Thesis for the degree of Licentiate of Engineering

On Prosodic Modification of Speech

Barbara Resch

KTH Electrical Engineering

Sound and Image Processing Laboratory
School of Electrical Engineering
KTH (Royal Institute of Technology)

Stockholm 2006

Resch, Barbara
On Prosodic Modification of Speech

ISBN 91-7178-267-2
TRITA-EE 2006:002
ISSN 1653 - 5146

Sound and Image Processing Laboratory
School of Electrical Engineering
KTH (Royal Institute of Technology)
SE-100 44 Stockholm, Sweden
Telephone + 46 (0)8-790 7790

Abstract

Prosodic modification has become of major theoretical and practical interest in the field of speech processing research over the last decades. Algorithms for time and pitch scaling are used both for speech modification and for speech synthesis. The thesis consists of an introduction providing an overview and discussion of existing techniques for time and pitch scaling and of three research papers in this area.

In paper A a system for time synchronization of speech is presented. It performs an alignment of two utterances of the same sentence, where one of the utterances is modified in time scale so as to be synchronized with the other utterance. The system is based on Dynamic Time Warping (DTW) and the Waveform Similarity Overlap and Add (WSOLA) method, a technique for time scaling of speech signals. Paper B and C complement each other and present a novel speech representation system that facilitates both time and pitch scaling of speech signals. Paper A describes a method to warp a signal with time-varying pitch to a signal with constant pitch. For this an accurate continuous pitch track is needed. The continuous pitch track is described as a B-spline expansion with coefficients that are selected to maximize a periodicity criterion. The warping to a constant pitch corresponds to the first stage of the system presented in paper C, which describes a two-stage transform that exploits long-term periodicity to obtain a sparse representation of speech. The new system facilitates a decomposition into a voiced and unvoiced component.

List of Papers

The thesis is based on the following papers:

[A] B. Resch, W.B. Kleijn, "Time synchronization of speech", in *Proc. Maveba*, 2003, pp. 215-218, Florence, Italy

[B] B. Resch, M. Nilsson, A. Ekman and W.B. Kleijn, "Estimation of the instantaneous pitch in speech", to be submitted to *IEEE Transactions on Audio, Speech and Language Processing, 2006*

[C] M. Nilsson, B. Resch, M.Y. Kim and W.B. Kleijn, "A canonical representation of speech" to be submitted to *IEEE Transactions on Audio, Speech and Language Processing, 2006*

Summary of the contributions of the author of the thesis to papers A-C:

[A] Did the major part in the theoretical derivations, conducted all experiments and did the major part in writing the paper.

[B] Did the major part in the theoretical derivations, conducted all experiments and did the major part in writing the paper.

[C] Helped with the theoretical derivations, experiments and writing the paper.

Acknowledgements

I would like to thank everybody who supported me during my studies. Many people have contributed in various ways to make the last three years a memorable time. First, I would like to thank my supervisor Professor Bastiaan Kleijn. I have learnt a great deal from your guidance during my time at the Sound and Image Processing Lab.

I would like to thank all my current and past colleagues: Professor Arne Leijon, Anders Ekman, David Zhao, Elisabet Molin, Harald Pobloth, Jan Plasberg, Jonas Lindblom, Jonas Samuelsson, Kahye Song, Karolina Smeds, Martin Dahlquist, Mattias Nilsson, Moo Young Kim, Peter Nordqvist, Renat Vafin, Sriram Srinivasan and Volodya Grancharov for creating such a friendly and enjoyable atmosphere that paves the way for fruitful technical discussions, but also a lot of fun and out-of-office activities. I also would like to thank the Lab's guest researchers who contributed to the nice working environment, Geoffrey Chan, Christian Feldbauer, Professor Peter Kabal, Jesús De Vicente Peña, Davor Petrinovic and Shenghui Zhao. Dora Söderberg - thank you for all the help with administrative issues, and being a 'normal' person in our nerdy environment. I would like to express my gratitude especially to the ones who worked together with me for some time in my main research project: Anders, Christian, Jonas L, Kahye, Mattias, Moo Young and Professor Peter Kabal. Thanks for all the help and interesting discussions we had. Especially my thanks to Mattias cannot be overstated, who helped me immensely in innumerable discussions during the last year. Sincere thanks also to Sriram for your time discussing my work.

Christian, Jonas S, Mattias and Sriram - thank you for valuable comments on the introduction of this thesis. Anders - thank you for the proofreading of paper B. I thank all my colleagues for the moral support, and would like to mention especially Jan, Jonas L, Jonas S, Mattias and Sriram, who encouraged me in difficult times.

I feel deeply indebted to all my friends here in Stockholm and in Austria (or where ever they are) for being there for me, spending time with me, vitalizing me in my spare time with their presence. Finally, I'm sincerely grateful to my family for all the love, care and support they give me.

Contents

Part I

Introduction

Chapter 1

Introduction

Speech is a natural way of communication between people. It is much more than just the information that is hidden in the words that are said. When listening to speech, we perceive not only what is said, but also how it is said. The way of speaking conveys a lot of information that is automatically processed in our brains to give an overall impression of the message we hear.

When speech is processed, special attention has to be paid to the way things are said. In most speech coding algorithms, the goal is to reconstruct the speech, as similar to the original as possible, not just to make the message understood. In speech recognition, the way of speaking often has a major influence on the performance. If the speech to be recognized differs too much from the speech used for training the recognizer, the performance of the recognizer degrades. In speech synthesis, the way of speaking is clearly one of the most important factors to be considered. The term prosody refers to distinctive variations of pitch, timing and stress in spoken language. The task in speech synthesis is to artificially generate speech with a given linguistic context (the words), spoken with a certain prosody. The prosody depends on the linguistic context, but also on the speaker, and many other factors (e.g., the emotional state of the speaker) that determine how we perceive and interpret the synthesized speech.

State-of-the-art speech synthesis systems are based on concatenative synthesis [1–7]. When building such a system, utterances from speakers that are to be featured by the system are recorded that contain all the different sounds of a language. These utterances are then segmented into their phonetic units that can be put together to form new sentences. In practical systems, the segments to be connected consist of more than one phonetic unit, (e.g, three units, triphones) to capture the effect of coarticulation. Existing synthesis systems can be broadly divided into two

classes [8]: Traditional concatenative synthesis methods, e.g. [4–7] and unit selection synthesis methods, e.g. [1, 2].

For traditional concatentive synthesis systems, the prosody of the segments is modified such that the segments that are concatenated fit together, and also according to the target contours, determined by the prosody of the sentence to be synthesized. The operations that are required to modify the prosody are to change the time scale, the pitch scale and the energy contour of the segments. Unit selection synthesis methods do generally have a larger database than traditional concatentive synthesis systems and search for segments matching the desired linguistic data structure that are then used without applying signal modification. The main limiting factors for the performance of traditional concatentive synthesis systems is the naturalness of the target contour and the quality of the signal modification algorithms. In the unit selection methods, the availability of segments to express any linguistic data structure in the speech database constitutes the main problem, since the number of distinct prosodic and phonetic contexts that can occur is extremely large. This becomes even more a problem in the increasing field of synthesizing expressive speech, e.g., [1, 9], which requires larger deviations in both pitch and time scale. Hence, speech synthesis systems with an unrestricted vocabulary and flexibility in the synthesis such as to express emotions, demand the use of signal modification algorithms.

Algorithms for time and pitch-scaling find use not only in speech synthesis systems, but also in many other applications. For these applications the algorithms for the prosodic modification are normally applied to larger portions of speech, as opposed to the segments in concatenative synthesis. The principles of the algorithms for time and pitch scaling are the same.

In sound engineering both time and pitch scaling techniques are used for the post processing of recorded audio data. Time scaling can be used to adjust a recorded audio track in duration to a certain movie sequence. It can also be applied for post-synchronization from outdoor recordings, where the desired speech is disturbed by noise from the environment, with clean studio speech.

Algorithms for prosodic modification cannot only be of use in professional studio environments, but also in devices from consumer electronics, which can be enriched with extra features that are appreciated by the users. Recently the use of audio books has become more popular, where one can buy a CD or tape with the recording of somebody reading a book. By providing the functionality of time scaling the audio material, the user can adjust the speed of the played sound according to his wishes and needs.

The same functionality is also useful for books and programs for foreign language learning. For a beginner, it will be much easier to listen to slowly spoken text; for a more advanced learner it might be a good challenge to try to understand fast spoken text. For language learning, the application of pitch scaling as well as time scaling can be beneficial in the learning process. To learn correct pronunciation for the new language, it can be advantageous to listen to native text spoken with a pitch similar to the student's own pitch. In addition to the aforementioned applications, time and pitch scaling are of use in chat programs, where the users may want to disguise their voice.

The research in the field of prosodic modifications started in the 60s with the phase vocoder [10]. Since then many different approaches have been presented in the literature. Whereas the research in the beginning was more oriented towards presenting algorithms that made it possible to perform time and pitch scaling, the main focus of current research is to develop methods that sound more natural than existing methods. Most of the methods are explicitly or implicitly based on a simple model of speech production, the source filter model.

In chapter 1 we describe properties of speech, relate them to the source filter model and discuss several signal representations that are useful to understand the remainder of the thesis. A more detailed specification of the task of time and pitch scaling is given in chapter 2, followed by an overview of different methods for time and pitch-scaling presented in the literature. Finally in chapter 3, we describe the contributions of the thesis, we present the general ideas and give a short summary of the findings and outcomes of the papers included in the second part of this thesis.

Chapter 2

Background

In this chapter, we provide the necessary background for the thesis. The aim of the chapter is twofold, to provide basic knowledge about speech, and its properties that are relevant to the topic of speech modification, and to complement the papers that are included in the thesis, by presenting methods and techniques that could not be described in sufficient depth in the papers due to lack of space.

We discuss the properties of the speech signal from the point of view of speech production, both in the physiological sense (section 1), and on the basis of a simple signal processing oriented model (section 2). In section 3 we present various ways to represent and describe the speech signal that are relevant to the remainder of the thesis.

1 Properties of speech

The properties of speech are best understood when considering the mechanisms behind speech production. Speech basically consists of a sequence of different sounds out of a set of sounds. The set of sounds that are used in one specific language is referred to as phonemes. The generating power for producing the speech signal comes from the lungs, from where an air stream is released towards the vocal chords. When the vocal chords oscillate in the air stream, the air flow gets chopped into a sequence of nearly periodic pulses (glottal pulses) and the sound that is produced is said to be *voiced*. When the vocal chords are not vibrating, because they are too tense or slack, the sound is said to be *unvoiced*. The air flow from the vocal chords is passed on through the so-called vocal tract, consisting of the oral and nasal cavities as shown in Figure 2.1.

vowel	F1	F2	F3	example
iy	270	2290	3010	beet
ih	390	1990	2550	bit
eh	530	1840	2480	bet
ae	660	1720	2410	bat
ah	520	1190	2390	but
aa	730	1090	2240	hot
ao	570	840	2410	bought
uh	440	1020	2240	foot
uw	300	870	2240	boot
er	490	1350	1690	bird

Table 2.1: Formant frequencies of vowels [11].

Let us consider voiced speech in more detail. The short-time (about 20-30 ms) frequency spectrum of the pulse train emerging from the oscillation of the vocal chords consists of the first harmonic located at the fundamental frequency, and a number of overtones, which are harmonics at integer multiples of the fundamental frequency. The vocal tract can be brought into different shapes that form resonances at certain frequencies. These resonances are called *formants*, and they cause amplifications of those harmonics that lie close to the resonances. The location of the resonance frequencies determines the sound characteristic, and discerns between different vowels. Normally, there are three resonances of significance below 3500 Hz. Table 1 shows the formant frequencies of the phonemes that belong to the category of vowels. During voiced sounds the spectral envelope of the speech signal is mainly determined by the shape of the vocal tract. The spectral shape of the glottal pulses shows a low-pass characteristic (-12 dB per octave), the radiation from the lips causes a highpass ($+6$ dB per octave) effect.

For unvoiced sounds the air flow from the lungs becomes turbulent at a constriction somewhere in the vocal tract, producing a noise-like sound. The location of the constriction depends on the sound being produced. E.g., for the phoneme /f/ the constriction is near the lips, for /sh/ it is at the back of the oral tract. The spectral envelope of unvoiced sounds generally shows high-pass spectral characteristics.

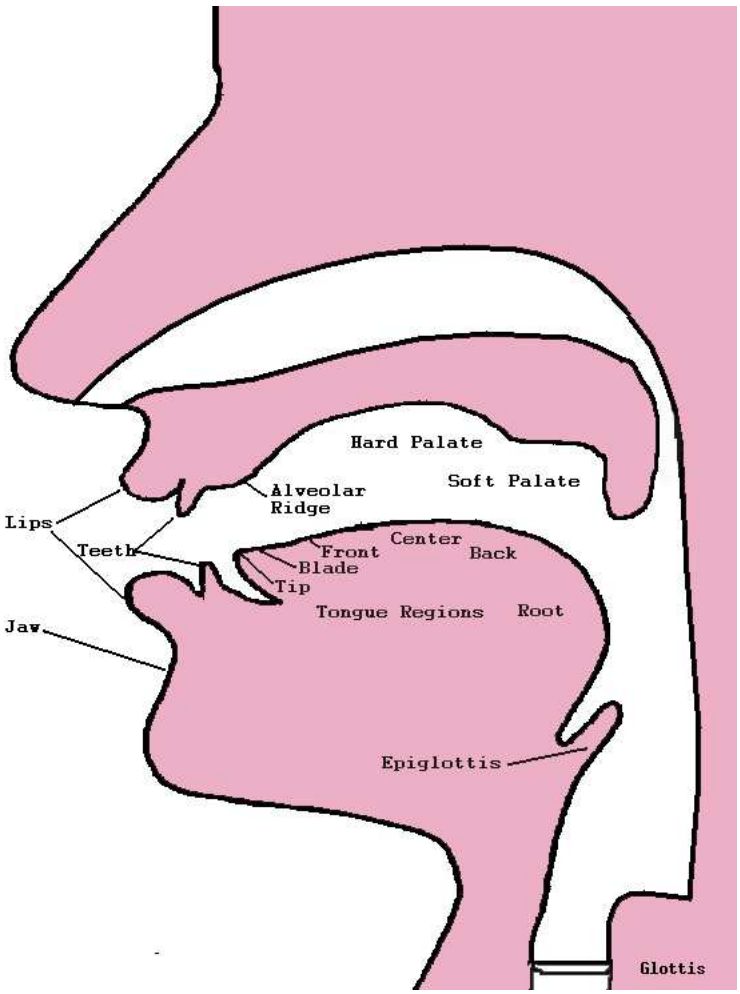

Figure 2.1: Speech production: The vocal tract

2 The source-filter model of speech production

When we perform modification of the time-scale or the pitch of speech we use methods from signal processing to modify certain properties of the speech signal. It is therefore reasonable to look at the speech production process from a signal processing point of view. In this section we give a brief description of a simple speech production model that is commonly

used in speech processing tasks. A more elaborate discussion on speech production models can be found in [11, 12].

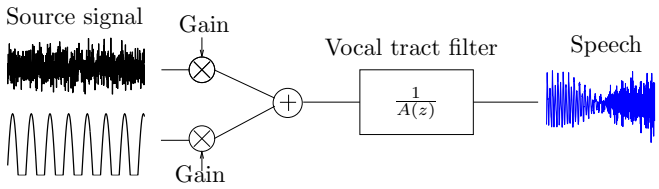

Figure 2.2: The source filter model for the production of speech

A signal processing model, as shown in Figure 2.2 is often used to describe the speech production mechanism. The speech signal is modelled as an excitation signal (source signal) that is filtered through a time-varying filter, which models the resonances of the vocal tract. In such a simplistic model the tilt of the spectrum stemming from the radiation from the lips can be included either in the source signal, or as used in autoregressive modelling (see section 3.1) in the filter part of the model. Similarly, the low-pass characteristics of the glottal pulses can also be captured by the filter, assuming a spectrally flat excitation signal. The source signal is created as a weighted sum of a periodic signal, consisting of a pulse train of glottal pulses and a noise component. The frequency of the source signal is referred to as the pitch, or the fundamental frequency of speech. Due to the spectral properties of the glottal pulse train, the spectrum of voiced speech is harmonic with its components at multiples of the fundamental frequency.

For most speech processing algorithms the speech signal is assumed to be stationary over short segments of time (typically 20 - 30 ms). This assumption can be motivated by the fact that the shape of the vocal tract is varying slowly. Many speech processing methods, e.g., [11, 12] therefore operate on segments of the aforementioned duration. The samples of a segment are often weighted by a smooth windowing function [13] to avoid distortions caused by the abrupt start and end of the segment.

3 Signal representations

In the following, we briefly present techniques related to the source filter model of speech, and speech processing in general. For a more detailed discussion, we refer to [11, 12, 14].

3.1 Autoregressive modelling of speech

The resonances of the vocal tract can be modelled using autoregressive (AR) modelling, where it is assumed that the values of the speech signal depend on a weighted sum of previous samples plus noise. The transfer function of the vocal tract is then described by an all-pole function $H(z) = \frac{1}{A(z)}$. In that way, the excitation signal is modelled as spectrally flat and the spectral tilt of the excitation signal stemming from the vocal chords and the spectral tilt caused by the radiation from the mouth are also modelled by $H(z) = \frac{1}{A(z)}$.

The autoregressive coefficients of $A(z)$ are obtained by linear prediction analysis, where the current samples of a speech signal are predicted as a linear combination of the past values,

$$\tilde{x}(n) = \sum_{k=1}^{p} a_k x(n-k) \tag{2.1}$$

where a_k are the filter coefficients belonging to the filter $A(z)$ and p is the order of the LP analysis. By filtering the speech signal by the inverse filter of $H(z)$, the short-time correlation can be removed, and thus, the signal is whitened. The whitened signal and the filter coefficients can be used as a representation of the original signal for the purpose of coding. Such a procedure is often referred to as Linear Predictive (LP) Coding [15, 16].

The filter coefficients a_k are estimated by minimization of the mean squared error between the speech signal $x(n)$ and the predicted speech signal $\tilde{x}(n)$. This can done under the assumption that the speech is stationary over short time intervals (20 - 30 ms) and that the autocorrelation function can be estimated accurately, using the autocorrelation method. The special structure of the equation system to be solved allows for an efficient handling using the Levinson Durbin algorithm, e.g. [12].

3.2 The Discrete Time Fourier Transform and the Discrete Fourier transform

The Discrete Time Fourier Transform (DTFT), e.g., [12, 13], can be seen as a tool for the analysis of the spectral properties of discrete signals. The DTFT of a discrete signal $x(n)$ is defined as a complex valued function:

$$X(f) = \sum_{n \in \mathbb{Z}} x(n) e^{-j2\pi f n}, \tag{2.2}$$

where f is a real variable denoting the frequency (normalized by the sampling frequency), $\mathbb{Z}$ denotes the set of integers and $j = \sqrt{-1}$.

In practice the DTFT is evaluated at discrete frequencies $f_k = k/N$ that are equally spaced in the range of 0 to 1. The resulting transform is referred to as Discrete Fourier Transform (DFT) and defined for finite-length signals.

3.3 The Discrete Cosine Transform

Similar to the Fourier transform, the discrete cosine transform (DCT), e.g., [13, 17], can be used for analyzing the spectral properties of a signal. In case of the DCT, the basis sequences are real and consist of cosines of varying frequencies. Since the cosine functions are periodic and symmetric, the DCT implies assumptions on the periodicity and symmetry of the original signal. The different ways of how to form a periodic sequence from a finite length sequence, determine the exact construction of the DCT basis sequences. Figure 2.3 shows how the periodic continuation for the four most common DCT transforms is done. In total there are 16 orthonormal transforms for real sequences. The DCT is often preferred over the DFT since it provides better energy compaction, with just a few of the transform coefficients representing the majority of the energy in the sequence, which is closer to the Karhunen-Loeve-Transformation (KLT) [1] at a lower computational cost.

3.4 The Short Time Fourier Transform

The short time Fourier transform (STFT) can be seen as a way to represent the signal both in the time domain and frequency domain. It is used by most of the frequency domain methods for time and pitch scaling throughout section 3. A summarized description of the STFT is therefore beneficial to understand the general principles of these methods.

The short time Fourier transform (STFT) transforms short segments of the signal that are windowed. The positioning of the windows is given by the analysis time instants t_a, which are often regularly spaced. The STFT for a signal is a function of the time instant t_a, and the frequency f,

$$X(t_a, f) = \sum_{n \in \mathbb{Z}} w_a(n) x(t_a + n) e^{-j2\pi f n}. \tag{2.3}$$

The analysis window $w_a(n)$ is of finite support and normally symmetric. The Fourier transform of two signals that are multiplied in the time domain, corresponds to a convolution of the spectra of the signals in the frequency domain. Thus, the chosen windowing function has a significant influence on the STFT. More specifically, a trade-off has to be made

[1] The Karhunen-Loeve-Transformation (KLT) for a random vector X is a data dependent transform that diagonalizes the covariance matrix of X. It transforms a random vector into a vector of uncorrelated components.

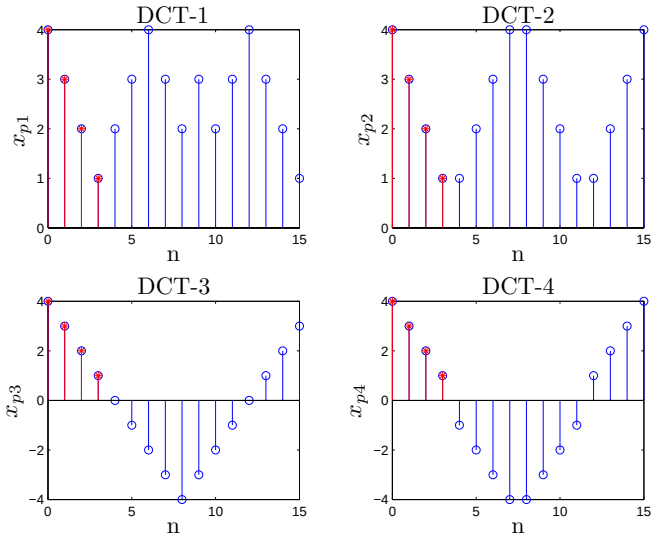

Figure 2.3: Periodic extension of the four-point sequence $x(n)$ (plotted with stars) for the DCT types, DCT-1, DCT-2, DCT-3 and DCT-4.

between spectral resolution, defined by the width of the main-lobe of the window, and spectral leakage, defined by the attenuation of the side-lobes.

Figure 2.4 shows an example where the STFT is obtained for a segment of voiced speech of 20 ms, comparing the usage of a smooth Hamming window to a rectangular window. The Hamming window has a wider main-lobe and lower side-lobes compared to the rectangular window. The Figure shows how the spectra of the different windows are positioned at the harmonics of the signal as weighted images due to the convolution.

In practice, the DFT is used instead of the DTFT yielding a discrete resolution in frequency. One can obtain a graphical representation of the time-frequency properties of a signal, by plotting the absolute magnitude of a sequence of STFT. The analysis instances are regularly spaced and the time is generally represented on the horizontal axis, the frequency on the vertical axis, as shown in Figure 3.1. This yields the so-called spectrogram. The Fourier transform can be replaced by a discrete cosine transform as described in section 3.3, to represent the signal in the frequency domain based on cosines instead of exponential functions.

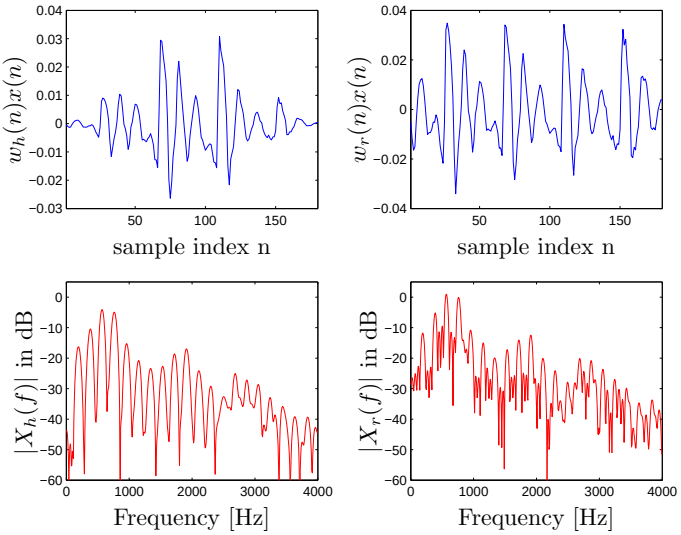

Figure 2.4: Influence of the analysis windowing function $w_a(n)$ on the STFT. Example on a segment of voiced speech of 20 ms, sampled at 8 kHz. Left: Hamming window, $w_a(n) = w_h(n)$. Right: Rectangular window, $w_a(n) = w_r(n)$

3.5 The Modulated Lapped Transform

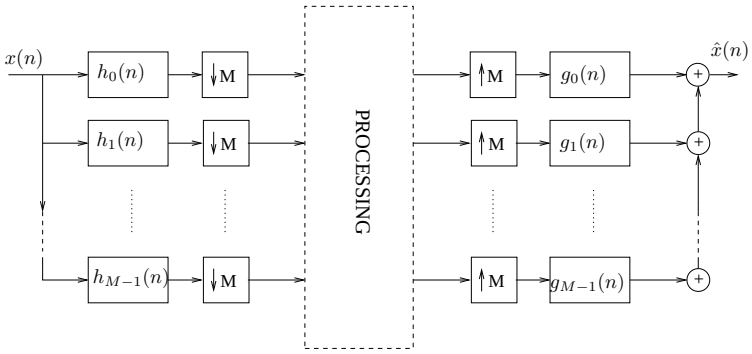

Figure 2.5: Typical filter bank for subband signal processing.

The speech representation system presented in paper C consists of several transformation stages, whereof one is realized as a modulated lapped transform (MLT). On account of this, we include a summarized

and rather intuitive discussion of the MLT here. For a more detailed and rigorous description we refer to [18]. A modulated lapped transform (MLT) is a special kind of filter bank as shown in Figure 2.5. Such a filterbank can be used to decompose the signals into several subbands by means of band pass filters. It consists of two stages, the analysis stage represented by the analysis filters, defined by their impulse responses $h_0(n)$, $h_1(n)$ ... $h_{M-1}(n)$, and the synthesis stage, represented by the synthesis filters defined by their impulse responses $g_0(n)$, $g_1(n)$ ... $g_{M-1}(n)$. In each channel on the analysis stage the signal is downsampled by a factor M, and upsampled by the same factor M on the synthesis side.

Under certain conditions one can design a critically sampled filterbank that yields perfect reconstruction. The MLT is such a filterbank, where the analysis and synthesis filters are cosine modulated low-pass filters of filter length $2M$. The idea of the MLT was first proposed by [19, 20], later discussed more in detail and first referred to as "MLT" in [18].

3.6 B-Spline interpolation

In this section we introduce the concept of interpolation using B-splines, which is used in paper B to express a continuous representation for a pitch track, and also the speech signal. For a more detailed description of B-splines, we refer to [21–24].

A discrete signal $s_d(k)$ can be interpolated using a B-spline expansion by a sum of shifted weighted polynomial basis functions β^n of finite support,

$$s(x) = \sum_{k \in \mathbb{Z}} c_k \beta^n(x - k), \tag{2.4}$$

where c_k are the coefficients that describe the signal unambiguously. Whereas the discrete signal $s_d(k)$ is only defined on $k \in \mathbb{Z}$, the continuous signal $s(x)$ is defined for all real x. The signal $s(x)$ consists of piecewise polynomial functions that are smoothly patched together, such that the continuity of the function and its derivatives up to order $n - 1$ are guaranteed.

The basis functions β^n are bell shaped, symmetrical functions, that can be constructed by the $(n + 1)$-fold convolution of the spline of order 0, β^0:

$$\beta^n(x) = \underbrace{\beta^0(x) * \beta^0(x) ... * \beta^0(x)}_{n+1} \tag{2.5}$$

where $*$ denotes the convolution operator [2].

[2] The continuous-time convolution of the functions $f(x)$ and $g(x)$ is defined as $h(x) = f(x) * g(x) = \int_{-\infty}^{\infty} f(x - y)g(y)dy$

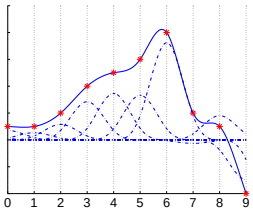

Figure 2.6: Interpolation with splines. The interpolation (solid line) of the data points is the sum of shifted weighted cubic B-splines.

The spline of order 0 has the shape of a rectangular pulse:

$$\beta^0(x) = \begin{cases} 1 & \text{for } -\frac{1}{2} < x < \frac{1}{2}, \\ \frac{1}{2} & \text{for } |x| = \frac{1}{2}\ , \\ 0 & \text{otherwise}\ . \end{cases} \tag{2.6}$$

The derivative of a spline expansion with basis functions of order n as in equation (2.4) is easily expressed as a spline expansion with basis functions of order $n-1$,

$$\frac{\partial s(x)}{\partial x} = \sum_{k \in Z} (c_k - c_{k-1}) \beta^{n-1}(x - k + 1/2). \tag{2.7}$$

This relation is useful especially in a framework with optimization using a gradient based method as presented in paper B.

When using a B-spline expansion for the interpolation of discrete signals, one has to determine the coefficients c_k such that the interpolating function goes through the given discrete data points. This can either be done using a matrix framework [21], or by means of filtering if the available data points are equally spaced [22, 23].

Figure 2.6 shows an example of B-spline interpolation. The shifted basis functions are located at the positions of the sample points, often referred to as knots.

Chapter 3

Prosodic modification of speech

1 Time scaling

The objective of time scale modification is to alter the speaking rate without changing the spectral content of the original speech. Considering the source-filter model of speech, this means that the time evolution of the excitation signal and the vocal tract filter needs to be time scaled.

Time scale modification can be performed uniformly, changing the rate by a certain factor, or non-uniformly according to the prosody or the sound characteristics of different parts of speech. A time scaling function, the so-called time warping function, assigns time instants in the original signal (analysis instants) to corresponding time instants in the new signal (synthesis instants). Non-uniform time scaling can increase the intelligibility of the resulting time-scaled speech [25]. Finding a mapping to increase the intelligibility is a nontrivial task and has been dealt with in [26, 27].

Non-uniform time scaling finds use also in concatenative speech synthesis, where the properties of the segments to concatenate have to be modified according to linguistic constraints. The time scaling methods discussed in the following subsections are suitable for both uniform and non uniform time-scaling.

Figure 3.1 shows an example of a speech segment that is time-scaled. The figure shows the time domain waveform, and the corresponding spec-

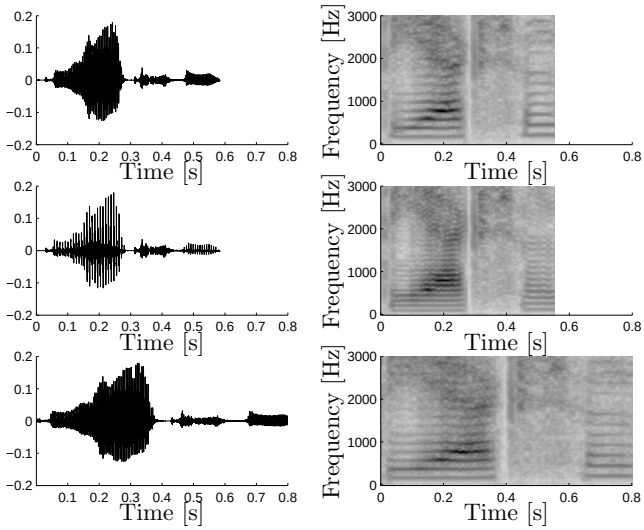

Figure 3.1: An example for pitch and time scaling on the speech samples 'the action' using Pitch Synchronous Overlap and Add (PSOLA) [4]. First row: Original and the corresponding spectrogram. Second row: After pitch scaling to 75 % of the original fundamental frequency and the corresponding spectrogram. Third row: After time scaling to 1.3 times slower and the corresponding spectrogram.

trograms. In the time-scaled signal (third row) both the harmonic structure and the formant tracks are stretched in time compared to the original.

2 Pitch scaling

Pitch corresponds to the fundamental frequency of the vibration of the vocal chords, and thus, to the fundamental frequency of the excitation signal. It is only defined for voiced, or partly voiced speech. The objective of pitch scale modification is to alter the fundamental frequency without affecting the spectral envelope or the formant structure of the signal. Decomposing the speech signal into a source signal and the filter coefficients describing the vocal tract is therefore essential for pitch scaling.

Pitch scaling can be done uniformly or non-uniformly. Uniform pitch scaling means that the pitch is altered by a constant factor over the whole speech signal. Thereby the intonation does not change. By changing the pitch in a non-uniform manner, the intonation of the speech utterance

can be modified. This finds application mainly in concatenative speech synthesis, where a certain intonation is imposed on the speech, depending on the linguistic context.

Figure 3.1 shows an example of a speech signal that is pitch scaled. The signal in the second row is obtained by applying a down scaling of the pitch of the signal in the first row. In the pitch-scaled signal the fundamental frequency and thus the harmonic structure is modified, whereas the formant tracks are preserved.

3 Algorithms for prosodic modifications

In the following, we discuss different methods for time and pitch scaling that were proposed in the literature. The methods can be divided into two groups, parametric methods, and non-parametric methods. In parametric methods the speech signal is represented by a set of parameters. Performing pitch or time scaling is done by changing these parameters, and synthesizing the signal with the modified parameters. In non-parametric models the time and pitch scaling is done on the speech itself. The modification can be carried out on segments of the speech signal in the time domain or in the frequency domain.

3.1 Non-parametric methods

Within the group of non-parametric methods another subdivision can be made, whether the modification is carried out in the frequency or the time domain. Time domain methods are commonly less computationally complex, since the operation to transform the signal into the frequency domain is omitted. A detailed description of different non-parametric methods can be found in [28, 29].

Time domain based methods

Time domain methods operate on segments of the original speech waveform, which are synthesized in an overlap-add manner. The concept of time scaling using overlap and add (OLA) synthesis was first introduced by [30]. It operates by concatenating short windowed segments, which are extracted from the original signal at time instants given by the time warping function. Repeating the same segment two or more times in the synthesis is equivalent to time stretching, whereas the concatenation of segments that are non-adjacent corresponds to time compression. When the segments are concatenated, the periodic structure of the signal has to be preserved, corresponding to retaining the local pitch of the speech

signal. The different OLA methods provide different strategies to solve this problem.

In the Synchronized Overlap and Add method (SOLA) [31] the positioning of the segment in the synthesized signal is chosen to give maximum correlation to the previously synthesized segment. The Waveform Similarity Overlap and Add method (WSOLA) [32] is based on the same principle, but varying the point of extraction from the original signal, such that the periodicity is preserved. In [33] it is the duration of the segment that is synthesized that is altered to avoid discontinuities. The methods of [31–33] method operate directly on the waveform, and are single purpose methods, to stretch or compress the waveform. They make use of the periodic structure, i.e., the local pitch, without the need of an explicit pitch estimation.

In the time domain Pitch Synchronous Overlap and Add (TD-PSOLA) method [4, 34] the length of the segments is proportional to the local pitch period. The windowed segments are of the length of a multiple of the local pitch period. In this manner, when overlapping the synthesized segments, the periodicity is preserved. For PSOLA, the speech needs to be labelled with pitch marks, that mark the position of the pitch pulse in each pitch period. For unvoiced parts the pitch marks are evenly spread with a fixed spacing. A reliable and accurate pitch estimation algorithm is needed to obtain these pitch marks. Unlike the single purpose methods discussed before, the PSOLA method is suitable for both time and pitch modification. For the purpose of pitch scaling, the pitch synchronous weighted segments are added with more overlap for increased pitch, or less overlap for lowered pitch. For TD-PSOLA, this procedure is performed on the speech signal directly, without separating the signal into a residual signal and filter coefficients. Due to the pitch-synchronized processing, the shape of the waveform, stemming from the pitch pulse on the source signal filtered through the vocal tract, is preserved, and thus, the formant structure unaltered.

Time domain based methods are of low complexity compared to frequency domain based methods and often result in high-quality speech modification for moderate time scaling factors. A major limitation of all OLA methods is the introduction of tonal noise when stretching unvoiced parts [28]. The repetition of unvoiced segments leads to a long-term correlation that causes undesired periodicity. In [4] it was therefore suggested to time reverse every second segment during unvoiced regions. This gives an improvement during purely unvoiced regions, but is not applicable during speech regions, where the source is a mixture of a periodic signal and noise, such as voiced fricatives.

1. Compute the STFT for all analysis time instants and calculate the instantaneous frequencies in each channel.
2. Calculate the instantaneous phases for the corresponding synthesis instants.
3. Calculate the modified signal by the inverse STFT of the short time signals and overlap add.

Table 3.1: Algorithm for time scaling in the phase vocoder. From [28].

Frequency domain based methods

Frequency domain methods are mostly based on or strongly influenced by the phase vocoder. The phase vocoder is a well established tool for time and frequency scale modifications. The concept was first introduced by Flanagan and Golden [10]. Methods based on the phase vocoder,e.g., [10, 36, 37], work in the frequency domain of speech. The short time Fourier spectra of overlapping segments of the signal are modified for both time and pitch scaling. Thereby the sequence of the STFT signals has to fulfill consistency constraints, since the segments are overlapping. More precisely, one has to take care that the phases of the Fourier transform associated with each bin in the modified STFT are preserved in successive segments.

A time scaling method based on the short time Fourier transformation was proposed by Portnoff [37]. In [28] a slightly modified and simpler scheme is presented, where time scaling is performed in the frequency domain by modification of the evolution of amplitudes and frequencies of the STFT of the signal. The cut-off frequency for the analysis window $w(n)$ is chosen to be less than half the spacing between the pitch harmonic such that the shifted and weighted images of the main lobe of the window are non-overlapping. This is referred to as narrow-band analysis condition [37]. One can derive an expression for the so-called instantaneous phase and instantaneous frequency that specifies the phase and frequency in a continuous manner for each channel from the STFT from two successive time instants that are close enough to assume the phase as a slowly evolving function [28]. The basic algorithm is summarized in table 3.1 [28].

This kind of processing does not make any distinction between voiced and unvoiced sounds, which causes tonal effects in unvoiced regions for large stretching factors. Aside from that, phase vocoder time scaling often produces chorusing effects (subjective sensation that several people are speaking at the same time) [28], reverberation and transient smearing (slight loss of percussiveness) [36]. In [38] and [36] these phenomena are

1. Perform a source filter decomposition
2. Perform a STFT on the source signal
3. Perform linear interpolation on the real and imaginary parts of the short time spectra from the analysis instants
4. If needed (if step 3 corresponds to a downsampling), perform spectral copying or folding to regenerate high-frequency components
5. Compute the instantaneous frequencies and phases at the synthesis instants
6. Calculate the short time modified signal by the inverse STFT and overlap add
7. Apply the synthesis filter of the source-filter decomposition to the pitch scaled source signal

Table 3.2: Algorithm for pitch Scaling in the phase vocoder. From [28].

discussed and analyzed in detail providing insights into their causes, and techniques to reduce these effects are presented.

Pitch scaling on the short time Fourier transform requires the separation of the speech signal into a source signal, representing the excitation, and a time-varying filter, representing the vocal tract according to the speech production model (section 2). In the phase vocoder the source signal is pitch-scaled using a resampling procedure. The resampling is done in the frequency domain by linear interpolation of the short-time Fourier coefficients given at the analysis instants to obtain the spectra at the the synthesis instants. The algorithm is summarized in table 3.2. A detailed description of the phase vocoder implementation for pitch scaling can be found in [28].

Frequency Domain Pitch Synchronous Overlap and Add (FD-PSOLA) [4] is used only for pitch scaling and operates on the short time Fourier spectra of windowed pitch-synchronous segments (pitch cycles). The modification is carried out in the frequency domain of the residual, by interpolation between the STFT as for the phase vocoder. Since the STFT spectra are pitch synchronous the phase does not need not be adjusted as it is the case for the phase vocoder. Similar as in the TD-PSOLA method, the pitch scaling causes a time scaling that has to be compensated by the

TD-PSOLA time scaling algorithm.

In [39] a new one-stage phase vocoder technique for pitch-shifting has been introduced that is based on peak detection in the short time Fourier transform, where the frequency axis is divided into 'regions of influence' dominated by each peak. These regions are shifted to a new location independently to other peaks, in contrast to conventional phase vocoder pitch scaling where all the peaks are scaled in frequency with a constant factor. In doing so, the computational cost of the new method is independent of the amount of the modification, and the method facilitates more complex frequency domain modifications such as harmonizing, partial stretching and so on. In this technique there is no source-filter separation and the modifications are performed directly on the speech signal. This results in a shift of the formant frequencies in the modified signal. Therefore this method is referred to as a method for pitch shifting, chorusing, harmonizing and other exotic effects.

In general, frequency-based methods tend to have problems due to the limited time-frequency resolution and accumulated phase and amplitude errors [40]. It should be mentioned for completeness that phase vocoder techniques are popular methods not only for speech modification but also for modification of other audio signals, such as music. They provide a powerful tool for spectral modifications of computer music. In these applications the disadvantages as unnatural sound quality and reverberation are not necessarily objectionable, but often even desired.

3.2 Parametric methods

Parametric methods are based on a certain underlying model of speech. The parameters are estimated from the original speech, and used explicitly for the modification of speech. The modified speech signal is produced in a synthesis process using the parameters of the model that have been changed. Most parametric methods described here can be considered as frequency domain based methods [41–47], since the underlying signal models generally represent the signal in the frequency domain. A time domain parametric method is described in [48].

In the linear predictive vocoder [49] speech is modelled by a convolution of a pulse train with a time-varying filter, representing the vocal tract. This method does not provide high-quality modifications.

Sinusoidal models [50–53] represent voiced speech as a sum of sinusoids with slowly varying amplitudes, instantaneous frequencies and phases. Quatieri and McAulay [41, 54] present a sinusoidal speech anal-

ysis/synthesis system, which is applied to both time scale modification and pitch scaling in [52]. The signal is generated according to a speech production model [11] as an excitation waveform passing through a vocal tract. The excitation is modelled by a sum of sine waves and the vocal tract is realized as a time-varying filter. The amplitudes, frequencies and phases of the excitation are obtained from a 512 point STFT by using a peak picking algorithm for both voiced and unvoiced speech. The window duration is set to 2.5 times the speaker's measured average pitch with a minimum width of 20 ms. The frequencies of the detected sinusoids are not constrained to be harmonic. A frame-to-frame peak matching algorithm as described in [52] is used to match spectral peaks in an 'optimal' sense in adjacent frames, introducing a 'birth' and 'death' of sinusoidal components. During unvoiced speech, the minimum window length of 20 ms provides a sufficiently dense sampling such that the frequency tracks are numerous and dense in frequency with frequent events of 'birth' or 'death' of singular tracks. During voiced speech the tracks are less dense, regularly spaced in frequency and long living over several frames. The tracker is able to adapt quickly to transitory behavior, such as onsets. The time varying filter is represented by the so called system amplitude and phase along the frequency tracks of the excitation, which are estimated using homomorphic deconvolution [12]. For time scaling the excitation frequencies and amplitudes and vocal tract parameters are modified in time for each frequency track. For pitch scaling, the frequency of the excitation function is scaled, and the system amplitudes and phases, corresponding to the vocal tract, are found at the new locations of the frequency tracks.

A further improvement of this system is presented in [42], which is referred to as shape-invariant modification. The basic idea is to preserve the original waveform shape in modified speech, as in contrast to [41, 54], where the time scaling of the excitation phases leads to waveform dispersion. In the newer approach of [42] the excitation is represented in terms of pitch pulse locations, where the sine waves are in phase and added coherently. The excitation phase function is created directly with respect to the new time scale to avoid dispersion. The method requires accurate pitch estimation, and it is reported in [42] that pitch errors lead to discontinuities, perceived as glitches in the reconstructed speech. Speech is only modified when pitch can be estimated, and is left unchanged for purely unvoiced regions.

A sinusoidal model that is combined with analysis by synthesis/overlap and add synthesis (ABS/OLA) is presented in [45]. The signal is expressed as a sum of overlapped short-time signals, represented as a sum of sinusoidal components. The method uses a successive approximation-based analysis by synthesis procedure rather than peak picking to determine

the model parameters. Thereby the mean squared error between the modelled and the real speech is minimized by successively adding one more sinusoidal component to the modelled speech at each iteration. In contrast to [41,42,54] the number of sinusoids is not limited by the number of identifiable peaks. The computational load of the analysis-by-synthesis procedure, including an exhaustive frequency search is considerably higher than for [41, 42, 54]. An extension of the ABS/OLA method is presented in [46], accounting especially for modification of unvoiced speech. Unlike the previous method, where voiced and unvoiced speech are treated the same way leading to artifacts such as 'tonal' effects in time-scale expansion, a binary voiced/unvoiced decision is first made and the phases belonging to unvoiced or noise like segments are randomized.

Harmonic and noise models (HNM), initially proposed in [43] are conceptually similar to sinusoidal methods. In this two-fold representation, the signal is decomposed as a sum of a deterministic component of harmonically related sinusoids, and a stochastic component accounting for everything that is not described by the harmonic components. A further development of such a representation was used in [44] for speech modification. The sinusoidal parameters are estimated from pitch synchronous segments in the time domain using a least squares approach. The complex amplitudes of the harmonics within segments are expressed as linear functions of time. The stochastic part is defined as the residual signal obtained by subtraction of the sinusoidal components from the original speech. Its frequency contents are modelled by a time varying all-pole filter and the time domain structure by an energy envelope function. The synthesis is done by overlapping and adding pitch synchronous segments, similarly to the PSOLA method [4]. For time-scaling, two different schemes for the deterministic and stochastic part are applied. The deterministic part is generated as the sum of the harmonics with a time-scaled version of the linear function for the amplitudes. The stochastic part is generated as a white Gaussian noise filtered by the all-pole filter and modulated by a time-scaled energy envelope function. This approach allows for large time-stretching factors without generating undesired periodicity in unvoiced sounds. The application of the harmonic plus noise model in concatenative synthesis is discussed in [6].

A parametric method for time-scaling based on a nonlinear oscillator model operating in the time domain of speech is presented in [48]. The speech is first subjected to LP analysis and additional low-pass filtering to obtain the so called glottis-signal. Short-time stationary segments of the glottis signal are interpreted as attractors in the state space of an underlying dynamical system and a nonlinear predictor is trained online for each block of the glottis signal. The modified glottis signal is synthesized by self-

oscillation of this system, followed by the LP synthesis and a high-pass filter. An extension of the method is presented in [56,57], to address the issue that the low dimensional oscillator cannot capture the noise-like components of the speech signal properly. A second nonlinear predictor is introduced trained on the signal obtained by subtraction of the synthesized speech and the original speech that models the time domain structure of the noise.

Chapter 4

Contributions of the present work

The three papers that are presented in this thesis focus on different aspects of speech modification. Paper A presents a system for the synchronization of speech that allows to perform an alignment of two utterances of the same sentence, where one of the utterances is modified in time scale so as to be synchronized with the other utterance. Papers B and C present two parts for a system that facilitates a novel speech representation that is suitable for time and pitch scaling of speech signals. In paper B a pitch estimation method is developed to determine a continuous track of the instantaneous pitch. Given such a pitch track, the speech signal can be warped to a signal with normalized pitch, as required for the system that is presented in paper C. This system performs a decomposition of the speech signal into a voiced and an unvoiced part, which can be modified in different ways for the purpose of time and pitch scaling, yielding a natural quality of the modified speech.

1 Time Synchronization of Speech

In paper A we present a system for time synchronization of speech that time-aligns two utterances of the same sentence. One of the utterances is defined to be the so-called reference utterance, whose time-scale is used as a reference for the other utterance. A time warping vector is obtained by time alignment of the two utterances, that specifies the non-uniform time scaling of the second utterance, such that it is synchronized with the reference utterance. Such a time synchronization system can be used for foreign language learning, in speech therapy, or for post synchronization of

recordings for audio-for-video applications.

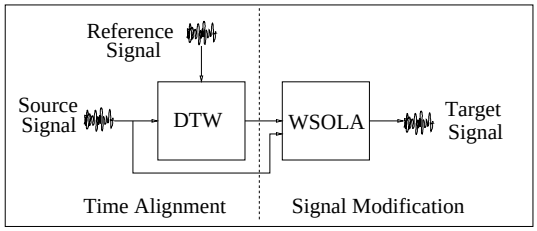

Figure 4.1: Block diagram of the time synchronization system. The first block consists of Dynamic Time Warping (DTW) and the second of Waveform Similarity Overlap and Add (WSOLA).

1.1 System description

A block diagram of the basic system is depicted in Figure 4.1. The signal to be modified is denoted as the *source signal*, the signal that serves as a reference is denoted as a *reference signal*, and the modified signal is denoted as a *target signal*. The system consists of two stages, in the first stage the alignment between the source and reference signal is determined, in the second stage the non-uniform time scaling is performed according to the alignment from stage one.

The alignment in the first stage, represented by a time warping vector is found by use of Dynamic Time Warping (DTW) [58]. By applying local constraints the curvature of the time warping vector can be controlled. For the time synchronization system the task requires a large flexibility in the alignment, since the utterances spoken by different persons may show considerable differences in local speaking rate and also phonetic differences in the pronunciation of words. However, it has been shown that the time alignment stemming from Dynamic Time Warping with large flexibility is often not suited for the following second modification stage, since it produces an artificially sounding target signal. In paper A we propose an accumulated local penalty constraint to obtain a time-alignment that yields a more natural quality for the target signal. The local penalty constraint makes such curvatures that cause unnatural sound quality of the target signal less likely, but does not prevent them from occurring. Therefore, we introduce a post-processing step to smoothen the time warping vector to assure a natural quality for the synthesis of the target signal. The time-scale modification in

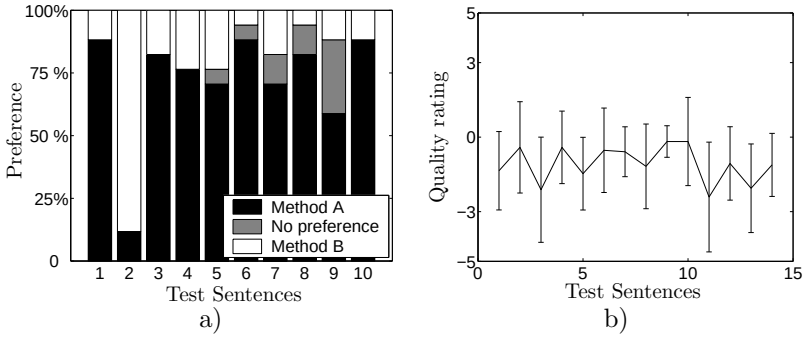

Figure 4.2: Results from our listening test for different test sentences (horizontal axis). a) A-B preference test. Method A: Time synchronization system. Method B: Natural synchronization b) Quality rating comparing the modified and original utterances from +5 (time synchronized version sounds much better) to -5 (time synchronized version sounds much worse)

the second stage is performed using the WSOLA algorithm [32] according to the specification of the time warping vector. To decrease the periodicity that appears when realizing long time stretching factors in unvoiced regions, we time-reverse every third segment in the synthesis that is classified as unvoiced.

1.2 Results

A listening test was carried out to evaluate the performance of the time synchronization system. The test was conducted on 17 listeners, using 10 different utterances of the TIMIT [59] database. The results of the listening test are shown in Figure 4.2. Graph a) presents the results of a preference test, where the listeners were asked to judge the accuracy of the synchronization. The reference system in this test is obtained by human synchronization from recordings were the speakers were asked to speak synchronously to the TIMIT utterances. The second part of the test aimed at evaluating the quality of the target speech comparing to the original unmodified speech as reference utterance in terms of naturalness. Graph b) shows that the modified sentences are perceived as being of good quality on average.

2 Estimation of the Instantaneous Pitch of Speech

The focus of paper B is on the estimation of the instantaneous pitch of speech. We present a method to optimize a continuous track of the instantaneous pitch, which facilitates changes in the pitch between adjacent pitch cycles. Such accurate pitch estimation is required for pitch synchronous processing of speech, as in e.g., speech synthesis [4,44,47], speech coding [60–63] and speech enhancement [64].

2.1 System description

Starting from a harmonic model of a signal with constant fundamental frequency, one can generate a signal with time-varying instantaneous fundamental frequency. The instantaneous fundamental frequency is referred to as pitch. A continuous warping function defines the warping between the two different domains of the signal with the time-varying and the constant pitch. The pitch can be shown to be the time derivative of this warping function. In the paper we describe a method to find the inverse warping function that can be used to warp a signal with time-varying pitch into a signal with constant pitch. Figure 4.3 shows an example of a signal with time varying frequency along the vertical axis, that is warped into a a signal with constant instantaneous fundamental frequency along the horizontal axis and the corresponding warping function.

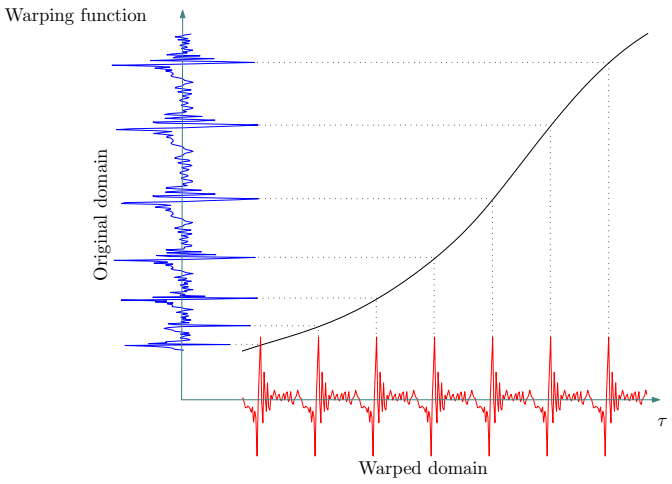

Figure 4.3: Example for the warping function to map the signal to a signal with constant pitch

We obtain the warping function by numerical optimization of a waveform similarity criterion that is evaluated in the warped domain, where the signal is of constant pitch. There we minimize the difference of the energy of the warped signal and the energy of the warped signal shifted for one normalized pitch period using a gradient descent algorithm [65]. The sought warping function is expressed in terms of B-splines [21–23] (see section 3.6). The processing is done on the LP-residual of the signal, where the short time correlation is removed, emphasizing the long term correlation corresponding to the pitch of the signal. Furthermore we apply a multi-stage procedure, where the result of the optimization of the current stage is used as an initial estimate of the next stage to increase the robustness. The different stages consist of low-pass filtering the signal prior to the optimization where the cut-off frequency is increased between succeeding stages.

2.2 Results

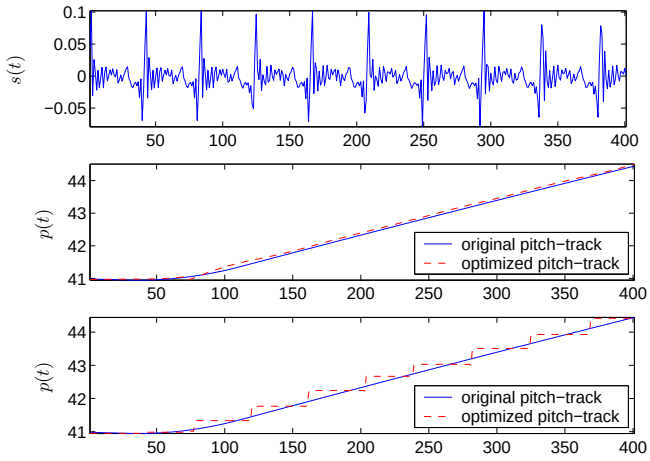

Figure 4.4: Example with an artificial signal with slowly changing pitch. Upper plot: Artificial signal $s(t)$. Middle and Lower plot: Real and optimized instantaneous pitch, where the warping function is modelled by third-order splines (middle plot) and first-order splines (lower plot).

The method is tested both on artificial speech-like signals, where the true pitch track is known, and on real speech signals, namely 20 sentences from the TIMIT database [59]. We evaluate objective measures such as the coefficient of correlation and the ratio of voiced-to-unvoiced component obtained by modulation filtering. In our experiments we test two different

models of the warping function. In one case the warping function is modelled by a piecewise linear function, corresponding to a piecewise constant instantaneous pitch, in the other case the warping function is modelled as a smooth function, corresponding to a smooth instantaneous pitch. Figure 4.4 shows an example where the pitch is evaluated on an artificial speech-like signal, comparing the two cases of piecewise constant and smooth modelling of the pitch. The experiments on the artificial signals and on speech indicate that a piecewise linear model for the pitch is sufficient to obtain good results. The experiments on speech show the superiority of the proposed methods compared to the reference methods, Yin [66] and Praat [67], in terms of the objective measures.

3 A Canonical Representation of Speech

The last paper of this thesis presents a representation for speech that efficiently uses the short-term and long-term dependencies of speech to derive a representation that is both complete and compact. The representation is based on a decomposition of the speech signal into a voiced and an unvoiced component. This decomposition together with the completeness and compactness properties make the representation attractive for speech modification such as time and pitch scaling, and also speech coding.

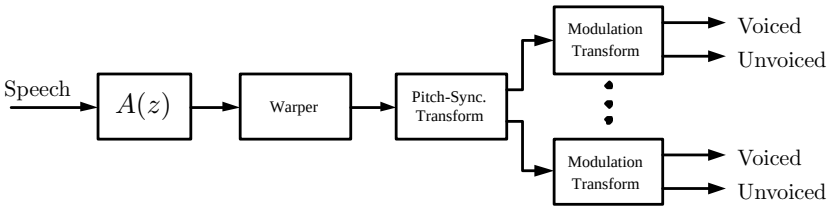

Figure 4.5: Block diagram of the analysis stage.

3.1 System description

The analysis stage of the system is shown in Figure 4.5. In the first stage the signal is subjected to a linear prediction (LP) analysis, where the incoming speech signal is decomposed into a set of time-varying filter coefficients that describe the properties of the vocal tract and a residual signal, as obtained by filtering the speech signal with the filter $A(z)$ to obtain the excitation signal. The resulting excitation signal is warped to a constant pitch using a procedure as described in paper B, and transformed pitch

synchronously to the frequency domain by a modulated lapped transform (MLT) (see section 3.5). The following stage consists of a second transform that operates on the frequency channels of the output of the first transform over time. It is realized as non overlapping discrete cosine transform (DCT) (see section 3.3) with a square window. This transform separates the slowly and faster changing components in the frequency domain and corresponds thus to the decomposition into a voiced and unvoiced component. The support of the first transform is constant, given by the time synchronous structure and defined as two pitch cycles. The second transform has a time varying support that is chosen to maximize the energy concentration of the resulting coefficients. Doing so the second transform performs a segmentation over time, where regions that have similar properties of the spectral fine structure are merged. The synthesis is done analogously to the analysis by applying the inverse transforms in inverse order, yielding perfect reconstruction if the signal is not modified.

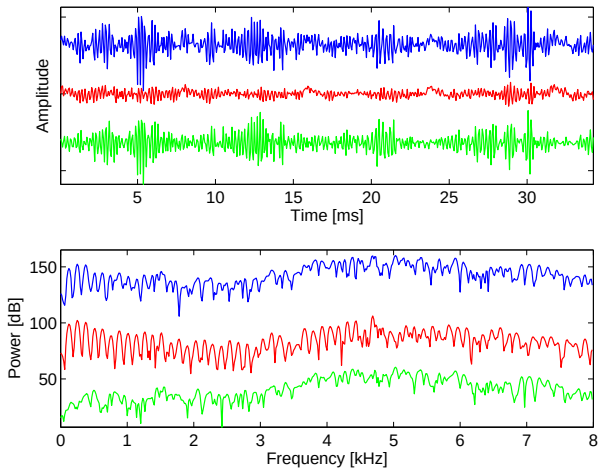

Figure 4.6: Voiced and unvoiced decomposition of a voiced fricative /z/-sound. Upper and lower subplots show the original (top), voiced (middle), and unvoiced (bottom) signal waveforms and power spectra respectively.

3.2 Results

Paper C demonstrates the application of the presented speech representation in the field of speech coding and speech modification. Especially

the decomposition into voiced and unvoiced components that allows for different processing strategies for the different speech segments make it attractive for these applications. Figure 4.6 shows an illustrative example of the voiced and unvoiced decomposition of a voiced fricative. In voiced fricatives (e.g., /j/ in the word judge) the excitation signal contains both a harmonic and a noise-like component. These components are successfully decomposed as can be seen in the figure, where the strong harmonic characteristics are captured in the voiced components up to 5 kHz, whereas the overall signal is dominated by the unvoiced component above 1 kHz.

A listening test has been carried out that shows the applicability of the proposed system for coding and prosodic modification. The phase of the unvoiced component was randomized prior to the synthesis of the signal to remove artificial periodicity that occurs when time-stretching to a large extent. The results of the test show that the reconstructed randomized speech is of high quality, close to the quality of the original signal.

References

[1] A. Iida, N. Campbell, F. Higuchi, and M. Yasumura, "A corpus-based speech synthesis system with emotion," *Speech Commun.*, vol. 40, pp. 161–187, Apr. 2003.

[2] B. Bozkurt, O. Ozturk, and T. Dutoit, "Text design for TTS speech corpus building using a modified greedy selection," in *Proc. Eurospeech, Geneva, Switzerland*, pp. 277–280, Sept. 2003.

[3] P. Rutten, G. Coorman, J. Fackrell, and B. V. Coile, "Issues in corpus based speech synthesis," *IEE Seminar State of the Art in Speech Synth.*, pp. 16/1 –16/7, Apr. 2001.

[4] E. Moulines and F. Charpentier, "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Commun.*, vol. 9, pp. 453–467, Dec. 1990.

[5] T. Dutoit, V. Pagel, N. Pierret, F. Bataille, and O. van der Vrecken, "The MBROLA project: towards a set of high quality speech synthesizers free of use for non commercial purposes," in *Proc. Int. Conf. on Spoken Language Processing*, vol. 3, pp. 1393 –1396, Oct. 1996.

[6] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 21 – 29, Jan. 2001.

[7] A. Acero, "Source-filter models for time-scale pitch-scale modification of speech," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, vol. 2, pp. 881–884, May 1998.

[8] J. Santen, A. Kain, E. Klabbers, and T. Mishra, "Synthesis of prosody using multi-level unit sequences," *Speech Commun.*, vol. 46, pp. 365–375, 2005.

[9] D. S. G. Vine and R. Sahadani, "Synthesis of emotional speech using RP-PSOLA," in *IEE Colloquium on the Stae of the Art in Speech Synthesis*, pp. 8/1 –8/6, Apr. 2000.

[10] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell System Technical J.*, pp. 1493–1509, Nov. 1966.

[11] L. R. Rabiner and R. W. Schaefer, *Digital processing of speech*. Englewood Cliffs,New Jersey: Prentice-Hall, 1978.

[12] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*. Upper Saddle River, New Jersey: Prentice-Hall, 2001.

[13] A. Oppenheim, R. Schaefer, and J. Buck, *Discrete -time signal processing*. Upper Saddle River: Prentice-Hall, 1999.

[14] T. F. Quatieri, *Speech Signal Processing*. Upper Saddle River: Prentice-Hall, 2002.

[15] W. B. Kleijn and K. K. Paliwal, "An introduction to speech coding," in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), pp. 1–44, Elsevier Science Publishers, 1995.

[16] P. Kroon and W. B. Kleijn, "Linear prediction based on analysis-by-aynthesis coding," in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), pp. 79–113, Elsevier Science Publishers, 1995.

[17] G. Strang, "The discrete cosine transform," *SIAM Review*, vol. 41, pp. 135–147, 1999.

[18] H. Malvar, "Lapped transforms for efficient transform/subband coding," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 38, pp. 969 – 978, 1990.

[19] J. Princen, A. Johnson, and A. Bradley, "Subband/transform coding using filter bank designs based on time domain aliasing cancellation," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, vol. 12, pp. 2161 – 2164, 1987.

[20] J. Princen and A. Bradley, "Analysis/synthesis filter bank design based on time domain aliasing cancellation," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 34, pp. 1153 – 1161, 1986.

[21] C. de Boor, *A practical guide to splines*. New York: Springer-Verlag, 2001.

[22] M. Unser, A. Aldroubi, and M. Eden, "B-spline signal processing: Part I - theory," *IEEE Trans. Signal Processing*, vol. 41, no. 2, pp. 821–833, 1993.

[23] M. Unser, A. Aldroubi, and M. Eden, "B-spline signal processing: Part II - efficiency design and applications," *IEEE Trans. Signal Processing*, vol. 41, no. 2, pp. 834 – 848, 1993.

[24] M. Unser, "Splines: a perfect fit for signal and image processing," *IEEE Signal Processing Mag.*, vol. 16, pp. 22 – 38, Nov. 1999.

[25] E. Janse, S. Nooteboom, and H. Quené, "Word-level intelligibility of time-compressed speech: prosodic and segmental factors," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, vol. 41, pp. 273–529, Oct. 2003.

[26] S. Lee, H. D. Kim, and H. S. Kim, "Variable time-scale modification of speech using transient information," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, vol. 2, pp. 1319 –1322, Apr. 1997.

[27] M. Covell, M. Withgott, and M. Slaney, "MACH1: nonuniform time-scale modification of speech," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, vol. 1, pp. 349 –352, Apr. 1998.

[28] E. Moulines and J. Laroche, "Non-parametric techniques for pitch-scale and time-scale modification of speech," *Speech Commun.*, vol. 16, pp. 175–205, Feb. 1995.

[29] E. Moulines and W. Verhelst, "Time-domain and frequency-domain techniques for prosodic modification of speech," in *Speech Coding and Synthesis* (W. B. Kleijn and K. K. Paliwal, eds.), pp. 519–555, Elsevier Science Publishers, 1995.

[30] R. Crochiere, "A weighted overlap-add method of short-time Fourier analysis/synthesis," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 28, pp. 99 –102, Feb. 1980.

[31] S. Roucos and A. Wilgus, "High quality time-scale modification of speech," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, vol. 10, pp. 493 – 496, Apr. 1985.

[32] W. Verhelst and M. Roelands, "An overlap-add technique based on waveform similarity (WSOLA) for high quality time-scale modification of speech," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, vol. 2, pp. 554 –557, Apr. 1993.

[33] J. Laroche, "Autocorrelation method for high-quality time/pitch-scaling," in *Proc. IEEE Workshop Applications of Signal Processing to Audio and Acoust.*, pp. 131 –134, Oct. 1993.

[34] C. Hamon, E. Mouline, and F. Charpentier, "A diphone synthesis system based on time-domain prosodic modifications of speech," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, vol. 1, pp. 238 –241, May 1989.

[35] T. Dutoit and H. Leich, "MBR-PSOLA: Text-to-speech synthesis based on an MBE re-synthesis of the segments database," *Speech Commun.*, vol. 13, pp. 435–440, Dec. 1993.

[36] J. Laroche and M. Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 323–332, May 1999.

[37] M. Portnoff, "Time-scale modification of speech based on short-time fourier analysis," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 29, pp. 374 – 390, June 1981.

[38] J.Laroche and M. Dolson, "Phase-vocoder: about this phasiness business," in *Proc. IEEE Workshop Applications of Signal Processing to Audio and Acoust.*, p. 4pp, 1997.

[39] J. Laroche and M. Dolson, "New phase vocoder technique for pitch-shifting, harmonizing and other exotic effects," in *Proc. IEEE Workshop Applications of Signal Processing to Audio and Acoust.*, pp. 91–94, Oct. 1999.

[40] A. J. S. Ferreira, "An odd-DFT based approach to time-scale expansion of audio signals," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 441 –453, July 1999.

[41] T. Quatieri and R. J. McAulay, "Speech transformations based on a sinusoidal representation," in *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 34, pp. 1449 –1464, Apr. 1986.

[42] T. Quatieri and R. J. McAulay, "Shape invariant time-scale and pitch modification of speech," *IEEE Trans. Signal Processing*, vol. 40, pp. 497 –510, Mar. 1992.

[43] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 36, pp. 1223–1235, Aug. 1988.

[44] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic+noise model," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, vol. 2, pp. 550 – 553, Apr. 1993.

[45] E. B. George and M. J. T. Smith, "Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 389–406, Sept. 1997.

[46] M. W. Macon and M. A. Clements, "Sinusoidal modeling and modification of unvoiced speech," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 557–560, Nov. 1997.

[47] D. O'Brien and A. I. C. Monaghan, "Concatenative synthesis based on a harmonic model," *IEEE Trans. Speech Audio Processing*, vol. 9, pp. 11 –20, Jan. 2001.

[48] G. Kubin and W. B. Kleijn, "Time-scale modification of speech based on a nonlinear oscillator model," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, vol. i, pp. I/453 –I/456, Apr. 1994.

[49] B. S. Atal and S. L. Hanauer, "Speech analysis and synthesis by linear prediction of the speech wave," *JASA*, vol. 50, pp. 637–655, 1971.

[50] P. Hedelin, "A tone oriented voice excited vocoder," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, vol. 6, pp. 205 –208, Apr. 1981.

[51] L. Almeida and F. Silva, "Variable-frequency synthesis: An improved harmonic coding scheme," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, vol. 9, pp. 437 –440, Mar. 1984.

[52] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 34, pp. 744–754, Aug. 1986.

[53] R. J. McAulay and T. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis* (W. Kleijn and K. Paliwal, eds.), (Amsterdam), Elsevier Science Publishers, 1995.

[54] T. Quatieri and R. J. McAulay, "Speech transformations based on a sinusoidal representation," *IEEE Trans. Acoust. Speech, Signal Processing*, vol. 34, pp. 1449 – 1464, Dec. 1986.

[55] D. O'Brian and A. Monaghan, "Shape invariant pitch and time-scale modification of speech based on a harmonic model," in *Improvements in Speech Synthesis* (E. K. et al., ed.), John Wileys and Sons, Ltd, 2002.

[56] E. Rank and G. Kubin, "Towards an oscillator-plus-noise model for speech synthesis," in *ISCA Tutorial and Research Workshop on Non-linear Speech Processing*, May 2003.

[57] E. Rank and G. Kubin, "An oscillator-plus-noise model for speech synthesis," *Speech Commun.*, p. in press, 2005.

[58] H. F. Silverman and D. P. Morgan, "The application of dynamic programming to connected speech recognition," *IEEE ASSP Mag.*, vol. 7, pp. 6–25, July 1990.

[59] "DARPA-TIMIT," *Acoustic-phonetic continuous speech corpus, NIST Speech Disc 1-1.1*, 1990.

[60] R. Taori, R. Sluijter, and E. Kathmann, "Speech compression using pitch synchronous interpolation," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Processing*, vol. 1, pp. 512 – 515, 1995.

[61] W. B. Kleijn and J. Haagen, "Waveform interpolation for speech coding and synthesis," in *Speech Coding and Synthesis* (W. Kleijn and K. Paliwal, eds.), pp. 175 – 205, Amsterdam: Elsevier Science Publishers, 1995.

[62] P. Veprek and A. Bradley, "Speech compression by vector quantization of epochs," in *Proc. IEEE Symp. Signal Processing and Its Applications*, vol. 1, pp. 491 – 494, Aug. 1999.

[63] N. R. Chong-White and I. Burnett, "Accurate, critically sampled characteristic waveform surface construction for waveform interpolation decomposition," *IEE Electronic Letters*, vol. 36, no. 14, pp. 1245–1247, 2000.

[64] T. S. Y. Kuroiwa, "An improvement of LPC based on noise reduction using pitch synchronous addition," in *Proc. IEEE Int. Symp. Circuits and Systems*, vol. 3, pp. 122 –125, 1999.

[65] J. Nocedal and S. J. Wright, *Numerical optimization.* New York: Springer-Verlag, 1999.

[66] A. de Cheveigne and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.*, vol. 111, no. 4, pp. 834 – 848, 2002.

[67] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International 5*, vol. 9/10, pp. 341–345, 2001.