HELSINKI UNIVERSITY OF TECHNOLOGY
Faculty of Electronics, Communications and Automation
Department of Signal Processing and Acoustics

Tuomo Raitio

Hidden Markov Model Based Finnish Text-to-Speech System Utilizing Glottal Inverse Filtering

Master's Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology.

Espoo, May 30, 2008

Supervisor: Professor Paavo Alku

HELSINKI UNIVERSITY OF TECHNOLOGY

ABSTRACT OF THE MASTER'S THESIS

Author:	Tuomo Raitio
Name of the thesis:	Hidden Markov Model Based Finnish Text-to-Speech System Utilizing Glottal Inverse Filtering
Date:	May 30, 2008 **Number of pages:** 89 + 5
Faculty:	Electronics, Communications and Automation
Professorship:	S-89
Supervisor:	Prof. Paavo Alku

In this work, a new hidden Markov model (HMM) based text-to-speech (TTS) system utilizing glottal inverse filtering is described. The primary goal of the new TTS system is to enable producing natural sounding synthetic speech in different speaking styles with different speaker characteristics and emotions. In order to achieve these goals, the function of the real human voice production mechanism is modeled with the help of glottal inverse filtering embedded in a statistical framework of HMM.

The new TTS system uses a glottal inverse filtering based parametrization method that enables the extraction of voice source characteristics separate from other speech parameters, and thus the individual modeling of these characteristics in the HMM system. In the synthesis stage, natural glottal flow pulses are used for creating the voice source, and the voice source characteristics are further modified according to the adaptive all-pole model generated by the HMM system in order to imitate the natural variation in the real voice source.

Subjective listening tests show that the quality of the new TTS system is considerably better compared to a traditional HMM-based speech synthesizer. Moreover, the new system is clearly able to produce natural sounding synthetic speech with specific speaker characteristics.

Keywords: speech synthesis, synthetic speech, TTS, HMM, glottal inverse filtering

TEKNILLINEN KORKEAKOULU DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Tuomo Raitio
Työn nimi:	Äänilähteen käänteissuodatusta hyödyntävä Markovin piilomalleihin perustuva suomenkielinen puhesynteesijärjestelmä
Päivämäärä:	30.5.2008 **Sivuja:** 89 + 5
Tiedekunta:	Elektroniikka, tietoliikenne ja automaatio
Professuuri:	S-89
Työn valvoja:	Prof. Paavo Alku

Tässä työssä esitetään uusi Markovin piilomalleihin (hidden Markov model, HMM) perustuva äänilähteen käänteissuodatusta hyödyntävä suomenkielinen puhesynteesijärjestelmä. Uuden puhesynteesimenetelmän päätavoite on tuottaa luonnolliselta kuulostavaa synteettistä puhetta, jonka ominaisuuksia voidaan muuttaa eri puhujien, puhetyylien tai jopa äänen emootiosisällön mukaan. Näiden tavoitteiden mahdollistamiseksi uudessa puhesynteesimenetelmässä mallinnetaan ihmisen äänentuottojärjestelmää äänilähteen käänteissuodatuksen ja HMM-mallinnuksen avulla.

Uusi puhesynteesijärjestelmä hyödyntää äänilähteen käänteissuodatusmenetelmää, joka mahdollistaa äänilähteen ominaisuuksien parametrisoinnin erillään muista puheen parametreista, ja siten näiden parametrien mallintamisen erikseen HMM-järjestelmässä. Synteesivaiheessa luonnollisesta puheesta laskettuja glottispulsseja käytetään äänilähteen luomiseen, ja äänilähteen ominaisuuksia muokataan edelleen tilastollisen HMM-järjestelmän tuottaman parametrisen kuvauksen avulla, mikä imitoi oikeassa puheessa esiintyvää luonnollista äänilähteen ominaisuuksien vaihtelua.

Subjektiivisten kuuntelukokeiden tulokset osoittavat, että uuden puhesynteesimenetelmän laatu on huomattavasti parempi verrattuna perinteiseen HMM-pohjaiseen puhesynteesijärjestelmään. Lisäksi tulokset osoittavat, että uusi puhesynteesimenetelmä pystyy tuottamaan luonnolliselta kuulostavaa puhetta eri puhujien ominaisuuksilla.

Avainsanat: puhesynteesi, synteettinen puhe, TTS, HMM, äänilähteen käänteissuodatus

Acknowledgements

This Master's thesis has been carried out at the Department of Signal Processing and Acoustics at the Helsinki University of Technology. The work has also been contributed by the Department of Speech Sciences at the University of Helsinki.

Firstly, I would like to thank my supervisor professor Paavo Alku for providing me the opportunity to do this thesis on this special topic. His professional knowledge and ideas with encouraging feedback and motivation have been essential for this work. I would also like to thank docent Martti Vainio from the University of Helsinki for providing this productive collaboration. I am also grateful to Mr. Antti Suni from the University of Helsinki for the indispensable work with the HMM system and for the valuable comments for developing the speech synthesizer. I also wish to thank Mr. Hannu Pulakka for the help with the subjective listening tests and the analysis of the results.

Finally, I would like to thank all the people at the Laboratory of Acoustics and Audio Signal Processing for giving me the most enjoyable and inspiring working environment. I would also like to thank all the encouraging people who supported me during this work.

Otaniemi, May 30, 2008

Tuomo Raitio

Contents

Abbreviations

AC	Alternating Current
ACF	Autocorrelation Function
AMDF	Average Magnitude Difference Function
CCR	Comparison Category Rating
ClQ	Closing Quotient
CMOS	Comparison Mean Opinion Score
CQ	Closed Quotient
DAP	Discrete All-Pole
DC	Direct Current
DP	Dynamic Programming
ERB	Equivalent Rectangular Bandwidth
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
GIF	Glottal Inverse Filtering
GV	Global Variance
HMM	Hidden Markov Model
HRF	Harmonic Richness Factor
HSMM	Hidden Semi-Markov Model
IAIF	Iterative Adaptive Inverse Filtering
JND	Just Noticeable Difference
LP	Linear Prediction
LPC	Linear Predictive Coding
LSF	Line Spectral Frequency
LSP	Line Spectrum Pair
MSD-HMM	Multi-Space Probability Distribution Hidden Markov Model
NAQ	Normalized Amplitude Quotient
OQ	Open Quotient

PSP	Parabolic Spectral Parameter
PDA	Pitch Detection Algorithm
SQ	Speed Quotient
TTS	Text-to-Speech
VOT	Voice Onset Time
ZCR	Zero-Crossing Rate

List of Figures

List of Tables

Chapter 1

Introduction

The ultimate goal of text-to-speech (TTS) synthesis is to create natural sounding speech from arbitrary text. Moreover, the current trend in TTS research calls for systems that enable producing speech in different speaking styles with different speaker characteristics and even emotions. In order to fulfill these stringent general requirements, two major synthesis techniques have attracted increasing interest in the speech research community during the past decade. These two alternatives are the unit selection technique and the hidden Markov model (HMM) based approach. The former has been shown to yield synthetic speech of highly natural quality. However, unit selection techniques do not allow for easy adaptation of the TTS system to different speaking styles and speaker characteristics. In order to obtain various voice characteristics in text-to-speech systems based on the selection and concatenation of acoustical units, a large amount of speech data is required. It is difficult and laborious to collect and segment the speech units, and the implementation of the TTS system requires databases of extensive sizes, which severely limit the use of this TTS technique for example in handheld devices. HMM-based techniques, in turn, benefit from better adaptability and a clearly smaller memory requirement. However, the current HMM systems often suffer from degraded naturalness in quality. It can be argued that a potential reason for the reduced naturalness in the current HMM-based TTS systems can be explained by the use of signal generation techniques which are oversimplified to properly mimic natural speech pressure waveforms.

A large part of what can be characterized as naturalness in speech emerges from different voice characteristics as well as their context dependent changes. Therefore, it is justified in speech synthesis to search for methods aiming at accurate modeling of different voice characteristics as well as prosodic features of speech. Towards these goals, HMM-based synthesizers have been developed with special emphasis on voice characteristics such as speaker individualities, speaking styles, and emotions. Moreover, some recent studies have

introduced improved signal generation techniques for parametric HMM-based TTS systems that have been shown to improve the quality of synthetic speech compared to traditional methods. However, the quality of the TTS systems using these techniques still remains far from the quality of natural speech.

In the real human voice production mechanism, the excitation of voiced speech is represented by the glottal volume velocity waveform generated by the vibrating vocal folds. This excitation signal, the glottal source, has naturally attracted interest in speech synthesis, and many techniques have been proposed to mimic the glottal source of natural speech. Artificial models for the glottal source have been used in order to improve the quality of the synthesis. However, current models for the glottal source are oversimplified as well, and the resulting quality of the synthesis has not been satisfactory. To overcome the problems due to oversimplified glottal source models, the idea of utilizing glottal flow pulses extracted from natural speech with the help of glottal inverse filtering has been proposed. However, previous studies based on glottal flow pulses extracted from natural speech are limited to special purposes such as the generation of isolated vowels, and the benefits from combining automatic glottal inverse filtering with an HMM-based speech synthesizer have not been utilized.

The human voice production and especially the voice source has been an active research topic at the Department of Signal Processing and Acoustics at the Helsinki University of Technology. One particular outcome of the research has been the glottal inverse filtering method developed by professor Paavo Alku in the early 1990s. The developed method has been studied and verified to yield reasonable estimates of the glottal source, and it has been used at the department and by other researchers for estimating the glottal source. Also speech synthesis has been a topic of special interest among the people who nowadays work at the department, but the previous research has been focused mainly on formant synthesis based techniques, and recently the activity of the research on speech synthesis has been minimal.

Phonetics and linguistics have been widely studied at the Department of Speech Sciences at the University of Helsinki. Also speech technology has been an active research topic. Lately, an HMM-based speech synthesizer was adopted to study Finnish speech synthesis with a special emphasis on the modeling of Finnish prosody. The research with the synthesizer has been focused mostly on modeling the linguistic features of speech with the HMM system, but speech synthesis algorithms have not been widely studied.

Since the information about the voice source characteristics is considered valuable in HMM-based speech synthesis, a collaboration between the two departments was started in 2007. The objective was to use the glottal inverse filtering method to reveal the voice source characteristics, and utilize that information in HMM-based speech synthesis. Thus, a new

HMM-based speech synthesis system was created in co-operation with the two departments.

In this thesis, a new HMM-based speech synthesis system that utilizes glottal inverse filtering is presented. The new TTS system aims to produce natural sounding synthetic speech capable of conveying different styles of speaking as well as emotions. In order to achieve these goals, the function of the real human voice production apparatus is modeled with the help of glottal inverse filtering embedded in an statistical framework of HMM.

The thesis is organized as follows. Basic information about speech production, perception, and synthesis is presented in Chapter 2. The methods used in speech synthesis in general and especially the methods utilized in the new TTS system are presented in Chapter 3. The new HMM-based TTS system is presented and fully described in Chapter 4, and the evaluation of the constructed TTS system and the obtained results are described in Chapter 5. Discussion about the new synthesizer and the utilized methods with final conclusions are presented in Chapter 6.

Chapter 2

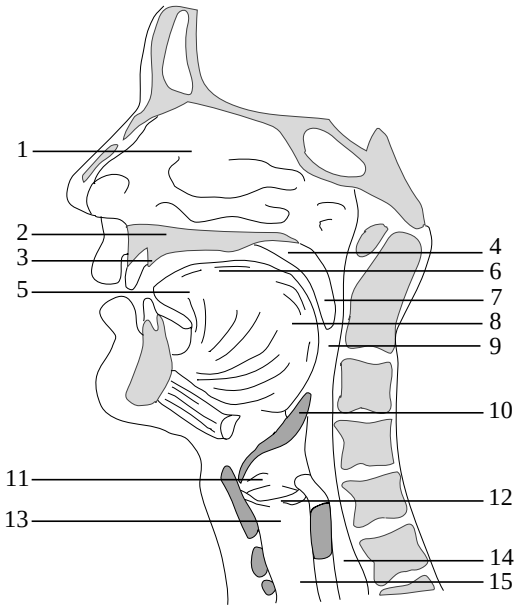
Background

This chapter describes speech production and perception from the perspective of speech synthesis, followed by a representation of the source-filter theory, a model that most speech synthesizers are based on. A general description of speech synthesis systems is given at the end of the chapter.

2.1 Fundamentals of Speech Production

Speech is produced by regulating the airflow from lungs through throat, nose and mouth. The air in the lungs is pressed upon chest and lung tissues, resulting in an airflow to trachea and larynx. At larynx the airflow is modulated by vocal folds, which creates the main excitation for voiced speech. Pharynx connects the larynx to oral and nasal cavities, which are collectively called the vocal tract. The volume and dimensions of the pharynx and oral cavity can be adjusted, functioning as an acoustic time-varying filter. Finally sound is radiated to surrounding air at lips and nostrils. The speech production mechanism is illustrated in Figure 2.1.

The produced speech sounds can be basically classified into three categories. Firstly, voiced speech sounds are produced by using the air pressure to get the vocal folds into vibratory motion. This generates a periodic signal rich in harmonics. Voiced sounds form the main part of most West European languages. In English, for example, 78% of phonemes are voiced (Catford 1977). Secondly, unvoiced sounds are produced by constricting the airflow somewhere in the vocal tract. This creates a continuous turbulent airflow characterized by a noise-like waveform without a harmonic structure. The continuous unvoiced speech sounds are called fricatives. Thirdly, unvoiced stop consonants are produced by completely stopping the airflow in the vocal tract. The release of the increased pressure creates a transient noise burst. Speech sounds are often a combination of both voiced and



1. Nasal cavity
2. Hard palate
3. Alveolar ridge
4. Soft palate (velum)
5. Tongue tip
6. Dorsum
7. Uvula
8. Radix
9. Pharynx
10. Epiglottis
11. False vocal folds
12. Vocal fold
13. Larynx
14. Esophagus
15. Trachea

Figure 2.1: Speech production mechanism. (Karjalainen 2000)

unvoiced components.

2.1.1 Vocal Folds

The vocal folds are two elastic tissue structures situated horizontally at larynx. The opening between the vocal fold is called the *glottis* (Flanagan 1972a). The alignment and the tension of the vocal folds can be adjusted by the surrounding muscles and cartilages, which enables switching between respiration and different phonation modes. During respiration the vocal folds are widely separated (abducted), but during phonation they are close to each other (adducted). When the vocal folds are adjusted properly and the airflow through glottis is of sufficient velocity, the vocal folds start to self-oscillate.

The behavior of the vocal folds in phonation is illustrated in Figure 2.2. As the air from the lungs is pushed upwards to the closed vocal folds, the subglottal pressure is increased. This gradually forces the vocal folds to open, increasing the airflow between the vocal folds. The airflow causes an underpressure at the glottis, which draws the vocal folds together, contracting the open glottal area. Finally the glottis closes up as the vocal folds hit together, creating the main excitation in phonation. After the closure, the subglottal pressure begins to increase again, starting a new period. Figure 2.3 shows how the airflow varies in time in phonation.

The periodic signal generated by the vibrating motion of the vocal folds is called the glottal flow, glottal volume velocity waveform, or simply the voice source. The rate at which

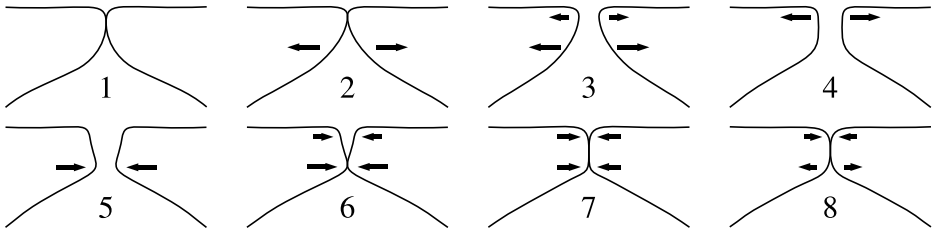


Figure 2.2: Diagram showing an idealized cycle of vocal fold vibration. (Based on Story (2002))

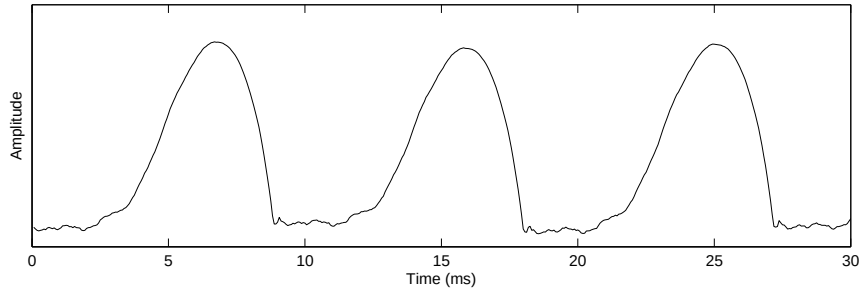


Figure 2.3: Glottal volume velocity waveform estimated from a sustained vowel [a] produced by a male speaker using normal phonation.

the vocal folds vibrate defines the fundamental frequency or f_0 of the speech. The average fundamental frequency of speech is 120 Hz for men and 200 Hz for women (Karjalainen 2000), and in normal speech the fundamental frequency varies approximately from 50 to 500 Hz (Hess 1983). However, the fundamental frequency can be greatly varied from 33 to 3100 Hz in arbitrary utterance (Hess 1983).

In addition to the control of fundamental frequency, the vocal folds can be adjusted to vibrate in different phonation modes. This affects the characteristics of the voice source. Normal speech is typically categorized into three phonation modes: breathy, normal (modal) and pressed. The main distinction among different phonation modes is the degree of adduction. If the adduction is loose, the phonation is called breathy. On the contrary, if the adduction is intense, the phonation is called pressed. In normal phonation the adduction is between breathy and pressed. The phonation type affects the waveform and the spectrum of the voice source. In breathy phonation, the fundamental frequency component is emphasized, whereas in pressed phonation the higher harmonics are emphasized. The spectral envelope of the voice source is called the *spectral tilt* (Klatt & Klatt 1990). Additionally, speech sounds can be produced in different registers, such as vocal fry and falsetto. In vo-

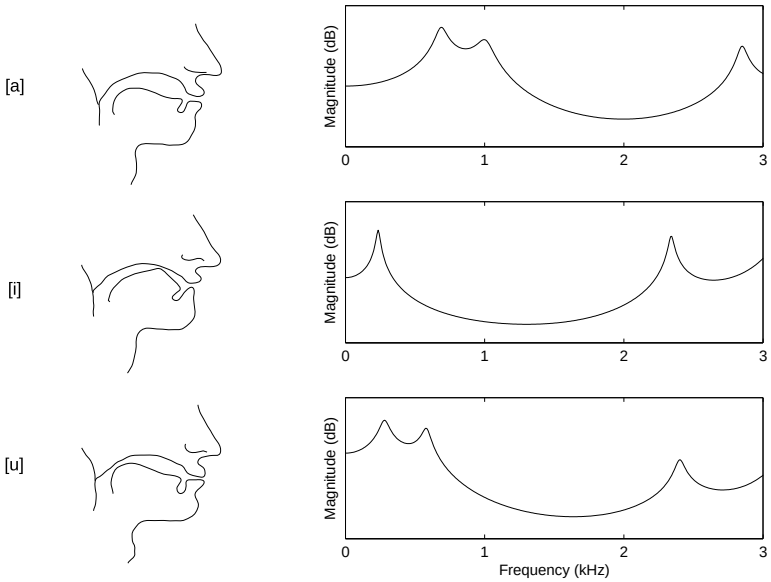


Figure 2.4: Illustration of the profile of the vocal tract and the resulting spectral envelope in phonation of vowels [a], [i] and [u].

cal fry, the vocal folds are loosely closed which permits the air to bubble through glottis, creating a low-pitched creaky sound. In falsetto, the vocal folds are only partly vibrating, creating a breathy high-pitched sound.

2.1.2 Vocal Tract

The vocal tract extends from the glottis up to the lips. It consists of four cavities: the larynx, the pharynx, the oral cavity and the nasal cavity. The length of the vocal tract is normally about 17 cm in men and 15 cm in women (Karjalainen 2000, Claes, Dologlou, ten Bosch & van Compernolle 1998). The function of the vocal tract is to shape the spectral characteristics of the source. It functions as a time-varying filter that creates moving resonances or *formants*. The shape and dimensions of the vocal tract defines the properties of this filter. Different sounds are formed by modifying the vocal tract profile by changing the position of the tongue, lips, jaw and velum. In vowels, the oral tract is open, but in nasal sounds the oral cavity is blocked and the velum lowers down and couples the closed oral tract to the nasal tract. An illustration of the profile of the vocal tract and the resulting spectrum envelope in phonation of different vowels in show in Figure 2.4.

2.2 Basics of Speech Perception

The perception of speech is a special function of hearing. Speech perception has been widely studied with psychoacoustical and physiological methods. The basic acoustical cues of speech perception are rather well know, but the speech-specific higher-level hearing mechanism is yet widely unknown. This section will present the fundamental properties of hearing and the most important perceptual characteristics of speech.

2.2.1 Hearing

Hearing is the ability to perceive sounds by detecting the pressure variations in the air. The outer ear is the first organ that takes part in the perception of sounds. It consists of pinna, ear canal and eardrum. The pinna gathers and focuses sound energy, and has a great effect on spatial hearing. The ear canal extends from pinna to eardrum, and functions as an acoustic filter. It amplifies frequencies around 3 kHz (Gulick, Gescheider & Frisina 1989), which is an important region for the speech perception. The eardrum transforms the acoustic wave motion to mechanical vibrations. The three ossicles in the middle ear transfer the vibrations of the eardrum to oval window. The purpose of the middle ear is to efficiently transfer mechanical energy to waves in fluid. The transmission of sound through the middle ear is most efficient at frequencies from 500 to 4000 Hz (Moore 1997). The sound is transferred to electrical signals in cochlea in the inner ear. The cochlea is filled with liquid, which moves in response to the vibrations coming from the oval window. As the liquid moves, hair cells are set in motion, which then convert the vibrations to neuronal firings. The cochlea can be considered a bank of filters whose outputs are ordered tonotopically. The auditory nerve gathers sound information decoded to electrical signals for further processing. The overview of the human hearing system is presented in Figure 2.5.

The range of audible frequencies extends from 20 Hz to 20 000 Hz. The hearing is most sensitive at 4 kHz, and the sensitivity decreases towards both extremes. However, the full bandwidth of hearing is not used in speech, since the speech signal can be band-limited to about 10 kHz with only minor effects on its perception (Kleijn & Paliwal 1995). Furthermore, the intelligibility of speech can be obtained with much narrower bandwidth. The perceived pitch of a sound is generally proportional to the logarithm of the frequency. Therefore in many applications it is more convenient to describe the frequency related quantities, such as pitch or formant frequencies, with perceptually weighted auditory scales instead of physical frequency. One of the earliest auditory scales was the pitch ratio scale measured in mels (Stevens, Volkmann & Newman 1937). The mel-scale is obtained by asking subjects to adjust the frequency of a tone to be half or twice as high as that of a tone given for comparison. Two other important auditory scales are the critical band rate measured in barks

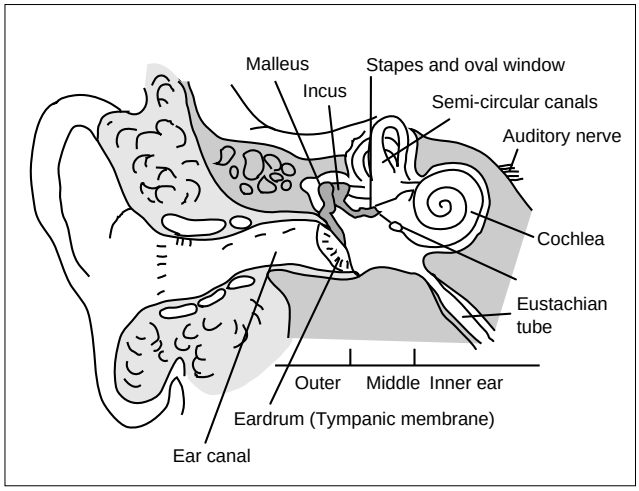


Figure 2.5: Human hearing system. (Karjalainen 2000)

(Zwicker, Flottorp & Stevens 1957) and the Equivalent Rectangular Bandwidth (ERB) rate (Moore & Glasberg 1974), which are both based on measuring the frequency resolution.

2.2.2 Perception of Voiced Speech Sounds

Voiced speech sounds are described by the periodic excitation of the vocal folds and the formant structure resulting from the profile of the vocal tract. Two or three first formants are used to distinguish between different vowels (Pickett 1999). The other formants remain rather constant regardless of changes in articulation. Individual formants can be described by center frequency, amplitude and bandwidth. The just noticeable differences (JNDs) for first and second formant frequencies have been measured to be from 3 to 5 percent of the center frequency (Flanagan 1972a). The formant amplitude JNDs are estimated to be 1.5 dB and 3 dB for first and second formants, respectively (Flanagan 1972a). Changes in formant bandwidth (–3 dB) of order 20–40 percent have been found to be just noticeable (Flanagan 1972a).

Nasal sounds are described by a nasal murmur, whose spectrum is dominated by the low frequency components. The spectrum is determined mostly by the main resonance of the nasal cavity. The spectrum of nasal murmur vary little among different nasal sounds, since the size and shape of the nasal cavity cannot be altered. Nasal sounds incorporate also antiformants which reduce the energy at certain regions. Nasal sounds are considerably lower in intensity than vowels due to blocked oral cavity.

The source waveform and spectrum of voiced sounds can be varied in excitation intensity, fundamental frequency and phonation type. The spectral slope of the excitation can vary from –15 dB per octave of breathy phonation to –9 dB per octave of pressed phonation

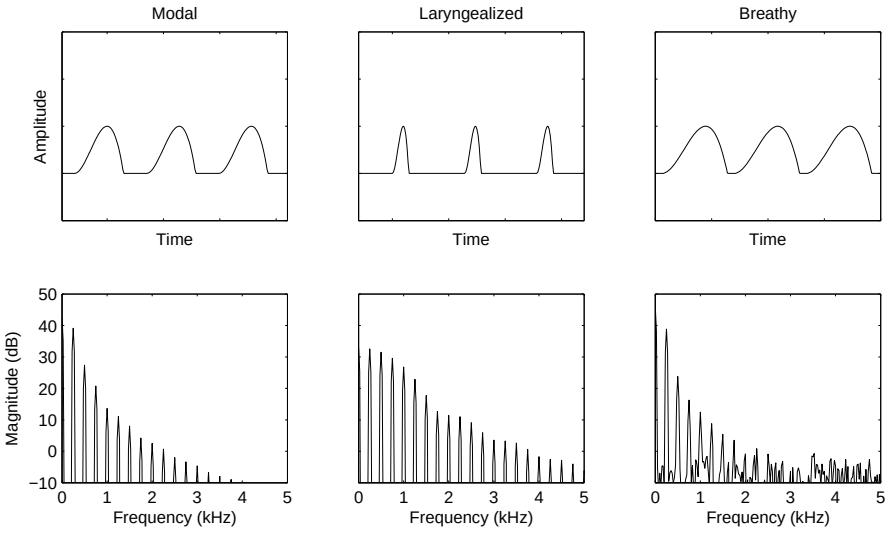


Figure 2.6: Glottal flow and its spectrum in different phonation modes. On the left, an idealized glottal flow in modal (normal) phonation and its spectrum is presented. The spectral slope of the excitation is about –12 dB per octave. In the middle, laryngealized (pressed) phonation is illustrated where the spectral slope is about –9 dB per octave. On the right, breathy phonation is visualized where the spectral slope is –15 dB per octave and the higher harmonics are replaced by aspiration noise.

(Pickett 1999). The spectral slope can also vary due to increased vocal effort. In pressed (laryngealized) phonation, the glottal pulse waveform is narrower, the fundamental frequency component is reduced, and there may be diplophonic irregularities in fundamental period. Breathy phonation is characterized by increased open glottal period, increased amplitude of the fundamental component, and a tendency of higher harmonics to be replaced with aspiration noise (Klatt & Klatt 1990). Figure 2.6 shows how the glottal waveform and its spectrum vary in different phonation modes.

Temporal fine structure is also known to exist in the glottal source. The shape and the periodicity of the glottal pulse is subject to various perturbations, for example jitter and shimmer. Although the magnitude of jitter in normal voices is found to be slightly less than the detectability threshold, the fine structure of the glottal flow may be of significant perceptual importance (Klatt & Klatt 1990).

2.2.3 Perception of Unvoiced Speech Sounds

Unvoiced speech sounds can be basically categorized either as fricatives or stop consonants. Although fricatives and stop consonants may also incorporate a voiced component, the

properties of voiced fricatives and stop consonants differ greatly from pure vowels, and their distinctive features are somewhat the same whether the voiced part is present or not.

Fricatives are described by a continuous aperiodic noise. The spectral characteristics of the noise vary according to the articulatory configuration. The duration of these sounds is relatively long, though the length depends on many contextual factors. For example, the duration of the fricative /s/ can range from 20 to 200 ms (Klatt 1974, Klatt 1976). The spectral envelope, energy, and temporal characteristics vary according to individual phoneme. In the case of voiced fricatives, the glottal vibrations modulate the continuous aperiodic noise.

Stop consonants are mainly described by a low-energy interval called a stop gap followed by a transient noise burst. The spectral characteristics of the noise burst vary according to the articulatory configuration. However, it is doubtful whether the spectral content of this burst is sufficient for phonetic identification (Kent & Read 1992). Typically the bursts are no longer than 5 – 50 ms in duration (Kent & Read 1992), and they are one of the shortest acoustic events that are analyzed in speech. The burst can also be aspirated, in which case the burst is accompanied by a fricative-like longer noise tail. The burst is preceded by the stop gap, which corresponds to the low energy period due to the articulatory occlusion. For voiceless stops, the stop gap is virtually silent. For voiced stops in other than word-initial position, the stop gap usually contains a low frequency band of energy called the *voice bar*. The duration of the stop gap is usually 50 – 150 ms (Kent & Read 1992). Stop consonants are also characterized by a delay in voicing relative to the beginning of the noise burst. This delay is called the voice onset time (VOT) (Lisker & Abramson 1964). The voice onset time is a major feature in distinguishing voiced and unvoiced stop consonants. For unvoiced stop consonants the VOT is between 25 and 100 ms, whereas for voiced stop consonants VOT can vary between –20 and 20 ms (Kent & Read 1992), in which case the voicing can start before the noise burst.

During the transition from voiced to unvoiced and from unvoiced to voiced sounds, the formant frequencies are shifted due to change in the vocal tract shape. This functions as an additional acoustic cue for the identification of the phoneme. For example, experiments with synthetic speech show that stop consonants are identified without the noise burst if the formant transitions are properly specified (Kent & Read 1992). In natural speech, however, the influence of formant transitions is not so clear. The formant transition are about 50 ms in duration (Kent & Read 1992).

2.2.4 Acoustic Effects of Context and Speaker

The properties of individual speech sounds are largely dependent on the context of the phoneme in syllables, words, and phrases. In continuous speech, the speech sounds are

produced in rapid succession, and the boundaries between individual sounds are blurred. This effect, the interaction of speech segments, is called the coarticulation. Coarticulation enables fast and smooth production of speech, but in speech technology it is challenging to model and take the coarticulation into account when performing tasks such as speech synthesis or speech recognition.

In addition to the coarticulation, speech is greatly affected by the message intended to be conveyed. All the modifications in the intonation, stress, and rhythm of speech fall in the category of prosody. Prosody has many functions, such as syntax, indication of utterance type, and expression of interaction, attitude or emotion. The prosodic features vary according to the prosodic functions. Prosody is formally defined as the suprasegmental features of speech that are conveyed by the parameters of fundamental frequency, intensity, and duration. In addition to these parameters, the spectrum pattern of speech is also varied in terms of prosody. A large part of the prosodic variation is generated by changing the characteristics of the glottal source signal, such as fundamental frequency, intensity, and voice source spectrum. Since anomalies in prosody are easily perceived, the correct modeling of prosody and thus the glottal source signal is especially important in order to create natural sounding speech. For an extensive summary of speech prosody, see for example (Pickett 1999).

Speakers vary substantially according to gender, age and other individual differences. First, due to the differences in physical properties of the speaker, such as the size and shape of vocal folds and vocal tract, the speech sounds are produced differently. Second, the individual differences in speaking style, such as language, accent, speech rate, and dialect, affect the use of the speech production organs. The message of speech is usually well understood despite the great variability, but the speaker characteristics as such give much useful information. This is a property of natural speech, and therefore the preservation or alteration of speaker characteristics in speech is an important objective. For example, in speech synthesis, this requires the correct modeling of the speech production mechanism as well as the higher-level speaking characteristics.

2.3 Source-Filter Theory

The source-filter theory of speech production states that speech signal can be represented in terms of source and filter characteristics (Fant 1960). In human speech production the primary sound source is the excitation of the vibrating vocal folds. The periodic vibration generates a rich harmonic spectrum, whose energy declines with increasing frequency. The average rate of decline is 12 dB per octave (Flanagan 1972a, Kent & Read 1992), but it can be greatly varied according to phonation mode. The vocal tract modifies the excitation spectrum with a transfer function with formants or antiformants. Finally the sound radiates

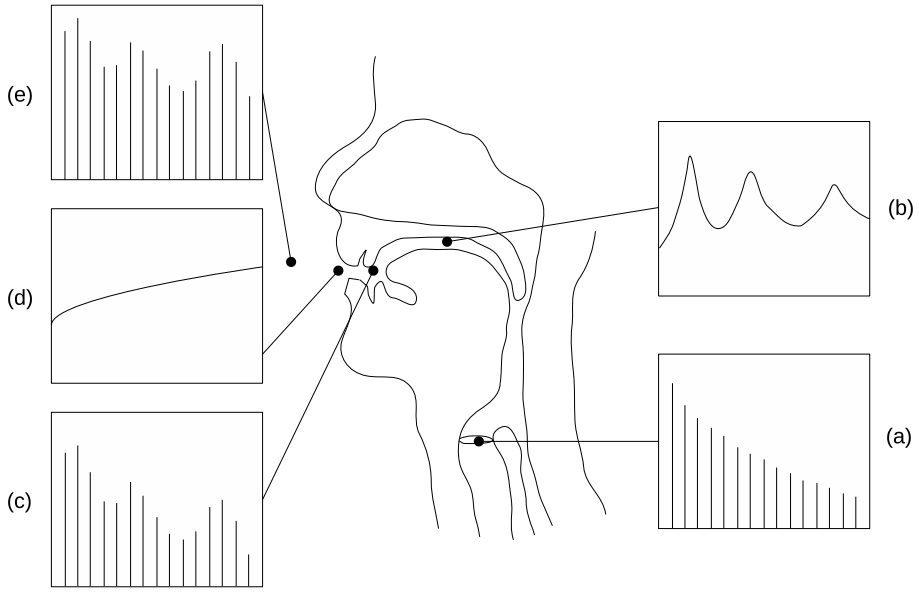


Figure 2.7: Source-filter model of speech production. (a) Speech is initiated by the vibrations of the vocal folds. This generates a rich periodic spectrum, which energy declines with increasing frequency. (b) The vocal tract modifies the glottal excitation by creating resonances. (c) Spectrum of the signal before lip radiation. (d) The radiation of sound from lips and nostrils to surrounding air creates an effect that enhances the higher frequencies of the signal. This is called the lip radiation. (e) The spectrum of the speech signal.

to the surrounding air at lips and nostrils. This causes a frequency dependent effect called lip radiation, which acts as a high-pass filter. The magnitude of this effect is approximately 6 dB per octave (Flanagan 1972a), but it is usually approximated by a simple differentiation operation (Markel & Gray 1980). The source-filter theory is summarized in Figures 2.7 and 2.8.

Assuming a linear time-invariant system, the above model can be described in Z-transform notation by the equation

$$S(z) = E(z)G(z)V(z)L(z), \tag{2.1}$$

where $S(z)$ is the speech signal, $E(z)$ the impulse excitation, $G(z)$ the glottal shaping model, $V(z)$ the vocal tract model, and $L(z)$ the lip radiation model (Markel & Gray 1980). The impulse excitation $E(z)$ does not represent a physical signal, but is rather used as a mathematical input to the glottal model filter. Transfer functions $G(z)$ and $V(z)$ are usually

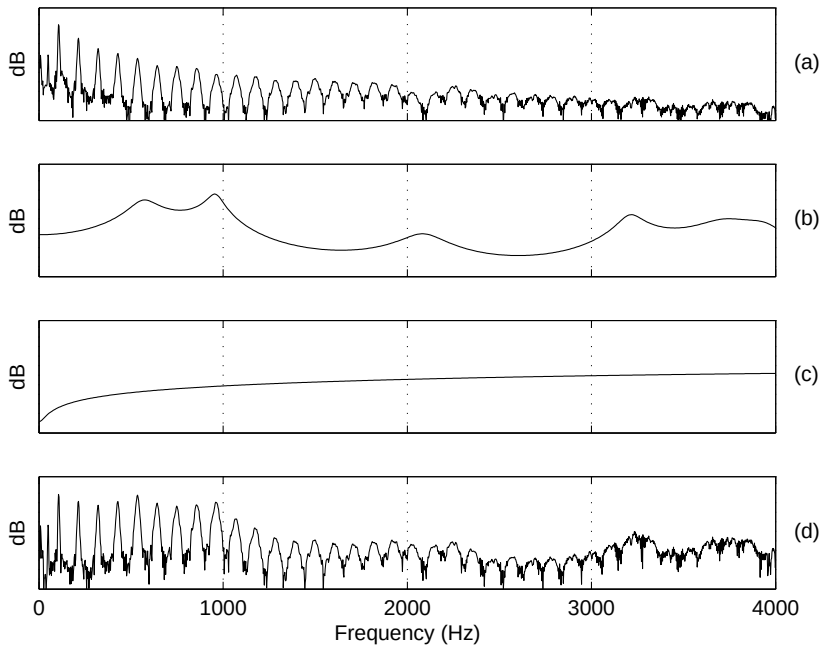


Figure 2.8: Spectrum of the component in the source-filter theory. (a) Spectrum of the glottal excitation. (b) Amplitude response of the vocal tract filter. (c) Amplitude spectrum of the lip radiation. (d) Spectrum of the speech signal.

described with all-pole linear filters, and $L(z)$ is given by a differencing filter

$$L(z) = 1 - \rho z^{-1}, \tag{2.2}$$

where, in the definition by Markel & Gray (1980), ρ is set to 1. Lip radiation $L(z)$ is the only numerator in Equation 2.1, but it is nearly canceled by one of the denominator terms (Markel & Gray 1980). Thus the model can be described as

$$S(z) = E(z)\frac{1}{A(z)}, \tag{2.3}$$

where the all-pole filter is defined as

$$A(z) \simeq \frac{1}{G(z)V(z)L(z)}. \tag{2.4}$$

The detailed derivation of this model is presented for example in Fant (1960) and Flanagan (1972a).

The source-filter theory is a linear mathematical model with many simplifying assumptions. Therefore some aspects of the theory are not really valid. For instance, it has been

observed that interaction between source and vocal tract can occur in natural speech (Klatt & Klatt 1990). Moreover, the all-pole model is not perfectly appropriate for modeling antiformants, which are present in nasal sounds. However, the (all-pole) source-filter model is sufficient for most applications since the benefits of the linear model are much greater than the disadvantages.

2.4 Overview of Speech Synthesis

Speech synthesis is the artificial generation of speech. Speech synthesis has various useful applications, such as telecommunication services, man-machine communication, language education, aid to persons with disabilities, research on speech production and perception, and many others. Depending on the application, different implementations of a speech synthesizer may be used. Today a text-to-speech (TTS) system is maybe the most common and the most versatile solution. The ultimate goal of such a system is to read any text and convert it to speech. However, there are various criteria for evaluating the resulting speech or the system as a whole, and various approaches can be used to meet the required specifications. In this section, first a short overview of history and development of speech synthesis is presented. For more extensive summary of history and development of speech synthesis, see for example Klatt (1987), Flanagan (1972a), and Flanagan (1972b).

2.4.1 History of Speech Synthesis

The earliest attempts to produce artificial speech were made more than two hundred years ago (Flanagan 1972a). The early mechanical implementations of speech synthesizers modeled the physiology of the speech production organs. For example, in 1791 von Kempelen presented a speaking machine which consisted of bellows, a vibrating reed, and a rubber tube modeling the vocal tract.

As the electrical technology evolved, interest in speech synthesis increased. The first formant synthesizer was built by Stewart in 1922 (Klatt 1987). It consisted of two resonant circuits, which were excited by a buzzer. This early synthesizer was able to generate a static vowel with two lowest formants. The first electrical device that could produce continuous speech was the Voder developed at the Bell Telephone Laboratories in 1939. It was based on the idea of a vocoder, a voice coder, which could analyze speech into slowly varying parameters and then reconstruct an approximation of the original speech from the parameters. The first dynamically controlled formant synthesizers were introduced in 1953. Walter Lawrence's PAT and Gunnar Fant's OVE I, and especially their improved later versions, could generate intelligible speech. Shortly after that the first articulatory speech synthesizer was introduced in 1958. Consequently, the speech analysis and synthesis techniques split

into two paradigms: modeling of the speech production mechanism itself, and modeling only the speech signal (Sproat & Olive 1995). This division stands even today, though much co-operation exists between the fields. The first full text-to-speech system was developed in 1968 by Noriko Umeda, and in 1972 John Holmes demonstrated that synthetic speech could be so natural sounding that the average person could not tell the difference between the synthetic and the original sentence (Klatt 1987).

Since the late 1970s, many commercial speech synthesis and text-to-speech products have been introduced, with MITalk (Allen, Hunnicut & Klatt 1987) being probably the best known TTS system. In the mid 1980s the concept of high quality TTS synthesis appeared, mostly due to new technologies. Modern synthesizers have largely moved from electronic circuitry to simulation on a digital computer. The methods used in speech synthesis technology today are very sophisticated as the latest findings from research on information technology, signal processing, acoustics, speech production, and linguistics are applied directly to speech synthesizers. The quality of speech synthesis has improved to a level of great intelligibility, but the naturalness is yet a problem. Nevertheless, more natural sounding speech synthesizers are constantly developed based on various different methods. In the next two sections, TTS architecture and speech synthesis methods are considered in more detail.

2.4.2 General TTS Architecture

If the input to a speech synthesizer is given as text, the system is called a text-to-speech (TTS) synthesizer. However, in the case of speech synthesizers with limited vocabulary, such as machines playing prerecorded samples, the definition is not unambiguous. According to the more specific definition by Dutoit (1997), text-to-speech means "the production of speech by machines, by way of the automatic phonetization of utter".

A general functional diagram of a TTS system is shown in Figure 2.9. A TTS synthesizer consists of two main components, called the high-level and low-level synthesis. The high-level synthesis converts the text input to a form that corresponds to the desired acoustic phonation of the utterance. This means converting the text input into a phonetic or some other linguistic representation and predicting the desired prosody. In the process, the input text is first normalized into plain letters, and the structural properties of the text are analyzed. After that, the text is converted to a phonetic level, which is called the letter-to-sound conversion (Pickett 1999). Varying amount of linguistic analysis is performed on the text in order to predict the prosodic features of the utterance, such as phrasing and accentuation patterns. Based on the prosodic analysis and the structural information, actual f_0 contour and phone durations are predicted for the utterance, typically using statistical methods. From the linguistic and prosodic information, the low-level synthesis generates

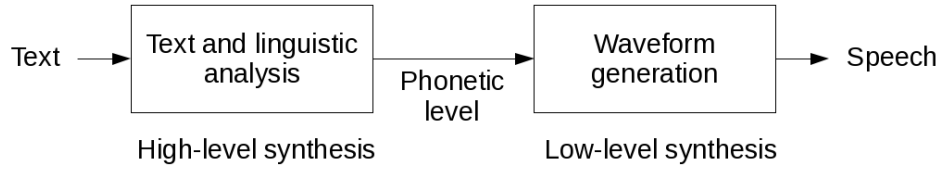


Figure 2.9: General functional diagram of a TTS system.

the speech waveform. For the waveform generation, today's TTS systems commonly employ techniques based either on the source-filter theory or modification and concatenation of prerecorded speech samples. Speech synthesis methods are considered in the next section in more detail.

2.4.3 Speech Synthesis Methods

Once the high-level synthesis of a TTS system has completed its task, the low-level synthesis starts generating the speech waveform. The waveform generation can be accomplished in many ways, and the synthesis methods can be categorized according to various criteria. A basic division can be made according to whether the speech is completely artificially generated from parameters, or are real speech samples used in the process. This property greatly affects the functioning of the synthesizer. Formant synthesis, articulatory synthesis, and linear predictive coding (LPC) based synthesis can be placed to the first category, whereas concatenative synthesis belongs to the latter.

Formant Synthesis The most basic acoustic speech synthesis technique, formant synthesis, employs the source-filter theory of speech production described in Section 2.3. The vocal tract model consists of individually adjustable formant filters connected in serial, parallel, or often both. Different phonemes are constructed by adjusting the center frequency, bandwidth, and gain of each filter. If the adjustment is made at certain time intervals, for example every 5 ms, continuous speech can be generated. The source can be modeled with voice pulses or noise. A basic speech synthesis model based on the source-filter theory is shown in Figure 2.10.

Formant synthesis received a big boost in 1980 with Dennis Klatt's publication of a sophisticated formant synthesizer with a complete computer program for speech synthesis. Today, the quality of formant synthesizers is inferior compared to the latest synthesis methods, such as concatenative and LPC-based methods, but formant synthesis has many applications in reading machines for the blind and in speech perception experiments for creating stimuli (Pickett 1999).

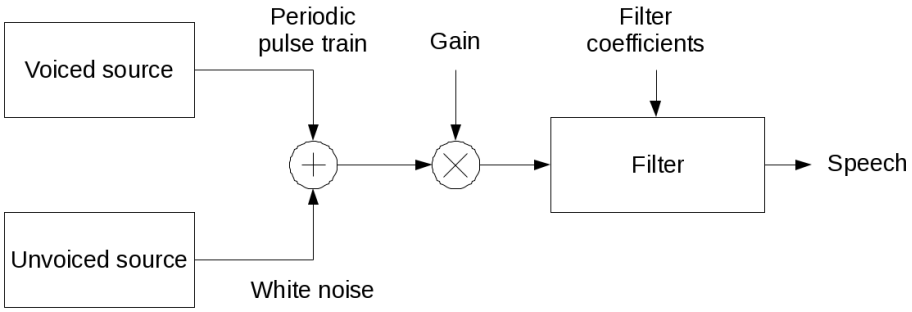


Figure 2.10: Speech synthesis model based on the source-filter theory.

Articulatory Synthesis Articulatory synthesis tries to model the natural speech production process as accurately as possible. Thus it is theoretically the best method for high quality speech synthesis, but it is also by the same token the most difficult in terms of implementation and computational load. Because of the limitations of the current speech production models and computational power, articulatory synthesis has not achieved as much success as other speech synthesis methods. However, it has many useful applications in basic speech research, and it might have a promising future since better articulatory models are steadily developed and computational resources are increasing.

Concatenative Synthesis In concatenative synthesis, prerecorded samples of real speech are smoothly combined to create an arbitrary synthetic utterance. Common unit lengths are word, syllable, demisyllable, phoneme, diphone, and triphone. Because the natural characteristics of the speech are preserved in the units, concatenative synthesis is capable of generating highly intelligible and natural synthetic speech. However, the discontinuities in concatenation points can cause distortion despite the use of various smoothing algorithms. Also, the set of speech units is always limited. It is highly impractical or impossible to store all the necessary units for various speakers in various contexts. This constraint makes the concatenative speech synthesis less flexible: it can imitate the specific speaker with only one voice quality. Another constraint is the need for vast storage for all the recorded units, but with the cost of computer storage decreasing, and with the development of fast database access techniques, this problem is not as serious as it used to be. Today the concatenative speech synthesis is probably the most widely used and most natural sounding, but due to the mentioned limitations, it might not be the best solution.

LPC-Based Synthesis In linear predictive coding (LPC) based speech synthesis, source-filter theory of speech production is utilized the same way as in formant synthesis, but in

the LPC-based synthesis the filter coefficients can be automatically estimated from a short frame of speech instead of finding the parameters for individual formant filters. With an appropriate excitation, the filter coefficients can be used to synthesize speech. The excitation is either periodic source signal or noise, depending on whether the synthesized speech segment is voiced or unvoiced (see Figure 2.10). Linear prediction (LP) is a widely used method in speech technology, and is more closely discussed in Section 3.1. Though the quality of a basic LPC vocoder is considered poor, high quality synthetic speech can be produced with more sophisticated LPC-based synthesis methods. The type of excitation signal is especially important for the quality of synthetic speech, as will be shown later in this thesis.

HMM-Based Synthesis One widely applied method in speech synthesis is the use of hidden Markov models (HMMs). HMM is a statistical model, which can be used for modeling the speech parameters extracted from a speech database, and then generating the parameters according to text input for creating the speech waveform. HMM-based speech synthesis systems are able to produce speech in different speaking styles with different speaker characteristics and even emotions. They also benefit from better adaptability and clearly smaller memory requirement. However, the HMM-based TTS systems often suffer from degraded naturalness in quality compared to concatenative based speech synthesizers. Nonetheless, the HMM-based TTS systems are developing fast, and much work is carried out for finding techniques to enhance the quality and naturalness of synthetic speech. The current prevalent platform for HMM-based speech synthesis is the HTS system developed in Japan (HTS 2008). It is widely used among speech synthesis researchers and developers, and lately numerous HMM-based TTS systems have been introduced for various languages. The hidden Markov models are generally described in Section 3.5 and the new HMM-based TTS system is further described in Chapter 4.

Chapter 3

Methods and Algorithms

In this chapter, the most essential tools for speech synthesis in general and especially for this work are presented. The more detailed description of the methods implemented in the new synthesizer are presented in Chapter 4.

3.1 Linear Prediction

Linear predictive coding (LPC) is one of the most widely applied techniques in speech technology. Although linear prediction (LP) has been applied in many fields for a long time, the first researchers to directly apply it to speech analysis and synthesis were Saito & Itakura (1967) and Atal & Schroeder (1967). Today linear prediction has various applications in speech technology, for example in speech coding, synthesis, analysis, and recognition, both for commercial and research purposes.

The basic idea behind linear prediction is that a sample of data can be predicted by a linear combination of previous samples. In speech technology, however, the goal of LPC is not really to predict any samples, but to represent the spectral envelope of the speech signal. Therefore, speech can be represented as a combination of a filter and an excitation signal, which is equivalent to the source-filter model of speech production. The importance of LPC lies both in its accuracy of estimating the speech parameters and in its relative speed of computation.

3.1.1 Derivation of LPC filter

A sample $\hat{x}_n$ can be stated as a linear combination of m past samples. This can be formulated as

$$\hat{x}_n = -\sum_{i=1}^{m} a_i x_{n-i}, \tag{3.1}$$

where a_i $(1 \leq i \leq m)$ are the predictor coefficients, m the model order and the minus sign has been added for convenience. Thus the error signal, or the residual, can be stated as

$$e_n = x_n - \hat{x}_n = x_n + \sum_{i=1}^{m} a_i x_{n-i} = \sum_{i=0}^{m} a_i x_{n-i}, \tag{3.2}$$

where a_0 is 1. The optimal predictor coefficients a_i $(1 \leq i \leq m)$ can be obtained by minimizing the square of e_n. This least squares minimization leads to the following so called normal equations:

$$\sum_{k=1}^{m} a_k \sum_{n} x_{n-i} x_{n-k} = \sum_{n} x_{n-i} x_n, \quad 1 \leq i \leq m \tag{3.3}$$

(Markel & Gray 1980, Rabiner & Schafer 1978). Several algorithms have been developed to solve Equation 3.3, but in speech processing, two specific solutions are commonly used. These are referred to as the *covariance method* and the *autocorrelation method*, of which only the latter is guaranteed to yield a stable filter. In the autocorrelation method, it is assumed that the error is minimized over an infinite interval, $-\infty < n < \infty$, and the signal is zero outside the time interval $0 \leq n \leq N-1$. In the autocorrelation method, Equation 3.3 can be rewritten as m simultaneous linear equations:

$$\begin{pmatrix} R(0) & R(1) & R(2) & \dots & R(m-1) \\ R(1) & R(0) & R(1) & \dots & R(m-2) \\ R(2) & R(1) & R(0) & \dots & R(m-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ R(m-1) & R(m-2) & R(m-3) & \dots & R(0) \end{pmatrix} \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_m \end{pmatrix} = - \begin{pmatrix} R(1) \\ R(2) \\ R(3) \\ \vdots \\ R(m) \end{pmatrix}, \tag{3.4}$$

where

$$R(k) = \sum_{m=0}^{N-1-k} x_n x_{n+k} \tag{3.5}$$

In matrix form, Equation 3.4 can be written as

$$\mathbf{Ra} = -\mathbf{r}. \tag{3.6}$$

Various approaches can be used to solve the coefficients **a** from Equation 3.6, but since the matrix **R** is a symmetric Toeplitz matrix, the coefficients are most efficiently solved with the recursive Levinson-Durbin algorithm (Markel & Gray 1980, Rabiner & Schafer 1978).

3.1.2 Properties of Linear Prediction

Linear prediction coefficients can be represented as a digital filter, whose power spectrum represents the spectral envelope of the analyzed signal. The resulting finite impulse re-

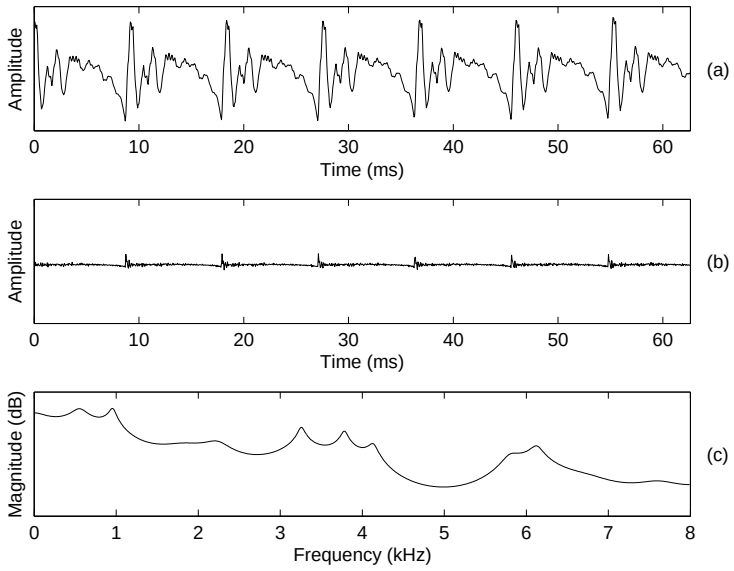


Figure 3.1: Illustration of LPC analysis. (a) Speech signal. (b) Residual signal. (c) Spectral envelope of the speech estimated with 20th order LPC.

sponse (FIR) filter is called the inverse filter, and is denoted in the Z-domain as

$$A(z) = \sum_{k=0}^{m} a_k z^{-k}. \tag{3.7}$$

The linear speech production model introduced in Section 2.3 states that the speech can be thought as a result of impulse excitation, glottal shaping model, vocal tract model, and lip radiation model, i.e. $S(z) = E(z)G(z)V(z)L(z)$. The idea of LPC analysis is to separate the excitation and the filter using the speech analysis model described as

$$E(z) = S(z)A(z). \tag{3.8}$$

Thus, the excitation $E(z)$, or the residual signal, will become an impulse train with additive white noise, whereas the filter $A(z)$ is an estimate of the overall effect of $G(z)V(z)L(z)$, the all-pole spectral model of speech. An illustration of LPC analysis is shown in Figure 3.1.

The inverse of the prediction filter is called a synthesis filter. In speech synthesis, the synthesis filter is excited by an appropriate excitation signal to create speech. The synthesis model is described as

$$S(z) = E(z)\frac{1}{A(z)}. \tag{3.9}$$

3.1.3 Line Spectrum Pair (LSP) Decomposition

LPC is frequently used for transmitting the spectral envelope of speech, and therefore it has to be tolerant to quantization and transmission errors. Since the LP coefficients as such are very sensitive to errors, various coefficient representations have been developed to make the transmission of coefficients more robust. One of the most efficient and widely used coefficient representations is the line spectral frequencies (LSFs), which are the roots of the LSP polynomials. LSP polynomials were first introduced by Itakura (1975), but it was Soong & Juang (1984) who got them to awareness of the general public.

The spectral envelope of a speech signal can be represented by an LP polynomial $A(z) = \sum_{k=0}^{m} a_k z^{-k}$, where a_k are the model coefficients, m the model order and $a_0 = 1$. The line spectrum pair (LSP) polynomials for $A(z)$ are defined as

$$\begin{aligned} P(z) &= A(z) + z^{-m-1}A(z^{-1}) \\ Q(z) &= A(z) - z^{-m-1}A(z^{-1}). \end{aligned} \tag{3.10}$$

The original polynomial $A(z)$ can be reconstructed by

$$A(z) = \frac{1}{2}\left[P(z) + Q(z)\right]. \tag{3.11}$$

The result of such a decomposition is that if $A(z)$ has all roots within the unit circle, then the roots of the polynomials $P(z)$ and $Q(z)$

1. are on the unit circle.
2. are simple (they do not overlap).
3. are interlaced.

(Soong & Juang 1984). These properties are illustrated in Figure 3.2. Since the roots are interlaced, the stability of the filter is guaranteed if and only if the locations of the roots on the unit circle are monotonously increasing. Moreover, line spectral frequencies have a well-behaved dynamic range (Soong & Juang 1984), that is, a slight variation to the root locations on the unit circle does not affect much the filter characteristics. Thus, given a set of line spectral frequencies, the corresponding LPC filter can be reconstructed with great robustness and stability check. Another benefit from using LSFs is its better interpolation characteristics compared to LPC coefficients. In speech coding and synthesis the parameters are transmitted frame-wise, possibly causing large changes in the values, which may be heard as undesired transients. To avoid this, interpolation is used between adjacent frames to smooth the parameters. The interpolation characteristics of LSFs in terms of spectral distortion and stability have been found at least equal (Umezaki & Itakura 1986, Atal, Cox

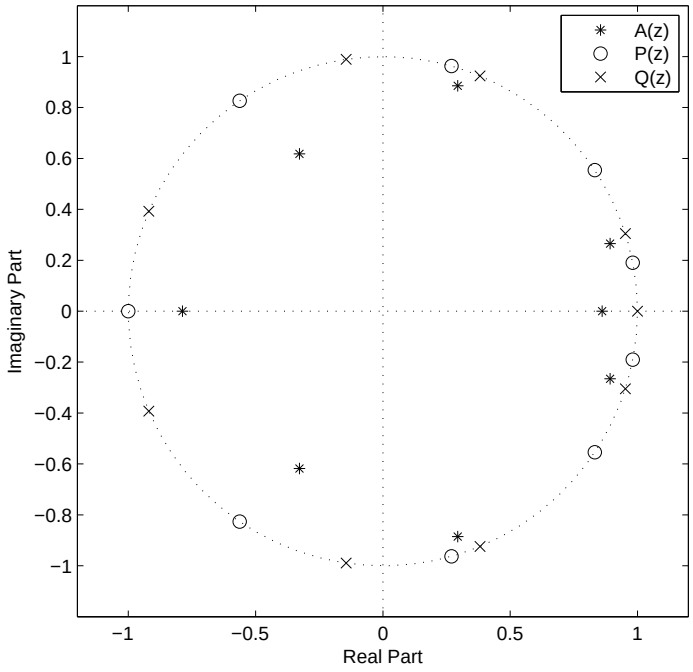


Figure 3.2: Illustration of the unit circle and the roots of the polynomials $A(z)$, $P(z)$, and $Q(z)$. The order of the polynomial $A(z)$ is eight. The trivial zeros of $P(z)$ and $Q(z)$ are at -1 and $+1$, respectively.

& Kroon 1989) or superior (Paliwal & Kleijn 1995) compared to any other representation, including the original LPC coefficients.

In the calculation of LSFs, a major task is to find the roots of the LSP polynomials. In the process, first, the trivial zeros of the polynomials $P(z)$ and $Q(z)$, depicted in Table 3.1, are removed. Since there is no general formula for solving the roots of a polynomial of order greater than four, numerical methods must be used for solving the remaining roots. However, it is known that the roots of the LSP polynomials lie always on the unit circle. This information can be utilized in order to make the numerical search more efficient. More-

Table 3.1: Trivial zeros of the LSP polynomials.

	$(m+1)$ even	$(m+1)$ odd
$P(z)$	none	$z=-1$
$Q(z)$	$z=+1,-1$	$z=+1$

over, the symmetric properties of the polynomials $P(z)$ and $Q(z)$ can be utilized in solving the roots. Usually, the polynomials are transformed to Chebyshev polynomials (Kabal & Ramachandran 1986), which reduce the order of the root solving problem to half.

3.2 Fundamental Frequency Estimation

Fundamental frequency (f_o) estimation is one of the most important problems in speech processing. Although many solutions have been proposed, and many of them work well in their specific context, none of the presently available methods can be expected to give perfectly satisfactory results across wide range of speakers, applications, and operating environments. There are many reasons for the difficulties in fundamental frequency estimation. Although the periodicity of the speech signal derives from the vibrations of the vocal folds, the estimation algorithms must cope with a mixed excitation consisting of voiced and unvoiced components. The characteristics of the voiced component can vary greatly, and the fundamental frequency is changing continuously with time, often with each glottal period. The voice onsets and offsets, subharmonics of fundamental frequency, formant structure, and the wide dynamic range of speech make the fundamental frequency estimation more challenging.

Fundamental frequency estimators, or pitch detection algorithms (PDAs), usually consist of three components: a pre-processing stage, the f_0 estimation, and a post-processing stage. The aim of the pre-processing stage is to remove interfering signal components, such as extraneous noise, vocal tract influence, and DC offset, and to transform the signal to better fit the later processing stages. The preprocessing methods include for example low-pass filtering, inverse filtering, cubing, and peak or center clipping (Talkin 1995, Rabiner 1977). The purpose of the post-processing stage is to correct the errors made in the f_0 estimation. A straightforward and very successful strategy is to use median filtering (Rabiner, Sambur & Schmidt 1975), which very effectively ignores outliers while preserving the fine structure of the f_0 contour and the sharpness of true step transitions. Another successful method is dynamic programming (DP). It is based on the concept of cost function which penalizes for large variation between two consecutive samples and rewards for close vicinity between them. Dynamic programming was first used in f_0 estimation by Bauer & Blankenship (1974), and later clearly outlined by Ney (1981) that it could be also used for f_0 smoothing. Also various heuristics can be used to correct the errors by utilizing prior knowledge of the speech signal or information from past f_0 estimates.

3.2.1 Time Domain Approaches

Time domain PDAs are based on the estimation of the fundamental period of the quasiperiodic speech signal. A straightforward way of finding the fundamental period is to examine how often events, for example peaks or valleys in the waveform repeat themselves. These methods are easy to implement and do not require much computing power, but are not very robust with complex speech spectra. The algorithm described by Gold & Rabiner (1969) is probably one of the most widely used methods of this type. Another related feature, the zero-crossing rate (ZCR) is a measure of how often waveform crosses zero per unit time. It gives information about the spectral content of the signal, but in f_0 estimation it has certain problems described for example in Gerhard (2003) and Kedem (1986). However, ZCR as such is a statistically informative feature (Gerhard 2003), and it can be used successfully for example in classification problems. Since f_0 estimation is closely related to the classification of speech into voiced or unvoiced segments, ZCR can be used as a supplementary feature in f_0 estimation.

Autocorrelation analysis is one of the most robust and reliable methods in fundamental frequency estimation. It is based on the fact that a periodic signal will be similar from one period to the next. An autocorrelation function (ACF) is the measure of similarity of the speech signal with itself as a function of time separation between them. In autocorrelation analysis, an ACF is computed from a windowed segment of speech signal. The analysis frame size is chosen to be at least twice the longest expected period (Rabiner 1977). For signal x_n and window size w, the ACF is defined as

$$r_n(\tau) = \sum_{j=n}^{n+w-1} x_j x_{j+\tau}, \tag{3.12}$$

where τ is the time delay (Paliwal & Kleijn 1995). However, for f_0 estimation purposes, it is more convenient to use a slightly different definition:

$$r_n(\tau) = \sum_{j=n}^{n+w-1-\tau} x_j x_{j+\tau}. \tag{3.13}$$

In Equation 3.13, the size of the analysis window decreases as τ increases. This has a tapering effect, so that the ACF will have smaller values with increasing τ. This attenuates the multiples of fundamental period peaks in ACF. Thus, the highest peak excluding the peak at zero depicts the fundamental period of the windowed signal, i.e. the relation of the signal with itself is strongest at time τ ($\tau \neq 0$). Usually the highest peak is found with an exhaustive search within a predefined range of lags. A segment of speech signal and its autocorrelation function is shown in Figure 3.3.

The computation of the autocorrelation function is quite time consuming, but many algorithms have been introduced to make the computation faster (see for example Rabiner &

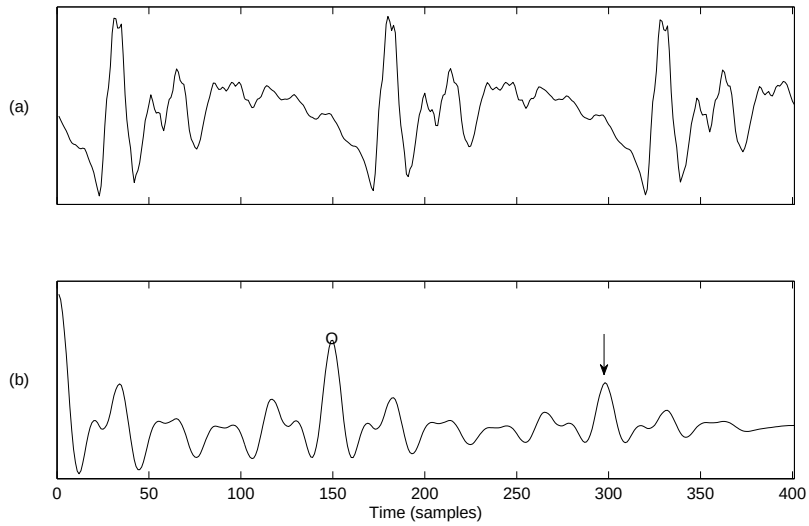


Figure 3.3: Speech signal (a) and its autocorrelation function (b). The peak corresponding to the fundamental period is depicted with a circle. The multiple of the fundamental period peak, depicted with an arrow, is attenuated due to the tapering effect.

Schafer (1978)). Since the autocorrelation function $r(\tau)$ is the inverse Fourier transform of the power spectrum $S(f)$, autocorrelation can be defined as

$$r(\tau) = \int_{-\infty}^{\infty} S(f) e^{j2\pi f\tau} df. \tag{3.14}$$

Autocorrelation function is usually calculated through the efficient fast Fourier transform (FFT).

Despite the robustness of the autocorrelation analysis in many contexts, it has some flaws that reduce its utility as a PDA. The autocorrelation function of a voiced speech usually displays a prominent peak at pitch period, but also other peaks are often present due to the formant structure. The strength of the other peaks can be reduced with various pre-processing techniques, but, nonetheless, errors due to other peaks are common for autocorrelation based PDAs. Another problem is the difficulty of selecting an appropriate window length. Since the window length should be two or three times the longest expected period, the optimal window size varies according to speaker. For high pitched speaker, the window size should be short (5–20 ms), whereas for low pitched speakers it should be long (20–50 ms). Autocorrelation analysis is also unreliable at speech segments with rapid changes in fundamental frequency.

Average magnitude difference function (AMDF) (Ross, Shaffer, Cohen, Freudberg & Manley 1974) is another way to express the similarity between periods, but the AMDF

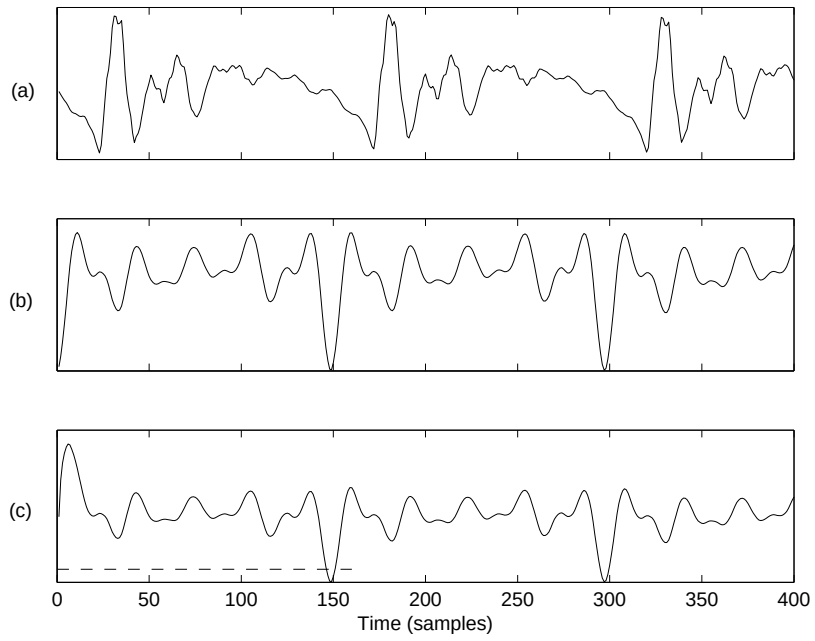


Figure 3.4: (a) Speech signal. (b) Average magnitude difference function (AMDF). (c) The cumulative mean normalized difference function. The threshold for the dip is depicted with a dashed line.

performs the comparison using differences rather than products. Probably the most popular method using AMDF is the YIN developed by de Cheveigne & Kawahara (2002). For signal x_n and window size w, the AMDF is defined as

$$d_n(\tau) = \sum_{j=1}^{w} (x_j - x_{j+\tau})^2 \tag{3.15}$$

(de Cheveigne & Kawahara 2002). In order to avoid the zero-lag dip and secondary dips due to resonances, YIN employs a cumulative mean normalized difference function:

$$d'_n(\tau) = \begin{cases} 1 & \text{if } \tau = 0, \\ d_n(\tau) / \frac{1}{\tau} \sum_{j=1}^{\tau} d_n(j) & \text{otherwise.} \end{cases} \tag{3.16}$$

The AMDF and the cumulative mean normalized difference function are shown in Figure 3.4. The fundamental period is found by setting an absolute threshold for the dip, and selecting the smallest value of τ being deeper than the threshold. Other improvements introduced in the YIN method include parabolic estimation and best local estimate, which further reduce errors.

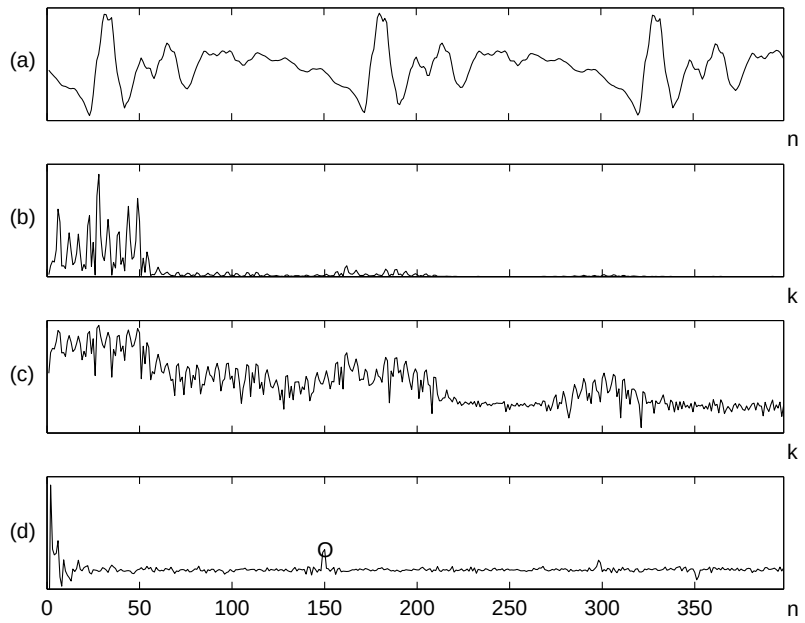


Figure 3.5: Stages in cepstrum analysis. (a) Speech signal. (b) Amplitude spectrum. (c) Logarithm of amplitude spectrum. (d) Cepstrum. The peak corresponding to the fundamental period is depicted with a circle. n and k represent discrete time samples and frequency bins, respectively. The length of the speech segment is 25 ms, and the spectrum is shown from 0 to 8000 Hz.

3.2.2 Frequency Domain Approaches

Since periodic signals tend to be composed of harmonically related partials, information about f_0 can be extracted by examining the partials. Frequency domain methods are based on detecting these partials. Most methods utilize Fast Fourier Transform (FFT) to convert the signal to a frequency spectrum. Other methods may use a comb-filter or a filter bank to find the partials (Gerhard 2003). One particularly useful method utilizes cepstrum to reveal signal periodicity. Cepstrum was first described by Bogert, Healy & Tukey (1963) and first used in speech analysis by Noll (1964). Cepstrum was originally defined as the Fourier transform of the logarithm of the amplitude spectrum of a signal, but today inverse Fourier transform is commonly used instead of Fourier transform. Because the spectrum of a periodic signal has regularly spaced peaks, or harmonics, the spectrum of that signal shows a peak at the period of the original waveform. The logarithm is taken to transform the original spectrum to such a form that the dynamics of the speech signal is properly

represented. In addition to the fundamental frequency estimation, cepstrum is also used for representing the spectral envelope of speech (see for example Imai (1983)). Figure 3.5 shows the different stages in cepstrum analysis.

Many other time and frequency domain methods exist, but the discussed methods are probably the most widely used in speech processing. For more information about f_0 estimation, see for example Hess (1983), Paliwal & Kleijn (1995), and Gerhard (2003).

3.3 Glottal Inverse Filtering

Glottal Inverse Filtering (GIF) is a procedure where the glottal source signal, the glottal volume velocity waveform, is estimated from a voiced speech signal. The basic idea of GIF is to separate the source and the filter based on the linear speech production model described in Section 2.3.

The estimation of the glottal source has many applications, for example in speech analysis, speech synthesis, and the study of laryngeal pathology. Since the presentation of the idea of glottal inverse filtering by Miller (1959) many different methods have been developed. Depending on the procedure that is used in recording the speech signal, inverse filtering methods can be divided into two categories. The first category consists of methods in which a specially designed pneumotachograph mask, a Rothenberg's mask (Rothenberg 1973), is used to record the speech signal. The second category consists of methods in which the speech signal is recorded with a microphone in free field outside the mouth. The microphone and other recording equipment must not cause phase distortion to the speech signal in order to get feasible results from inverse filtering. Also, for methods in the second category, the effect of lip radiation must be taken into account. After the recording, the vocal tract parameters can be estimated either by hand or automatically. The use of automatic estimation instead of estimation by hand is justified since the automatic estimation is fast and easy to use, and the estimation by hand may vary subjectively. This section will be primarily concerned with automatic inverse filtering methods applied to speech signals recorded outside the mouth.

According to the linear speech production model, the speech can be described by the equation

$$S(z) = E_g(z)V(z)L(z), \tag{3.17}$$

where $S(z)$ is the speech signal, $E_g(z)$ the glottal volume velocity waveform, $V(z)$ the vocal tract transfer function, and $L(z)$ the lip radiation model. The glottal volume velocity waveform $E_g(z)$ corresponds to $E(z)G(z)$ in Equation 2.1. In glottal inverse filtering, the speech signal is first analyzed to determine the parameters of the vocal tract transfer function. Then, the effect of the vocal tract can be canceled by filtering the speech signal through the inverse

model of the vocal tract transfer function, the inverse filter. Finally, the lip radiation effect is canceled from the inverse filtered signal. Glottal inverse filtering is conceptually defined as solving the glottal volume velocity $E_g(z)$ by the equation

$$E_g(z) = \frac{S(z)}{V(z)L(z)}. \tag{3.18}$$

Since the lip radiation $L(z)$ can be considered to be the same for all speech sounds, only the parameters of the vocal tract transfer function $V(z)$ are required.

3.3.1 Iterative Adaptive Inverse Filtering

Iterative adaptive inverse filtering (IAIF) is an automatic glottal inverse filtering method developed by Alku (1992). The only input required to the system is the acoustical speech signal recorded with a microphone. The method is completely automatic, and can be implemented to run real time. The main tool of the method is LPC, described in Section 3.1. The estimated glottal flow is obtained by canceling the effects of the vocal tract and the lip radiation. The IAIF method has been extensively studied by Pulakka (1995), and the results suggest that IAIF produces reasonable estimates of the glottal volume velocity waveform.

Figure 3.6 shows the block diagram of the IAIF method. The method consists of the following stages. First (block 1), the signal $s(n)$ is high-pass filtered to remove any distorting low-frequency fluctuation in the speech signal. The high-pass filter is a linear FIR filter with a cut-off frequency of 60 Hz. Second (block 2), a first order LPC is computed for the signal. This yields a first order preliminary estimate for the combined effect of the glottal flow and the lip radiation. This is denoted as $H_{g1}(z)$. Next (block 3), the estimated effects of the glottal flow and the lip radiation are canceled from the speech signal by inverse filtering. The resulting signal is then analyzed using a pth order linear prediction in order to obtain an estimate for the effect of the vocal tract. The vocal tract effect is denoted as $H_{vt1}(z)$. The order of the LPC analysis, p, depends on the sampling frequency, and for speech sampled at 16 kHz, appropriate values are typically between 16 and 24. Next (block 5), the effect of the vocal tract is canceled by inverse filtering the speech signal with the obtained model. The first estimate of the glottal flow is obtained (block 6) by canceling the effect of the lip radiation from the inverse filtered speech signal. The inverse of the lip radiation is modeled as integration. A new estimate for the contribution of the glottal flow on the speech spectrum, denoted by $H_{g2}(z)$, is computed in block 7. The order of the LPC analysis g is typically between 4 and 8 for speech sampled at 16 kHz. A new model of the vocal tract filtering effect is obtained through canceling the glottal contribution (block 8) and lip radiation (block 9), and LPC analysis of order p (block 10). Canceling the effects of the estimated vocal tract and the lip radiation by inverse filtering (block 11) and integrating

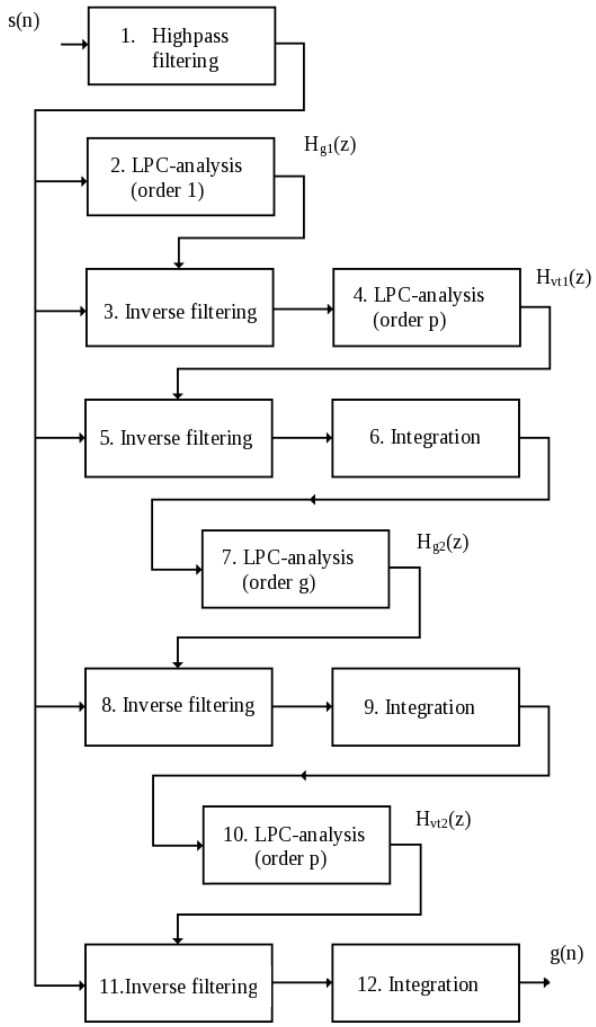


Figure 3.6: Block diagram of the IAIF method for estimating the glottal excitation $g(n)$ from the speech signal $s(n)$. The model for the vocal tract is estimated through iterative procedure (blocks 2–10). The estimated glottal flow is obtained by canceling the effects of the vocal tract (block 11) and the lip radiation (block 12) from the speech signal. (Alku et al. 1999)

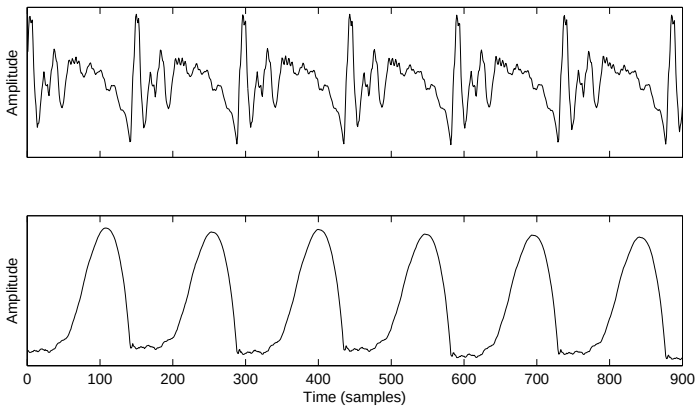


Figure 3.7: (a) Sustained vowel [a] produced by a male speaker using normal phonation. (b) Corresponding glottal flow estimated with IAIF.

(block 12) the speech signal, a final estimate for the glottal volume velocity waveform $g(n)$ is obtained. An example of a speech signal and corresponding glottal flow estimated with IAIF is shown in Figure 3.7.

3.4 Parametrization of Glottal Flow

Parametrization of the glottal volume velocity waveform is an essential part of voice source analysis. After the glottal flow has been estimated by some inverse filtering method, the source signal is parametrized by quantifying the obtained waveforms with properly selected numerical values. These quantities, the glottal flow parameters, aim to represent the most important features of the glottal flow in a compressed numerical form. Various parametrization methods focus on different features of the glottal flow, and the selection of the parametrization method for a certain purpose is crucial in order to extract the desired information.

Parametrization of the glottal flow can be applied in the following three partly overlapping areas. First, the most general application is the categorization of the voice source, i.e. dividing the speech sounds into various categories according to the different modes of voice source. Second, parametrization can be used in the study of vocal disorders. Third, the parametrization methods can also be applied to voice and speech synthesis.

The parametrization methods can be roughly divided into two categories: time domain and frequency domain methods. Time domain methods can be further divided into time-based and amplitude-based parameters. Time-based parameters can be used with any inverse filtering method, whereas for amplitude-based parameters, a properly calibrated Rosen-

berg's mask is required. Since the amplitude-based parameters are not relevant concerning the topic of this thesis, speech synthesis, these parameters are not further discussed. Amplitude-based parameters are discussed generally for example in Alku (2003). Additionally, a third category of time domain methods includes techniques that model the whole glottal waveform by fitting certain predefined mathematical functions to the glottal flow.

3.4.1 Time Domain Parameters

In time domain parametrization methods, both glottal flow and its derivative can be used to extract the desired parameters. One cycle of glottal flow and its derivative obtained by inverse filtering, with the most essential notations used in time domain parameterization methods are shown in Figure 3.8. The most widely used time domain glottal flow parameters are open quotient (OQ), speed quotient (SQ), and closing quotient (ClQ). Other time domain parameters used in voice source studies include closed quotient (CQ), which is sometimes used instead of ClQ, return quotient (RQ), and normalized amplitude quotient (NAQ). Using the notations in Figure 3.8, these parameters are defined as

$$\begin{aligned} OQ &= (t_o + t_{cl})/T \\ SQ &= t_o/t_{cl} \\ ClQ &= t_{cl}/T \\ CQ &= t_c/T \\ RQ &= t_{ret}/T \\ NAQ &= ac/(d_{peak}T). \end{aligned} \tag{3.19}$$

Many studies have been carried out to find out the relations of these six time domain parameters to various speech production features, such as loudness, pitch, phonation type, and gender, but the results are not unambiguous. The behavior of the time domain parameters are studied for example in Holmberg, Hillman & Perkell (1988), Sulter & Wit (1996), Price (1989), Alku, Bäckström & Vilkman (2002), and Bäckström, Alku & Vilkman (2002). A good general review of the time domain parameters is given by Alku (2003).

In addition to the previously discussed numerical measures, it is also possible to model the whole glottal volume velocity waveform by defining an artificial waveform that fits the glottal flow or its derivative obtained by inverse filtering. In this type of parametrization, an appropriate mathematical model for the glottal flow is defined, after which the model parameters are optimized in order to get the best possible match between the model and the real glottal flow. The models for the glottal flow have evolved much from the simple sawtooth waveforms and filtered impulse trains used in early models. Commonly the modern glottal flow models are composed of piecewise continuous functions composed of sinusoidal, exponential or polynomial terms. For example, a fairly primitive pulse model

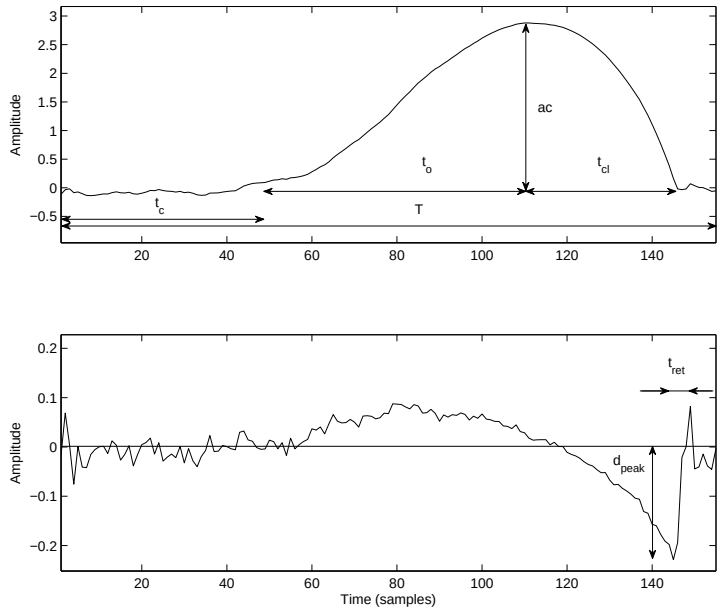


Figure 3.8: One cycle of glottal flow (upper) and its time derivative (lower). The following time domain notations are depicted: fundamental period (T), closed phase (t_c), opening phase (t_o), closing phase (t_{cl}), return phase (t_{ret}), AC flow (ac), and the negative peak amplitude of the derivative (d_{peak}). The original speech signal is a sustained vowel [a] produced by a male speaker using normal phonation.

used in the Klattalk synthesizer (Klatt 1987) is composed of a single third order polynomial

$$E(t) = at^2 - bt^3, \tag{3.20}$$

where t is time, and terms a and b are defined according to the desired amplitude and shape of the pulse. The closed period is simply padded with zeros. Figure 3.9 shows one period of the Klatt model and its derivative. The waveform model for the main excitation at the discontinuity at glottal closure is very simple, since the return phase of the derivative, which accounts for the degree of spectral tilt, cannot be controlled.

One of the most widely used glottal flow models is the Liljencrants-Fant model (LF model) (Fant, Liljencrants & Lin 1985). In this model, the glottal flow derivative is presented by sinusoidal and exponential terms defined uniquely by four parameters. An illustration of a typical LF model pulse and its derivative are presented in Figure 3.10. The first part of the glottal flow derivative is modeled with an exponentially increasing sinusoid that starts at the opening instant of the vocal folds, t_0, and ends at the instant of the maximum

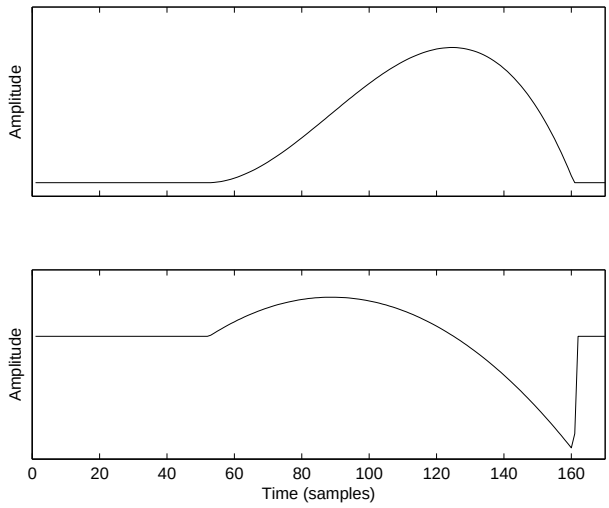


Figure 3.9: Illustration of the Klatt model for the glottal flow pulse (upper) and its time derivative (lower).

negative amplitude t_e. The second part is modeled with a function consisting of exponential terms. After reaching the value E_e, the pulse abruptly returns to zero with time constant t_a, which models the closure of the vocal folds after the abrupt flow termination. The time constant t_a is defined as the duration between t_e and the time when the tangent of the exponential at time t_e hits zero. The exponential part ends in the zero at time t_c. The LF model is defined as

$$E(t) = \begin{cases} E_0 e^{\alpha t} sin(\omega_g t), & t < t_e \\ \frac{-E_e}{\varepsilon t_a}\left(e^{-\varepsilon(t-t_e)} e^{-\varepsilon(t_c-t_e)}\right), & t_e < t < t_c \\ 0, & t_c < t < T, \end{cases} \tag{3.21}$$

where $\omega_g = \pi/t_p$ and $t_c = T = 1/f_0$. Parameters α and ε can be calculated from Equation 3.21 by assuming $E(t_e) = E_e$ and the energy balance $\int_0^T E(t) = 0$. Thus, parameters t_p, t_e, t_a, and E_e uniquely define the model. For detailed derivation of the model, see for example Fant et al. (1985). Another way of defining the model is to use the following notations

$$\begin{aligned} R_g &= 1/(2t_p) \\ R_k &= t_e/t_p - 1 \\ R_a &= t_a/t_0. \end{aligned} \tag{3.22}$$

(Fant, Kruckenberg, Liljencrants & Båvegård 1994). In addition, a basic waveshape parameter for the pulse can be defined as

$$R_d \approx (0.5 + 1.2R_k)(R_k/(4R_g) + R_a)/0.11. \tag{3.23}$$

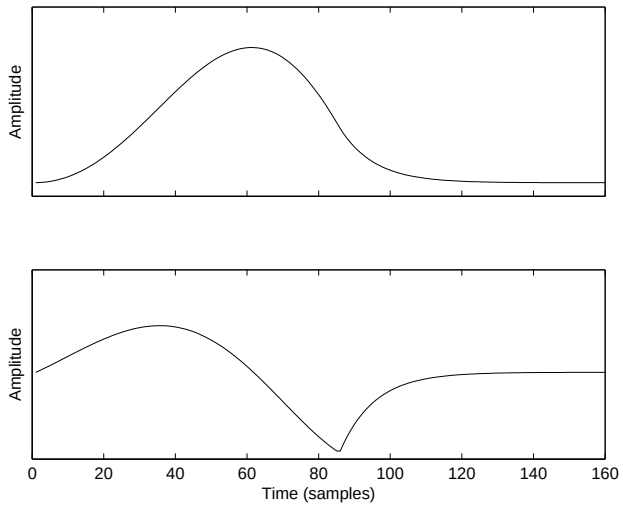


Figure 3.10: Illustration of a typical LF model pulse (upper) and its time derivative (lower).

LF model is commonly used in inverse filtering combined with automatic fitting of the model parameters. The fitting can be performed by matching the time domain pulse waveform or by comparing the spectrum of the model and the original pulse. The properties of LF model have been studied in numerous papers. Basic characteristics and frequency domain properties are described for example in Fant (1995). The new control schema for the model, which substantially simplifies the description of the voice source rules for example in text-to-speech synthesis, was presented in Fant et al. (1994) and further explained in Fant (1995). Importantly, the parameter R_d is able to represent the voice quality characteristics in an effective single numerical measure. In normal covariation of voice source parameters it is possible to define the LF model with unique value of R_d, or conversely, it is possible to predict the values of R_g, R_k, and R_a from R_d.

In addition to Klatt and LF models, several other models have been proposed, for example in Rosenberg (1971), Rothenberg, Carlson, Granström & Gauffin (1975), Fant (1979), Hedelin (1984), Ananthapadmanabha (1984), and Fujisaki & Ljungqvist (1986).

3.4.2 Frequency Domain Parameters

The time domain changes in the glottal flow, for example changes in the phonation type from breathy to pressed, correspond to the changes in the spectral decay of the power spectrum of the voice source. Therefore, the frequency domain methods are developed to parametrize the spectral decay of the power spectrum. The spectrum can be evaluated with FFT or with all-pole modeling, either pitch-synchronously or over several fundamental

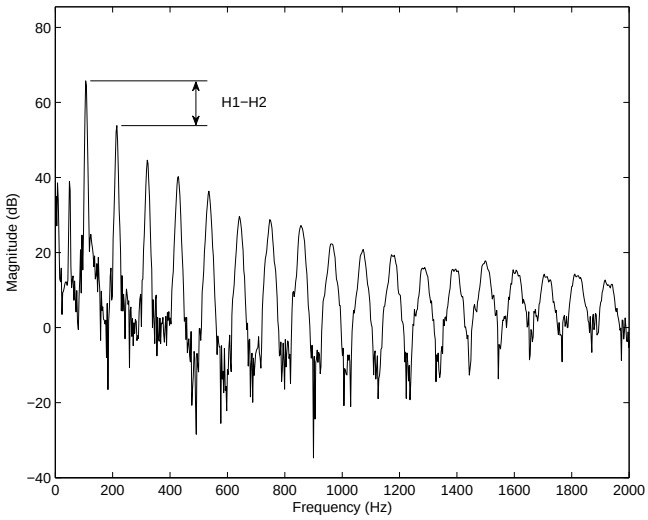


Figure 3.11: Spectral decay of the voice source spectrum quantified by H1–H2.

periods. For measuring the spectral decay, Childers & Lee (1991) have proposed a quotient called Harmonic Richness Factor (HRF). It is defined as the ratio between the sum of the amplitudes of harmonics above the fundamental and the amplitude of the fundamental, i.e.

$$HRF = \frac{\sum_{i \geq 2} H_i}{H_1}, \tag{3.24}$$

where H_i is the amplitude of the ith harmonic and H_1 is the amplitude of the fundamental. With this quotient, the vocal fry was characterized by a high HRF value (2.1 dB), modal voices with a medium HRF value (-9.9 dB), and breathy voices with a low HRF value (-16.8 dB). For the same purpose, Titze & Sundberg (1992) measured the difference between the amplitude of the fundamental and the second harmonic. This measure, usually denoted by H1–H2, has been widely used as a measure of vocal quality. The idea of H1–H2 is demonstrated in Figure 3.11. It has been shown that H1–H2 has a large correlation with CQ (Fant 1995), and a linear equation for the relation between H1–H2 and the LF model parameter R_d is derived in Fant (1995). Also linear regression (Howell & Williams 1988, Howell & Williams 1992) and Parabolic Spectral Parameter (PSP) (Alku, Strik & Vilkman 1997) have been proposed to model the spectral slope of the glottal flow.

3.4.3 Voice Source Models in Speech Synthesis

It has been known for some time that the voice source characteristics are especially important for the quality of speech. The earliest synthesizers used a train of impulses or trian-

gular pulses as an excitation signal, resulting in a harsh speech quality. More sophisticated speech synthesizers have tried to model the glottal volume velocity waveform, such as the KLGLOTT88 model in KLSYN88 synthesizer (Klatt & Klatt 1990). LF model pulses have been used in speech synthesis experiments for example by Carlson, Fant, Gobl, Granström, Karlsson & Lin (1989) and Carlson, Granström & Karlsson (1991), and for example Cabral, Renalds, Richmond & Yamagishi (2007) have combined the LF model to an HMM-based speech synthesizer. However, the use of artificial glottal flow pulses usually results in a somewhat buzzy quality due to strong harmonic structure at higher frequencies compared to speech with natural glottal flow signal. To overcome this problem, natural glottal flow pulses extracted from speech by inverse filtering have been proposed as a voice source. For example, natural glottal flow pulses have been used in creating natural sounding speech stimuli for speech research by Alku et al. (1999). Natural glottal flow pulses have also been used in formant speech synthesis by Matsui, Pearson, Hata & Kamai (1991).

3.5 Hidden Markov Models

Hidden Markov Models (HMMs) are statistical models which can be applied to modeling of various types of sequential data. For example in speech synthesis and recognition, HMMs have been used with great success. HMMs were first described in a series of publications in the late 1960s and early 1970s, but widespread understanding and application of the theory of HMMs to speech processing begun not until the late 1980s. Today HMMs are widely used in many fields, and the popularity is ever increasing.

A hidden Markov model can be described as a finite state machine which generates a sequence of time observations. A time observation is generated by first making a decision to which state to proceed, and then generating the observation according to the probability density function of the current state. The system modeled by an HMM is assumed to be a Markov process, in which the probability of a state transition depends only on the path of the past states. This characteristic is called the Markov property. Formally, HMM is a doubly stochastic process consisting of underlying stochastic process that is not observable (hidden), but can be observed through another set of stochastic processes that produce the sequence of observations. This means that the stochastic function of HMM is a result of two processes, one of which is the underlying hidden Markov chain having a finite number of states, and another being the set of random processes associated with each state. At discrete time instant, the process is assumed to be at some state and an observation is generated by the stochastic process of the current state. The underlying Markov chain changes states with time according to the state transition probability matrix. In principle, the underlying Markov chain can be of any order, and the outputs may be multivariate random processes.

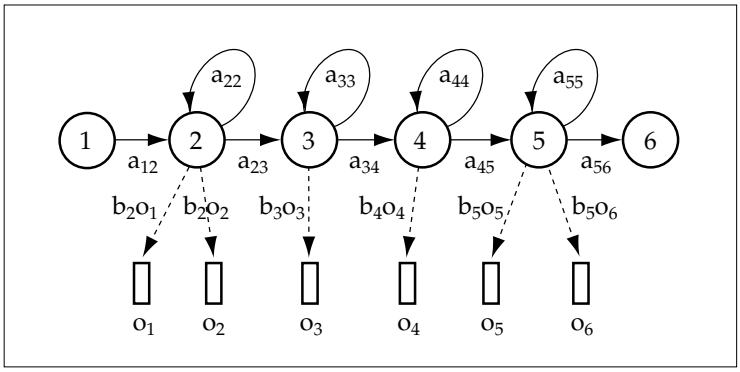


Figure 3.12: Example of an HMM structure. The states of the HMM are denoted with circles numbered from one to six. A state transitions probability from state i to state j is denoted as a_{ij}. An output probability density of state i is denoted as b_i, and the generated observation at time instant t is denoted as o_t. (Karjalainen 2000)

An illustration of a 6-state left-to-right HMM structure is shown in Figure 3.12, in which the state index increases or stays the same with each time step. Generally, left-to-right HMM structures are used for modeling systems whose properties evolve in a successive manner, such as speech and written language.

An N-state HMM is defined by a state transition probability distribution $\mathbf{A} = \{a_{ij}\}_{i,j=1}^{N}$, output probability distribution $\mathbf{B} = \{b_j(\mathbf{o})\}_{j=1}^{N}$, and initial state probability distribution $\Pi = \{\pi_i\}_{i=1}^{N}$, where a_{ij} is the state transition probability from state q_i to state q_j and $\mathbf{o}$ is the observation vector. A compact notation for the set of model parameters is represented as $\lambda = (\mathbf{A}, \mathbf{B}, \Pi)$.

There are basically three problems associated with HMMs:

1. Given the observation sequence $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T)$ and a model $\lambda = (\mathbf{A}, \mathbf{B}, \Pi)$, how to efficiently calculate $P(\mathbf{O}|\lambda)$, the probability of the observation sequence, given the model?

2. Given the observation sequence $\mathbf{O} = (\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_T)$ and the model λ, how to choose a corresponding optimal state sequence $\mathbf{Q} = (q_1, q_2, ..., q_T)$?

3. How to adjust the model parameters $\lambda = (\mathbf{A}, \mathbf{B}, \Pi)$ to maximize $P(\mathbf{O}|\lambda)$?

The first problem is used for finding the probability that the observed sequence was produced by the given model. On the other hand, it can be also used to score different models on how well they match the given observation sequence. The probability can be calculated by the equation

$$P(\mathbf{O}|\lambda) = \sum_{\text{all Q}} P(\mathbf{O}|\mathbf{Q}, \lambda) P(\mathbf{Q}|\lambda). \tag{3.25}$$

The direct calculation of $P(\mathbf{O}|\lambda)$ is straightforward, but it involves on the order of $2TN^T$ calculations. Thus, the problem is usually evaluated with the Forward-Backward algorithm (see for example (Rabiner 1989)), which requires only N^2T calculations. The most widely used criterion for optimal state sequence for problem 2 is to find the single best state sequence that maximizes the $P(\mathbf{Q}|\mathbf{O},\lambda)$. This can be solved with the Viterbi-algorithm (Viterbi 1967). The third problem is the most difficult one. No analytical solution is known for solving the model which maximizes the probability of the observation sequence. However, iterative algorithms, such as the Baum-Welch algorithm (Baum, Petrie, Soules & Weiss 1970), or equivalently Expectation-Maximization (EM) algorithm (Dempster, Laird & Rubin 1977)), and gradient based algorithms can be used for maximizing $P(\mathbf{O}|\lambda)$.

Hidden Markov Models can be extended with various features to make the use of them more versatile and efficient. For example autoregressive HMMs, inclusion of null transitions, state tying, state duration densities and various optimization criteria have been proposed. Useful features in HMM-based speech synthesis are described in Chapter 4. For more information about HMMs in general, see for example Rabiner (1989) and Rabiner & Juang (1993).

Chapter 4

HMM-Based Speech Synthesis System

In this chapter, a new HMM-based text-to-speech system is represented. First, a general overview of the synthesizer is given, after which the operational principles and the implementation of the synthesizer are described in more detail.

4.1 System Overview

In this work, a new HMM-based text-to-speech system utilizing glottal inverse filtering is implemented. The main goal of this new TTS system is to produce natural sounding synthetic speech capable of conveying different styles of speaking as well as emotions. In order to achieve this goal, the function of the real human voice production mechanism is modeled with the help of glottal inverse filtering embedded in an HMM framework. Automatic glottal inverse filtering is used in the parametrization stage in order to compute a parametric feature expression for the voice source and the vocal tract transfer function. The extracted parameters are fed into an HMM system for training and then generated from the trained HMM according to text input. In the synthesis stage, natural glottal flow pulses are used for generating the source signal for voiced sounds, and the spectral envelope of this glottal excitation is modified with an adaptive IIR filter to imitate the time-varying changes in the real voice source. The current implementation of the system is applied for Finnish, but, in principle, it can be extended to other languages as any data driven synthesizer.

The overview of the system is shown in Figure 4.1. The system consists of two major parts: training and synthesis. In the training part, speech parameters computed by glottal inverse filtering are extracted from sentences of a speech database. This parametrization stage is a major innovation in the new TTS system in comparison to previous HMM-based

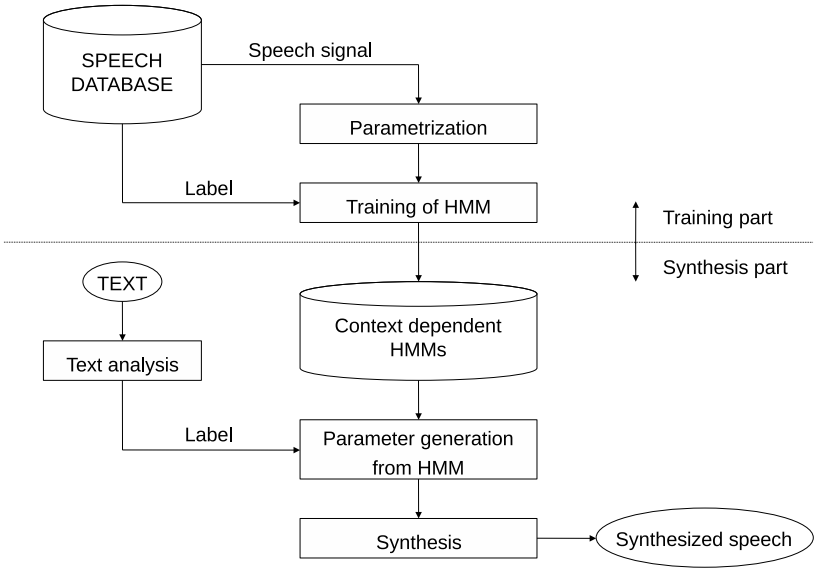


Figure 4.1: Overview of the HMM-based text-to-speech system.

synthesizers. The obtained speech parameters are then modeled in the framework of the HMM. In the synthesis part, the HMMs are concatenated according to the analyzed input text and speech parameters are generated from the HMM. The parameters are then fed into the synthesis module for creating the speech waveform.

4.2 Training Part

The goal of the training stage is to create a model of the speaker by parametrizing a large database of speech which syntax and phonemic content is labeled, and then training the HMM system with the parameters.

4.2.1 Speech Parametrization

The purpose of the parametrization stage is to compress the information of the speech signal to a few parameters which would describe the essential characteristics of the original speech signal as accurately as possible. A very efficient way of parametrization is to separate the speech signal to source and filter. In speech synthesis, the speech signal is usually separated artificially to source signal and filter coefficients that do not correspond to the real glottal flow and the vocal tract filter. This approach has the downside that it is very hard to model the real mechanisms of speech production due to the artificial nature of the signal and the parameters. In the case of separating the speech signal into quantities that correspond to

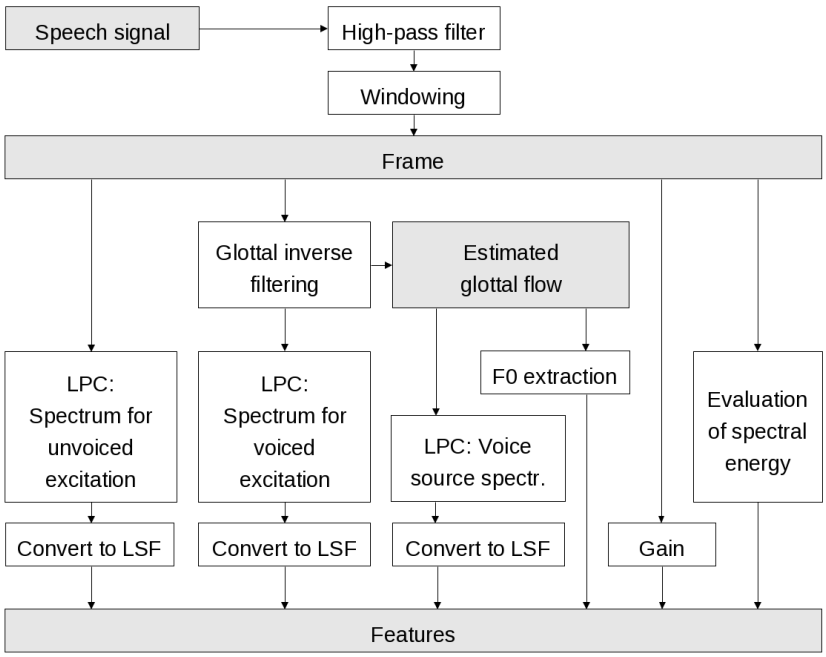


Figure 4.2: Flow chart of the speech parametrization stage.

real phenomena, it is easier to model the speech production mechanism in the framework of HMM, and thus produce more natural sounding synthetic speech. Therefore, glottal inverse filtering is chosen for the core of the new implemented system.

The flow chart of the parametrization stage is shown in Figure 4.2. First, the speech signal is high-pass filtered in order to remove any distorting low-frequency fluctuations. The high-pass filter is a linear phase FIR filter with a cut-off frequency of 60 Hz. The high-pass filtering is especially important for glottal inverse filtering, where even weak low-frequency components may cause extensive fluctuations in the estimated glottal flow. After the high-pass filtering, the signal is windowed with a rectangular window to 25-ms frames at 5-ms intervals. The mean of each frame is first removed to ensure zero DC component within the frame. The parameters are then extracted from each frame.

The core of the parametrization stage is the glottal inverse filtering that estimates the glottal volume velocity waveform from the speech pressure signal. An automatic inverse filtering method, iterative adaptive inverse filtering (IAIF) described in Section 3.3 is utilized in the system. The IAIF iteratively cancels the effects of the vocal tract and the lip radiation from the speech signal using adaptive all-pole modeling. The glottal inverse filtering is applied successively to the 25-ms rectangular frames, revealing the corresponding glottal volume velocity waveform. LPC algorithm used in the inverse filtering is imple-

mented using the autocorrelation method, and LU decomposition is used for solving the Normal Equations (3.6). In calculating the LPC, the frame is windowed using the Hann window, defined as

$$w(n) = 0.5 \left(1 - \cos\left(\frac{2\pi n}{N-1}\right)\right). \tag{4.1}$$

The filtering of each frame is initiated with samples preceding the actual frame in order to prevent the discontinuity due to the filtering delay. For canceling out the lip radiation effect, a leaky integrator is used, defined in the Z-domain as

$$H(z) = \frac{1}{1 - \rho z^{-1}}, \tag{4.2}$$

where ρ is a value near one. The value of ρ can be defined by the user, but in the experiments, value $\rho = 0.99$ was used.

The extracted features and the number of parameters per frame are presented in Table 4.1. The parameters can be divided into source and filter parameters. For creating the voice source, fundamental frequency, energy, spectral energy, and voice source spectrum are extracted. For creating the formant structure corresponding to the vocal tract filtering effect, spectra for voiced and unvoiced speech sounds are extracted. Separate spectra for voiced and unvoiced excitation are extracted since the vocal tract transfer function as such does not generate appropriate spectral envelope for unvoiced speech sounds. The extracted features are further explained in the following paragraphs.

The outputs of the glottal inverse filtering block are the estimated glottal flow and the LPC model of the vocal tract (denoted by Voiced spectrum in Table 4.1). A sufficient order of the LPC model for the vocal tract is approximately 20. The spectral envelope of the resulting glottal flow is parametrized with LPC (denoted by Voice source spectrum in Table 4.1) in order to model the spectral characteristics of the voice source in the synthesis stage. An appropriate degree of the LPC analysis for the glottal flow is between 8 and

Table 4.1: Speech features and the number of parameters.

Feature	Parameters per frame
Fundamental frequency	1
Energy	1
Spectral energy	5
Voice source spectrum	10
Voiced spectrum	20
Unvoiced spectrum	20

12. Further increasing the degree does not necessarily improve the quality of the synthesis, because the minor changes in the source spectrum may not originate from the real glottal phenomena, but might stem from the errors due to the framing or imperfect glottal inverse filtering. Moreover, the slightest changes in the voice source spectrum might not be consistent enough with the context in order efficiently train the HMM system. Additionally, an LPC model (denoted by Unvoiced spectrum in Table 4.1) is computed for unvoiced speech sounds directly from the speech frame. A sufficient order of the LPC model for the unvoiced spectrum is approximately 20.

All the obtained LPC models are converted to LSFs, a parametric representation of LPC information well-suited to be used in a statistical HMM system. The trivial zeros of the LSP polynomials are removed by deconvolution, and for finding the roots of the LSPs, Chebyshev transform is utilized in order to make the algorithm more efficient. LSFs of voiced and unvoiced spectrum are further converted to the mel scale in order to perceptually emphasize the learning of the low frequencies by the HMM algorithm. The conversion of frequency to mel-scale is defined as

$$f = 700 \left(e^{m/1127.01048} - 1 \right). \tag{4.3}$$

Since the fundamental frequency of speech originates from the glottal vibrations, it is easy to extract the f_0 of the frame from the glottal volume velocity waveform. However, the frame of obtained glottal flow for extracting the source and vocal tract characteristics is only 25 ms in duration, which makes the extraction of the fundamental frequency below 80 Hz unreliable. Therefore another glottal inverse filtering is performed with a 50-ms window, which enables the reliable extraction of f_0 values up to 40 Hz. The fundamental frequency extraction is performed by evaluating the autocorrelation function from the glottal, and finding the highest peak of the ACF. A predefined amount of samples is removed at the beginning of the ACF in order to avoid the selection of the zero lag peak. The resulting f_0 is also verified to fit into a predefined range of valid fundamental frequencies, or otherwise the frame is marked as unvoiced. In the present experiment, the range was set to cover the frequencies from 30 to 260 Hz. The voiced or unvoiced decision is additionally made according to the amount of low-frequency energy in the frame and the zero-crossing rate (ZCR). The low-frequency energy was evaluated in the range of 0–1000 Hz. If the energy is below or the ZCR value exceeds the predefined limits, the frame is determined as unvoiced. In order to reduce the occasional errors in fundamental frequency estimation, a 3-point median filtering is applied to the f_0 contour.

The energy of the speech is evaluated by the sum of squares of the samples in the original 25-ms frame. In addition, the spectral energy of five bands (0–1000 Hz, 1000–2000 Hz, 2000–4000 Hz, 4000–6000 Hz, and 6000–8000 Hz) is calculated from the speech frame

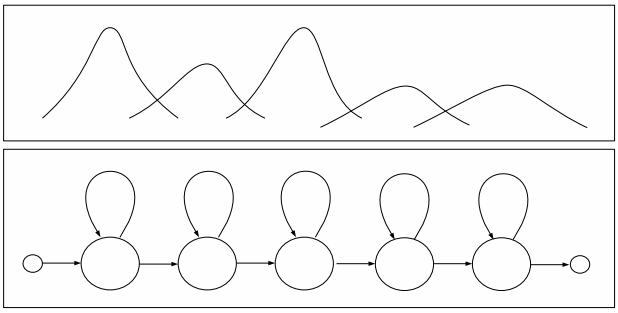


Figure 4.3: Illustration of a 7-state left-to-right context dependent HMM structure with 5 emitting states (lower), and the corresponding state duration model (upper).

with FFT for determining the unvoiced excitation.

4.2.2 Training of HMM

After the parametrization, the obtained speech features are modeled simultaneously in a unified framework of HMM. First, monophone HMM models are trained. A 7-state left-to-right HMM structure with 5 emitting states is used. All parameters excluding the fundamental frequency are modeled with continuous density HMMs by single Gaussian distributions with diagonal covariance matrices. The fundamental frequency is modeled by a multi-space probability distribution (MSD-HMM) (Tokuda, Masuko, Miyazaki & Kobayashi 1999). The conventional HMM modeling cannot be applied since the observation sequence of the fundamental frequency is composed of continuous values and discrete symbols that represent unvoiced frames. The state durations for each phoneme HMM are modeled with multi-dimensional Gaussian distributions (Yoshimura, Tokuda, Masuko, Kobayashi & Kitamura 1998). The HMM structure and its state duration model are illustrated in Figure 4.3. In the current system, each feature is modeled in an individual stream, and for the fundamental frequency three streams are used due to the MSD-HMM, resulting in a model of eight streams. The delta and delta-delta coefficients of each feature are calculated in order to enable smooth transitions between states in parameter generation, resulting in a feature order of 171 in total. Otherwise the training procedure is similar to that described in Tokuda, Zen & Black (2002).

After the training of the monophone HMMs, various contextual factors are taken into account and the monophone models are converted into context dependent models. As the number of the contextual factors increases, their combinations also increase exponentially. Due to the limited amount of training data, model parameters cannot be estimated with sufficient accuracy. To overcome this problem, the models for each feature are clustered

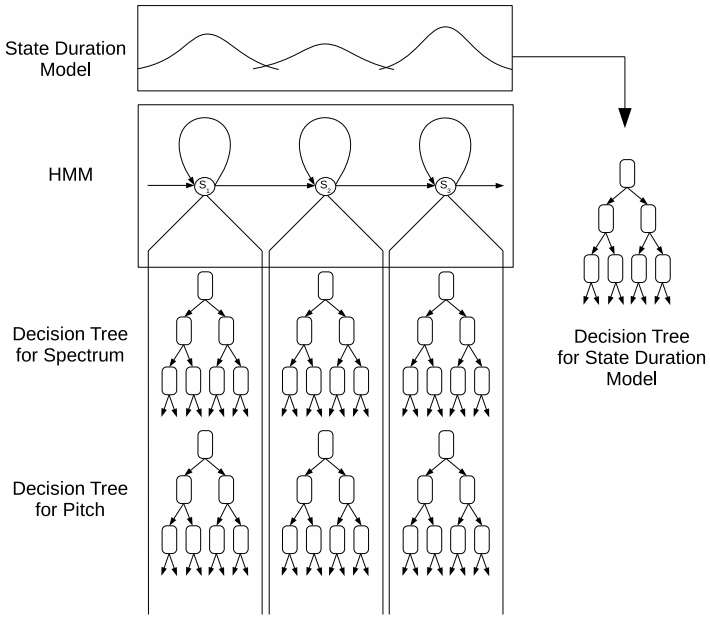


Figure 4.4: Illustration of the decision-tree based context clustering for spectrum, pitch, and state duration model. Context clustering is performed for all speech features.

independently by using a decision-tree based context clustering technique (Odell 1995). The clustering is also required in order to generate synthesis parameters for new observation vectors that are not included in the training material. The decision-tree based context clustering is illustrated in Figure 4.4. The contextual factors that are taken into account in the current model are described in Table 4.2.

4.3 Synthesis Part

In the synthesis part, the model created in the training part is used for generating the speech parameters according to text input. The parameters are then fed into the synthesis module for generating the speech waveform.

4.3.1 Speech Parameter Generation

In order to generate speech parameters according to the text input, first, phonological and high-level linguistic analysis are performed, where the text input is converted to a context-based label sequence. According to the label sequence and the decision trees generated by the training stage, a sentence HMM is constructed by concatenating context dependent HMMs. State durations of the sentence HMM are determined so as to maximize the like-

Table 4.2: Contextual factors used in the current implementation of the synthesizer. The included sources of the contextual factors are marked as LL = two unit left, L = left unit, C = current unit, R = right unit, RR = two units right.

Description	Context
Phoneme length	L, C, R
Phoneme identity	LL, L, C, R, RR
Mora index	C
Phoneme position in syllable	L, R
Phoneme position in word	L, R
Phoneme context	L, C, R
Mora index in syllable	L, C, R
Phoneme count in syllable	L, C, R
Syllable position in word	L, R
Syllable position in phrase	L, R
Syllable position in utterance	L, R
Syllable stress	C
Syllable accent	L, C, R
Syllable distance to focused syllable	L, R
Syllable distance to accentuated syllable	L, R
Syllable distance to stressed syllable	L, R
Syllable distance to break	L, R
Strength of nearest breaks	L, R
Content or functional word	L, C, R
Word distance to focused word	L, R
Word distance to accentuated word	L, R
Word accent	L, C, R
Word position in phrase	L, R
Word position in utterance	L, R
Syllable count in word	L, C, R
Syllable count in phrase	C
Phrase position in utterance	L, R
Word count in utterance	C
Phrase count in utterance	C
Does the utterance start a topic	C

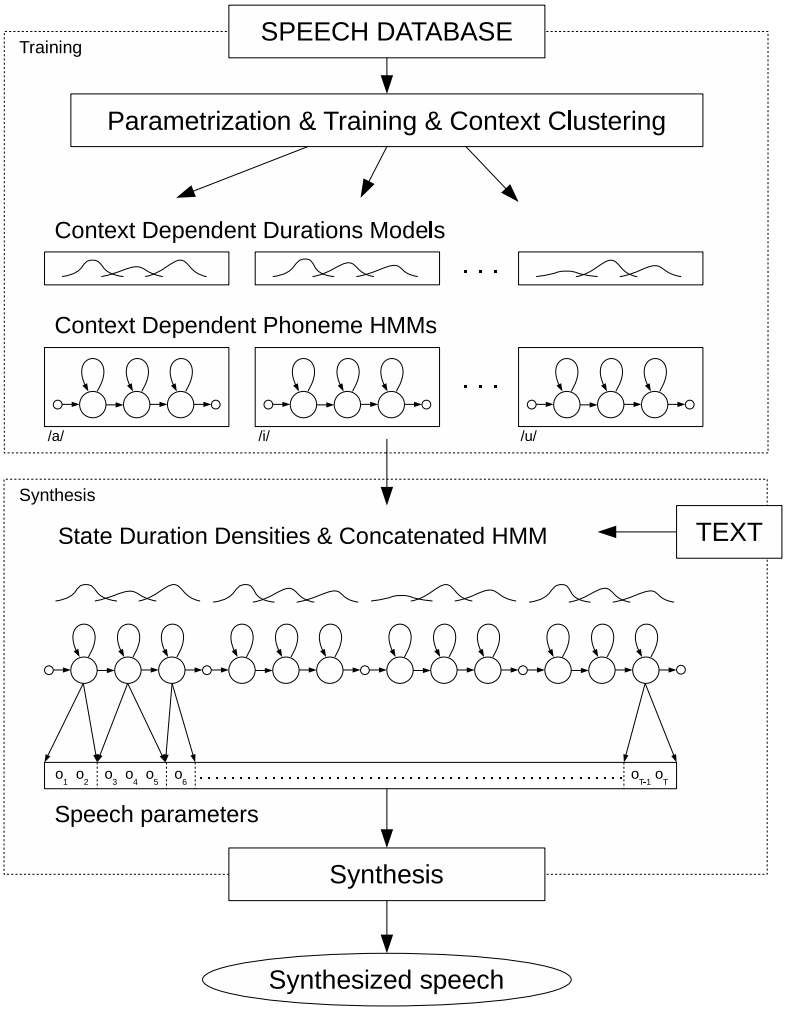


Figure 4.5: Illustration of the HMM-based generation process of speech parameters ranging from training stage to waveform generation.

lihood of the state duration densities. According to the obtained sentence HMM and state durations, a sequence of speech features are generated by using a speech parameter generation algorithm (Tokuda, Masuko, Yamada, Kobayashi & Imai 1995, Tokuda, Yoshimura, Masuko, Kobayashi & Kitamura 2000). Figure 4.5 illustrates the generation process of speech parameters ranging from training stage to waveform generation.

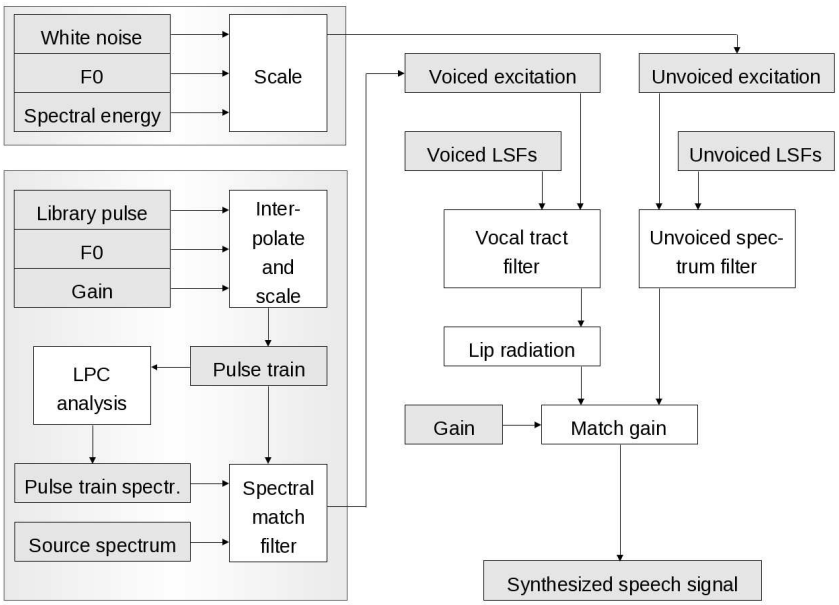


Figure 4.6: Flow chart of the synthesis stage.

4.3.2 Synthesis

The flow chart of the synthesis block is presented in Figure 4.6. The excitation signal consists of voiced and unvoiced sound sources. A natural glottal flow pulse is used as a library pulse for creating the voice source. In comparison to artificial glottal flow pulses, the use of natural glottal flow pulses helps in preserving the naturalness and quality of the synthetic speech. The library pulse was extracted from an inverse filtered frame of a sustained natural vowel produced by a male speaker. The extraction of the pulse was performed by cutting off the pulse at the beginning of the closed time. The main excitation, the discontinuity at the end of the open time, was left intact since it defines important properties of the excitation. The glottal flow pulse was further slightly modified in the time domain in order to remove some resonances that were present during the closed phase due to imperfect glottal inverse filtering. The beginning and the end of the pulse were also set to same level (zero) by subtracting a linear gradient from the pulse. An appropriate library pulse was selected by evaluating the resulting quality of the synthesized speech. The library pulse used for creating the voiced excitation and its derivative are shown in Figure 4.7.

By interpolating and scaling in magnitude this real glottal flow pulse, a pulse train comprising a series of individual glottal pulses with varying period lengths and energies is generated. A cubic spline interpolation technique is used for making the glottal flow pulse longer or shorter in order to change the fundamental frequency of the voice source. However, cubic spline interpolation has some undesirable effects on the glottal flow pulse when

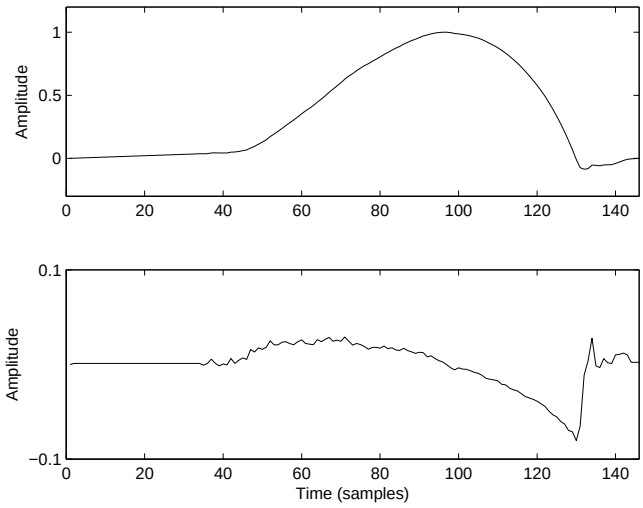


Figure 4.7: Library pulse used for creating the voiced excitation (upper) and its time derivative (lower).

the desired pulse length is much longer than the original pulse. Due to the interpolation some of the high-frequency components are lost, resulting in an unnatural sounding synthetic speech. This effect, however, can be avoided by selecting a glottal flow pulse of sufficient length so that it is not necessary to use too extreme interpolation. For making the glottal flow pulse shorter than the original one, there is no such a problem. However, the fundamental frequency of the selected library pulse was 110 Hz, and for generating low pitched synthetic speech, the library pulse must be interpolated to about twice its length. This processing results in the mentioned loss of high frequencies. However, the selected library pulse was otherwise considered suitable in terms of speech quality.

In order to mimic the natural variations in the voice source, the desired voice source all-pole spectrum ($H_{\text{orig}}(z)$) generated by the HMM is applied to the pulse train. This is achieved by first evaluating the LPC spectrum of the generated pulse train ($H_{\text{synth}}(z)$), and then filtering the pulse train with an adaptive IIR filter

$$H_{\text{match}}(z) = \frac{H_{\text{orig}}(z)}{H_{\text{synth}}(z)}, \tag{4.4}$$

which flattens the spectrum of the pulse train and applies the desired spectrum. An illustration of the procedure is shown in Figure 4.8. The LPC spectrum of the generated pulse train is evaluated by fitting an integer number of the modified library pulses to the 25-ms frame, and performing the LPC analysis without windowing. Before the reconstruction of this filter, the LPC spectrum of the generated pulse train is converted to LSFs, and both

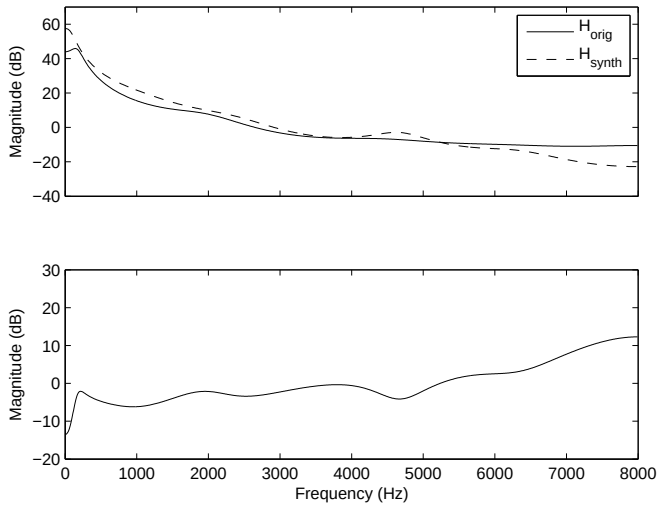


Figure 4.8: Illustration of the modification of the voice source spectrum. The 10th order LPC spectra of the estimated real glottal flow (solid line) and the interpolated library pulse (dashed line) are shown in the upper panel. The frequency response of the spectral matching filter is show in the lower panel. The speech segment is a vocal [e] spoken by a male with fundamental frequency of 77 Hz.

LSFs (the desired voice source all-pole spectrum is originally in the form of LSFs) are then interpolated to frame by frame basis with cubic spline interpolation, and finally converted back to LP coefficients. The filter coefficients are not updated for every sample, but only for every second, third, or fourth sample, depending on the setup. However, some artefacts can be detected due to more abrupt changes in the filter coefficients if the update interval is greatly increased. Since the estimated voice source spectrum may vary substantially in time, and the estimated spectrum of the synthetic pulse train may differ greatly from the real voice source spectrum, the resulting spectrum of $H_{\text{match}}(z)$ may occasionally be somewhat inappropriate. In order to avoid these occasional major changes in the spectrum of $H_{\text{match}}(z)$ and thus possible audible artefacts in the voice source, the LP coefficients of $H_{\text{orig}}(z)$ and $H_{\text{synth}}(z)$ are first damped with an exponential window by multiplying the coefficient vectors with a damping vector $D = (d^0\ d^1\ d^2\ \cdots\ d^{m-1})$, where d is a damping coefficient near one and m is the model order. Values $d = 0.98$ and $d = 0.99$ were used for damping the coefficients of $H_{\text{orig}}(z)$ and $H_{\text{synth}}(z)$, respectively.

The unvoiced sound source is represented by white noise. In order to incorporate an unvoiced component also when the speech sounds are voiced (e.g. breathy sounds), both voiced and unvoiced streams are produced concurrently throughout the frame. During

unvoiced speech sounds, the unvoiced excitation is the primary sound source, but during voiced speech sounds, the unvoiced excitation is much lower in intensity. The unvoiced excitation of white noise is controlled by the f_0 value and further weighted according to the energies of the five frequency bands. The practice of using the spectral energy for weighting the white noise was experimentally studied, and the best result were achieved by weighting the noise mostly by the two highest energy bands (4000–6000 Hz and 6000–8000 Hz). In fact, the accuracy of the unvoiced excitation and the resulting speech quality is very good for unvoiced segments, but the method does not perform very well for voiced sounds incorporating an unvoiced component. This is mostly due to the simple weighting procedure according to spectral energy, which is not able to properly distinguish between harmonics of the voice source and the unvoiced noise component. Therefore, spectral energy is unable to distinguish between normal and breathy voiced sounds and cannot generate appropriate noise source. Moreover, the unvoiced LPC spectrum does not entirely correspond to the spectrum of the noise component in voiced segments since the unvoiced spectrum describes both the voiced and unvoiced speech sounds. In order to make the incorporated noise component in voiced speech segments sound more natural, the noise component is modulated according to the glottal flow pulses. However, if the modulation is too intensive, the resulting speech sounds unnatural. Experiments with the modulation technique showed that a good compromise is achieved by having a 50% baseline for the noise component, and then modulating the remaining part of the noise.

A formant enhancement procedure (Ling, Wu, Wang, Qin & Wang 2006) is applied to the LSFs of voiced and unvoiced spectrum generated by the HMM to compensate for the averaging effect of the statistical modeling. LSFs are modified according to the following procedure. If the LSFs of a frame are defined as l_i, $i = 1,...,m$, where m is the model order, the new enhanced LSFs can be calculated from order 2 to order to order $m-1$ by

$$l'_i = l_{i-1} + d_{i-1} + \frac{d_{i-1}^2}{d_{i-1}^2 + d_i^2}\left[(l_{i+1} - l_{i-1}) - (d_i + d_{i-1})\right], \tag{4.5}$$

where

$$d_i = \alpha\,(l_{i+1} - l_i), \tag{4.6}$$

and $\alpha < 1$ and $i = 2,...,m-1$. Term α controls the degree of enhancement. The less α is, the more intense the enhancement will be. The formant enhancement technique is able to change otherwise muffled synthetic speech to a more clear articulation and better overall quality. However, if the method is applied too intensively, the strongest formants in the speech are heard as disturbing whistle sounds.

After the formant enhancement, the voiced and unvoiced LSFs generated by the HMM are interpolated to frame by frame basis with cubic spline interpolation. LSFs are then

converted to LP coefficients, and used for filtering the excitation signals. The update interval is critical to the speed of the filtering procedure, but some artefacts can be detected due to more abrupt changes in the filter coefficients if the update interval is greatly increased.

For voiced excitation, the lip radiation effect is modeled as a first-order differentiation operation $L(z) = 1 - \rho z^{-1}$, where $\rho = 0.99$. Finally the two filtered excitation signals are combined, and the gain of the signal is set according to the energy measure generated by the HMM. This is achieved by successively evaluating the energy of the synthesized signal E_{synth} within each 25-ms frame, and then evaluating a ratio

$$G = \sqrt{\frac{E_{\text{orig}}}{E_{\text{synth}}}}, \tag{4.7}$$

where E_{orig} is the energy measure generated by the HMM. The vector consisting of values G is then interpolated to frame by frame basis, and applied to the synthetic speech signal to obtain natural gain.

4.4 Other Experimented Methods

Various experiments were made with the new TTS system. The most successful are currently implemented to the synthesizer and described in previous sections. Other methods and experiments, that were not successful or require further development in order to be useful are describe here. Some of the promising methods are further discussed in terms of future plans in Chapter 6.

4.4.1 Voice Source Models

The voice source of the current TTS system is based on interpolating a natural glottal flow pulse extracted from speech through glottal inverse filtering. Various natural glottal flow pulses from different speakers were experimented in the synthesis. However, many other voice source models were also experimented during the development of the synthesizer. First, the Klatt model for the voice source was largely utilized due to easy implementation and simple fundamental frequency control. The open quotient of the pulse can be modified independently from the fundamental frequency, which enables studying the effect of the open quotient to synthesized speech. Many experiments were made with the open quotient of the Klatt model. Nevertheless, the Klatt model was too simplistic to provide naturalness to synthesized speech. One of the greatest drawbacks of the Klatt model is the inability to control the discontinuity at the glottal closure that greatly accounts for the degree of spectral tilt. In addition, LF model pulses were experimented with the synthesizer, but further methods for modifying the parameters of the model were not implemented.

4.4.2 Spectral Modification of Voice Source

In the current implementation, the spectral decay of the voice source is measured with an all-pole model. The spectrum envelope of the all-pole model describes both the spectral tilt and the more detailed spectrum of the voice source. The all-pole model is good in terms of accuracy and details, but the training of the complex parameter set to the HMM system is still under development. In earlier experiments, the spectral decay was measured with single number quantities, such as Harmonic Richness Factor (HRF) and the difference between the amplitude of the fundamental and the second harmonic (H1–H2). Both measures were evaluated from FFT spectrum of length 2048, and in the evaluation of HRF, first ten harmonics were measured. Both measures were extracted and experimented with analysis-synthesis method, and the two measures gave fairly consistent results with the spectral tilt. Both measures showed also a good correlation relative to each other. Since the both methods yielded similar results, only the H1–H2, which requires less computation, was used in further experiments. The appropriate mapping and applying of the measured spectral tilt to the real spectral tilt of the synthetic voice source was a challenging task. Since a single parameter is not capable of describing the detailed spectral behavior of the voice source, the spectral decay of the synthetic voice source was modified with various heuristics. For example, the open quotient of the Klatt model pulse was mapped to H1–H2 values, and the mapping was utilized in the synthesis. This created a variation in the voice source that imitated the natural behavior, but the procedure did not actually improve the quality of the synthesized speech, but created some audible artefacts. Moreover, the use of the H1–H2 value of the natural glottal source to modify Klatt model pulse is not warranted to perform well, since the relation between open quotient and H1–H2 is probably different between the Klatt model and natural glottal flow pulses. However, due to variation in the voice source, some segments were more natural to some extent than without the variation in the open quotient.

The H1–H2 parameters extracted from natural glottal source were also used for changing the spectral decay of the voice source created with natural glottal flow pulses. This was achieved by filtering the created voiced excitation signal with appropriate filters. For example, first order FIR and IIR filters were used in order to change the spectral decay. The mapping of the filter coefficients to H1–H2 values was based on comparing the spectral decay of the synthesized and natural glottal flow pulses at particular H1–H2 values and through listening the resulting synthetic speech. As a result, variation was created to the synthetic voice source, but the procedure created also some audible artefacts. The artefacts were produced most probably due to changes in the spectral slope of the voice source signal that were occasionally too intensive or without an appropriate context. Moreover, in natural speech there are most probably also other changes in the characteristics of speech wave-

form that accompany the changes in the spectral decay which make the change in spectral decay sound more natural. In addition, the frequency response of a first order FIR or IIR filter does not correspond to the real spectral changes in the voice source. Filters with linear frequency response were also used in the experiments for changing the spectral decay of the synthesized voice source, but better results were not achieved. A single number measure of spectral tilt describes the spectral behavior of the glottal flow only partially, since the spectral characteristics of the speech signal certainly varies in many dimensions. On the contrary, the ability of the all-pole model to describe the spectral characteristics of voice source is much more versatile, and therefore all-pole model was finally utilized.

4.4.3 Fundamental Frequency Control

The fundamental frequency of the synthetic voice source is modified by changing the period of the glottal flow pulse. In earlier experiments with the Klatt model, the period was changed according to Equation 3.20 such that the open quotient remained constant unless other methods were used for altering the OQ. The method for changing the fundamental period of the natural glottal flow pulse is interpolation. Although the use of interpolation for changing the fundamental frequency is not perfectly appropriate, it yields satisfactory results for the purpose. Two different methods for interpolation were experimented: cubic spline and linear interpolation. In theory, cubic spline interpolation is better since it incurs a smaller error than linear interpolation, but in practice the problem is not straightforward. The cubic spline interpolation incurs also a loss of higher frequencies, which is not desirable. In the case of linear interpolation the loss is not so severe, but the higher frequencies are composed of random interpolation errors, which is not necessarily desirable. However, the differences between the two methods are not perceptually very distinctive, and the cubic spline interpolation technique was selected on theoretical grounds.

4.4.4 Other Voice Source Modifications

Since only one natural glottal flow pulse is used in the synthesis, the variation between adjacent pulses is minimal. This creates a strong harmonic structure at higher frequencies, and might result in a buzzy sound quality. Although some variation emerges from the differences in the fundamental period, the resulting slight variation is clearly not sufficient to reduce the harmonicity at higher frequencies. Therefore random modification of the spectrum of each individual pulse was experimented. The spectrum was varied through filtering each pulse with a random filter whose frequency response at higher frequencies was varied several decibels. However, the effect of such procedure was not audible unless the variation was increased to a point where distinctive artefacts were perceived.

Jitter in the fundamental period was experimented by creating a random variation to the length of each glottal flow pulse. This effect was also not audible unless the amount of jitter was increased enough to cause distinctive artefacts. Since the experiment did not reveal any benefits from using jitter, it is not used in the current implementation.

Diplophony was also experimented by increasing the length of every second glottal pulse and decreasing the length of the other pulses at certain segments. The longer pulses were slightly emphasized by increasing the gain of the pulses. Various parameters were experimented in order to control the amount of diplophony, such as H1–H2 and fundamental frequency. Heuristic rules for the amount of diplophony were created, for example, the amount of diplophony was increased if the fundamental frequency was under certain limit and the H1–H2 was high, indicating creaky voice quality. The effect of diplophony was clearly audible at certain segments, such as the end of an utterance, but the effect did not notably improve the naturalness of the synthetic speech. However, the created diplophony did not incur any artefacts either. Thus, artificial diplophony could be used in certain specific speaking styles in order to enhance the impression of creaky voice.

4.4.5 Unvoiced Excitation

Various methods were experimented in order to create a natural unvoiced sound source. The currently used method has the flaw that it cannot create a natural noise component to voiced speech segments. This problem derives from the inability of the spectral energy to distinguish between noise and harmonics of the voice source. To overcome this problem, band-pass voicing analysis was experimented. The signal was filtered to five frequency bands with pass-bands of 0–1000 Hz, 1000–2000 Hz, 2000–4000 Hz, 4000–6000 Hz, and 6000–8000 Hz. The voicing strength of each band was estimated using normalized correlation coefficient around the fundamental period. The normalized correlation coefficient is defined as

$$c_T = \frac{\sum_{n=0}^{N-1} x_n x_{n+T}}{\sqrt{\sum_{n=0}^{N-1} x_n x_n \sum_{n=0}^{N-1} x_{n+T} x_{n+T}}}, \tag{4.8}$$

where x_n is the speech signal at sample n, N is the pitch analysis window, and T is the fundamental period. The pass-band voicing strengths describe the relation between the amount of harmonics and noise for each band. Experiments with synthetic voice source and white noise show that the band-pass voicing strengths can distinguish between pure voiced and noisy speech sounds. The band-pass voicing strengths were also used in synthesis, where a voicing strength for each band controlled the gain of band-pass filtered noise. In addition, an adaptive band-pass filter based on the band-pass voicing strengths for controlling the white noise was experimented. However, the band-pass voicing strengths were not really robust in estimating the amount of voicing, and the use of band-pass voicing strength

for controlling the noise source did not improve the quality of the synthesis. Nevertheless, the band-pass voicing strengths were not extensively studied and might be worth further experiments.

Various experiments were also made with noise source type. In one experiment, the spectral energy was used to weight the gain of each frequency band individually. However, since the spectral energy cannot measure the amount of voicing for each band, the results were not any better than in the current implementation. In another experiment, high-pass filtered noise was used instead of white noise in order to reduce the artefacts emerging from the strong noise component at the harmonics of the voice source. However, low frequency noise is required in some speech sounds, and therefore the method is not really practical.

The noise source was also modified according to the H1–H2 measure. If the H1–H2 showed high values, indicating possible breathy voice, the gain of the noise source was increased. This method showed some improvements to the quality of the synthetic speech. Since the H1–H2 does not necessarily correspond to increased amount of noise, but only indicates the type of phonation, the method created also some artefacts due to inconsistent noise in synthetic speech if the relation between H1–H2 and noise gain was too strong. Nevertheless, some improvements could be achieved by using the H1–H2 to slightly control the noise source.

4.4.6 Parameter Smoothing

The statistical HMM system has the property that the generated speech parameters are always smooth. On the contrary, in direct analysis-synthesis, the extracted parameters vary unnaturally rapidly and contain errors. Thus, the quality of the resulting synthetic speech is unsatisfactory unless the parameters are smoothed. Since the analysis-synthesis is an extremely useful tool for experimenting with the parametrization and synthesis, smoothing methods are required. The parameters were generally smoothed in time by convolving the parameter vectors with a Gaussian function defined as

$$f(n) = Ae^{-\frac{n^2}{2\sigma^2}}, \tag{4.9}$$

where A normalizes the sum of the discrete values of the function to one, and σ controls the width of the peak. A convolution with such function corresponds to low-pass filtering. The length of the smoothing vector and the parameter σ were chosen for each parameter independently to achieve an appropriate amount of smoothing. In the case of multidimensional speech features, such as the LSFs and spectral energy, each parameter vector was smoothed in time. For smoothing the fundamental frequency, a special procedure was used in order to avoid the smoothing of the boundaries between voiced and unvoiced frames. Thus, only

the voiced parts were smoothed separately. Similar procedure was used for smoothing the H1–H2, since H1–H2 is measured only when the frame is voiced.

4.4.7 Other Experiments

Other experiments that do not constitute any larger topic are described here. Although most of the experiments are documented, many experiments made with the details of the TTS system are left out in order to point out only the most relevant experiments and findings.

The glottal inverse filtering is never perfect, and often there are some resonances left on the glottal volume velocity waveform. Some of the resonances are caused by nearly real valued roots of the vocal tract transfer function. The resonances can be diminished by reducing the distance of the real roots from the origin. Thus, a real root scaling algorithm was developed that sought for nearly real valued roots and scaled the found ones in order to reduce the resonances. The limit for the imaginary part was set to 0.01, and the found real valued roots were scaled to about half the original distance from the origin. Although the resonances might have been diminished a bit, the processing incurred some artefacts to the spectrum of the vocal tract. Therefore, the scaling of the real roots were excluded from the inverse filtering block.

The evaluation of the spectrum of the estimated glottal flow with LPC has the disadvantage that the estimation is focused for the high-energy part of the signal. Thus, the fine structure of the declining spectrum of the glottal flow is hard to estimate accurately with low-order LPC. In order to estimate the spectral envelope of the glottal flow more accurately, higher frequencies of the glottal flow were emphasized before LPC analysis. Correspondingly, the pre-emphasis was also applied to the estimation of the synthetic glottal flow pulses. Thus, the effect of the pre-emphasis would be canceled when matching the desired spectrum to the spectrum of the synthetic voice source. However, the results with analysis-synthesis were not satisfactory due to increased differences between the synthetic and the natural glottal flow spectra. The use of pre-emphasis resulted in distorted voice source and thus the quality of the resulting synthetic speech was poor. Nevertheless, the pre-emphasis method could be useful for enhancing the training of the HMM system since over-emphasized parameters should be easier to learn.

The spectral shaping of the voice source through filtering changes not only the spectrum but the phase of the signal as well. Although human hearing is not sensitive at perceiving the changes in phase, the filtering might incur audible changes that are not present in the real voice source. To study the effect of the changes in phase due to filtering, the filtering was applied both to normal and time reversed excitation signal. It turned out that the direction of the filtering did have an effect on the characteristics of the resulting speech. However, either of the tested methods did not provide clearly better quality, but only a slight difference was

observed between the methods. Although the time reversal has an effect on the resulting quality of the synthetic speech, there is no physical justification to use it. Therefore, the filtering of the synthetic voice source is performed in a traditional way. Nevertheless, the changes in phase due to filtering raise interesting questions about the validity of the artificial modification of the voice source.

For enhancing the robustness of the fundamental frequency estimation, a pitch tracking algorithm was implemented. The algorithm selected the two highest peaks from the autocorrelation function and selected the one that best fit to the series of old values. The new sample was simply assumed to be the average of the previous samples. The number of previous samples was varied in the experiments, and an appropriate number of previous samples was considered to be from 10 to 20. In addition, a statistical selection algorithm was tested that takes the height of the autocorrelation peaks into account. Thus, the algorithm selected the higher peak more probably than the lower peak. This procedure is justified with the fact that if the highest peak is always selected, the f_0 curve may end up following the wrong peak due to occasional errors. If the highest peak is selected statistically, the algorithm will most probably follow the correct peak. In direct analysis-synthesis, the statistical method produces some errors, but when applied to the HMM system, the errors average out. The performance of the algorithm was two-sided. If the f_0 curve would originally contain much errors, the algorithm reduced the amount of errors efficiently. However, if the f_0 curve was rather smooth, the algorithm made some errors that impaired the quality. Since the robustness of the fundamental frequency estimation was considered sufficient without the pitch tracking algorithm, it was not included to the implemented system. However, the pitch tracking algorithm might be useful when processing speech of certain speaking styles that make the estimation of the fundamental frequency otherwise difficult. Moreover, if linear or polynomial fitting of the f_0 curve for selecting the new peak would be applied to the algorithm, the robustness of the pitch tracking algorithm would probably be notably better.

Autocorrelation based method is currently used for estimating the fundamental frequency. However, the AMDF-based method has been shown to yield more robust results. The AMDF algorithm was not implemented to the system since the accuracy of the autocorrelation method was considered sufficient. However, in further development, the inclusion of AMDF-based f_0 estimation algorithm is not excluded.

In the current implementation, separate parameters for voiced and unvoiced spectra are used. The increased amount of parameters makes it more laborious to generate the speech parameters from the trained HMM. Moreover, filtering of the voiced and unvoiced excitation signals with separate filters is computationally laborious. In the current implementation, the vocal tract filter is not appropriate for filtering the unvoiced excitation, and therefore separate filters for voiced and unvoiced excitation signals must be used. However, by

modifying the unvoiced excitation, the vocal tract filter could be applied to the unvoiced excitation as well, and the two excitation signals could be first summed together and then filtered with only one filter. The most significant difference between the vocal tract spectrum and the unvoiced spectrum is the lip radiation incorporated in the unvoiced spectrum. Therefore, by integrating the unvoiced excitation signal, the vocal tract spectrum could be used for filtering the unvoiced excitation. However, the results from integrating the unvoiced spectrum show that the quality of the resulting speech is degraded due to other differences between the two spectra. For example, the low frequencies are significantly different. At present, the two separate spectra are justified with a better quality of the synthesized speech, but in further development it is desirable to incorporate only one spectrum.

4.5 Implementation Issues

The parametrization and synthesis stages are constructed as stand-alone programs that can be run independently from the HMM system. The development of the speech synthesis system began with speech analysis and synthesis experiments with MATLAB (MathWorks Inc. 2008), but the final system was implemented in C in order to enable fast processing and compatibility with the HMM system. The HMM system used in the synthesizer is based on the HMM-based speech synthesis system (HTS) developed in Japan (HTS 2008). The HTS is a package built on top of the hidden Markov model toolkit (HTK) developed in the UK (HTK 2008). HTS and HTK consist of a set of source libraries and tools available in C source form, and they are both under a free software license. At the University of Helsinki, the HTS has been further modified to meet the requirements of the new synthesizer. The frontend for the phonological analysis and feature extraction has also been developed at the University of Helsinki. The synthesizer includes a text user interface for analyzing and synthesizing speech with different setups. The described system is implemented for audio sampled at 16 kHz, but other sampling rates can be used as well with minor changes. The development of the new TTS system continues, but the main structure of the synthesizer is expected to remain rather unchanged.

Chapter 5

Evaluation of the Text-to-Speech System

The evaluation of a TTS system is a diverse issue. Obviously, the most important aspect in the evaluation is the quality of the synthesized speech. Since the speech quality is a very multidimensional term, its evaluation is problematic. The quality of a TTS system can be assessed in terms of the overall speech quality, or the quality can be determined in terms of several different aspects, such as intelligibility or naturalness. Since the intelligibility of TTS systems today is adequate for most applications, it is often the naturalness that is of primary concern in evaluation. A large number of possible deficiencies can cause synthetic speech to sound unnatural to varying degrees. For example, artefacts or deficiencies in intonation, stress, accent, duration, tempo, and voice quality features all affect the perceived naturalness of the synthetic speech. However, the evaluation of every individual feature might not be the best approach in order to assess the naturalness of speech. A simple and quite reliable way assessing the naturalness of synthetic speech is to present a pair of test sentences synthesized by each system to be compared, and asking the test subjects to judge, which one they would prefer (Klatt 1987). This method does not distinguish the individual aspects that makes one method better or worse than others, but the method can be assumed to yield reliable estimates of how the TTS systems would perform in practical use.

There are several other aspects in the evaluation of a TTS system that do not concern the quality of speech, but the use of the system in a specific application. These properties, such as the ability to adapt to different speaker individuals or speaking styles, and the requirements for the used platform, heavily depend on the used synthesis method. Although the main issue in choosing a TTS system is the quality of speech, certain fundamental restrictions may prevent the use of some speech synthesis methods in a specific application. Moreover, flexibility is becoming an ever more important criterion in a TTS system.

5.1 Subjective Evaluation

In order to evaluate the quality or naturalness of the synthetic speech, subjective listening tests are required. In order to obtain preliminary data about the quality of the new TTS system, two subjective listening tests were conducted. First, the quality of the new system was compared both to natural speech and to synthetic speech generated by a traditional HMM-based TTS system. Second, the new TTS system was compared with a traditional HMM-based TTS system. By performing two individual tests, the performance of the new TTS system can be reliably compared to both natural speech and other TTS systems.

5.1.1 Test Setup

The implemented TTS system was trained with a prosodically annotated database of 600 phonetically rich sentences spoken by a 39-year-old Finnish male speaker, comprising approximately one hour of speech material. The speech was sampled at 16 kHz. A 20th-order LPC was used in parametrizing the spectra of voiced and unvoiced speech, and a 10th-order LPC was used in parametrizing the voice source spectrum. Features described in Chapter 4 were extracted together with their delta and delta-delta features from the speech database.

For evaluation purposes, a de facto standard HTS model structure described in (Yoshimura, Tokuda, Masuko, Kobayashi & Kitamura 1999, Tokuda et al. 2002) was used as a baseline system. This previously developed HMM system uses the mel-cepstral analysis technique (Imai 1983) for spectrum modeling and a simple impulse train excitation model for excitation generation. Instead of using more sophisticated excitation models, the simple one was selected for the comparison because its quality is generally known among the speech synthesis community. The training procedures for both TTS systems were similar.

The spectrograms of a Finnish utterance and corresponding synthetic versions generated by the baseline and the new system are presented in Figure 5.1. The differences between the utterances can be clearly seen from the spectrograms. For example, synthetic speech generated by the new system has clearly more distinct formants and formant transitions than the baseline system.

The speech samples of the implemented TTS system were slightly high-pass filtered in order to compensate for the slight emphasis on the lowest frequencies. The cut-off frequency of the filter was 79 Hz, and the attenuation at 50 Hz was –20 dB. After the filtering, the speech samples sounded slightly more like the natural speech samples. The processing is considered appropriate since the implemented TTS system is only in a developmental stage. For the listening test, the energy between the test samples was normalized in order to avoid differences in loudness.

The listening test was conducted at the Department of Signal Processing and Acoustics

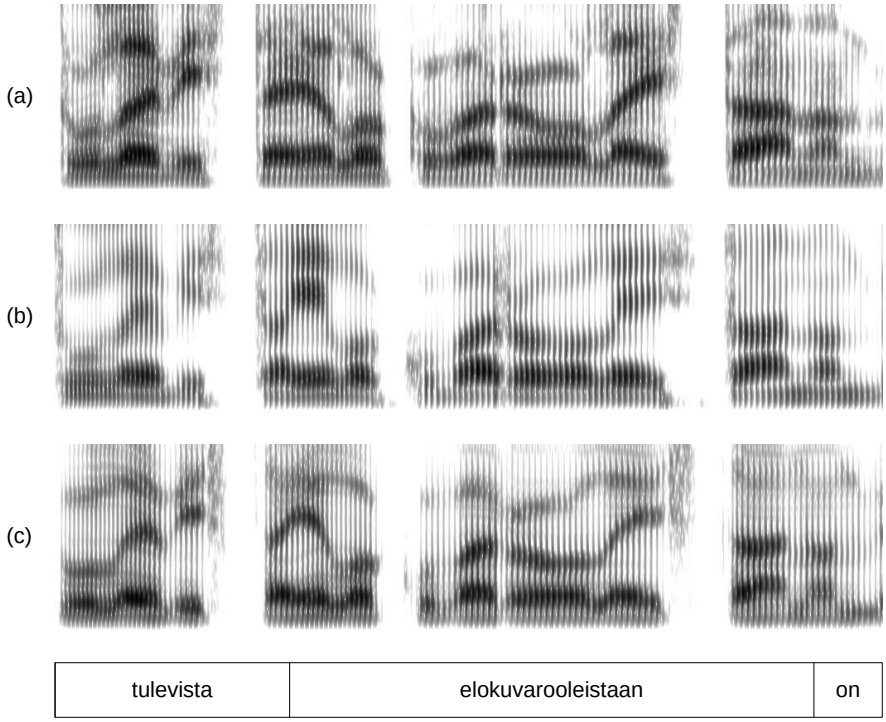


Figure 5.1: Spectrograms of (a) natural speech, (b) synthetic speech generated by the baseline system, (c) synthetic speech generated by the new system.

at the Helsinki University of Technology in Finland. The test sessions took place between March 10th and March 18th 2008. The listening environment was an acoustically modified multipurpose room with low background noise level. The subjects listened the speech samples through Sennheiser HD 580 headphones. The listening test software GuineaPig 3 (Hynninen & Zacharov 1999) was used in the test. Altogether 11 naive listeners, 9 men and 2 women, participated in the test. All subjects were native speakers of Finnish between 24 and 31 years of age.

5.1.2 Comparison Category Rating Test

In the first part of the subjective evaluation, a Comparison Category Rating (CCR) test was carried out. The test resembled the ITU-T Comparison Category Rating test (ITU 1996) with minor changes. Although the CCR test is designed for slightly different purposes, the test was considered suitable for obtaining preliminary data. In the CCR test, the listeners

are presented with a pair of speech samples on each trial, and asked to assess the quality of the second sample compared to the quality of the first one on the 7-point Comparison Mean Opinion Score (CMOS) scale. In effect, the listeners provide two judgments with one response: "Which sample has better quality?" and "By how much?". The CMOS scale is presented in Table 5.1. Corresponding Finnish descriptions were used in the test. The test user interface is shown in Figure 5.2.

The test sample pairs consisted of natural speech, synthetic speech generated by the implemented system, and synthetic speech generated by the baseline system. Ten randomly chosen sentences from held-out data were used for generating the test samples. The sentences are presented in Appendix A. All sample pairs are presented twice, exchanging the order of the samples for the second time. Ten null pairs, where the two samples are the same, were included in the test in order to assess the reliability of the given answers by each listener. The test consisted a total of 70 speech sample pairs (10 sentences, 3 methods, 2 orders, plus 10 null pairs). The subjects could play the sample pairs as many times as they wanted.

The sample pairs were randomized individually for each test subject with a block randomization method. The sample pairs were first divided into blocks, where each condition occurs exactly once, and the sentences are evenly distributed over the blocks. Then, the order of the blocks was randomized so that the same sentence is not presented twice in a row in the final order of presentation.

The subjects were given written instructions for the test, and further oral instructions were given when necessary. The test consisted of a practice session of five sample pairs selected randomly from the test sample pairs. During the practice session the listeners were allowed to adjust the volume to a comfortable listening level. During the test session the volume was kept constant. All subjects did the test individually with their own pace, but they were encouraged to rate the overall quality of the speech samples and told that a

Table 5.1: Rating scale used in the CCR test.

3	Much Better
2	Better
1	Slightly Better
0	About the Same
-1	Slightly Worse
-2	Worse
-3	Much Worse

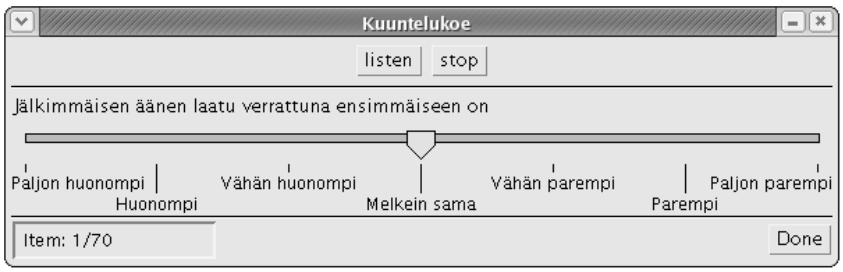


Figure 5.2: User interface used in the CCR test. The sample pair is played by pressing *listen* and the currently playing sample pair can be stopped by pressing *stop*. After listening the sample pair once or more, the quality of the second sample compared to the first sample is graded by setting the slider to the desired verbal description. The button *Done* is pressed to proceed to the next trial. The number of completed trials out of total trials is shown on the lower left corner.

detailed consideration would not necessarily yield a better result. The CCR test took from 20 to 30 minutes per listener.

Results

The ranking of the three methods with 95% confidence intervals according to the CCR test is shown in Figure 5.3. The ranking of the methods was evaluated by averaging the scores of the CCR test for each method. The 95% confidence intervals based on the 1-sided t-test were calculated by the following equations (3GPP 2003):

$$\begin{aligned} \text{upper limit} &= \text{CMOS}_{\text{test}} + \frac{t_{N-1,\alpha/2}\, s_{\text{test}}}{\sqrt{N}} \\ \text{lower limit} &= \text{CMOS}_{\text{test}} - \frac{t_{N-1,\alpha/2}\, s_{\text{test}}}{\sqrt{N}}, \end{aligned} \tag{5.1}$$

where $\text{CMOS}_{\text{test}}$ is the averaged CMOS score for the method in question, $t_{N-1,\alpha/2}$ is the inverse value from Student's t-distribution with $N-1$ degrees of freedom and probability of $\alpha/2$, s_{test} is the sample standard deviation, and N is the number of answers per method. For 95% confidence intervals $\alpha = 0.05$. The preferences between the methods were found to be statistically significant. Although the 1-sided t-test is not precisely the right method for testing the means of more than two datasets, it yields results accurate enough considering that the means of the three methods are significantly different with a large margin. If the margin would be considerably smaller, more accurate statistical methods, such as multiple comparison procedures, would be required. Bar plots of the scores and mean scores with confidence intervals are presented in Figure 5.4.

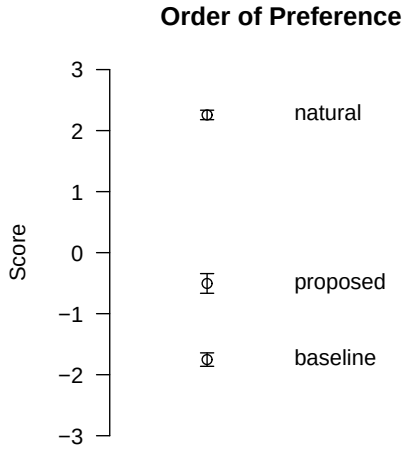


Figure 5.3: Ranking of the CCR test for the following speech samples: natural speech (natural), proposed system (proposed), baseline system with an impulse train excitation model (baseline). The mean score has no explicit meaning, but the distances between the scores are essential. The 95% confidence intervals are presented for each score.

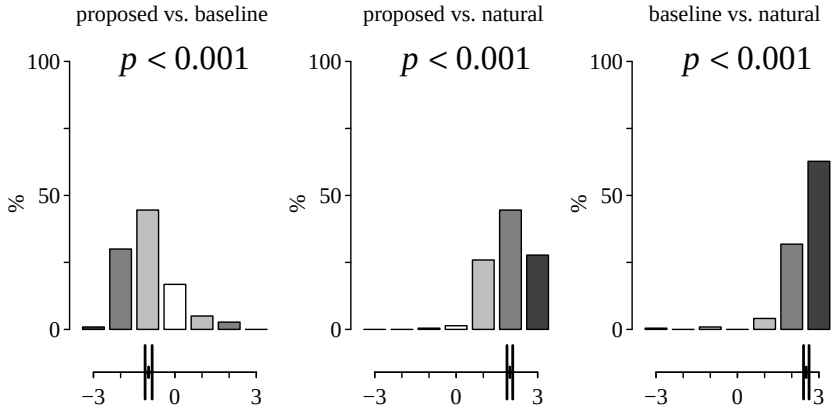


Figure 5.4: Bar plots of the scores and mean scores with confidence intervals for the following speech samples: natural speech (natural), proposed system (proposed), baseline system with an impulse train excitation model (baseline).

The consistency and the reliability of the listeners were assessed by comparing the two given grades for each same sample pair, and inspecting the grades given for the null pairs. In optimal case, the two scores for the same sample pair would be the same in both occasions, and the score for the null pairs would be zero (About the Same). Overall, the consistency was fairly good since the difference between the same sample pairs were mostly zero or one, with only few exceptions. The reliability of the listeners was generally good since nearly all listeners graded all the null pairs zero. The differences between the scores of the same sample pairs and the scores for the null pairs are presented in Figures A.1 and A.2 in Appendix A. The distribution of the given grades is quite uniform for all the listeners, which indicates that the CMOS scale was appropriate for the test. The distributions of the given grades by each listener are presented in Figure A.3 in Appendix A.

5.1.3 Pair Comparison Test

In the pair comparison test, only the synthetic sounds generated by the two HMM-based TTS systems were involved. A pair comparison test method was used, where subjects listened to samples referred to as A and B, and selected the one they would rather listen to. They were also given an option to choose that the samples sounded about the same, indicating no preference between the two samples. The user interface of the pair comparison test is shown in Figure 5.5.

Ten randomly chosen sentences from held-out data (different from the ones used in the CCR test) were used for generating the test samples for each method. The sentences are presented in Appendix B. The test subjects made all the possible comparisons for each sample in both orders. Furthermore, 4 null pairs, where the samples A and B are the same, were included in order to assess the reliability of the given answers by each listener. The pair comparison test included a total of 24 trials, consisting of 20 comparisons between the methods and 4 null pair trials. The samples were presented to the listener in random order, and the order was different for each listener. The subjects could listen the samples as many times as they wanted before giving the answer.

The subjects were given written instructions for the test, and further oral instructions were given when necessary. The test consisted of a practice session of five comparisons selected randomly from the test samples. During the practice session the listeners were allowed to adjust the volume to a comfortable listening level. During the test session the volume was kept constant. All the subject did the test individually with their own pace, but they were encouraged to rate the overall quality of the speech samples and told that a detailed consideration would not necessarily yield a better result. The pair comparison test test took from 10 to 15 minutes per listener.

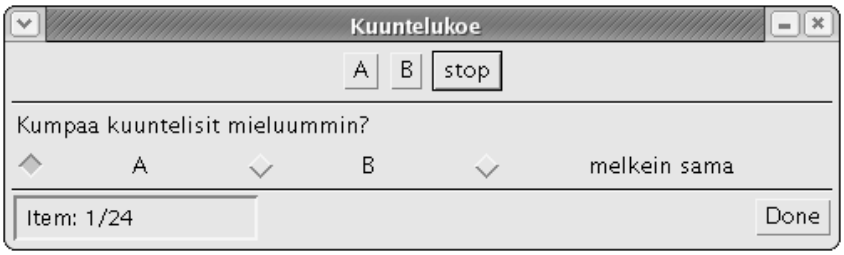


Figure 5.5: User interface used in the pair comparison test. The samples are played by pressing *A* or *B*, and the currently playing sample can be stopped by pressing *stop*. After listening the both samples at least once, the question "Which one would you rather listen to?" is answered by selecting one of the three alternatives: sample A, sample B, or no preference for either of the samples. The button *Done* is pressed to proceed to the next trial. The number of completed trials out of total trials is shown on the lower left corner.

Results

The preference scores of the synthesis methods with 95% confidence intervals are shown in Figure 5.6. The confidence intervals based on the binomial distribution are calculated using the following formulas (NIST/SEMATECH 2008):

$$\begin{aligned} \text{upper limit} &= \frac{\hat{p}+\frac{z^2_{\alpha/2}}{2n}+z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z^2_{\alpha/2}}{4n^2}}}{1+\frac{z^2_{\alpha/2}}{n}} \\ \text{lower limit} &= \frac{\hat{p}+\frac{z^2_{\alpha/2}}{2n}-z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}+\frac{z^2_{\alpha/2}}{4n^2}}}{1+\frac{z^2_{\alpha/2}}{n}}, \end{aligned} \tag{5.2}$$

where n is the number of samples, $\hat{p}$ is the proportion of items in the category in question, and $z_{\alpha/2}$ is the upper critical value from the normal distribution that is exceeded with probability $\alpha/2$, where $\alpha = 0.05$ for 95% percent confidence intervals. The preferences between the two methods were found to be statistically significant.

The consistency and the reliability of the listeners were assessed by the following ways. First, the number of answers to A, B, and "no preference" were counted for each listener. This gives information about the behavior of each subject. Ideally, the number of answers to A and B should be equal, since each comparison is made twice, and the second time the order of samples in A and B is reversed. Mostly the answers were equally distributed, but for one subject the answers for B were highly emphasized for some reason. The distribution of the answers to A, B, and "no preference" for each subject is presented in Figure B.1 in Appendix B. Second, the answers to the null pair trials were studied. Ideally, no preference should be addressed for either of the samples. Generally, there was no preference for either

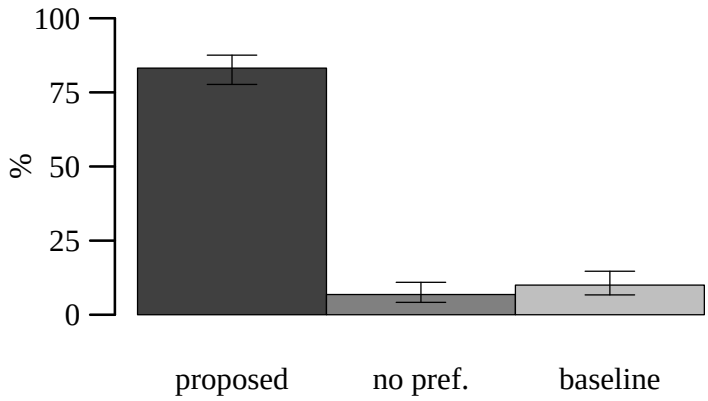


Figure 5.6: Results of the pair comparison test applied for the proposed system (proposed) and the baseline system with an impulse train excitation model (baseline). The bars indicate the percentage of the total number of answers to the question "Which one would you rather listen to?". The center bar (no pref.) indicates no preference for either of the methods. The 95% confidence intervals are presented for each bar.

of the samples in the case of null pair trials, but one subject did not show such behavior, but in almost every case answered either A or B. The answers to the null pair trials are presented in Figure B.2 in Appendix B. Third, the answers of each subject to all non-null sample pairs were compared, and the proportion of sample pairs receiving the same answer twice was calculated. Since each comparison is presented twice in the test, ideally the subject should always choose the same answer, A, B, or "no preference" for the sample in the two occurrences of the sample pair. Generally, this kind of consistency was fairly good (about 80%), but two subjects failed to be consistent in their answers. The consistency of the answers for each subject is presented in Figure B.3 in Appendix B. In addition, the answers to different methods by each listener, and the answers to different methods by each sentence are presented in Figures B.4 and B.5 in Appendix B, respectively.

5.1.4 Result Analysis

The results of the CCR test show that the proposed new TTS system utilizing glottal inverse filtering has a considerably better quality than the previously developed HHM-based method. Compared to natural speech, the quality of the new system is clearly worse. However, since the prosodic features of the synthetic speech were generated directly from the HMM, the evaluated degradation in quality partly results from the prosodic discrepancies

between the synthetic and natural speech samples.

The results of the second test show that the new system is almost always preferred over the baseline system. Since the same prosody model was used for both systems, the results are comparable. However, during the preliminary evaluation, it was noted that the baseline system sounded better with the new prosody model than with the original one. Thus, the baseline system does not necessarily represent the reference system (Yoshimura et al. 1999, Tokuda et al. 2002), but might be slightly better in quality. Nevertheless, since the results show obvious preference for the new system over the baseline system, this issue is not of great concern.

After the two listening tests, the subjects were asked to describe possible artefacts they noticed in the synthesized speech in order to obtain information about the most salient aspects that degrade the quality of synthetic speech. The listeners described the baseline system with terms such as creaky, machine-like, dry, rough, and robotic. The quality of the new system compared to the baseline system was described as more natural and human-like with clear characteristics of a person. However, the proposed systems was described also as too emphasized on the low frequencies, which was, according to some subjects pleasing, but at the same time it made the speech less clear. All the synthetic speech samples were described as somewhat metallic and machine-like, but very intelligible. The quality of the prosody and the transients of the synthetic speech samples were criticized compared to natural speech, and occasional other artefacts were also reported.

The comments from the listeners suggest that the synthetic speech of the proposed new system is much more natural sounding than the synthetic speech of the baseline system. Moreover, the comments show that the new system is able to produce synthetic speech with specific speaker characteristics. However, there are many aspects in the new system that were criticized, such as the emphasis on the low frequencies, metallic sound, and the artefacts in prosody.

5.2 Computational and Implementation Considerations

A TTS system must be computationally feasible in order to be of practical use. In order to implement a real time TTS system, a lot of computing power and memory is often required. Personal computers now mostly meet the requirements of current TTS systems, but applications on memory and processing power constrained devices, such as mobile phones and other handheld devices, are much more challenging. Moreover, the need for TTS systems for the low resource devices is continuously increasing.

One of the advantages of the HMM-based TTS system is its low memory requirement. The whole TTS system takes less than ten megabytes of space. Compared to concatenative

TTS systems, which may take hundreds of megabytes of space, the difference in memory requirement is remarkable. The HMM-based speech parameter generation algorithm is also fast, and combined with text analysis, the generation of parameters from text input can be run almost real time on a personal computer. In addition, there are also many possibilities to make the HMM-based speech parameter generation faster than at present. However, the most laborious part of the HMM-based speech synthesis is the waveform generation from the parameters. The waveform generation consists of two main tasks: excitation generation and filtering. The generation of the excitation signal in a conventional HMM-based speech synthesis system is very straightforward: the excitation consists of either impulses or white noise. On the contrary, in the implemented TTS system, the generation of voiced excitation covers the interpolation of natural glottal flow pulses and further modification of the voice source spectrum through filtering with an adaptive IIR filter. Thus, the generation of the excitation signal is computationally more demanding than in a conventional speech synthesizer. After the signal generation, the voiced and unvoiced excitation signals are both filtered with adaptive FIR filters, which is the most laborious part of the synthesis.

The current implementation of the HMM-based TTS system cannot synthesize speech in real time. The required synthesis time largely depends on the update interval of the filter coefficients. If the coefficients are updated for every second samples, which yields quality indistinguishable from the synthesis with continuous updating of the filter coefficients, the synthesis takes currently about twice the duration of the resulting speech. If only every eighth coefficient is used, the synthesis time is less than the duration of the resulting speech, but also some minor artefacts can be detected.

The implementation of the TTS system is not entirely optimized in terms of computational efficiency, but it is rather a platform for experimenting with the new synthesis technique. There are several issues concerning the implementation that could be developed in order to make the synthesis computationally more efficient. Since most of the processing time is used for filtering the voice source and excitation signals, the natural focus of development would be the filtering algorithms. The filtering process also comprises the conversion of LSFs to LPC polynomial on each filter update. The implementation of the conversion function is not optimized for recurrent use in *for*-loops, and thus the optimization would speed up the filtering. Additionally, the update interval could be made adaptive to avoid unnecessary computation. Thus, the filter coefficients would be updated only if the changes in LSFs would be large enough to cause artefacts.

Through the optimization, the synthesis could be made run with a considerably smaller delay than at present. Moreover, if the text analysis, speech parameter generation, and waveform generation would be implemented to run concurrently, continuously generating synthetic speech, near real time practical implementation would be possible.

Chapter 6

Discussion

This chapter concludes the thesis with a discussion of the most important, current and future issues concerning the new TTS system. The utilization of glottal inverse filtering and modeling of the voice source characteristic in a HMM-based TTS system are discussed. Alternative methods of implementation are considered, and future plans for further developing the TTS system are described. Finally, the conclusions of the thesis are presented.

6.1 Discussion and Proposed Improvements

Although the experimental results show that the proposed new system is able to generate natural sounding speech, the full potential of the new system is not entirely used in the current implementation. There are several aspects concerning the design and implementation of the TTS system that are not optimal in any terms, and, as noted before, the current implementation is rather a platform for experimenting and further developing the new TTS method. The components of the TTS system that have major contribution on the quality of the synthesized speech are discussed in the next few sections. Possible improvements are proposed and the estimated benefits are evaluated.

6.1.1 Glottal Inverse Filtering

The aim of the new HMM-based TTS system is to create natural sounding speech in different speaking styles with different speaker characteristics and even emotions. These goals are achieved partly through the ability of the HMM system to model these characteristics, but foremost through the ability of the training stage to distinguish and parametrize these features from natural speech. Since a large part of what can be characterized as naturalness in speech emerges from different voice source characteristics as well as their context dependent changes, the core of the new TTS system is the glottal inverse filtering, which

enables the parametrization of the glottal source characteristics.

The parametrization of speech signal is traditionally performed through a decomposition to source and filter. The decomposition can be performed in various ways, for example through basic all-pole modeling, through glottal inverse filtering, or with any other method. The resulting source and filter need not to represent any real mechanism of speech production. The purpose of the decomposition is merely to represent the speech signal in terms of source and filter characteristic, and thus reduce the information required for representing the speech signal. In this respect, the approach of using inverse filtering does not yield any more accurate results than other decomposition methods. However, the decomposition through glottal inverse filtering yields information about the real functioning of the vocal folds and the vocal tract filter. This enables the modeling of glottal source and vocal tract filter parameters individually, and the further analysis and modification of the characteristics are possible based on the knowledge of the speech production mechanism.

While the HMM system is a statistical method for describing the speech signal, it can not distinguish the voice source characteristics from source and filter based on traditional decomposition. In other words, if an HMM system is trained with traditional source and filter parameters, the context dependent changes of the voice source characteristics are spread randomly to both source and filter characteristic, and therefore the desired parameters are statistically smoothed out. Thus, the voice source characteristics cannot be utilized in a traditional HMM-based TTS system. On the contrary, in the new HMM-based TTS system that utilizes glottal inverse filtering, separate voice source and vocal tract characteristics are fully available for modeling in order to imitate the natural speech production mechanism, and thus produce natural sounding synthetic speech. Moreover, the individual modeling of different speech characteristic enables the easy adaptation and alteration in speaking style, speaker characteristics and emotion.

6.1.2 Spectral Modeling

In the current TTS system, linear prediction is used for estimating the spectral envelope of the speech signal, and further the spectral envelope of the voice source. However, it has been known for some time that linear prediction suffers from various drawbacks that are especially evident during voiced segments of speech. Specifically, the peaks of the LPC spectral envelope are biased towards the pitch harmonics, which causes bias to the estimated formant structure. To overcome these problems, discrete all-pole (DAP) modeling (El-Jaroudi & Makhoul 1991) could be utilized for evaluating the spectral envelope of speech instead of LPC. Generally, DAP modeling gives better spectral envelopes than linear prediction. Moreover, DAP modeling enables the spectral weighting of the analysis, which could be utilized in order to obtain better spectral model for the most important frequencies

in speech.

6.1.3 Library Pulse

The use of a natural glottal pulse for creating the voiced excitation helps in preserving the naturalness and quality of the synthetic speech. However, since there is a great variation in the shape and spectrum of the glottal flow pulses in natural speech, the use of a single glottal library pulse is not really justified. Experiments made with the TTS system show that the use of a single glottal library pulse is unable to mimic the dynamics of the glottal flow pulses that exist in natural continuous speech. It would be preferable to use more than one library pulse in order to create desired variability, and probably the quality of the synthesized speech would improve as the number of different glottal flow pulses would increase. However, there are several difficulties in such an approach. Firstly, experiments made with the system show that the selection of the library pulse has a significant effect on the quality of the synthesized speech. The characteristics of the synthesized speech are substantially different with different glottal flow pulses even if the pulses may seem very similar to each other in the time domain. Thus, it is challenging to create a set of real glottal flow pulses that would be suitable for creating the voice source and would lead to natural sounding synthetic speech.

However, the approach of using more than one natural glottal flow pulses for creating the voiced excitation is attractive. One method of implementing the pulse library is to extract several pulses as similar to each other as possible, and then randomly use the pulses in synthesis. This might create the desired variation in the voice source, and the machine-like or metallic characteristics due to strong harmonic structure at higher frequencies would probably be diminished. However, on the grounds of small scale experiments, there is a large possibility that the differences between the library pulses are heard as artefacts in the resulting synthesized speech. Another approach, and maybe even more challenging one, is to extract separate glottal flow pulses from speech of different fundamental frequency, phonation type or intensity. This would bring out the characteristics of natural speech, and would at the same time partly solve the problem of changing the fundamental frequency of the voice source. The extraction of the glottal flow pulses and the construction of an appropriate library for synthesis would require much work. However, the results of a successful pulse library and algorithm for using the pulses could yield substantial enhancements to the naturalness of the synthesized speech. To reduce the artefacts due to the differences between adjacent glottal flow pulses, pulse modifying algorithms could be utilized. A simple method would be to gradually change the waveform of the glottal flow pulse from one pulse to another. Nevertheless, extensive experiments with more than one library pulse were not performed.

It is challenging to create natural sounding synthetic speech by using different glottal flow pulses, because it is likely that the differences between individual pulses cause artefacts to some extent. Moreover, it is not only the variation that is required between the adjacent glottal flow pulses, but there is also rules based on the physical functioning of the vocal folds that define the properties of the glottal source. Thus, more information about the physical functioning of the vocal folds would be beneficial in order to fully utilize the pulse library. Alternatively, other methods for creating the voice source could be used, for example methods used primarily in articulatory synthesis, but unfortunately there are no proper methods for physical modeling of the voice source.

6.1.4 Fundamental Frequency Modification

Despite the fairly natural synthetic speech, the interpolation of the glottal flow pulse according to the fundamental frequency is far from the natural behavior of the glottal flow. The interpolation is a compromise to alter the fundamental frequency due to the lack of proper methods for modeling the behavior of the vibrating vocal folds. The interpolation procedure has many disadvantages. Firstly, although the original time properties of the glottal flow pulse are shifted in proportion to each other, it is different from the natural behavior of the glottal flow pulse. For example, although the open time would be longer in low-pitched pulses than in high-pitched ones, in does not mean that the abrupt glottal closure should be different between the pulses. By using interpolation, all the different properties of the glottal flow pulse are changed simultaneously without any physical basis. Secondly, the cubic spline interpolation changes the frequency content of the glottal flow pulse. It is not exactly known how the spectral characteristics of a glottal flow pulse should behave when the fundamental frequency is changed. However, the cubic spline interpolation is a compromising method for changing the fundamental frequency.

There are alternative methods for interpolation in order to change the fundamental period of the voice source. Since the closed time of the glottal flow pulse is kept somewhat constant regardless of other changes in the voice source, it is easy to change the fundamental period through alteration of the closed time. However, preliminary test with the technique showed that the quality was not improved. This is probably due to the constancy of other properties of the glottal flow pulse, which results in unnatural synthetic speech when the fundamental period is different from the length of the original pulse. Another approach is to interpolate only some parts of the original pulse, for example excluding the main excitation. Thus, most of the properties of the glottal flow pulse would change according to the fundamental frequency, but yet the main excitation would provide the desired higher frequencies. This method has not been experimented, but the obvious defect of the method is the lack of variation in the main excitation.

The implementation of a pulse library consisting of glottal flow pulses of different fundamental period could be the best solution for changing the fundamental frequency of the voice source. If the different glottal flow pulses describe the characteristics of the voice source at different fundamental frequencies, the resulting synthetic speech should be more natural, if the differences between adjacent pulses would not produce audible artefacts. If the library consisting of pulses with different fundamental periods is dense enough, the fine adjustment of the fundamental period can be performed with a slight interpolation without greatly affecting the properties of the glottal flow pulse. As noted in the previous section, the difficulty is to find and extract such pulses that would describe only the desired properties, and would be otherwise similar to each other in order to avoid artefacts.

6.1.5 Spectral Modification of Voice Source

The all-pole model used for spectral modification of the voice source is good in terms of accuracy and details, but there are some problems in the training of the complex parameter set to the HMM system. At present, the parameters generated by the HMM system are somewhat oversmoothed, and thus do not create as much variation to the voice source as desired. Since the inverse filtering procedure is not perfect, the voice source incorporates also some residual spectrum from the vocal tract. This might lead to a situation where the voice source spectrum describes the residual of the vocal tract more than the behavior of the voice source. Informal observation of the decision trees of the voice source spectrum revealed, that the context clustering was performed mostly based on the phone identity, suggesting that some traces of the vocal tract were left on the voice source spectrum. The residual on the voice source spectrum is not a problem since the quality of the synthesized vowel might improve due to the phone-specific modification of the voice source, but the problem is that the HMM might not be able to model the voice source spectrum that describes the phonation type as efficiently as desired. It seems that the HMM system is somewhat unable to find appropriate linguistic context for the changes in the voice source spectrum. However, this problem might arise from many reasons, such as too small training material, or secondly the training material might be spoken too monotonously without great variations in the voice source spectrum. In addition, there might also be other reasons concerning the HMM system that make the training of the voice source spectrum parameters less effective. Although improvements should be made in order to model the spectrum of the voice source, some features of speech are currently well modeled. For example, one repeatedly seen splitting criterion was syllable position in the utterance, showing that the HMM system can learn the characteristics of the utterance-final creaky voice, typical in Finnish. However, further development for the modification of the voice source spectrum is required since the appropriate variation in the voice source spectrum is one of the most important features of the

new synthesizer. An especially interesting topic is how the spectrum of the voice source would relate to the higher level phonological factors.

6.1.6 Impression of Breathiness

The synthesized unvoiced speech sounds are currently fairly natural, but the unvoiced component incorporating the voiced component sounds rather poor. This problem derives from the inability of the spectral energy to distinguish between noise and harmonics of the voice source. Since the natural impression of breathiness is important in order the create natural sounding synthetic speech, methods for measuring and adding breathiness must be developed. Thus, the relation between voiced and unvoiced components should be measured. There are several algorithms to perform the task, such as band-pass voicing strength (see Section 4.4.5) and harmonic-to-noise ratio. Harmonic-to-noise ratio measures the ratio between the magnitude of the harmonics and the magnitude of interharmonic noise. Thus, by measuring the relation between voiced and unvoiced components at different bands, an appropriate amount of noise could be added to each band. The band-pass voicing strength has been already experimented, but further studies are required in order to create a natural impression of breathiness.

The impression of breathiness could be expected to be more natural if the noise component was modulated according to the voiced component. However, there is no solid information about the mechanism of the modulation. The modulation according to the glottal flow pulses was experimented, but large improvements were not achieved. Experiments also showed, that some time characteristics of the turbulent noise were not critically important, since there were no audible differences regardless of whether the noise was present at glottal open time or closed time. The higher frequencies that originate from the glottal flow pulses in natural speech follow the periodicity of the glottal flow, but the discrimination between the noise originating from the vocal folds or turbulent flow is difficult.

6.1.7 HMM System

Methods for the HMM modeling are rapidly developing, especially in the field of speech synthesis. The current implementation of the HMM system does not represent the state-of-the-art of the HMM synthesizers, but is a basic implementation for experimenting with the new synthesizer. Thus, many improvements could be incorporated within the HMM system which would probably improve the quality of the synthesized speech. For example, the introduction of hidden semi-Markov models (HSMM) (Zen, Tokuda, Masuko, Kobayashi & Kitamura 2004) and speech parameter generation considering global variance (GV) (Toda & Tokuda 2007) have been proposed to enhance the performance of the HMM system, just

to name a few. With the release of HTS version 2.0 (Zen, Nose, Yamagishi, Sako, Masuko, Black & Tokuda 2007) various improvements and new features have been included in the HMM system. Moreover, speaker adaptation and adaptive training have been introduced in the HTS version 2.0 in order to enable flexible speech synthesis.

6.2 Future Work

Most of the main topics of further development were mentioned in previous section, but no explicit directions of the development were discussed. Generally, the aim of the TTS system is to enable generating highly natural synthetic speech capable of conveying different speaker characteristics. The basic blocks that enable these goals are implemented, but further development is required in order to fully utilize the capability of the new TTS system. For example, the forthcoming development will be focused on improving the use and shaping of the natural glottal flow pulses, and enhancing the use of the voice source characteristics obtained by glottal inverse filtering.

6.3 Conclusions

In this thesis, a new HMM-based text-to-speech system utilizing glottal inverse filtering was described. Subjective listening tests showed that the quality of the proposed new TTS system was considerably better compared to a traditional HMM-based TTS system with an impulse train excitation model. The information about the voice source characteristic obtained through glottal inverse filtering and the use of natural glottal flow pulses clearly improved the quality of the synthesized speech. Moreover, individual modeling of the voice source characteristics in the framework of HMM enables flexible speech synthesis with arbitrary speaker's voice, various speaking styles, and emotional expressions. The new method has the potential to produce highly natural sounding synthetic speech. The development of the new TTS system continues in order to fully utilize the new techniques introduced in this work.

Bibliography

3GPP (2003). Minimum performance requirements for noise suppresser application to the adaptive multi-rate (AMR) speech encoder, 3rd Generation Partnership Project, 3GPP TS 26.077 V5.0.1.

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

Alku, P. (2003). Parameterisation methods of the glottal flow estimated by inverse filtering, *Voice quality: Functions, analysis and synthesis*, pp. 81–88.

Alku, P., Bäckström, T. & Vilkman, E. (2002). Normalized amplitude quotient for parametrization of the glottal flow, *J. of the Acoustical Society of America* **112**(2): 701–710.

Alku, P., Strik, H. & Vilkman, E. (1997). Parabolic spectral parameter – A new method for quantification of the glottal flow, *Speech Communication* **22**: 67–79.

Alku, P., Tiitinen, H. & Näätänen, R. (1999). A method for generating natural-sounding speech stimuli for cognitive brain research, *Clinical Neurophysiology* **110**: 1329–1333.

Allen, J., Hunnicut, S. & Klatt, D. (1987). *Text-to-Speech: The MITalk System*, Cambridge University Press.

Ananthapadmanabha, T. V. (1984). Acoustic analysis of voice source dynamics, *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR), Royal Institute of Technology, Stockholm* **2-3**: 1–24.

Atal, B., Cox, R. & Kroon, P. (1989). Spectral quantization and interpolation for CELP coders, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 1, pp. 69–72.

Atal, B. S. & Schroeder, M. R. (1967). Predictive coding of speech signals, *Proc. Conf. Communication and Processing* pp. 360–361.

Bäckström, T., Alku, P. & Vilkman, E. (2002). Time-domain parameterization of the closing phase of glottal airflow waveform from voices over a large intensity range, *IEEE Transactions on Speech and Audio Processing* **10**(3): 186–192.

Bauer, W. R. & Blankenship, W. A. (1974). DYPTRACK – A noise-tolerant pitch tracker, Tech. Rep. NASL-S-210, 525, Dept. of Defence (NSA), U.S.A., Unclassified Report.

Baum, L. E., Petrie, T., Soules, G. & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains, *Annals of Mathematical Statistics* **41**(1): 164–171.

Bogert, B. P., Healy, M. J. R. & Tukey, J. W. (1963). The frequency analysis of time-series for echoes, *Proc. Symp. Time Series Analysis* pp. 209–243.

Cabral, J. P., Renalds, S., Richmond, K. & Yamagishi, J. (2007). Towards an improved modeling of the glottal source in statistical parametric speech synthesis, *Sixth ISCA Workshop on Speech Synthesis* .

Carlson, R., Fant, G., Gobl, C., Granström, B., Karlsson, I. & Lin, Q. (1989). Voice source rules for text-to-speech synthesis, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 1, pp. 223–226.

Carlson, R., Granström, B. & Karlsson, I. (1991). Experiments with voice modelling in speech synthesis, *Speech Communication* **10**: 481–489.

Catford, J. C. (1977). *Fundamental Problems in Phonetics*, Edinburgh University Press, Edinburgh.

Childers, D. G. & Lee, C. K. (1991). Vocal quality factors: Analysis, synthesis, and perception, *J. of the Acoustical Society of America* **90**(5): 2394–2410.

Claes, T., Dologlou, I., ten Bosch, L. & van Compernolle, D. (1998). A novel feature transformation for vocal tract length normalization in automatic speech recognition, *IEEE Transactions on Speech and Audio Processing* **6**(6): 549–557.

de Cheveigne, A. & Kawahara, H. (2002). YIN, a fundamental frequency estimator for speech and music, *J. of the Acoustical Society of America* **111**(4).

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B* **39**(1): 1–38.

Dutoit, T. (1997). *An Introduction to Text-to-Speech Synthesis*, Kluwer Academic Publishers.

El-Jaroudi, A. & Makhoul, J. (1991). Discrete all-pole modeling, *IEEE Transactions on Signal Processing* **39**(2): 411–423.

Fant, G. (1960). *Acoustic Theory of Speech Production*, Mouton, The Hague.

Fant, G. (1979). Glottal source and excitation analysis, *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR), Royal Institute of Technology, Stockholm* **1**: 85–107.

Fant, G. (1995). The LF-model revisited. Transformations and frequency domain analysis, *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR), Royal Institute of Technology, Stockholm* **2-3**: 119–156.

Fant, G., Kruckenberg, A., Liljencrants, J. & Båvegård, M. (1994). Voice source parameters in continuous speech. Transformation of LF-parameters, *Proc. International Conference on Spoken Language Processing* **3**: 1451–1454.

Fant, G., Liljencrants, J. & Lin, Q. (1985). A four-parameter model of glottal flow, *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR), Royal Institute of Technology, Stockholm* **4**: 1–13.

Flanagan, J. L. (1972a). *Speech Analysis, Synthesis and Perception*, Vol. 1, second edn, Springer-Verlag.

Flanagan, J. L. (1972b). Voices of men and machines, *J. of the Acoustical Society of America* **51**(5A): 1375–1387.

Fujisaki, H. & Ljungqvist, M. (1986). Proposal and evaluation of models for the glottal source waveform, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 11, pp. 1605–1608.

Gerhard, D. (2003). Pitch extraction and fundamental frequency: History and current techniques, Tech. Rep. TR-CS 2003-06.

Gold, B. & Rabiner, L. R. (1969). Parallel processing techniques for estimating pitch periods of speech in the time domain, *J. of the Acoustical Society of America* **46**(2B): 442–448.

Gulick, W. L., Gescheider, G. A. & Frisina, R. D. (1989). *Hearing. Physiological Acoustics, Neural Coding, and Psychoacoustics*, Oxford University Press, New York.

Hedelin, P. (1984). A glottal LPC-vocoder, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 9, pp. 21–24.

Hess, W. (1983). *Pitch Determination of Speech Signals. Algorithms and Devices*, Springer-Verlag.

Holmberg, E. B., Hillman, R. E. & Perkell, J. S. (1988). Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice, *J. of the Acoustical Society of America* **84**(2): 511–529.

Howell, P. & Williams, M. (1988). The contribution of the excitatory source to the perception of neutral vowels in stuttered speech, *J. of the Acoustical Society of America* **84**(1): 80–89.

Howell, P. & Williams, M. (1992). Acoustic analysis and perception of vowels in children's and teenagers' stuttered speech, *J. of the Acoustical Society of America* **91**(3): 1697–1706.

HTK (2008). Hidden Markov model toolkit. http://htk.eng.cam.ac.uk, referenced 30 May 2008.

HTS (2008). HMM-based speech synthesis system. http://hts.sp.nitech.ac.jp, referenced 30 May 2008.

Hynninen, J. & Zacharov, N. (1999). GuineaPig – A generic subjective test system for multichannel audio, *Proc. 106th Audio Engineering Society (AES) Convention* .

Imai, S. (1983). Cepstral analysis synthesis on the mel frequency scale, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 8, pp. 93–96.

Itakura, F. (1975). Line spectrum representation of linear predictor coefficients of speech signals, *J. of the Acoustical Society of America* **57**(S1): S35.

ITU (1996). Methods for subjective determination of transmission quality, International Telecommunication Union, Recommendation ITU-T P.800.

Kabal, P. & Ramachandran, R. V. (1986). The computation of line spectral frequencies using Chebyshev polynomials, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 6, pp. 1419–1426.

Karjalainen, M. (2000). *Kommunikaatioakustiikka*, Report 51, Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, Otamedia Oy, Espoo.

Kedem, B. (1986). Spectral analysis and discrimination by zero-crossings, *Proceedings of the IEEE* **74**(11): 1477–1493.

Kent, R. D. & Read, C. (1992). *The Acoustic Analysis of Speech*, Singular Publishing Group.

Klatt, D. H. (1974). Duration of [s] in English words, *Journal of Speech and Hearing Research* **17**: 41–50.

Klatt, D. H. (1976). Linguistic uses of segmental duration in English: Acoustic and perceptual evidence, *J. of the Acoustical Society of America* **59**(5): 1208–1221.

Klatt, D. H. (1987). Review of text-to-speech conversion for English, *J. of the Acoustical Society of America* **82**(3): 737–793.

Klatt, D. H. & Klatt, L. C. (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers, *J. of the Acoustical Society of America* **87**(2): 820–857.

Kleijn, W. B. & Paliwal, K. K. (1995). An introduction to speech coding, *in* W. B. Kleijn & K. K. Paliwal (eds), *Speech Coding and Synthesis*, Elsevier, chapter 1.

Ling, Z.-H., Wu, Y.-J., Wang, Y.-P., Qin, L. & Wang, R.-H. (2006). USTC system for Blizzard Challenge 2006 – An improved HMM-based speech synthesis method, *Blizzard Challenge Workshop* .

Lisker, L. & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements, *Word* **20**: 384–422.

Markel, J. D. & Gray, A. H. (1980). *Linear Prediction of Speech*, second edn, Springer-Verlag.

MathWorks Inc. (2008). MATLAB version 7.4. http://www.mathworks.com, referenced 30 May 2008.

Matsui, K., Pearson, S. D., Hata, K. & Kamai, T. (1991). Improving naturalness in text-to-speech synthesis using natural glottal source, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 2, pp. 769–772.

Miller, R. L. (1959). Nature of the vocal cord wave, *J. of the Acoustical Society of America* **31**(6): 667–677.

Moore, B. C. J. (1997). *An Introduction to the Psychology of Hearing*, fourth edn, Academic Press.

Moore, B. C. J. & Glasberg, B. R. (1974). Suggested formulae for calculating auditory-filter bandwidths and excitation patterns, *J. of the Acoustical Society of America* **74**(3): 750–753.

Ney, H. (1981). A dynamic programming technique for nonlinear smoothing, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 62–64.

NIST/SEMATECH (2008). e-Handbook of statistical methods. http://www.itl.nist.gov/div898/handbook, referenced 30 May 2008.

Noll, A. M. (1964). Short-time spectrum and "cepstrum" techniques for vocal-pitch detection, *J. of the Acoustical Society of America* **36**(2): 296–302.

Odell, J. (1995). The use of context in large vocabulary speech recognition. PhD thesis, Cambridge University.

Paliwal, K. K. & Kleijn, W. B. (1995). Quantization of LPC parameters, *in* W. B. Kleijn & K. K. Paliwal (eds), *Speech Coding and Synthesis*, Elsevier, chapter 12.

Pickett, J. M. (1999). *The Acoustics of Speech Communication. Fundamentals, Speech Perception Theory, and Technology*, Allyn and Bacon.

Price, P. J. (1989). Male and female voice source characteristics: Inverse filtering results, *Speech Communication* **8**: 216–277.

Pulakka, H. (1995). Analysis of human voice production using inverse filtering, high-speed imaging, and electroglottography. Master's thesis, Helsinki University of Technology.

Rabiner, L. & Juang, B.-H. (1993). *Fundamentals of Speech Recognition*, Prentice Hall.

Rabiner, L. R. (1977). On the use of autocorrelation analysis for pitch detection, *IEEE Trans. Acoustics, Speech, and Signal Processing* **25**(1): 24–33.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE* **77**(2): 257–286.

Rabiner, L. R., Sambur, M. R. & Schmidt, C. E. (1975). Application of nonlinear smoothing algorithm to speech processing, *IEEE Trans. Acoustics, Speech, and Signal Processing* **23**: 552–557.

Rabiner, L. R. & Schafer, R. W. (1978). *Digital Processing of Speech Signals*, Prentice-Hall.

Rosenberg, A. E. (1971). Effect of glottal pulse shape on the quality of natural vowels, *J. of the Acoustical Society of America* **49**(2B): 583–590.

Ross, M. J., Shaffer, H. L., Cohen, A., Freudberg, R. & Manley, H. J. (1974). Average magnitude difference function pitch extractor, *IEEE Trans. Acoustics, Speech, and Signal Processing* **22**(5): 353–362.

Rothenberg, M. (1973). A new inverse-filtering technique for deriving the glottal air flow waveform during voicing, *J. of the Acoustical Society of America* **53**(6): 1632–1645.

Rothenberg, M., Carlson, R., Granström, B. & Gauffin, J. (1975). A three-parameter voice source for speech synthesis, *Speech Communication* **2**: 235–243.

Saito, S. & Itakura, F. (1967). The theoretical consideration of statistically optimum methods for speech spectral density, *Report No. 3107, Electrical Communication Laboratory, N.T.T., Tokyo* .

Soong, F. K. & Juang, B.-H. (1984). Line spectrum pair (LSP) and speech data compression, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 9, pp. 37–40.

Sproat, R. & Olive, J. (1995). An approach to text-to-speech synthesis, *in* W. B. Kleijn & K. K. Paliwal (eds), *Speech Coding and Synthesis*, Elsevier, chapter 17.

Stevens, S. S., Volkmann, J. & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch, *J. of the Acoustical Society of America* **8**(3): 185–190.

Story, B. H. (2002). An overview of the physiology, physics and modeling of the sound source for vowels, *Acoustical Science and Technology* **23**(4): 195–206.

Sulter, A. M. & Wit, H. P. (1996). Glottal volume velocity waveform characteristics in subjects with and without vocal training, related to gender, sound intensity, fundamental frequency, and age, *J. of the Acoustical Society of America* **100**(5): 3360–3373.

Talkin, D. (1995). A robust algorithm for pitch tracking (RAPT), *in* W. B. Kleijn & K. K. Paliwal (eds), *Speech Coding and Synthesis*, Elsevier, chapter 14.

Titze, I. & Sundberg, J. (1992). Vocal intensity in speakers and singers, *J. of the Acoustical Society of America* **91**(5): 2936–2946.

Toda, T. & Tokuda, K. (2007). A speech parameter generation algorithm considering global variance for HMM-based speech synthesis, *IEICE Transactions on Information and Systems* **E90-D**(5): 816–824.

Tokuda, K., Masuko, T., Miyazaki, N. & Kobayashi, T. (1999). Hidden Markov models based on multi-space probability distribution for pitch pattern modeling, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 1, pp. 229–232.

Tokuda, K., Masuko, T., Yamada, T., Kobayashi, T. & Imai, S. (1995). An algorithm for speech parameter generation from continuous mixture HMMs with dynamic features, *Proc. Eurospeech* **1**: 757–760.

Tokuda, K., Yoshimura, T., Masuko, T., Kobayashi, T. & Kitamura, T. (2000). Speech parameter generation algorithms for HMM-based speech synthesis, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 3, pp. 1315–1318.

Tokuda, K., Zen, H. & Black, A. W. (2002). An HMM-based speech synthesis system applied to English, *Proceedings of 2002 IEEE Workshop on Speech Synthesis* pp. 227–230.

Umezaki, T. & Itakura, F. (1986). Analysis of time fluctuating characteristics of linear predictive coefficients, *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 11, pp. 1257–1260.

Viterbi, A. (1967). Error bounds for convolutional codes and an asymptotically optimum decoding algorithm, *IEEE Transactions on Information Theory* **13**(2): 260–269.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. & Kitamura, T. (1998). Duration modeling in HMM-based speech synthesis system, *Proc. International Conference on Spoken Language Processing* **2**: 29–32.

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T. & Kitamura, T. (1999). Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis, *Proc. Eurospeech* pp. 2374–2350.

Zen, H., Nose, T., Yamagishi, J., Sako, S., Masuko, T., Black, A. & Tokuda, K. (2007). The HMM-based speech synthesis system (HTS) version 2.0, *Sixth ISCA Workshop on Speech Synthesis* pp. 294–299.

Zen, H., Tokuda, K., Masuko, T., Kobayashi, T. & Kitamura, T. (2004). Hidden semi-Markov model based speech synthesis, *Proc. International Conference on Spoken Language Processing* **2**: 1397–1400.

Zwicker, E., Flottorp, G. & Stevens, S. S. (1957). Critical band width in loudness summation, *J. of the Acoustical Society of America* **29**(5): 548–557.

Appendix A

Details of the CCR Test

Table A.1: Sentences used in the CCR test.

1	Useimmat olivat nelissäkymmenissä.
2	Tulosten julkistamisen yhteydessä tulisi aina käydä ilmi paitsi otantamenetelmä, myös relevantin kadon osuus.
3	Tällä hetkellä kesäjuhlan taiteellisena johtajana tunnen olevani lähinnä rahankerjuuosaston päällikkö, ja selittelystä vastaava toimihenkilö.
4	Sotamuistot ovat vain alkuja pitkille kertomuksille.
5	Siihen taas poliisi ei nähnyt minkäänlaisia perusteita.
6	Nykyhetkestä katsoen tuohon ajatuskaavaan luuduttiin liian tiukasti, liian pitkäksi aikaa.
7	Niissä on lämpöä ja melankoliaa.
8	Niinpä kapinaa riennettiin kukistamaan.
9	Myös puolison sisaren miestä sanotaan langoksi.
10	Minusta tämä on tyhmyyttä.

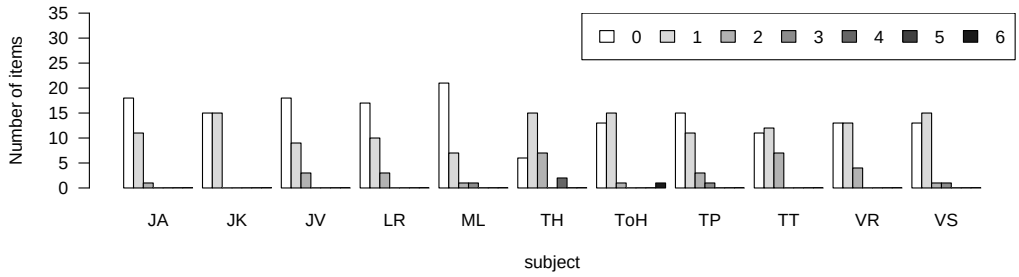


Figure A.1: Differences between the scores of the same sample pairs for each subject. Ideally, the difference should be zero.

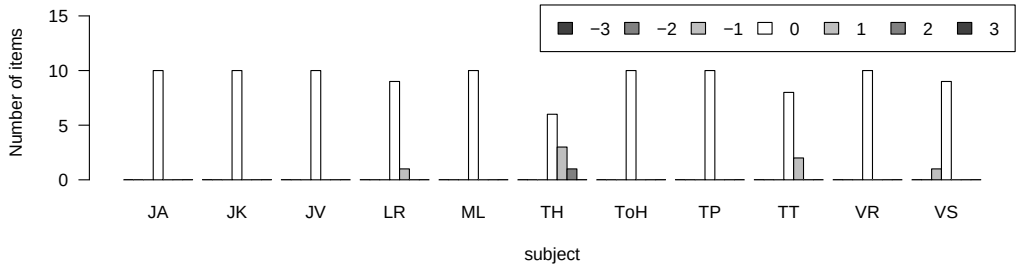


Figure A.2: Scores for the null pairs for each subject. Ideally, the score for the null pair should be zero.

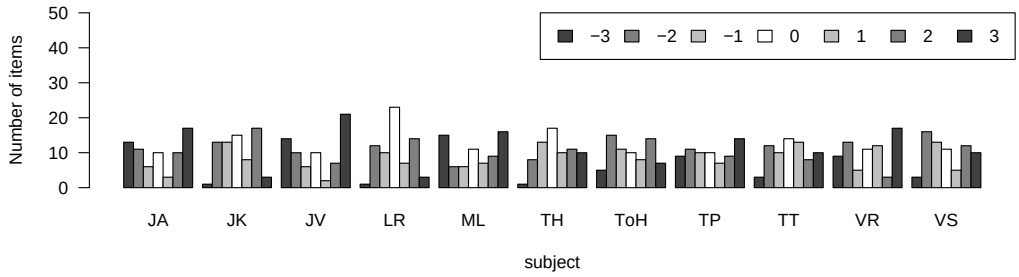


Figure A.3: Distribution of the given scores for each subject. Ideally, subjects should have utilized the entire scale.

Appendix B

Details of the Pair Comparison Test

Table B.1: Sentences used in the pair comparison test.

1	Tällainen tori oli nostajien mielestä hyvä nimenomaan laajapohjaisuudessaan.
2	Siitä maksetaan myös hyviä hintoja.
3	Se toi tuulahduksen toisenlaisesta aikakaudesta tämän päivän politiikkaan.
4	Ruotsalaiset itseasiassa tartuttivat huolensa meihin.
5	Nykyiset äänestäjät eivät Suomen sisäpolittiikkaa heilauta.
6	Maailma pyörii sittenkin, hyvää yötä.
7	Maksajia ei ilmaantunut.
8	Keväällä sisäinen jännitys laukesi.
9	Kesällä saattaa olla lämmintäkin.
10	Ja jälleen kerran hänen potilaansa kokosi voimansa ja alkoi toipua.

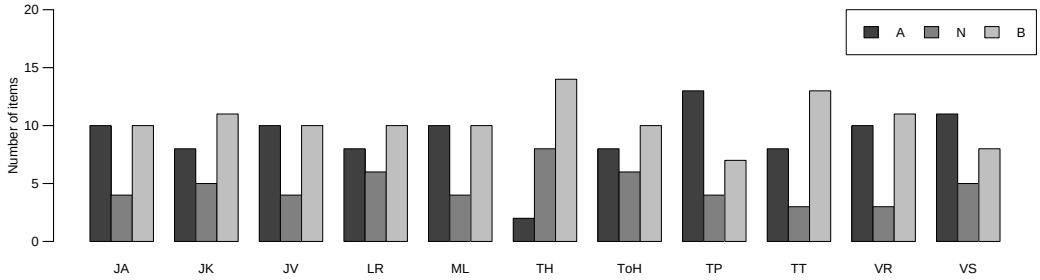


Figure B.1: Distribution of the answers to A, B, and "no preference" (N) for each subject. Ideally, the number of answers to A and B should be equal.

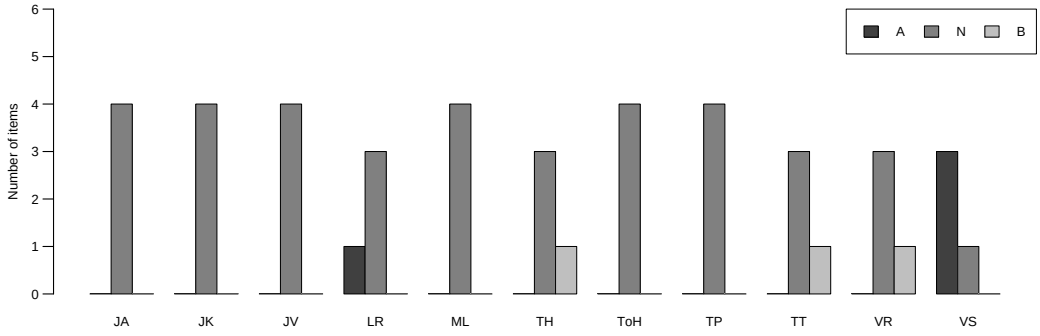


Figure B.2: Answers to the null pairs trials for each subject. Ideally, no preference should be addressed for either of the samples.

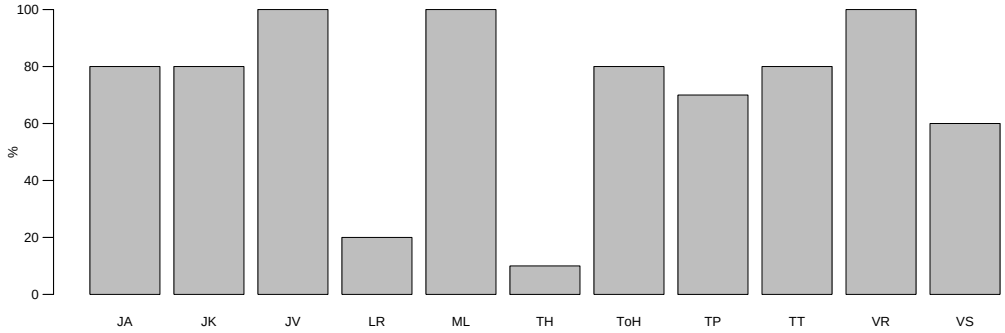


Figure B.3: Consistency of the answers for each subject measured by the proportion of a sample receiving the same answer on both occurrences of the same sample pair.

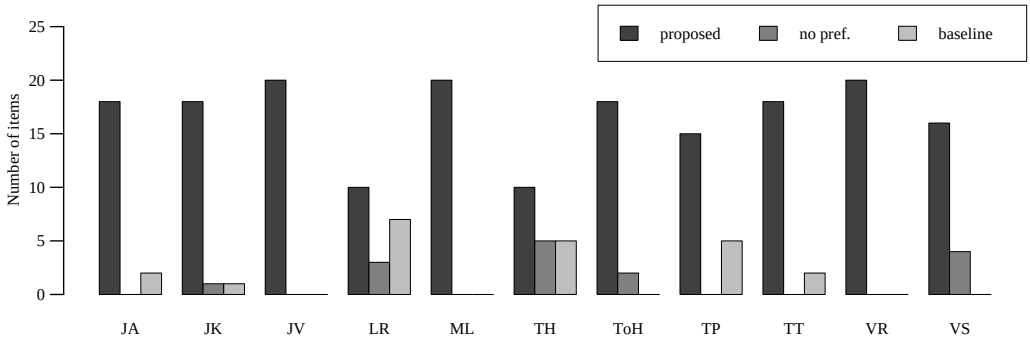


Figure B.4: Answers to different methods by each subject.

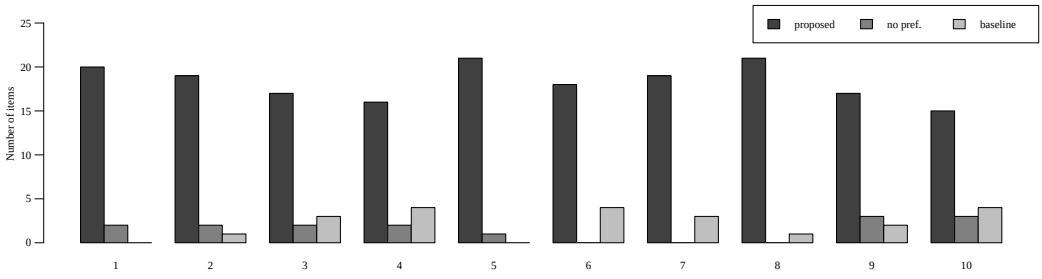


Figure B.5: Answers to different methods by each sentence.