	0.01
MFCC14	-0.01	0.0193	-0.20	1.03	0.05	MFCC14	-0.04	0.0227	-0.16	1.13	0.08
MFCC15	-0.02	0.0164	-0.09	0.53	0.02	MFCC15	-0.03	0.0185	0.10	1.44	0.03
MFCC16	-0.05	0.0178	-0.19	0.77	0.02	MFCC16	-0.07	0.0173	-0.14	0.57	0.01
MFCC17	-0.03	0.0141	-0.07	0.48	0.01	MFCC17	-0.05	0.0171	-0.15	0.42	0.00
MFCC18	-0.03	0.0133	-0.06	0.39	0.00	MFCC18	-0.03	0.0149	-0.15	0.65	0.01
MFCC19	-0.03	0.0132	-0.12	0.27	0.01	MFCC19	-0.01	0.0135	-0.15	0.21	0.00
MFCC20	-0.04	0.0112	0.06	0.31	0.01	MFCC20	-0.04	0.0126	-0.04	0.26	0.01
MFCC21	-0.02	0.0107	0.03	0.31	0.02	MFCC21	-0.02	0.0116	0.15	0.30	0.02
MFCC22	-0.02	0.0101	0.07	0.36	0.01	MFCC22	-0.04	0.0096	-0.02	0.25	0.00
MFCC23	-0.01	0.0080	0.05	0.29	0.01	MFCC23	-0.01	0.0087	0.06	0.27	0.00
MFCC24	-0.02	0.0075	-0.10	0.40	0.01	MFCC24	-0.02	0.0078	-0.02	0.29	0.02
Mean	0.12	0.1973	-0.02	0.76	0.05	Mean	0.14	0.1791	-0.12	0.74	0.06
Min	-0.06	0.0075	-0.93	-0.22	0.00	Min	-0.07	0.0078	-0.98	-0.24	0.00
Max	1.71	3.4505	1.02	2.60	0.29	Max	2.47	3.0434	0.76	2.49	0.33

Table A7: The statistical properties of the HSM vocoder parameters for the "Neutral" databases.

HSM	**Male**	**Neutral**				**HSM**	**Female**	**Neutral**			
	Mean	**Var**	**Skew**	**Kurt**	**Neg**		**Mean**	**Var**	**Skew**	**Kurt**	**Neg**
F0	94.95	417.1892	0.96	1.90	0.24	F0	156.20	1071.7951	0.82	1.35	0.15
LSF H				**Cov.d**	**0.44**	**LSF H**				**Cov.d**	**0.43**
LSF1	0.07	0.0005	1.05	1.05	0.08	LSF1	0.08	0.0003	1.00	3.01	0.24
LSF2	0.13	0.0013	0.48	-0.16	0.01	LSF2	0.13	0.0008	0.76	0.85	0.02
LSF3	0.21	0.0010	0.33	0.28	0.00	LSF3	0.21	0.0011	0.60	0.46	0.00
LSF4	0.29	0.0013	0.73	1.93	0.19	LSF4	0.28	0.0015	0.92	1.95	0.18
LSF5	0.43	0.0014	0.28	0.74	0.07	LSF5	0.41	0.0012	0.70	1.94	0.16
LSF6	0.53	0.0020	-0.20	0.34	0.03	LSF6	0.52	0.0012	0.19	1.43	0.21
LSF7	0.63	0.0021	0.20	0.30	0.02	LSF7	0.63	0.0017	-0.16	1.10	0.19
LSF8	0.72	0.0016	0.14	0.59	0.06	LSF8	0.72	0.0021	0.11	0.72	0.04
LSF9	0.84	0.0011	0.40	1.45	0.15	LSF9	0.83	0.0018	0.17	1.12	0.18
LSF10	0.95	0.0012	0.43	1.28	0.08	LSF10	0.94	0.0012	0.62	1.60	0.21
LSF11	1.05	0.0014	0.40	1.23	0.07	LSF11	1.06	0.0010	0.50	1.03	0.12
LSF12	1.16	0.0014	0.32	1.01	0.04	LSF12	1.15	0.0011	0.26	0.84	0.08
LSF13	1.27	0.0013	0.32	1.41	0.08	LSF13	1.26	0.0012	0.18	0.83	0.07
LSF14	1.37	0.0016	0.29	0.66	0.03	LSF14	1.36	0.0012	0.31	1.11	0.14
LSF15	1.46	0.0022	-0.24	0.66	0.06	LSF15	1.48	0.0012	0.30	1.10	0.18
LSF16	1.57	0.0025	-0.37	0.79	0.06	LSF16	1.59	0.0011	0.34	0.88	0.05
LSF17	1.68	0.0023	-0.37	1.39	0.17	LSF17	1.69	0.0012	0.41	0.68	0.05
LSF18	1.79	0.0019	-0.01	2.00	0.12	LSF18	1.79	0.0013	0.19	0.40	0.04
LSF19	1.90	0.0016	0.22	2.26	0.05	LSF19	1.90	0.0013	0.02	0.60	0.01
LSF20	2.01	0.0015	0.45	2.15	0.05	LSF20	2.01	0.0013	0.09	0.74	0.01
LSF21	2.13	0.0015	0.66	2.35	0.02	LSF21	2.12	0.0013	0.09	0.32	0.03
LSF22	2.24	0.0015	0.62	1.74	0.01	LSF22	2.23	0.0013	0.02	0.11	0.00
LSF23	2.35	0.0016	0.49	0.98	0.00	LSF23	2.34	0.0014	0.02	0.07	0.00
LSF24	2.46	0.0015	0.34	0.53	0.01	LSF24	2.45	0.0017	0.08	0.01	0.01
LSF25	2.57	0.0014	0.15	0.36	0.00	LSF25	2.57	0.0017	-0.10	0.02	0.02
LSF26	2.67	0.0013	0.14	0.18	0.00	LSF26	2.68	0.0014	-0.25	0.16	0.02
LSF27	2.78	0.0014	0.04	0.02	0.02	LSF27	2.79	0.0013	-0.16	-0.36	0.01
LSF28	2.89	0.0014	-0.13	0.17	0.01	LSF28	2.90	0.0011	-0.37	-0.17	0.01
LSF29	3.00	0.0012	-0.28	-0.03	0.01	LSF29	3.01	0.0009	-0.36	-0.06	0.00
LSF30	3.09	0.0004	-1.26	2.17	0.08	LSF30	3.10	0.0004	-1.08	1.24	0.05
Mean	1.54	0.0015	0.19	0.99	0.05	Mean	1.54	0.0012	0.18	0.79	0.08
Min	0.07	0.0004	-1.26	-0.16	0.00	Min	0.08	0.0003	-1.08	-0.36	0.00
Max	3.09	0.0025	1.05	2.35	0.19	Max	3.10	0.0021	1.00	3.01	0.24
LSF S V				**Cov.d**	**0.55**	**LSF S V**				**Cov.d**	**0.60**
LSF1	0.21	0.0019	0.45	0.60	0.03	LSF1	0.23	0.0025	0.50	0.01	0.01
LSF2	0.32	0.0043	0.33	-0.15	0.00	LSF2	0.32	0.0040	0.23	-0.28	0.01
LSF3	0.52	0.0044	-0.18	0.02	0.00	LSF3	0.51	0.0038	0.04	-0.13	0.00
LSF4	0.61	0.0053	-0.10	-0.19	0.00	LSF4	0.63	0.0061	-0.10	-0.21	0.00
LSF5	0.76	0.0041	-0.03	-0.07	0.00	LSF5	0.76	0.0070	0.23	-0.26	0.01
LSF6	0.90	0.0033	0.00	0.09	0.00	LSF6	0.90	0.0053	-0.01	-0.01	0.00
LSF7	1.04	0.0035	0.13	-0.02	0.01	LSF7	1.06	0.0031	-0.06	0.18	0.02
LSF8	1.18	0.0045	0.02	-0.31	0.02	LSF8	1.18	0.0035	0.10	-0.06	0.01
LSF9	1.33	0.0044	-0.02	-0.14	0.00	LSF9	1.32	0.0041	0.17	-0.03	0.01
LSF10	1.46	0.0064	-0.09	-0.55	0.04	LSF10	1.48	0.0047	-0.21	-0.10	0.00
LSF11	1.61	0.0089	-0.39	-0.38	0.04	LSF11	1.65	0.0038	-0.30	0.23	0.02

Table A7: The statistical properties of the HSM vocoder parameters for the "Neutral" databases.

HSM	Male Mean	Neutral Var	Skew	Kurt	Neg
LSF12	1.77	0.0095	-0.53	0.09	0.00
LSF13	1.96	0.0072	-0.47	0.78	0.03
LSF14	2.15	0.0055	-0.31	0.90	0.04
LSF15	2.33	0.0043	-0.23	0.77	0.02
LSF16	2.48	0.0036	-0.25	0.35	0.01
LSF17	2.63	0.0034	-0.44	0.68	0.07
LSF18	2.78	0.0035	-0.26	0.38	0.00
LSF19	2.95	0.0034	-0.49	0.26	0.00
LSF20	3.08	0.0010	-1.34	2.58	0.13
Mean	1.60	0.0046	-0.21	0.28	0.02
Min	0.21	0.0010	-1.34	-0.55	0.00
Max	3.08	0.0095	0.45	2.58	0.13
LSF S UV				**Cov.d**	**0.50**
LSF1	0.10	0.0022	1.47	3.50	0.15
LSF2	0.24	0.0032	0.82	1.75	0.07
LSF3	0.42	0.0031	0.98	2.37	0.35
LSF4	0.58	0.0026	0.69	1.07	0.09
LSF5	0.73	0.0028	0.85	2.66	0.32
LSF6	0.89	0.0035	0.92	2.48	0.41
LSF7	1.05	0.0031	0.97	2.47	0.56
LSF8	1.21	0.0032	0.67	2.21	0.35
LSF9	1.37	0.0028	0.09	1.97	0.43
LSF10	1.52	0.0027	-0.24	2.26	0.48
LSF11	1.68	0.0027	-0.65	4.51	0.63
LSF12	1.84	0.0023	-1.04	6.05	0.40
LSF13	2.00	0.0021	-0.90	3.98	0.33
LSF14	2.16	0.0020	-0.92	3.01	0.47
LSF15	2.32	0.0014	-0.42	2.69	0.24
LSF16	2.48	0.0011	-0.32	2.09	0.15
LSF17	2.64	0.0010	-0.47	1.96	0.17
LSF18	2.79	0.0009	-0.66	2.00	0.12
LSF19	2.95	0.0007	-0.73	1.85	0.06
LSF20	3.09	0.0004	-1.23	2.84	0.05
Mean	1.60	0.0022	-0.01	2.69	0.29
Min	0.10	0.0004	-1.23	1.07	0.05
Max	3.09	0.0035	1.47	6.05	0.63

HSM	Female Mean	Neutral Var	Skew	Kurt	Neg
LSF12	1.79	0.0042	-0.07	-0.36	0.03
LSF13	1.95	0.0051	-0.21	-0.23	0.01
LSF14	2.11	0.0054	-0.36	-0.08	0.02
LSF15	2.28	0.0061	-0.14	0.04	0.00
LSF16	2.46	0.0071	-0.32	-0.08	0.00
LSF17	2.63	0.0049	-0.78	0.98	0.10
LSF18	2.79	0.0037	-0.72	1.50	0.05
LSF19	2.97	0.0024	-0.56	1.08	0.05
LSF20	3.08	0.0007	-1.09	1.71	0.02
Mean	1.60	0.0044	-0.18	0.20	0.02
Min	0.23	0.0007	-1.09	-0.36	0.00
Max	3.08	0.0071	0.50	1.71	0.10
LSF S UV				**Cov.d**	**0.53**
LSF1	0.11	0.0020	1.71	4.59	0.41
LSF2	0.25	0.0030	0.39	1.45	0.08
LSF3	0.42	0.0026	0.74	2.23	0.35
LSF4	0.58	0.0027	0.92	1.61	0.20
LSF5	0.73	0.0024	0.85	2.26	0.30
LSF6	0.88	0.0030	0.68	2.17	0.31
LSF7	1.05	0.0027	0.57	2.27	0.45
LSF8	1.21	0.0022	0.92	3.61	0.46
LSF9	1.37	0.0024	0.84	3.98	0.80
LSF10	1.53	0.0022	0.52	3.34	0.73
LSF11	1.69	0.0016	0.04	2.42	0.37
LSF12	1.84	0.0023	-0.79	1.99	0.41
LSF13	2.00	0.0024	-1.38	3.76	0.87
LSF14	2.16	0.0016	-1.10	2.97	0.38
LSF15	2.32	0.0017	-1.06	2.85	0.48
LSF16	2.48	0.0018	-1.37	4.30	0.97
LSF17	2.64	0.0012	-1.14	3.81	0.35
LSF18	2.79	0.0009	-0.67	2.02	0.15
LSF19	2.95	0.0007	-0.65	1.69	0.09
LSF20	3.09	0.0004	-1.11	1.72	0.02
Mean	1.60	0.0020	-0.05	2.75	0.41
Min	0.11	0.0004	-1.38	1.45	0.02
Max	3.09	0.0030	1.71	4.59	0.97

Table A8: The statistical properties of the HSM vocoder parameters for the "Sad" databases.

HSM	Male Mean	Sad Var	Skew	Kurt	Neg
F0	92.84	427.3991	0.31	0.23	0.04
LSF H				**Cov.d**	**0.46**
LSF1	0.07	0.0005	1.13	1.66	0.10
LSF2	0.13	0.0012	0.51	-0.22	0.00
LSF3	0.21	0.0011	0.42	0.61	0.00
LSF4	0.30	0.0015	0.69	1.32	0.04
LSF5	0.44	0.0014	0.24	0.59	0.06
LSF6	0.54	0.0021	-0.19	0.27	0.01
LSF7	0.63	0.0020	0.24	0.36	0.01
LSF8	0.73	0.0016	0.34	0.55	0.04
LSF9	0.85	0.0012	0.40	1.05	0.06
LSF10	0.96	0.0013	0.31	0.89	0.03
LSF11	1.05	0.0015	0.26	0.91	0.02
LSF12	1.16	0.0015	0.33	0.80	0.01
LSF13	1.28	0.0014	0.28	0.92	0.03
LSF14	1.37	0.0018	0.16	0.45	0.01
LSF15	1.48	0.0023	-0.15	0.28	0.04
LSF16	1.59	0.0023	-0.30	0.73	0.05
LSF17	1.69	0.0019	-0.17	1.49	0.10
LSF18	1.80	0.0018	0.13	1.63	0.06
LSF19	1.91	0.0018	0.21	2.42	0.08
LSF20	2.02	0.0018	0.36	2.27	0.05
LSF21	2.14	0.0017	0.46	2.77	0.06
LSF22	2.24	0.0016	0.40	2.75	0.09
LSF23	2.36	0.0017	0.25	1.90	0.07
LSF24	2.47	0.0015	0.15	0.81	0.02
LSF25	2.58	0.0014	0.12	0.71	0.00
LSF26	2.69	0.0013	-0.09	0.32	0.00
LSF27	2.79	0.0013	-0.23	0.60	0.00
LSF28	2.90	0.0011	-0.25	0.73	0.02
LSF29	3.01	0.0010	-0.33	0.17	0.00
LSF30	3.10	0.0004	-1.38	3.15	0.14
Mean	1.55	0.0015	0.14	1.10	0.04
Min	0.07	0.0004	-1.38	-0.22	0.00
Max	3.10	0.0023	1.13	3.15	0.14
LSF S V				**Cov.d**	**0.59**
LSF1	0.21	0.0019	0.51	0.64	0.01
LSF2	0.33	0.0044	0.19	-0.35	0.04
LSF3	0.53	0.0046	-0.25	0.01	0.00
LSF4	0.62	0.0051	-0.16	-0.03	0.01
LSF5	0.78	0.0036	-0.08	-0.03	0.00
LSF6	0.92	0.0031	-0.08	0.16	0.00
LSF7	1.05	0.0035	0.01	0.02	0.00
LSF8	1.19	0.0043	-0.09	-0.08	0.01
LSF9	1.35	0.0046	-0.10	-0.20	0.00

HSM	Female Mean	Sad Var	Skew	Kurt	Neg
F0	169.59	4159.1191	0.94	0.00	0.00
LSF H				**Cov.d**	**0.42**
LSF1	0.08	0.0005	0.61	0.30	0.00
LSF2	0.13	0.0010	0.69	0.04	0.02
LSF3	0.21	0.0012	0.55	0.46	0.00
LSF4	0.30	0.0018	0.72	0.88	0.05
LSF5	0.43	0.0012	0.60	1.10	0.06
LSF6	0.53	0.0016	0.02	0.42	0.03
LSF7	0.64	0.0020	-0.11	0.11	0.00
LSF8	0.73	0.0020	0.24	0.42	0.01
LSF9	0.84	0.0016	0.24	0.49	0.02
LSF10	0.96	0.0012	0.35	0.50	0.02
LSF11	1.07	0.0011	0.30	0.32	0.01
LSF12	1.16	0.0013	0.24	0.59	0.02
LSF13	1.27	0.0014	0.42	0.86	0.03
LSF14	1.38	0.0014	0.21	0.57	0.03
LSF15	1.49	0.0014	0.12	0.31	0.00
LSF16	1.59	0.0016	0.07	-0.11	0.01
LSF17	1.69	0.0018	-0.12	0.00	0.00
LSF18	1.80	0.0019	-0.27	0.22	0.00
LSF19	1.91	0.0019	-0.40	0.39	0.01
LSF20	2.02	0.0017	-0.34	0.48	0.01
LSF21	2.13	0.0016	-0.19	0.05	0.02
LSF22	2.24	0.0014	-0.19	0.12	0.01
LSF23	2.35	0.0014	-0.12	-0.02	0.05
LSF24	2.46	0.0014	-0.03	-0.27	0.02
LSF25	2.57	0.0012	-0.14	-0.37	0.03
LSF26	2.68	0.0012	-0.20	-0.14	0.00
LSF27	2.80	0.0011	-0.32	0.04	0.01
LSF28	2.91	0.0010	-0.47	0.14	0.03
LSF29	3.01	0.0008	-0.48	0.39	0.03
LSF30	3.10	0.0003	-1.18	1.65	0.03
Mean	1.55	0.0014	0.03	0.33	0.02
Min	0.08	0.0003	-1.18	-0.37	0.00
Max	3.10	0.0020	0.72	1.65	0.06
LSF S V				**Cov.d**	**0.61**
LSF1	0.23	0.0026	0.51	0.11	0.01
LSF2	0.34	0.0045	0.05	-0.38	0.03
LSF3	0.53	0.0043	-0.03	-0.12	0.00
LSF4	0.64	0.0061	-0.11	-0.19	0.00
LSF5	0.77	0.0052	0.03	-0.20	0.00
LSF6	0.92	0.0037	-0.22	0.06	0.00
LSF7	1.07	0.0027	-0.10	0.15	0.01
LSF8	1.18	0.0035	0.09	0.07	0.00
LSF9	1.34	0.0041	-0.02	-0.12	0.02

Table A8: The statistical properties of the HSM vocoder parameters for the "Sad" databases.

HSM	**Male**	**Sad**				**HSM**	**Female**	**Sad**			
	Mean	**Var**	**Skew**	**Kurt**	**Neg**		**Mean**	**Var**	**Skew**	**Kurt**	**Neg**
LSF10	1.49	0.0072	-0.31	-0.58	0.07	LSF10	1.50	0.0044	-0.29	-0.11	0.01
LSF11	1.64	0.0072	-0.64	0.10	0.00	LSF11	1.65	0.0045	-0.31	-0.28	0.04
LSF12	1.80	0.0068	-0.51	0.72	0.02	LSF12	1.79	0.0057	-0.36	-0.33	0.04
LSF13	1.98	0.0064	-0.34	0.96	0.02	LSF13	1.95	0.0069	-0.57	-0.15	0.00
LSF14	2.15	0.0054	-0.33	1.12	0.07	LSF14	2.12	0.0061	-0.77	0.43	0.00
LSF15	2.32	0.0048	-0.45	1.53	0.16	LSF15	2.30	0.0051	-0.68	0.78	0.02
LSF16	2.49	0.0041	-0.49	1.54	0.12	LSF16	2.47	0.0043	-0.68	0.99	0.02
LSF17	2.65	0.0030	-0.33	0.83	0.07	LSF17	2.64	0.0030	-0.59	1.50	0.09
LSF18	2.80	0.0029	-0.42	0.87	0.07	LSF18	2.80	0.0029	-0.39	0.69	0.03
LSF19	2.96	0.0028	-0.41	0.43	0.01	LSF19	2.97	0.0023	-0.52	0.77	0.05
LSF20	3.08	0.0009	-1.34	2.41	0.09	LSF20	3.08	0.0007	-1.22	2.38	0.09
Mean	1.62	0.0043	-0.28	0.50	0.04	Mean	1.61	0.0041	-0.31	0.30	0.02
Min	0.21	0.0009	-1.34	-0.58	0.00	Min	0.23	0.0007	-1.22	-0.38	0.00
Max	3.08	0.0072	0.51	2.41	0.16	Max	3.08	0.0069	0.51	2.38	0.09
LSF S UV				**Cov.d**	**0.53**	**LSF S UV**				**Cov.d**	**0.59**
LSF1	0.10	0.0018	1.74	5.09	0.34	LSF1	0.11	0.0018	1.84	5.58	0.52
LSF2	0.25	0.0023	1.09	3.64	0.36	LSF2	0.26	0.0021	0.70	2.68	0.17
LSF3	0.42	0.0025	1.28	3.61	0.60	LSF3	0.42	0.0019	1.14	4.40	0.82
LSF4	0.57	0.0020	0.97	2.10	0.17	LSF4	0.57	0.0020	1.07	2.45	0.25
LSF5	0.73	0.0023	0.45	2.57	0.36	LSF5	0.73	0.0019	0.48	2.81	0.42
LSF6	0.88	0.0026	0.47	2.95	0.48	LSF6	0.88	0.0024	0.24	2.88	0.51
LSF7	1.05	0.0026	0.52	2.93	0.39	LSF7	1.05	0.0018	0.26	2.98	0.45
LSF8	1.21	0.0025	0.09	2.98	0.68	LSF8	1.20	0.0017	0.32	4.12	0.57
LSF9	1.37	0.0018	0.18	2.71	0.38	LSF9	1.37	0.0021	-0.21	3.59	0.62
LSF10	1.53	0.0019	-0.26	2.68	0.38	LSF10	1.53	0.0016	-0.15	3.32	0.47
LSF11	1.68	0.0020	-0.74	3.49	0.55	LSF11	1.69	0.0016	-0.77	2.62	0.37
LSF12	1.84	0.0020	-0.82	3.72	0.50	LSF12	1.84	0.0017	-1.08	2.58	0.32
LSF13	2.00	0.0021	-1.33	4.66	0.79	LSF13	2.00	0.0016	-1.34	4.01	0.50
LSF14	2.16	0.0017	-1.40	5.32	0.75	LSF14	2.16	0.0015	-1.36	4.30	0.51
LSF15	2.32	0.0015	-1.03	4.10	0.57	LSF15	2.32	0.0014	-1.25	4.74	0.35
LSF16	2.48	0.0010	-0.59	3.63	0.31	LSF16	2.48	0.0011	-0.91	4.59	0.32
LSF17	2.64	0.0008	-0.45	2.27	0.13	LSF17	2.64	0.0008	-0.74	3.24	0.18
LSF18	2.79	0.0008	-0.76	3.02	0.11	LSF18	2.80	0.0007	-0.74	3.57	0.22
LSF19	2.95	0.0007	-0.93	3.32	0.11	LSF19	2.95	0.0006	-0.60	2.04	0.05
LSF20	3.09	0.0004	-1.30	3.46	0.09	LSF20	3.09	0.0004	-1.09	1.70	0.04
Mean	1.60	0.0018	-0.14	3.41	0.40	Mean	1.60	0.0015	-0.21	3.41	0.38
Min	0.10	0.0004	-1.40	2.10	0.09	Min	0.11	0.0004	-1.36	1.70	0.04
Max	3.09	0.0026	1.74	5.32	0.79	Max	3.09	0.0024	1.84	5.58	0.82

Table A9: The statistical properties of the HSM vocoder parameters for the "Angry" databases.

HSM	**Male**	**Angry**				**HSM**	**Female**	**Angry**			
	Mean	**Var**	**Skew**	**Kurt**	**Neg**		**Mean**	**Var**	**Skew**	**Kurt**	**Neg**
F0	119.43	1408.0007	0.91	0.39	0.00	F0	165.51	1767.9013	0.87	0.89	0.04
LSF H				**Cov.d**	**0.50**	**LSF H**				**Cov.d**	**0.47**
LSF1	0.08	0.0007	0.67	0.12	0.00	LSF1	0.09	0.0006	0.89	1.43	0.09
LSF2	0.14	0.0015	0.42	-0.36	0.03	LSF2	0.15	0.0014	0.65	0.01	0.01
LSF3	0.21	0.0013	0.52	0.29	0.00	LSF3	0.22	0.0017	0.72	0.39	0.00
LSF4	0.30	0.0013	0.75	2.00	0.16	LSF4	0.29	0.0018	0.72	0.80	0.05
LSF5	0.44	0.0016	0.16	0.12	0.00	LSF5	0.42	0.0014	0.47	0.85	0.07
LSF6	0.54	0.0025	-0.13	-0.25	0.00	LSF6	0.53	0.0018	-0.01	0.24	0.00
LSF7	0.63	0.0023	0.29	0.05	0.00	LSF7	0.63	0.0024	0.01	-0.11	0.00
LSF8	0.73	0.0020	0.22	0.28	0.01	LSF8	0.73	0.0027	0.26	0.11	0.02
LSF9	0.85	0.0013	0.24	0.64	0.06	LSF9	0.84	0.0022	0.45	0.73	0.05
LSF10	0.95	0.0014	0.38	0.80	0.01	LSF10	0.95	0.0016	0.41	0.73	0.04
LSF11	1.04	0.0015	0.21	0.58	0.02	LSF11	1.06	0.0014	0.30	0.41	0.02
LSF12	1.15	0.0016	0.22	0.42	0.02	LSF12	1.16	0.0015	0.26	0.77	0.07
LSF13	1.26	0.0018	0.07	0.18	0.01	LSF13	1.26	0.0016	0.21	0.46	0.03
LSF14	1.36	0.0022	-0.05	0.09	0.01	LSF14	1.36	0.0017	0.10	0.26	0.03
LSF15	1.47	0.0024	-0.19	0.08	0.00	LSF15	1.48	0.0015	0.01	0.50	0.04
LSF16	1.57	0.0022	-0.36	0.44	0.03	LSF16	1.58	0.0014	0.09	0.66	0.02
LSF17	1.68	0.0018	-0.40	1.09	0.06	LSF17	1.68	0.0016	0.02	0.32	0.01
LSF18	1.79	0.0017	-0.21	0.77	0.04	LSF18	1.78	0.0018	-0.03	0.09	0.01
LSF19	1.90	0.0016	-0.06	0.73	0.05	LSF19	1.89	0.0022	-0.31	0.15	0.00
LSF20	2.02	0.0015	0.10	0.54	0.00	LSF20	2.00	0.0021	-0.40	0.59	0.06
LSF21	2.14	0.0013	0.12	0.06	0.01	LSF21	2.12	0.0019	-0.28	0.79	0.03
LSF22	2.24	0.0012	0.16	0.33	0.00	LSF22	2.23	0.0015	-0.12	0.45	0.01
LSF23	2.36	0.0013	0.10	-0.07	0.00	LSF23	2.34	0.0015	0.12	0.38	0.00
LSF24	2.46	0.0013	0.00	-0.04	0.01	LSF24	2.46	0.0015	0.14	0.14	0.00
LSF25	2.57	0.0014	-0.08	0.00	0.01	LSF25	2.57	0.0013	0.12	-0.04	0.01
LSF26	2.68	0.0015	-0.24	-0.10	0.00	LSF26	2.68	0.0011	-0.12	-0.24	0.00
LSF27	2.79	0.0015	-0.47	0.22	0.00	LSF27	2.79	0.0011	-0.25	-0.13	0.01
LSF28	2.90	0.0013	-0.57	0.72	0.02	LSF28	2.90	0.0010	-0.42	0.25	0.00
LSF29	3.01	0.0012	-0.53	0.38	0.00	LSF29	3.01	0.0008	-0.52	0.42	0.02
LSF30	3.09	0.0004	-1.18	1.56	0.06	LSF30	3.10	0.0003	-1.22	1.91	0.03
Mean	1.55	0.0016	0.01	0.39	0.02	Mean	1.54	0.0015	0.08	0.44	0.02
Min	0.08	0.0004	-1.18	-0.36	0.00	Min	0.09	0.0003	-1.22	-0.24	0.00
Max	3.09	0.0025	0.75	2.00	0.16	Max	3.10	0.0027	0.89	1.91	0.09
LSF S V				**Cov.d**	**0.62**	**LSF S V**				**Cov.d**	**0.59**
LSF1	0.22	0.0024	0.51	0.49	0.00	LSF1	0.23	0.0028	0.48	-0.04	0.00
LSF2	0.34	0.0047	0.35	-0.14	0.00	LSF2	0.33	0.0045	0.25	-0.22	0.00
LSF3	0.53	0.0054	-0.09	-0.28	0.00	LSF3	0.52	0.0046	0.06	-0.30	0.01
LSF4	0.62	0.0062	0.04	-0.30	0.00	LSF4	0.63	0.0076	0.06	-0.44	0.01
LSF5	0.78	0.0048	-0.10	-0.15	0.00	LSF5	0.77	0.0075	0.29	-0.19	0.02
LSF6	0.92	0.0034	0.00	0.15	0.00	LSF6	0.91	0.0056	0.03	-0.11	0.00
LSF7	1.04	0.0034	0.17	-0.11	0.00	LSF7	1.07	0.0038	0.14	0.05	0.00

Table A9: The statistical properties of the HSM vocoder parameters for the "Angry" databases.

HSM	**Male**	**Angry**				**HSM**	**Female**	**Angry**			
	Mean	**Var**	**Skew**	**Kurt**	**Neg**		**Mean**	**Var**	**Skew**	**Kurt**	**Neg**
LSF8	1.16	0.0048	0.13	-0.27	0.00	LSF8	1.18	0.0040	0.18	-0.08	0.00
LSF9	1.32	0.0061	0.03	-0.48	0.03	LSF9	1.32	0.0049	0.11	-0.34	0.01
LSF10	1.45	0.0077	-0.13	-0.77	0.16	LSF10	1.47	0.0051	-0.14	-0.15	0.00
LSF11	1.61	0.0076	-0.39	-0.44	0.05	LSF11	1.63	0.0047	-0.11	0.13	0.00
LSF12	1.78	0.0073	-0.34	-0.01	0.00	LSF12	1.76	0.0057	0.12	-0.26	0.02
LSF13	1.97	0.0069	-0.38	0.13	0.00	LSF13	1.92	0.0080	-0.10	-0.66	0.11
LSF14	2.15	0.0047	-0.50	0.72	0.02	LSF14	2.08	0.0087	-0.41	-0.30	0.02
LSF15	2.32	0.0037	-0.26	0.61	0.01	LSF15	2.27	0.0072	-0.35	0.16	0.00
LSF16	2.48	0.0036	-0.22	0.25	0.00	LSF16	2.46	0.0059	-0.43	0.46	0.01
LSF17	2.64	0.0036	-0.35	0.33	0.01	LSF17	2.63	0.0036	-0.45	0.72	0.02
LSF18	2.79	0.0035	-0.53	0.65	0.05	LSF18	2.79	0.0028	-0.29	0.53	0.04
LSF19	2.95	0.0034	-0.61	0.61	0.01	LSF19	2.97	0.0024	-0.52	0.56	0.01
LSF20	3.08	0.0010	-1.36	2.47	0.15	LSF20	3.08	0.0007	-1.27	2.71	0.11
Mean	1.61	0.0047	-0.20	0.17	0.03	Mean	1.60	0.0050	-0.12	0.11	0.02
Min	0.22	0.0010	-1.36	-0.77	0.00	Min	0.23	0.0007	-1.27	-0.66	0.00
Max	3.08	0.0077	0.51	2.47	0.16	Max	3.08	0.0087	0.48	2.71	0.11
LSF S UV				**Cov.d**	**0.46**	**LSF S UV**				**Cov.d**	**0.49**
LSF1	0.12	0.0023	1.27	3.19	0.22	LSF1	0.12	0.0026	1.24	2.25	0.12
LSF2	0.24	0.0033	0.73	2.17	0.28	LSF2	0.25	0.0038	0.55	0.69	0.01
LSF3	0.43	0.0038	0.65	2.07	0.32	LSF3	0.42	0.0041	0.36	0.97	0.16
LSF4	0.59	0.0028	0.19	1.49	0.13	LSF4	0.57	0.0038	0.60	0.79	0.02
LSF5	0.73	0.0036	0.05	0.74	0.02	LSF5	0.72	0.0040	0.31	0.75	0.05
LSF6	0.89	0.0040	0.08	0.61	0.02	LSF6	0.88	0.0040	0.04	1.14	0.06
LSF7	1.05	0.0035	0.20	0.97	0.08	LSF7	1.04	0.0035	0.29	0.95	0.08
LSF8	1.20	0.0041	-0.04	1.15	0.20	LSF8	1.19	0.0036	0.15	1.17	0.12
LSF9	1.36	0.0043	-0.70	1.83	0.35	LSF9	1.35	0.0039	-0.26	1.72	0.22
LSF10	1.52	0.0048	-1.35	3.22	0.61	LSF10	1.51	0.0037	-0.47	1.72	0.20
LSF11	1.67	0.0049	-1.51	3.95	0.57	LSF11	1.67	0.0033	-0.72	1.54	0.20
LSF12	1.84	0.0040	-1.30	3.51	0.32	LSF12	1.82	0.0042	-1.01	1.80	0.18
LSF13	2.00	0.0029	-1.03	2.00	0.25	LSF13	1.98	0.0051	-1.48	3.36	0.57
LSF14	2.16	0.0022	-1.00	2.47	0.25	LSF14	2.14	0.0042	-1.76	4.86	0.81
LSF15	2.32	0.0017	-0.58	1.89	0.17	LSF15	2.31	0.0031	-1.35	3.48	0.34
LSF16	2.48	0.0014	-0.64	1.95	0.20	LSF16	2.47	0.0025	-1.20	3.95	0.49
LSF17	2.64	0.0011	-0.68	2.07	0.15	LSF17	2.63	0.0016	-1.01	3.25	0.27
LSF18	2.80	0.0010	-0.68	2.05	0.09	LSF18	2.79	0.0012	-0.64	1.49	0.09
LSF19	2.96	0.0008	-0.76	1.52	0.04	LSF19	2.95	0.0009	-0.79	2.42	0.10
LSF20	3.09	0.0004	-1.28	2.58	0.10	LSF20	3.09	0.0004	-1.22	2.45	0.10
Mean	1.60	0.0028	-0.42	2.07	0.22	Mean	1.59	0.0032	-0.42	2.04	0.21
Min	0.12	0.0004	-1.51	0.61	0.02	Min	0.12	0.0004	-1.76	0.69	0.01
Max	3.09	0.0049	1.27	3.95	0.61	Max	3.09	0.0051	1.24	4.86	0.81