SLOVAK UNIVERSITY OF TECHNOLOGY BRATISLAVA

Faculty of Electrical Engineering and Information Technology

Department of Radioelectronics

Ing. Anna Madlová

Some Parametric Methods of Speech Processing

PhD Thesis

Supervisor: doc. Ing. František Židek, PhD.

Scientific Field: 26-13-9 Electronics

May 2001

Annotation

The thesis gives a comprehensive study of some parametric methods of speech processing with the emphasis on the sinusoidal model with harmonically related component sine waves. Improvements of known methods and new algorithms are devised for achieving better synthetic speech quality. Various approaches are implemented for the harmonic model with autoregressive and cepstral parametrization. Proposed methods are compared with respect to the spectral measure, the perceived speech quality, and the computational complexity.

Acknowledgement

I would like to express thanks to my supervisor doc. Ing. František Židek, PhD. for providing an interesting idea and his helpful comments in the beginning of my research.

I am very grateful to Ing. Robert Vích, DrSc., Dr.-Ing. h. c. from the Institute of Radio Engineering and Electronics, Academy of Sciences of the Czech Republic, for giving me ideas for next research and a number of constructive comments on my work.

This work has partly been done within the framework of the grants “New Methods and Systems of Signal Processing in Electronics, Medicine, and Ecology” (N^{o} 95/5195/288, 1995-1998), “Signal Processors and Microcontrollers in Electronics, Radiocommunication, and Biomedical Engineering” (N^{o} 95/5195/297, 1995-1999), “Modern Methods and Systems for Signal Processing in Multimedial Communication, Medicine, and Ecology” (N^{o} 1/6097/99, 1999-2001), “Digital Processing of Audio, Video, and Biomedical Signals” (N^{o} 102/VTP/2000, started in 2000). The main part of the work has been done in practical cooperation with the Institute of Radio Engineering and Electronics, Academy of Sciences of the Czech Republic in Prague, Department of Digital Signal Processing, which had started in 1997.

List of Most Important Abbreviations

ABS, AbS	analysis-by-synthesis
AIC	Akaike information criterion
AR	autoregressive
ARMA	autoregressive moving average
BIC	modified Akaike information criterion
CAT	criterion of autoregressive transfer
CELP	code-excited linear prediction
CSM	composite sinusoidal model
FFT	fast Fourier transform
FPE	final prediction error
HAP	harmonic model with AR parametrization
HCP	harmonic model with cepstral parametrization
HNM	harmonic plus noise model
HSX	harmonic-stochastic excitation
H/S	harmonic/stochastic
IFFT	inverse fast Fourier transform
LPC	linear predictive coding
LSF	line spectrum frequencies
LSP	line spectrum pairs
MA	moving average
MBE	multiband excitation
MDL	minimum description length
MEM	maximum entropy method
MHC	multimode harmonic coder
MSE	mean squared error, mean square error
MVDR	minimum variance distortionless response
OLA	overlap-add, overlap-and-add
PCW	pitch-cycle waveform
RMS	root mean square
TD-PSOLA	time-domain pitch-synchronous overlap-add
TTS	text-to-speech
V/UV	voiced/unvoiced

List of Most Important Symbols

$\{a_k\}$	LPC (AR) filter coefficients
$\{A_m\}$	harmonic model amplitudes
B_n (n>0)	n-th formant bandwidth
$\{c_n\}$	(real) cepstrum
e_p	pitch frequency error in points of FFT
e_s	pitch period error in samples
F_0	fundamental (pitch) frequency
$\hat{F}_0$	pitch frequency estimate
F_n (n>0)	n-th formant frequency
f_{max}	maximum voiced frequency
$\{f_m\}$	pitch harmonics
$\{\varphi_m\}$	harmonic model phases
$\varphi^{\min}$	minimum phase
f_s	sampling frequency
G	LPC (AR) filter gain
$H(z)$, $H(e^{j\omega})$	discrete transfer function, and corresponding frequency response
$h(n)$	impulse response
$H_P(z)$	preemphasis filter transfer function
L	pitch period in samples
L_k	k-th frame pitch period in samples
L_P	synthesis frame length with pitch-synchronous beginning
L_S	synthesis frame length in samples
N	analysis frame length in samples
N_A	AR model order
N_C	number of cepstral coefficients
N_F	number of points of FFT
$P(e^{j\omega})$	frequency response of the vocal tract model
$P(z)$	transfer function of the vocal tract model
$r(m)$	autocorrelation function
$S(e^{j\omega})$	speech spectrum
S_F	spectral flatness measure
$s(n)$	speech signal
$s_y(l)$	synthetic speech signal
$\mathbf{T}_k$	k-th frame vector of speech parameters
$w(n)$	weighting window

Contents

1 Introduction

Parametric modelling of speech signals finds its use in speech analysis and synthesis, speech coding, speech recognition, and speaker verification and identification [1]-[5]. Parametric methods of speech processing might be divided into two classes: source-filter modelling with the filter representation of the vocal tract transfer function, and sinusoidal modelling where the source and the system features are included in the parameters of the sinusoidal model. In the source-filter model the vocal tract can be modelled either by a filter bank, or a realizable rational transfer function, or an approximation of a nonrealizable exponential function in homomorphic modelling.

The rational transfer function modelling is equivalent to the parametric spectrum estimation, or the system identification. It includes pole-zero, all-pole, and all-zero models. However, an all-pole model has been used almost exclusively in speech processing being known as a linear predictive coding (LPC) model. For these models the excitation is produced by the impulse train, and/or the random noise.

The homomorphic modelling uses the sequence of cepstral coefficients to parametrize the vocal tract system. It inherently includes both poles and zeroes in its approximation. The excitation is similar to that used in the LPC model.

The sinusoidal modelling is based on superposition of the sine waves comprising properties of the system (vocal tract) and the excitation (glottal) signal as well. If the frequencies of the component sine waves are restricted to be integer multiples of the fundamental (pitch) frequency, the model is called the harmonic model.

The performance of the models is evaluated with respect to their temporal and spectral properties. The model spectral properties are compared with those of traditional spectrum estimation methods using the fast Fourier transform (FFT).

2 Speech Production

Speech sounds are produced by causing modulation of the airflow through constrictions in the airways between the larynx and the lips. This modulation of the flow gives rise to the generation of sound. One type of modulation arises from vibration of the vocal folds, which causes quasiperiodic changes in the space between the vocal folds (the glottis), and hence modulation of the volume flow through the glottis. Another type of modulation is a consequence of turbulence in the flow and hence the generation of turbulence noise. Transient sound sources can also be produced by raising or lowering the pressure behind a closure in the airway, and then rapidly opening this constriction, causing an abrupt change in the pressure behind constriction. The acoustic process involved in production of speech sounds can be modelled as in Figure 2.1.

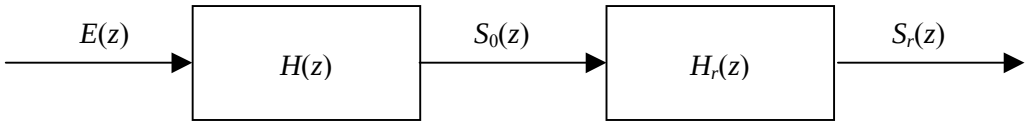

Figure 2.1 Representation of speech sound production as a source $E(z)$ filtered by a transfer function $H(z)$ to give the spectrum $S_0(z)$ at the mouth or nose, and a radiation characteristic $H_r(z)$ to give the spectrum of the radiated sound pressure $S_r(z)$.

One or more sources, with spectrum $E(z)$, form the excitation for an acoustic system with a transfer function $H(z)=S_0(z)/E(z)$, where $S_0(z)$ is the spectrum of the acoustic volume velocity at the mouth or nose, and a radiation characteristic $H_r(z)=S_r(z)/S_0(z)$, where $S_r(z)$ is the spectrum of the radiated sound pressure. Thus we have

$$S_r(z) = E(z) \cdot H(z) \cdot H_r(z). \tag{2.1}$$

When the vocal folds are appropriately positioned and the pressure is raised in the airways below the glottis, the folds are set into vibration and the airflow through the glottis is modulated periodically. The spectrum of this modulated flow is rich in harmonics. This periodic flow forms an acoustic source that provides excitation for the airways above the larynx. The frequency of vibration of the vocal folds during normal speech production is usually in the range 80-160 Hz for adult males, 170-340 Hz for adult females, and 250-500 Hz for younger children [6]. Time waves and spectra of a 24-ms frame of voiced speech (vowel "A") sampled at 8 kHz with different fundamental frequencies F_0 corresponding to male, female, and childish voices are shown in Figure 2.2 a), b), c).

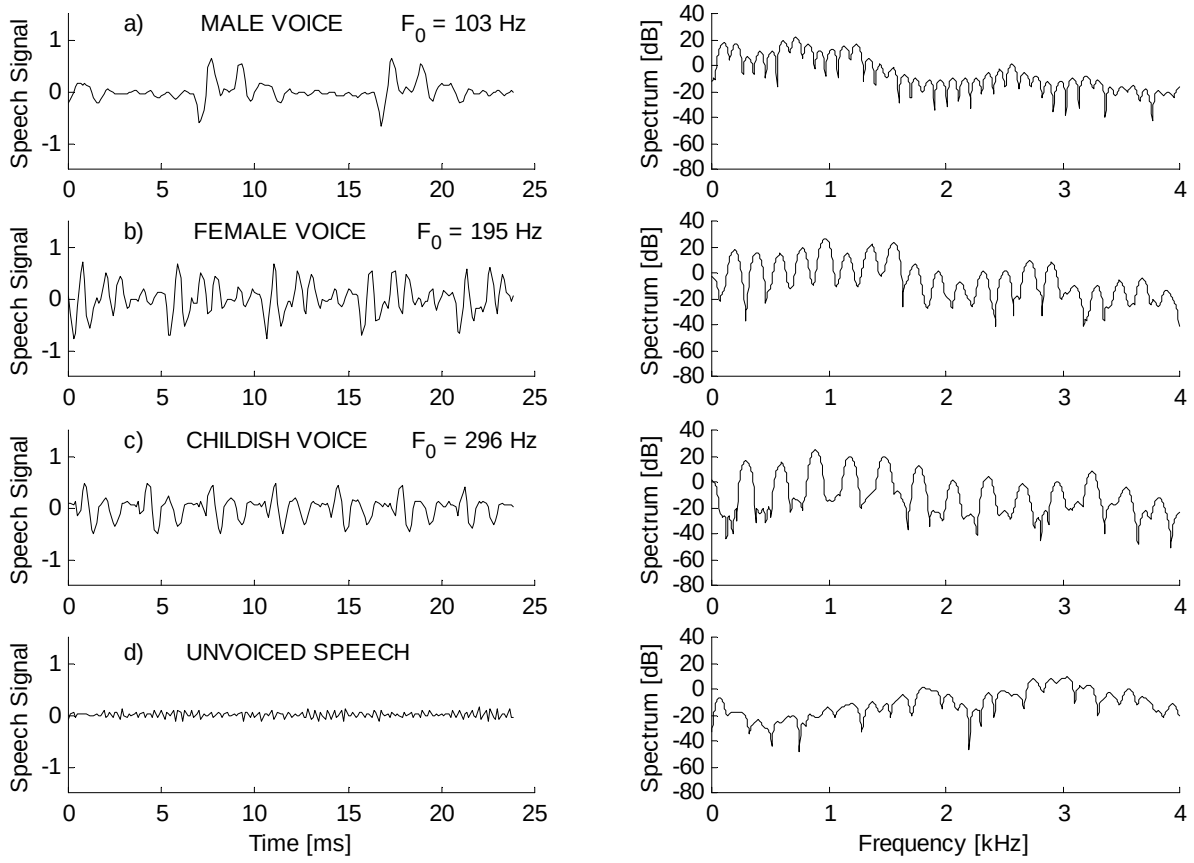

Figure 2.2 Comparison of time waveforms and spectra of voiced speech (vowel "A") for three different voices, and unvoiced speech (consonant "Š").

Vowel sounds are normally produced with a source at the glottis and with the airways above the glottis configured such that any narrowing in the airways is not sufficient to cause a buildup of pressure behind the constriction. The transfer function of the vocal tract for vowels has a relatively simple form for the special case in which the area function is uniform, the acoustic losses are neglected, and the radiation impedance at the mouth opening is assumed to be small.

For a cylinder tube of length x, this transfer function is

$$H(\omega) = \frac{1}{\cos(\omega \cdot x / v)}. \tag{2.2}$$

The transfer function has only poles, and these are at frequencies $F_1 = v/4x$, $F_2 = 3v/4x$, $F_3 = 5v/4x$,... These are called formant frequencies, and they are the natural frequencies of the vocal tract when it is closed at the glottis end. The velocity of the sound, v, at body temperature is 354 m/s, and the length of a typical adult male vocal tract is 0.17 m [6]. For

male speech the normal range of variation is F_1 = 180-800 Hz, F_2 = 600-2500 Hz, F_3 = 1200-3500 Hz, and F_4 = 2300-4000 Hz. The average distance between formants is 1000 Hz. Females have on the average 20% higher formant frequencies than males, but the relation between male and female formant frequencies is nonuniform and deviates from a simple scale factor. Formant bandwidths may vary considerably. Typical values are $B_n = 50\,(1 + F_n / 1000)$ Hz. Formant amplitudes vary systematically with the overall pattern of formant frequencies and the spectral properties of the voice source [7].

Consonants are produced with a relatively narrow constriction in some region of the airway above the larynx, whereas for vowels the airway is more open. Consonants can be classified along several dimensions depending on the articulator that is responsible for making the constriction in the vocal tract, the degree of constriction, the state of the glottis and the vocal folds when the constriction is formed, and whether or not pressure is built up behind the constriction [6].

Fricative consonants are produced by exciting the vocal tract with a steady airstream that becomes turbulent at some point of constriction. Some fricatives also have a simultaneous voicing component, in which case they have what is called mixed excitation. Those with simple unvoiced excitation are usually called unvoiced fricatives (e.g. "F", "S", "Š"), while those of mixed excitation are called voiced fricatives (e.g. "V", "Z", "Ž") [8]. Figure 2.2 d) shows the time wave and the spectrum of the unvoiced fricative consonant " Š".

Nasal consonants are voiced sounds produced by the glottal waveform exciting an open nasal cavity and closed oral cavity. The closed oral cavity is still acoustically coupled to the pharyngeal and nasal cavities, and will therefore affect the resulting spectral resonances by trapping energy at certain frequencies. This phenomenon gives rise to antiresonances in the overall vocal system. For nasals, formants occur approximately every 850 Hz instead of every 1000 Hz. The antiresonance produces a spectral zero in the frequency response that is inversely proportional to the length of the constricted oral cavity. Bandwidths of nasal formants are normally wider than those for vowels. This is due to the fact that the inner surface of the nasal cavity contains extensive surface area, resulting in higher energy losses due to conduction and viscous friction. Since the human auditory system is only partially able to perceptually resolve spectral nulls, discrimination of nasals based on the place of articulation is normally cued by formant transitions in adjacent sounds [8].

3 State of the Art

3.1 Source-Filter Speech Modelling

A rather useful model of speech production consists of a filter that is excited by either a quasiperiodic train of impulses (for voiced sounds) or a random noise source (for unvoiced sounds). The source-filter model realized by electrical circuits was first proposed by H. Dudley at Bell Laboratories in the 1930s. The output of the electrical excitation source was passed through a filter whose frequency response was adjustable. This variable filter was performed by a bank of bandpass filters covering the range of speech frequencies [1]. In 1960s, the formant synthesizers were proposed. Here, the resonant characteristics of the filter bank track the movements of the formants [4]. At present, two types of the source-filter model are useful for speech processing: the all-pole model known as the autoregressive (AR) model, and the pole-zero model known as the autoregressive moving average (ARMA) model [9]. Since the estimation of parameters for the AR model results in linear equations, it has a computational advantage over the ARMA techniques [10]. The AR model of a vocal tract is well known in speech processing as a linear predictive coding (LPC) model. For the LPC speech model, preemphasis should be performed prior to the analysis and postemphasis should be performed prior to the synthesis [11]. Preemphasis is a simple and effective way of accenting the higher formants, thus allowing more accurate formant tracking results [105]. The AR spectral estimate was originally developed for geophysical data processing, where it was termed the maximum entropy method (MEM) [12]-[18]. It has been used for many applications including LPC techniques in speech processing. For AR parameters determination in MEM [19], [20] the Burg algorithm is used. For speech modelling the autocorrelation and covariance methods of linear prediction analysis have been used more widely [9], [21]-[23]. However, as the autocorrelation method has a frequency domain interpretation [9], [24] a weighting window must be applied to speech data to reduce the spectral leakage associated with finite observation intervals [121]. The covariance method may yield unstable results; the Burg method and the autocorrelation method with Hamming window give comparable results, however, the Burg method has much higher computational complexity than the autocorrelation method [24], [25]. The Burg method belongs to a class of lattice methods guaranteeing stability of the LPC filter. In order to utilize this advantage several approaches were used to reduce computational costs of this class of methods [26]-

[28]. The computationally less expensive autocorrelation method does not yield unstable synthesis filters too, for it is looking only at the short term spectral behaviour, and a decreasing time sequence can show the same short term spectral behaviour as a growing sequence [24]. The AR model will be presented in Section 5.1.1. The autocorrelation method of the AR parameters determination will be described in Section 5.1.1.1. Methods of the AR model order selection together with the original results will be a subject of Section 5.1.1.2.

Another type of the source-filter speech model is the cepstral model using homomorphic signal processing [29]-[31] based on the idea of the log magnitude approximation filter [32], [33]. Padé approximation of the continued fraction expansion of the exponential function [30] is used to approximate a nonrealizable exponential function in homomorphic modelling. Other types of speech models use cepstral analysis on a perceptually warped frequency scale [34]-[36]. The cepstral model will be introduced in Section 5.1.2. The cepstral model parameters determination will be discussed in Section 5.1.2.1. The cepstral model order selection will be mentioned in Section 5.1.2.2.

Some original aspects of speech processing common to the source-filter model as well as the sinusoidal speech model (see Section 3.2) will be a subject of Chapter 4.

3.2 Sinusoidal Speech Modelling

3.2.1 Sinusoidal Model in Speech Coding and Synthesis

When compared with the source-filter model, a rather different approach represents a sinusoidal speech model. In its simplest form it needs neither an excitation that models the vocal cords activity, nor a filter that models the vocal tract system. It simply models the speech signal as a sum of sine waves with defined frequencies, amplitudes, and phases. The excitation and the transfer function of the vocal tract are inherently comprised in these sinusoidal model parameters. Perhaps, the first most detailed description of speech analysis/synthesis based on a sinusoidal model was presented in 1986 by R. J. McAulay, and T. F. Quatieri [37], [38], although some information about sinusoidal and harmonic coding and synthesis had been published a few years before also by other authors. In this model, first in every frame the amplitudes are computed from the local maxima of the magnitude spectrum, and the phases are determined from the phase spectrum at the corresponding frequencies. Then the frequencies of the local maxima in the consecutive frames are matched,

i.e. the frequencies of the current frame are connected with the nearest neighbour frequencies of the previous frame. For a given frequency track, the amplitudes are interpolated linearly, and a cubic function is used to unwrap and interpolate the phase such that the phase track is maximally smooth. However, this model cannot be used for speech synthesis or speech coding at low rates because of a very high number of sinusoidal parameters. The same authors later proposed a harmonic sine-wave model, i.e. a sinusoidal model with harmonically related sine waves [39], [40]. Section 5.2 introduces the harmonic model together with theoretical derivation of the number of composite harmonics and their amplitudes, the type and the length of the weighting window, and is normalization. Using a minimum-phase assumption not only for the vocal tract but also for the glottal pulse contribution in voiced speech, and using a random phase for unvoiced speech, there is only need for properly coding the sine-wave amplitudes. It can be done using an all-pole model or a cepstral model of the magnitude spectral envelope. The former is a subject of Section 5.2.1, the latter is a subject of Section 5.2.2. In [39], [40] for modelling of voiced fricatives and other speech sounds with mixed excitation the voicing probability is determined using pitch estimation based on a sinusoidal speech model. The sine-wave phases are made random above the voicing-adaptive cutoff frequency, which is determined by the voicing probability that is a measure of how well the harmonic set of sine waves fits the measured set of sine waves and is determined as a part of the pitch estimation process minimizing the mean squared error (MSE) [41]. A different approach is presented in Section 5.2.1.4, where the maximum voiced frequency is determined from the magnitude spectrum comparing the frequency distances between the pitch harmonics and the spectral local maxima. In [40] the authors propose the overlap-and-add (OLA) method with triangular, Hanning, or trapezoidal window instead of a computationally expensive matching algorithm with linear interpolation of amplitudes and cubic interpolation of phases. Sections 5.2.1.6 and 5.2.2.7 present a comparison of OLA with Hanning window and a concatenation of pitch-synchronous frames. The sinusoidal model [37]-[41] is suitable for prosodic modifications that are necessary in the text-to-speech (TTS) systems. A time-scale and pitch modification system that preserves shape-invariance property during voicing is done using a version of the sinusoidal analysis/synthesis system modelling and independently modifying the phase contributions of the vocal tract and the vocal chord excitation [42]. It improves the temporal structure of time-scale modified speech [38] determining pitch pulse locations, referred to as onset times. Illustration how this method can be applied to the TTS system was presented in [43]. Its further refinement [44] shows the treatment of the phase

information in the case of synthesis by concatenation, where it is necessary to assure phase continuity when concatenating two realizations of an allophone segmented from different words. In another modification to the shape-invariant sinusoidal speech model [45] the phases of the component sine waves used for excitation are made to add coherently at each glottal closure. It is useful especially for high degrees of modification, where the phase coherence is often lost using the method of [42]. To cope with this problem a variable order polynomial phase interpolation is used. Its order depends on the number of the excitation points in the synthesis frame. A new and simple approach to shape invariant time-scale modification operating entirely within the original sinusoidal model [37] was presented in [46]. The method is based upon a harmonic coding of each speech frame and makes no use of pitch-pulse onset times. Parameters of the cubic function interpolating phase of the time-scaled speech are chosen, not such that the smoothest possible frequency track is obtained but such that the shape of the track matches, as closely as possible, the shape of the original. No decoupling (into source and vocal tract models) of the speech production process is necessary. Instead of the excitation phase determination, phase coherence between the frames is kept by calculating the amount of time taken for the first harmonic to move from its measured phase to the phase adjusted by time-scaling while keeping its frequency constant. The target phase of each remaining harmonic is adjusted by the product of this amount of time and frequency. To keep track of previous phase adjustments when moving from frame to frame, this amount of time is accumulated before processing every new frame and it is used to adjust the target phase prior to the time-scaling of the frame. However, phase coherence was found to begin to break down for larger scaling factors (greater than 1.8). The method would then be of most use in concatenative speech synthesizers where scaling factors lie usually within the bounds handled by the algorithm. Although the time-scaling algorithm [46] needs no speech signal decomposition to a glottal and a vocal part, this decomposition must be incorporated in a pitch modification algorithm [47]. Here the time-scaling method [46] is extended to handle pitch modification. The speech is first inverse filtered using a simple algorithm to estimate the glottal wave. A pitch estimate is assigned to each frame of the estimated glottal wave. The fundamental frequencies of adjacent frames are matched and the original frequency track is computed. The integral of the new pitch-scaled frequency track is estimated by time-scaling the original frequency track by the pitch modification factor to give a new target phase value for the first harmonic. In a similar way as in time-scaling, the new target phase value for each of the remaining harmonics is adjusted by the accumulated amount of time descibed above.

Assuming that the glottal wave spectrum is relatively flat, the amplitude of each harmonic after pitch-scaling is left unchanged, i.e. the glottal wave spectrum is not resampled at the new harmonic frequencies but simply expanded/compressed to effect the desired pitch change. Although the glottal wave spectrum is altered, the shape of the time-domain glottal waveform is preserved and so voice quality should remain constant. Combining pitch and time-scale modification is straightforward. A single algorithm allows the independent modification of pitch and duration. For a given frame, the net scaling factor is given by a product of the time-scale and the pitch modification factors. The modified glottal wave subsequently serves as input to an LPC vocal tract filter. This algorithm also produces high quality results for scaling factors of the order required for concatenative speech synthesis avoiding the need for "pitch pulse onset time" estimation. In [123] the time-scale modification introduced in [46] was improved to process voiceless speech during time-scale expansion. "Noisy" sinusoids are split into two separate components each following a different frequency track modelled with a parabola. Using this approach, the tonal quality associated with time-scale expanded voiceless speech was eliminated even for large scaling factors.

In [48] the sinusoidal model was compared with a code-excited linear prediction (CELP) concluding with complementarity of the two methods. The authors state that "an ideal coder should be able to combine the noise-free quality of sinusoidal models with the robust analysis-by-synthesis procedures of CELP coders". The powerful features of the sinusoidal as well as the CELP coding algorithms are used in the hybrid speech coder proposed in [49]. The harmonic sinusoidal analysis is used to encode the periodic part of speech and to split the speech spectrum into two frequency regions of harmonic and random components. The unvoiced speech and the random part of voiced speech are coded using the CELP algorithm. The periodic part of speech waveform is obtained by applying an IFFT to the speech spectrum with the aperiodic part zeroed out. Subtracting the resulting periodic waveform from the original speech waveform, the aperiodic speech waveform is achieved. The spline envelope fitted to the sine-wave amplitudes in the spectral domain is modelled by the transfer function of an all-pole filter. A harmonic tracking algorithm is used for interpolating the sinusoidal parameters between their update points in adjacent frames to achieve high level of periodicity in voiced frames and remove the discontinuities across the frame boundaries.

The idea of speech modelling as a sum of sinusoids was addressed also in [50]. However, here the so called composite sinusoidal modelling (CSM) is considered as equivalent to line

spectrum frequencies (LSF) or line spectrum pairs (LSP) being only a useful representation of LPC parameters for speech coding purposes. CSM uses a rather complex mathematical procedure to compute amplitudes and frequencies from the sample autocorrelation function, while resetting the phases of sinusoids every pitch period.

An original approach to speech synthesis based on a sinusoidal representation was mentioned in [51], where numerical solutions of non-linear differential equations, which generate sinusoidal waves are used. For voiced sounds, these equations behave as a group of mutually synchronized oscillators; for voiceless sounds, they work as passive filters with input noise sources.

The harmonic model with an ARMA spectral envelope fitting accurately the harmonic short-time Fourier transform components was described in [52]. The narrowband components of speech are reproduced by sampling the pole-zero envelope at integer multiples of the fundamental. The residual is random-like and broadband, and its statistics are reproduced by exciting the pole-zero filter with white noise.

Similar to the harmonic speech model is a multiband excitation (MBE) model [53]. It uses an analysis-by-synthesis method, in which the excitation and vocal tract parameters are estimated simultaneously so that the synthesized spectrum is closest in the least squares sense to the spectrum of the original speech. Then, the voiced/unvoiced decisions are made based on the closeness of fit between the original and the synthetic spectrum at each harmonic of the estimated fundamental. Voiced speech is synthesized from the voiced envelope samples by summing the outputs of a band of sinusoidal oscillators running at the harmonics of the fundamental frequency. Unvoiced speech is synthesized from the unvoiced envelope samples by first synthesizing a white noise sequence. Its normalized Fourier transform is multiplied by the spectral envelope and then synthesized using the weighted OLA method. The final synthesized speech is generated by summing the voiced and unvoiced synthesized speech signals. A similar method called a hybrid harmonic/stochastic (H/S) model is mentioned in [54] and described in more detail in [55]. Possibilities offered by this model in the context of wide-band TTS synthesis based on segment concatenation are addressed here. When compared with MBE speech coding, the hybrid H/S model assigns frequency dependent voiced/unvoiced decisions associated to wide frequency bands rather than to individual harmonics. It computes samples by overlap-adding the IFFT of spectral frames, obtained by summing stochastic components (in the form of FFT bands with constant amplitudes and

random phases) and harmonic ones (in the form of the most significant samples of their Dirichlet kernels). A fast OLA/IFFT synthesis algorithm reduces computational load in comparison with the computationally expensive parameters interpolation approach. However, the resulting OLA synthesis frames are not pitch-synchronously related.

A modification of MBE called a variable-spectrum MBE model was proposed in [56]. Speech is described as a sum of periodic and noise components within a given time frame. The periodic part is a set of pitch harmonics with amplitude and frequency changing linearly within the time frame. The noise part is determined as a difference between the input signal and the synthesized periodic part. Its synthesis employs a white noise signal weighted by the noise spectrum envelope and transferred back to the time domain. The periodic and noise components are added and then combined using the OLA method.

Assumption that the speech signal is composed of a periodic and a stochastic part was also used in a harmonic plus noise model (HNM) [57]-[59]. When compared with the MBE [53] or hybrid H/S [55] models, the spectrum of the HNM is divided into only two bands. Although the notion of a maximum voiced frequency is introduced in [57], [58], its meaning is the same as the voicing transition frequency or the voicing-dependent cutoff frequency used in [39]-[41]. However, the upper band of the spectrum in the HNM is modelled using an all-pole filter driven by a white Gaussian noise instead of a sum of sine waves with random phases. The positions of the analysis instants are set at a pitch-synchronous rate (regardless of the exact position of glottal closure). Synthesis is also performed in a pitch-synchronous way using an OLA process. If the frame is voiced, the noise part is filtered by a high-pass filter with cutoff frequency equal to the maximum voiced frequency. Then, it is modulated by a triangular-like time-domain envelope synchronized with the pitch period. Thanks to the pitch-synchronous scheme of the HNM, time-scale and pitch-scale modifications are quite straightforward. In [60], [61] the HNM was used for voice conversion, i.e. modifying the speech signal of one speaker so that it sounds as if it had been pronounced by a different speaker. A methodology for representing the relationship between two sets of spectral envelopes is based on a Gaussian mixture model of the source speaker spectral envelopes. In [62], [63] a problem of removing phase mismatches at the frame boundaries is solved using the notion of centre of gravity. However, although using different mathematical foundation, its meaning is equivalent to the pitch onset time [39], [40], [42] determined by the same procedure of seeking the minimum of the MSE between the original speech frame and its sinusoidal resynthesis.

At AT&T Labs-Research, HNM was compared with a time-domain pitch-synchronous OLA (TD-PSOLA) method [64]. TD-PSOLA relies on the speech production model described by the sinusoidal framework, although the parameters of this model are not estimated explicitly. The analysis process consists of extracting short-time analysis signals by multiplying the speech waveform by a sequence of time-translated analysis windows. The analysis windows are located around glottal closure instants and their length is proportional to the local pitch period. During unvoiced frames the analysis time instants are set at a constant rate. During the synthesis process a mapping between the synthesis time instants and analysis time instants is determined according to the desired prosodic modifications. This process specifies which of the short-time analysis signals will be eliminated or duplicated in order to form the final synthetic signal. Results from the formal listening tests showed that HNM is a very good candidate for the next generation TTS. The score for HNM was consistently higher than for TD-PSOLA in intelligibility, naturalness, and pleasantness. The segment quality of synthetic speech was high, without smoothing problems and without "buzziness" observed with TD-PSOLA. An important point is that HNM is a pitch-synchronous system, which does not require glottal closure instants as in the case with TD-PSOLA. Although elimination/ duplication of short-time waveforms in TD-PSOLA is very simple and the computational load is very low, it introduces a tonal noise quality because of the repetition of segments, noticeable more during unvoiced frames and fricative voiced frames. Because of the non-parametric scheme of TD-PSOLA, limited smoothing possibilities are offered. Comparing TD-PSOLA and HNM regarding computational cost, it is clear that HNM has a much higher complexity than TD-PSOLA. Actually, this is the only drawback of HNM versus TD-PSOLA. In [124] the application of the HNM for concatenative TTS synthesis is described resuming the results of [57], [62], and [64].

A model similar to HNM is called a hybrid model [65] used for concatenation-based TTS synthesis. A pitch-synchronous analysis uses pitch-period detection and chaining described in [66]. However, although the pitch detection algorithm is described in detail here, it is rather time-consuming, sometimes giving bad location of pitch marks [67]. In the hybrid model [65] the Fourier transform-based analysis is performed over the overlapped segments with a double pitch-period periodicity. For 16-kHz sampling, the normalized energy of the even harmonics (i.e. harmonics of the pitch) in the frequency bands 0-2, 2-3, 3-4, 4-5 kHz is calculated. According to the results, the maximum voiced frequency is equal to one of the four values: 2,

3, 4, or 5 kHz. Then the harmonic part up to the maximum frequency is synthesized and subtracted from the original signal giving the noise part synthesized by a random excitation applied to an all-pole filter. The hybrid algorithm was tested on duration and pitch modifications of recorded sentences, and on the TTS synthesis system outperforming other speech models.

Also the authors of [68] discussed the adequacy of the sinusoidal model to the requirements of concatenative speech synthesis in TTS. They carried out a preference test between speech synthesized using a pitch synchronous LPC synthesizer and the sinusoidal synthesizer. The sinusoidal model used, is the harmonic speech model with cepstral parametrization of the spectral envelope according to primary description in [37]-[39], [41], [42]. In informal listening tests the sinusoidal synthetic speech was clearly preferred, especially in the case of male speech and synthetic prosody. For female speech a sampling frequency of 8 kHz might probably be too low to adequately reflect the characteristics of speech. The computational load of the sinusoidal synthesizer is about 10 times the computational load of the pitch synchronous LPC synthesizer due to computationally expensive frequency matching and interpolation algorithms of [37].

A new method for harmonics extraction in sinusoidal representation of a speech signal was introduced in [69]. The speech signal is decomposed into the harmonic components using a set of band-pass filters, and the harmonic frequencies are obtained as the instantaneous frequencies of these components.

A modification of the sinusoidal-based vocoder, called band-widened harmonic (BWH) vocoder [70], removes the “buzzy” quality due to strong periodicity. Here, controllable low-pass filtered random signals are combined into the harmonic amplitudes, which are linearly interpolated between frames. In this way the bandwidths of the harmonics could be properly widened and the “buzzy” quality is efficiently reduced, although it can still be heard mainly because of the linear interpolation between the previous and present spectrum.

In [71]-[73], the use of an LPC analysis along with gain adjustment, delay compensation, and all-pass phase correction was proposed for simultaneous representation of sinusoidal amplitude and phase parameters. The proposed algorithm results in improved phase matching for all categories of speech (voiced, unvoiced, onset, and transition) and yields improved reproduction of nasal and vowel sounds.

A method of fitting a spectral envelope parametrized by cepstral coefficients to a discrete set of spectral points using a log-spectral distortion criterion comprising a roughness penalty functional and a regularization (smoothing) parameter was introduced in [74]. This approach was extended in [75] by defining a statistically significant performance criterion, a penalized likelihood criterion, for measuring the envelope fit. The method is suitable for parametric modelling of the sinusoidal model spectral envelope in presence of measurement noise both for all-pole and cepstral parametrization.

Another approach to fitting an all-pole filter to harmonic spectrum is based on minimum variance distortionless response (MVDR) [76] determining an envelope fitting the spectral harmonic peaks rather than resolving the harmonics. The inverse MVDR filter is designed to minimize its output power subject to the constraint that its frequency response at the frequency of interest has unity gain. This distortionless constraint ensures that the MVDR filter will let the input signal components with the frequency of interest pass through undistorted, and the minimization of the output power ensures that the remaining frequency components in the signal are supressed in an optimal manner. The MVDR all-pole filter is stable and causal, and can be used in a manner similar to the way in which LPC filters are used in speech processing systems, e.g. source-filter speech models or sinusoidal speech models. Another approach to spectral envelope determination of the harmonic model with all-pole parametrization was implemented in the harmonic-stochastic excitation (HSX) vocoder [77]. It solves the similar problem as the MVDR that the minimization of the MSE tries to fit the synthesis filter frequency response to speech power spectrum and not to its envelope. Here, the speech spectrum is computed using pitch-synchronous Fourier transform with the length dependent not only on pitch period but also on voicing information, i.e. for strongly periodic signal it is longer to ensure better spectral resolution, for transition signal it is smaller. Harmonic frequencies are determined from amplitude spectrum by a peak-picking algorithm and a simple linear interpolation is used between frequencies of adjacent harmonics in the log scale. The interpolated log envelope representation is then brought to the linear spectral scale and the autocorrelation coefficients are found through the inverse Fourier transform of the envelope power spectrum. Filter coefficients are finally computed using classical LPC method. A somewhat different approach was used in [78]. Here, the pitch determination is performed in equidistant overlapping frames using the clipped autocorrelation function [79]. In [78] the staircase log spectral envelope is formed by the

maximum values within the intervals of the pitch-frequency length around the pitch harmonics. This staircase envelope is smoothed using the weighted moving average having the shape of the normalized Blackman window. Then, the inverse Fourier transform of the smoothed spectral envelope is treated as a real speech signal. Thus, the AR parameters are computed by the autocorrelation method from the time-domain signal corresponding to the speech envelope of the original speech signal. The method was first applied to childish voice resynthesis in [78] (see Section 5.2.1.7) as for high-pitch speakers (also for low-pitch speaker and higher AR order) enhancement of the spectral envelope using the proposed method is more evident. In Section 5.2.1.4 the proposed method is described in more detail and its application to male voice is presented.

The authors of [81] indicate the inadequacy of the minimum-phase assumption for modelling voiced speech because glottal pulses tend to have rather slow rising edges which are teminated by much sharper trailing edges. They introduce a method of determining both magnitude and phase information for a noncausal all-pole spectral envelope parametrization. In this model the vocal system represents the composite characteristics of the glottal pulse, vocal tract, and lip-radiation filters and maximum-phase nature of differentiated glottal pulses is matched well. The poles of the noncausal model are not constrained to be within the unit circle and a linear phase offset is used as additional parameter. Parameter estimation is performed using quasilinear and nonlinear least squares techniques. Noncausal models can be used in sinusoidal-model-based approaches without the difficulties of noncausal infinite impulse responses that occur in time-domain approaches. Although instability may occur when using the noncausal all-pole model for filtering the excitation signal, there is no such a problem when the noncausal model is used to encode both the magnitudes and phases of the measured harmonics. The speech is then reconstructed by evaluating the model at the harmonic frequencies and using the resulting magnitudes and phases in the bank of harmonic oscillators. In the sinusoidal transform coder [82], [83] noncausal all-pole modelling of the vocal tract is used to enhance the accuracy of phase representation. In addition, the Bark spectrum is used for perceptual coding of the sine-wave amplitudes because of its ability to achieve a uniform perceptual fit across the spectrum (Sounds separated by more than one Bark are resolvable as separate sounds). The MSE between the original waveform and its model fit for the noncausal all-pole model outperforms its minimum-phase counterpart with either all-pass compensation included or not.

A speech analysis/synthesis system described in [84] is based on the combination of an OLA sinusoidal model with an analysis-by-synthesis technique (ABS/OLA) to determine model parameters. Analysis as well as synthesis is performed by a constant frame rate, i.e. it is not pitch synchronous. Analysis-by-synthesis determines amplitude, frequency, and phase parameters minimizing the mean-square modelling error. The approximation of the original speech signal by adding a single component is updated recursively. In fact, it is approximation of the residual error left after approximating the speech segment by previous components. Since analysis-by-synthesis removes each component after determining its parameters, sidelobe effects, which have been observed to produce slight tonality in synthetic voiced speech using peak-picking analysis, are reduced. Time-scale, frequency-scale, and pitch-scale modifications preserving shape invariance, but more specifically suited to the OLA model, are described too. Pitch-scale modification is achieved by interpolating the complex phasor form of excitation amplitude/phase pairs uniformly at harmonic frequencies to produce a continuous excitation spectrum. Given a pitch-scale factor, the excitation spectrum is then resampled at modified harmonic frequencies. When compared with [37] where a direct summation of sinusoids is performed, the OLA model used in this system reduces synthesis computation by using the inverse FFT. An extension to the ABS/OLA sinusoidal speech modelling and modification algorithm [84] for unvoiced speech was presented in [85] using a perceptually motivated modulation of the sinusoidal component phases. It simply represents phase randomization implemented within the context of an OLA model by subdividing each synthesis frame and randomizing the phase offsets between components prior to synthesis of each subframe. The number of subframes is usually made proportional to the time-scale expansion factor. This refined model eliminates the problem of unnatural tonal artifacts that often arise in modification of unvoiced speech. The implementation of the ABS/OLA model within a TTS system is described in [86]. Apart from the phase randomization used in time-stretched speech [85], the pitch modulation compensation improving pitch-lowered speech is described here. However, for general use in the TTS system, the benefit of these extensions is counterbalanced with the problems of mis-estimation of other parameters. In [87]-[90] the ABS/OLA [84], [85] was used with pitch-synchronous analysis/synthesis [57], [58], [62], [63] and a regularized discrete cepstrum estimation of the spectral envelope [74], [75].

Enhancement of LPC spectrum for harmonic speech coding was presented in [91]. Since in sinusoidal and harmonic coders the amplitudes should be represented accurately at a set of

discrete frequencies, the goal is to fit a smooth curve through the desired spectral peaks. Then the LPC model is fitted to the smooth curve represented usually by a cubic spline. However, this biases the model fit in favour of the specific frequencies of interest, In sequel, it cannot be used for speech modification purposes necessary in TTS, where the spectral envelope must be preserved. Further improvement of this method uses a perceptual enhancement technique based on mel-warping the spectrum, i.e. mapping it linear at lower frequencies and logarithmic at higher frequencies [92].

Another improvement of the phase spectrum model considering the non-minimum-phase glottal pulse contribution to the vocal system phase was introduced in [93], [94]. It uses a phase only pitch-cycle waveform (PCW), i.e. a sum of harmonic sinusoids with unit amplitudes with the length of the integer part of the pitch period. The minimum-phase PCW is computed from the magnitude spectrum of the vocal tract system using the discrete cepstral coefficients. PCW corresponding to the measured harmonic phases is determined too. Temporal waveform alignment in the PCW domain is performed by finding the shift that maximizes the circular cross-correlation of both PCWs. The difference between the shifted measured PCW and minimum-phase PCW constitutes the residual waveform, which can be quantized with relatively small amount of bits. Harmonic phases over 1.5 kHz are directly obtained from the minimum-phase spectrum. Informal listening tests showed that naturalness of low pitched speakers can be significantly improved using this algorithm in a low bit rate multiband excitation coder.

A modified harmonic model able to produce also non-periodic pulse sequences in conjunction with a closed-loop analysis-by-synthesis scheme was described in [95]. It is effective for representing speech in transition regions such as voicing onsets, plosives, and non-periodic pulses. The excitation signal synthesis uses the classical harmonic model of the LPC residual for voiced and unvoiced speech. For the transitional speech, apart from amplitudes and phases representing the shape, pulse occurrence times and gains are used to preserve time domain information better. The reconstructed excitation signal is passed through the inverse short-term filter to obtain the reconstructed signal. A similar approach to transition speech synthesis combines a frequency-domain harmonic coding for periodic and “noise like” unvoiced segments with a time-domain waveform coder for transition signals [96]. It uses a multi-pulse excitation and a closed-loop analysis-by-synthesis search algorithm for time-domain waveform coding of transition segments, and a three-layer fully connected feed-forward

neural network classifier, trained by a large training set, to obtain one of the three classes – voiced speech, unvoiced speech, and transition speech. When switching from a transition frame to a harmonic frame, signal continuity is preserved using a phase obtained by maximizing the correlation of the shifted reconstructed harmonic excitation frame with the reconstructed transition excitation frame. To synchronize the reconstructed transition frame with the preceding harmonic one, the drift between the original signal and the reconstructed one is measured by the encoder, and the transition frame is extracted with the corresponding shift. No phase synchronization is required when switching to and from stationary unvoiced segments. Another similar enhanced frequency domain transition model is used in an analysis-by-synthesis multimode harmonic coder (AbS-MHC) [97]-[99] as an extension of [95]. The algorithm for the time domain closed-loop pitch estimation and classification has three stages. The first stage pre-classifies the input speech into one of two categories: the first category includes unvoiced speech and silence; while the second includes voiced speech and transition speech. The second stage is performed only on the voiced speech and the transition speech to perform the voiced/transition speech classification and determine the pitch. At the last stage, a pitch refinement and harmonic bandwidth estimation procedure is performed on subframes, which are declared as voiced. This procedure is similar to that described in [96]. Another enhancement of the analysis-by-synthesis sinusoidal model for low bit-rate speech coding was introduced in [100]. Classification of voiced/unvoiced and transition frames is similar as it was in [95]-[99]. However, the ABS/OLA [84] approach is used to model the LPC residual. For voiced frames, harmonically related frequencies represent the periodic part; non-harmonically related frequencies represent the aperiodic part above 1 kHz to account for the non-uniform frequency resolution of the human auditory system, the frequency resolution of which decreases with increasing frequency. The voiced residual is modelled as a sum of the harmonic and the non-harmonic models. In harmonic analysis, the frequency space consists of a set of non-overlapping frequency intervals, which are centred at the pitch harmonics. The frequency points generated during analysis do not have to be exact multiples of the pitch frequency, enabling harmonic analysis to capture periodic components even in frames that have changing pitch period. Non-harmonic analysis works on the error residual generated by the harmonic analysis. For the synthesis of the periodic part, a cubic phase model is used. The aperiodic part is synthesized by applying uniformly distributed random phases. Unvoiced analysis is the same as harmonic analysis with frequency intervals located at multiples of 100 Hz, with the exception that a magnitude envelope is used as in the case of transition

frames. The synthesizer applies uniformly distributed random phase to each frequency component.

Another modification of a sinusoidal model suitable in transitional speech segments such as speech onsets and voiced/unvoiced transitions was proposed in [101]. It is a generalized sinusoidal model called the exponential sinusoidal model, in which the amplitudes of the sinusoidal components are allowed to vary exponentially with time. The damping factor of each exponentially varying amplitude may be positive, negative, or zero. The models were evaluated in 25-ms frames with an overlap of 5 ms between consecutive frames. Modelled signals were generated by overlap-adding the modelled frames. Results have shown a considerable objective and subjective improvement, especially in transitional segments, compared to the basic sinusoidal model with a peak-picking procedure [37]. However, a drawback of the exponential sinusoidal model is the computationally expensive parameter estimation scheme.

Papers [102]-[104] use the trigonometric identity to represent the sinusoidal model as a sum of sine and cosine waves without explicit phases. In [102] syllable long speech segments (100 to 200 ms) are modelled with a single set of parameters. An instantaneous frequency is obtained by fitting the estimated pitch contour by a third order polynomial. The model is simplified by the fact that the higher instantaneous frequencies are assumed to be approximately harmonically related. Amplitude functions corresponding to the sine and cosine terms consist of connected polynomial pieces forming box-splines (B-splines) of degree three. Experiments gave better synthetic signal when compared with ABS/OLA [84] method. The model described in [103], [104] is called instantaneous amplitude (IA) model and it represents each component with two parametrized instantaneous amplitudes and one constant “centre” frequency. In this model the composite waves are not harmonically related, and the short analyzed segments are assumed to be stationary, so the instantaneous amplitudes are slowly time varying functions approximated by a Taylor series of finite terms. However, in practice they use a first order polynomial for all the amplitudes. Individual frequencies are determined by recursively finding the most significant frequency of the periodogram and computing amplitude parameters that minimize the time signal with this most significant component removed. The authors say that when compared with the classical model of [37], with the same complexity, the IA-model provides better modelling quality.

An improved harmonic coder for the wideband was presented in [106]. Speech frames are classified into fully unvoiced and mixed frames, which can contain a harmonic structure and a noise structure within the same spectral subband. Both unvoiced frames and voiced part of mixed frames are modelled as a sum of harmonically related sine waves. Then, for mixed frames, in all subbands, the ratio between the energy of the harmonic spectrum and the original one is evaluated. Considering spectral auditory masking effect, the comparison of this ratio to a threshold then discriminates subbands into "audible" ones and "inaudible" ones. In audible subbands, the error between the original and the harmonic spectra is modelled. The synthesized signal for a mixed frame is obtained by summing a voiced signal corresponding to the modelling of the original speech spectrum with a harmonic model, and of a noisy signal corresponding to the modelling of this error spectrum.

Comparison of three different AR orders for coding the parameters of the harmonic model are the subject of Section 5.2.1.5. Comparison of the AR and the cepstral coding of the harmonic model parameters using the same number of the AR or cepstral parameters is presented in Section 5.2.3. The same number of cepstral parameters is used also for comparison of the source-filter cepstral model and the harmonic model with cepstral parametrization in Section 5.2.4. It regards comparison with the model described in [30], [31] presented in [107], [108].

3.2.2 Sinusoidal Model in Other Speech Processing Applications

In [80] the use of the sinusoidal model for noise reduction is achieved by applying the hidden Markov model-based minimum MSE estimator to find the harmonic sinusoidal parameters of clean speech from speech corrupted by additive noise. The system needs two sets of training data: speech recorded under quiet conditions, and noise from the expected operating environment. It finds its use especially in speech compression and speech recognition in noisy mobile environments. This area will not be a subject of this thesis.

3.3 Research Goals of the Thesis

On the basis of the state-of-the-art evaluation, the main objectives of this thesis were determined as follows:

1. Use the harmonic model with AR and cepstral parametrization for speech analysis and synthesis.
2. Find new approaches to synthetic speech quality improvement.
3. Compare the proposed methods with respect to the spectral measure, the perceived speech quality, and the computational complexity.

4 Suggestions for All the Speech Models

Quality of speech synthesis depends highly on the choice and precision of speech signal parameters determined during analysis.

4.1 Pitch Precision in Time and Frequency Domain

A very important speech signal parameter independent on the type of the speech model is the fundamental frequency (or pitch), introduced in Chapter 2, and voicing mentioned in some of publications cited in Section 3.2.1. The pitch can be determined in time domain (in seconds or milliseconds, corresponding to samples), or in frequency domain (in Hertz, corresponding to points of FFT). Let us find relationship between the pitch period determination error in samples and the pitch frequency determination error in points of FFT. Let $\hat{F}_0$ be the estimate of the real pitch frequency F_0 in Hz. For sampling frequency f_s in Hz and N_F-point FFT the pitch period error e_s in samples is

$$e_s = \left(\frac{1}{\hat{F}_0} - \frac{1}{F_0} \right) \cdot f_s , \tag{4.1}$$

and the pitch frequency error e_p in points of FFT is

$$e_p = \left(\hat{F}_0 - F_0 \right) \cdot \frac{N_F}{f_s} . \tag{4.2}$$

Although number of samples and number of points of FFT are integer values, let us suppose them to be real values for derivation of more precise relation between e_s and e_p.

From (4.1) the value of $\hat{F}_0$ is

$$\hat{F}_0 = \frac{F_0 \cdot f_s}{e_s \cdot F_0 + f_s} . \tag{4.3}$$

Substituting (4.3) into (4.2) we obtain

$$e_p = \frac{- e_s \cdot F_0^2 \cdot N_F}{\left(e_s \cdot F_0 + f_s \right) \cdot f_s} . \tag{4.4}$$

In the similar way the value of $\hat{F}_0$ from (4.2) is

$$\hat{F}_0 = \frac{N_F \cdot F_0 + e_p \cdot f_s}{N_F}. \tag{4.5}$$

Using (4.5) in (4.1) the value of e_s is

$$e_s = \frac{-e_p \cdot f_s^2}{F_0 \cdot \left(N_F \cdot F_0 + e_p \cdot f_s\right)}. \tag{4.6}$$

Relationship of e_p as a function of e_s according to (4.4) for three different pitch frequencies F_0 = 100 Hz, F_0 = 200 Hz, F_0 = 320 Hz is depicted in Figure 4.1 for f_s = 8 kHz, N_F = 512, and in Figure 4.2 for f_s = 16 kHz, N_F = 1024. Using (4.6), intervals of the pitch period error e_s (in samples) corresponding to the maximum pitch frequency error e_p of ± 1 point of N_F-point FFT for the sampling frequency f_s and the pitch frequency F_0 are shown in Table 4.1 for N_F = 512, N_F = 1024, and N_F = 2048. Here we can see that for low-pitch male voice the pitch frequency error e_p of ± 1 point represents the pitch period error e_s greater than ± 1 sample (always several samples, sometimes even tens of samples). It means that if a frequency-domain pitch detector is used, its error of 1 point of FFT in the frequency domain corresponds to rather a great error in the time domain. Even increasing N_F to 2048 does not give much better precision of the frequency-domain pitch detector for male voice. For that reason the time-domain pitch detector is used in this work, where speech analysis is performed for f_s = 8 kHz, N_F = 512, and f_s = 16 kHz, N_F = 1024. The algorithm based on a clipped autocorrelation function [79] with binary voicing decision is utilized for all the experiments included in this thesis.

N_F	f_s [kHz]	e_p	F_0 [Hz]		
			100	200	320
512	8	$-1 \le e_p \le 1$	$-10.81 \le e_s \le 14.81$	$-2.9 \le e_s \le 3.39$	$-1.16 \le e_s \le 1.28$
	16	$-1 \le e_p \le 1$	$-38.1 \le e_s \le 72.73$	$-10.81 \le e_s \le 14.81$	$-4.45 \le e_s \le 5.41$
1024	8	$-1 \le e_p \le 1$	$-5.8 \le e_s \le 6.78$	$-1.5 \le e_s \le 1.63$	$-0.6 \le e_s \le 0.63$
	16	$-1 \le e_p \le 1$	$-21.62 \le e_s \le 29.63$	$-5.8 \le e_s \le 6.78$	$-2.33 \le e_s \le 2.57$
2048	8	$-1 \le e_p \le 1$	$-3.01 \le e_s \le 3.25$	$-0.77 \le e_s \le 0.8$	$-0.3 \le e_s \le 0.31$
	16	$-1 \le e_p \le 1$	$-11.59 \le e_s \le 13.56$	$-3.01 \le e_s \le 3.25$	$-1.19 \le e_s \le 1.25$

Table 4.1 Intervals of the pitch period error e_s (samples) corresponding to the maximum pitch frequency error e_p of ± 1 point of N_F-point FFT for the sampling frequency f_s and the pitch F_0.

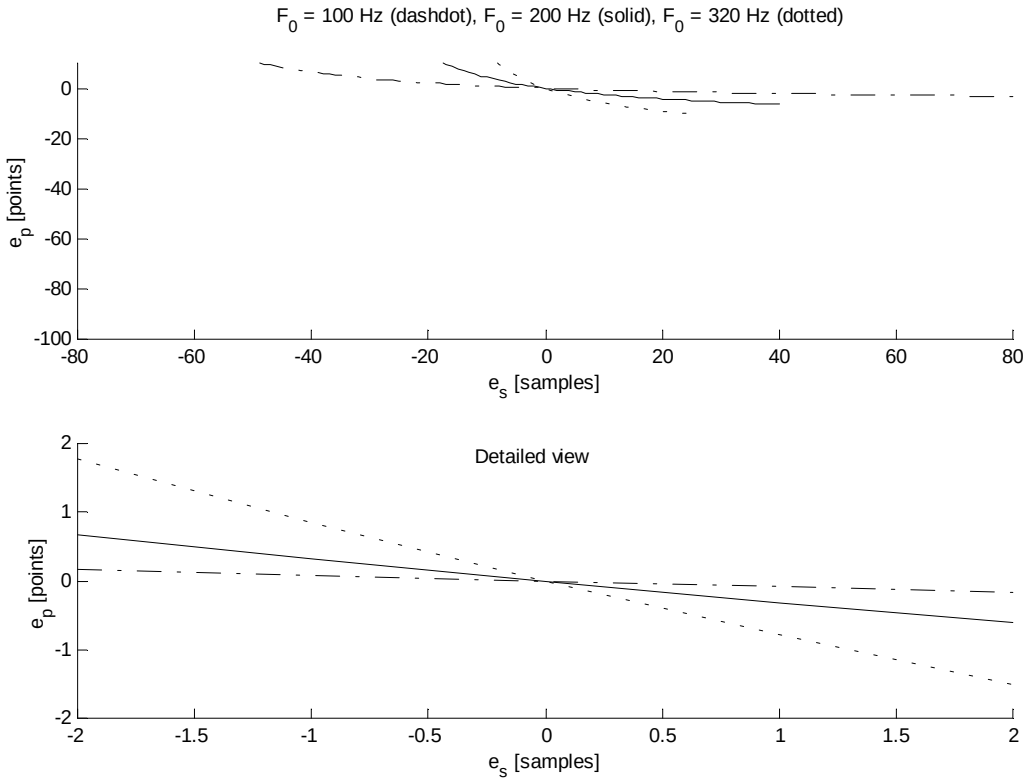

Figure 4.1 The pitch frequency error e_p as a function of the pitch period error e_s for f_s = 8 kHz, N_F = 512, and three different pitch frequencies F_0.

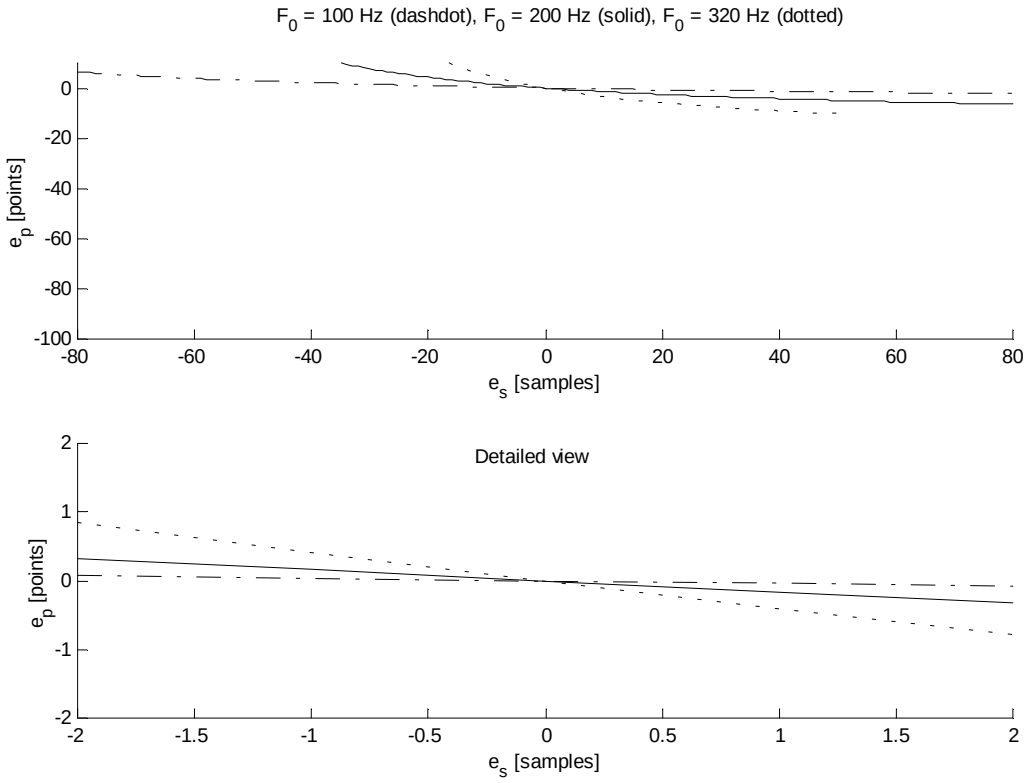

Figure 4.2 The pitch frequency error e_p as a function of the pitch period error e_s for f_s = 16 kHz, N_F = 1024, and three different pitch frequencies F_0.

4.2 Pitch Synchronization

If analysis is performed in equidistant speech frames, pitch synchronization must be used for good-quality synthesis. Pitch synchronization used in this thesis is inspired by the idea published in [31], however, the proposed algorithm is much more simple and it gives much less of pitch discontinuities at the frame boundaries with rapidly changing fundamental frequency. The speech signal is analyzed in the frame intervals of L_S samples with the frame length of $N = 2\,L_S$ and synthesized in equidistant frames of L_S samples (see Figure 4.3). The proposed method of pitch synchronization is illustrated in Figure 4.4.

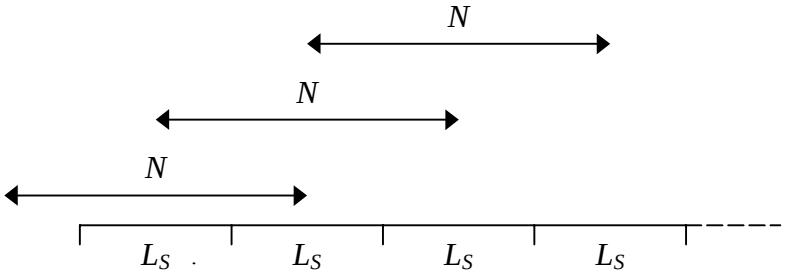

Figure 4.3 Positions of analysis frames (N samples) and equidistant synthesis frames (L_S samples).

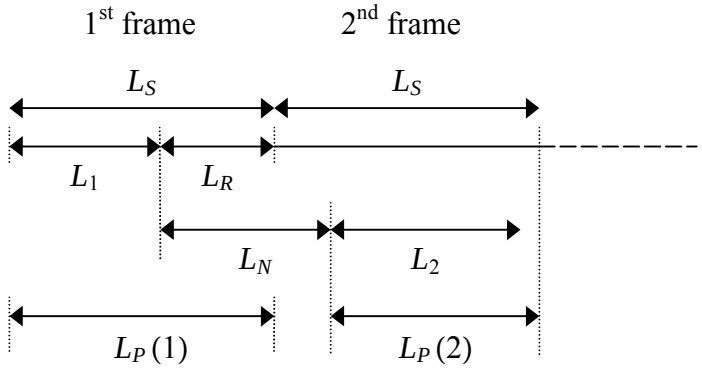

Figure 4.4 Illustration of pitch synchronization of equidistant analysis frames.

New synthesis frames of L_P samples with pitch-synchronous beginning are determined. In the first frame its length is initialized to be $L_P = L_S$, and it begins with the beginning of the first frame. Let L_k determines the pitch period in samples measured in the k-th frame. For speech frames classified by the pitch detector as unvoiced, the pitch period L_k is set to an implicit value, e.g. the mean value of the pitch period, or the pitch period of a previous voiced frame. In the first frame the number of pitch periods of the length L_1 will be $[\,L_P / L_1\,]$, where

[L_P / L_1] denotes the integral part of the number L_P / L_1. The remainder after division L_P / L_1 will be $L_R = L_P \,\text{modulo}\, L_1$, and it becomes a part of the next pitch period with the length of L_N samples. Its length should be between the values L_1 and L_2, so that following is satisfied:

$$L_R \cong L_1 \Rightarrow L_N \cong L_1 , \tag{4.7}$$

$$L_R \cong 0 \Rightarrow L_N \cong L_2 . \tag{4.8}$$

L_N consists of L_R comprised in the 1st frame containing the pitch period L_1, and ($L_N - L_R$) comprised in the 2nd frame containing the pitch period L_2. Proportionality can be ensured if L_N satisfies

$$\frac{L_R}{L_1} + \frac{L_N - L_R}{L_2} = 1 . \tag{4.9}$$

Expressing L_N from (4.9) gives

$$L_N = \left(1 - \frac{L_2}{L_1}\right) \cdot L_R + L_2 . \tag{4.10}$$

This relation is also valid for (4.7) and (4.8). Figure 4.5 shows L_N as a function of L_R for actual integer values of L_1, L_2, L_R. The next synthesis frame with a pitch-synchronous beginning has the length of $L_P = L_R + L_S - L_N$.

Next frames are treated in the same way as the first frame, however with the actualized value of L_P. For the k-th frame it means:

1st step: [L_P / L_k] frames of the length L_k,

2nd step: $L_R = L_P \,\text{modulo}\, L_k$,

3rd step: a frame of the length $L_N = \text{round}\left(\left(1 - \frac{L_{k+1}}{L_k}\right) \cdot L_R + L_{k+1}\right)$,

4th step: $L_P = L_R + L_S - L_N$.

Using the determined value of L_N, the vector of other speech parameters (e.g. autoregressive or cepstral) $\mathbf{T}_N$ in the frame of the length L_N, spanning from the k-th frame to the (k+1)-th frame, is given by

$$\mathbf{T}_N = \mathbf{T}_k + \frac{\mathbf{T}_{k+1} - \mathbf{T}_k}{L_{k+1} - L_k} \cdot (L_N - L_k), \tag{4.11}$$

where $\mathbf{T}_k$ and $\mathbf{T}_{k+1}$ represent vectors of these parameters in the k-th and (k+1)-th frames (see Figure 4.6).

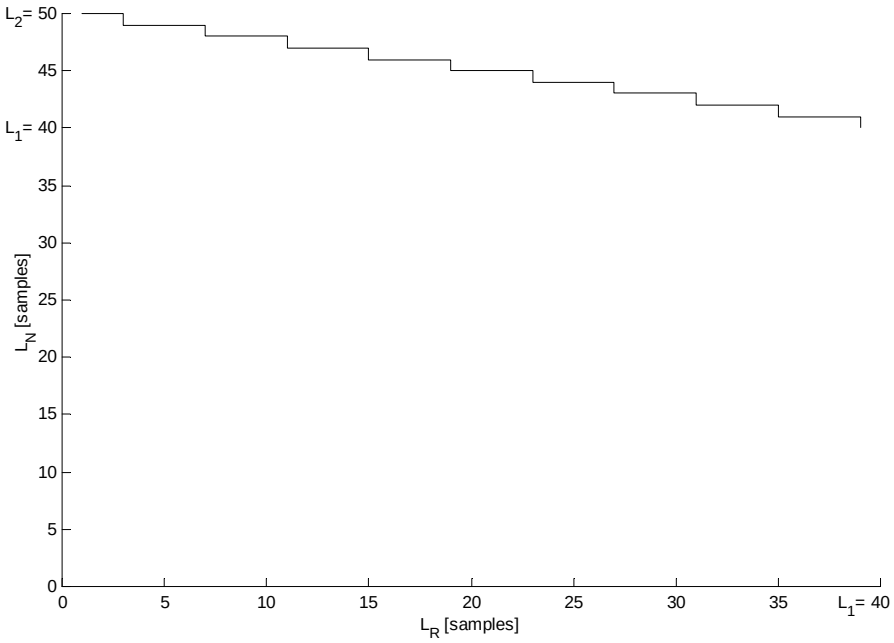

Figure 4.5 A pitch period L_N as a function of the remainder L_R for $L_1 = 40$, $L_2 = 50$.

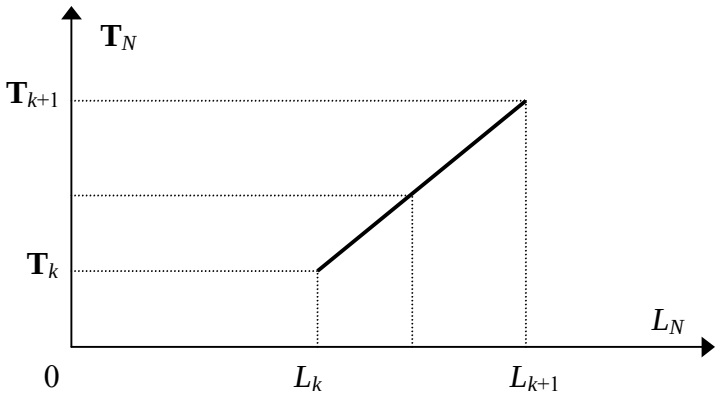

Figure 4.6 Speech parameters $\mathbf{T}_N$ as a function of the pitch period L_N.

5 Evaluation of the Speech Models

5.1 Source-Filter Model

Apart from the pitch and voicing comprised in the excitation, the source-filter model is represented by the parameters describing the transfer function of the vocal tract model. The principle is shown in Figure 5.1. Here $P(e^{j\omega})$ is the frequency response of the filter represented by the transfer function $P(z)$.

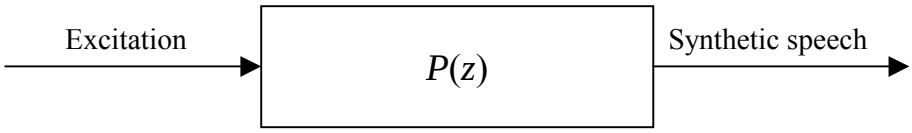

Figure 5.1 Principle of the source-filter speech model.

5.1.1 AR Model

The AR model mentioned in Section 3.1 has the frequency response given by

$$P\left(e^{j\omega}\right)=\frac{G}{1+\sum_{k=1}^{N_A} a_k \exp\left(-jk\omega\right)}, \tag{5.1}$$

where N_A is the order of the AR model, the gain G and the coefficients $\{a_k\}$ are the AR parameters or the LPC parameters. It is an all-pole model of a vocal tract given by the IIR filter

$$P(z)=\frac{G}{1+\sum_{k=1}^{N_A} a_k z^{-k}}. \tag{5.2}$$

5.1.1.1 AR Model Parameters Determination

The autocorrelation method addressed in Section 3.1 uses the Levinson-Durbin recursion applied to first N_A autocorrelation function values to compute the parameters $\{a_k\}$ and G describing the frequency response $P(e^{j\omega})$.

First, the autocorrelation function is computed using the formula

$$r(m) = \sum_{n=0}^{N-1-m} s(n) \cdot s(n+m), \tag{5.3}$$

where values $s(n)$ represent samples of the analyzed speech signal for $0 \le n \le N-1$.

The Levinson-Durbin algorithm recursively computes the parameter sets $\{a_{11}, G_1^2\}$, $\{a_{21}, a_{22}, G_2^2\}$, ... , $\{a_{N_A 1}, a_{N_A 2}, ..., a_{N_A N_A}, G_{N_A}^2\}$. An additional subscript has been added to each of the AR parameters to represent the order of the model. The final set of the order N_A is the desired solution. The recursive algorithm is initialized by

$$a_{11} = -\frac{r(1)}{r(0)}, \qquad G_1^2 = \left(1 - a_{11}^{\ 2}\right) \cdot r(0), \tag{5.4}$$

and the recursion for $k = 2, 3, ... , N_A$ is given by

$$a_{kk} = -\frac{r(k) + \sum_{i=1}^{k-1} a_{k-1,i} \cdot r(k-i)}{G_{k-1}^2}, \tag{5.5}$$

$$a_{ki} = a_{k-1,i} + a_{kk} \cdot a_{k-1,k-i} \qquad \text{for } i = 1, ... , k-1, \tag{5.6}$$

$$G_k^2 = \left(1 - a_{kk}^{\ 2}\right) \cdot G_{k-1}^2 . \tag{5.7}$$

The Levinson-Durbin algorithm provides the AR parameters for all the lower-order AR models fitting the data.

5.1.1.2 AR Model Order Selection

A rather crucial point in AR modelling of signals is the determination of the order of the AR process. In order to determine the most suitable model order N_A, one approach would be to try different values of N_A experimentally, and then choose the particular value which seems to be optimal in the sense that it satisfies some predetermined requirements for the particular situation. In general, it has been observed that, for a given value of record length N, small values of N_A yield spectral estimates with insufficient resolution, whereas for large values of N_A the estimates are statistically unstable with spurious details. Thus, it is expected that the value of the filter order N_A should be in the vicinity of some percentage of the record length N. Since the choice also depends on the statistical properties of the time series under analysis, it turns out that for the majority of practical measurements where the data can be considered

short-term stationary, the optimum value of N_A lies in the range from 0.05 to 0.2 N [109]. On the other hand an empirical rule for harmonic processes with noise constrains the order N_A to the range from $N/3-1$ to $N/2-1$ [110]. For speech signals a reasonable guess of the order can be made if one has a priori knowledge regarding the number of basic resonances one can expect in the data at hand. The resulting order estimate should serve as the lower bound; anything less than that will result in a poor model [111]. Since one formant occurs approximately every 1 kHz and one pole pair is necessary to model each formant, the model order is typically selected to be around twice the bandwidth of the signal in kilohertz [81].

Several information theoretic criteria are available for AR model order selection. The main idea of these criteria is that there should be a trade-off between a model fit and a model complexity. Thus all the criteria have one term measuring the model fit, the data term comprising estimated variance of the driving process represented by the gain G, and one term penalizing its complexity, the penalty term comprising the model order N_A.

The final prediction-error (FPE) criterion is defined as an estimate of the mean-square error in prediction expected when a predictive filter, calculated from one observation of the process, is applied to another independent observation of the same process. For a filter of the order N_A, the FPE is defined by

$$FPE(N_A) = \frac{N + N_A + 1}{N - N_A - 1} \cdot G_{N_A}^2, \tag{5.8}$$

where $G_{N_A}^2$ is the output error energy of the filter. Since $G_{N_A}^2$ decreases with N_A, while the other term increases with N_A, the $FPE(N_A)$ will have a minimum at some optimal value N_A.

The Akaike information criterion (AIC) is based on the minimization of the log-likelihood of the prediction-error variance as a function of the filter order N_A. The criterion is defined by

$$AIC(N_A) = \ln G_{N_A}^2 + \frac{2 \cdot N_A}{N}. \tag{5.9}$$

Here, again, the optimum value of N_A is the value for which the $AIC(N_A)$ has minimum.

In the criterion of autoregressive transfer (CAT) function the optimal filter order is obtained when the estimate of the difference in the mean-square errors between the true filter, which

exactly gives the prediction error, and the estimated filter, is minimum. This difference can be calculated, without explicitly knowing the exact filter, by using the formula [109]

$$CAT(N_A) = \frac{1}{N} \cdot \sum_{i=1}^{N_A} \frac{N-i}{N \cdot G_i^2} - \frac{N-N_A}{N \cdot G_{N_A}^2}. \tag{5.10}$$

Besides these, some other criteria have been developed, e.g. the modified Akaike information criterion (BIC) [112] (in [113] referred as the minimum description length (MDL) criterion)

$$BIC(N_A) = N \cdot \ln G_{N_A}^2 + N_A \cdot \ln N. \tag{5.11}$$

According to [110] the mentioned criteria give orders which produce acceptable spectra in the case of low noise but underestimate the order for higher noise levels.

The FPE, AIC, and CAT are asymptotically equal, i.e. for the data length *N* approaching infinity they select the same order. The BIC has higher penalty term, so it results in lower orders.

The male voice sampled at 8 kHz was analyzed using the FPE, AIC, CAT, and BIC criteria [114]. First, the speech signal was applied to the preemphasis filter

$$H_P(z) = 1 - 0.9 \cdot z^{-1}. \tag{5.12}$$

The criteria were performed on the frames of 24 ms stationary parts of the vowels "A" (193 frames), "E" (195 frames), "I" (198 frames), "O" (192 frames), "U" (192 frames), nasals "M" (629 frames), "N" (757 frames), and unvoiced fricative "Š" (894 frames) spoken by Jiljí Kubec with the mean pitch frequency of about 110 Hz. These frames were made from segments of 384 samples of "A", 386 samples of "E", 389 samples of "I", 383 samples of "O", 383 samples of "U", 820 samples of "M", 948 samples of "N", and 1085 samples of "Š" by shifting a window of 192 samples with the shift of one sample. Spectra of the optimal orders according to the order selection criteria were compared with the conventionally used model of the 8th order for 8 kHz sampling (twice the bandwidth of 4 kHz as discussed at the beginning of this section). Model parameters were computed using the autocorrelation method described in Section 5.1.1.1. The RMS log spectral measure [116] between these spectra and a reference spectrum was used as a comparison criterion. A smoothed logarithmic FFT spectrum was used as a reference. It was performed by homomorphic filtering using convolution of the log spectrum with the Blackman window specified by following equation [29]

$$w(n) = 0.42 - 0.5 \cdot \cos\left(\frac{2\pi \cdot n}{M-1}\right) + 0.08 \cdot \cos\left(\frac{4\pi \cdot n}{M-1}\right), \tag{5.13}$$

for $0 \le n \le M-1$.

The filter length M is given by [115]

$$M = \left[\frac{3 \cdot N_F}{L} + 0.5\right], \tag{5.14}$$

where N_F is the FFT length and L is the pitch period in samples (brackets represent integral part of the number).

Experiments have shown that the FPE, AIC and CAT criteria have almost always higher values than the BIC criterion. Comparison of minimum, maximum, and median values of these criteria for all the mentioned frames is in Table 5.1. The FPE, AIC and CAT criteria give nearly the same minimum, maximum, and median values, however, there exist some frames where their values differ. Dependence of AIC and BIC on the AR model order for a frame of the vowel "A" and nasal "M" is shown in upper part of Figure 5.2 and Figure 5.3. Lower parts of these figures represent the AR spectra of the optimum AIC and BIC orders together with the smoothed periodogram for the same speech frames. Comparison of the spectral measure of the 8th order AR model and the optimum order model according to the four criteria can be seen in Table 5.2.

sound	FPE order			AIC order			CAT order			BIC order		
	min	max	median	min	max	median	min	max	median	min	max	median
"A"	9	20	9	9	20	9	9	20	9	9	9	9
"E"	11	11	11	11	11	11	11	11	11	6	11	11
"I"	16	26	16	16	26	16	16	26	16	9	11	9
"O"	8	14	8	8	14	8	8	14	8	8	8	8
"U"	8	20	20	8	20	20	8	20	20	8	8	8
"M"	11	38	20	11	38	20	11	37	18	4	14	11
"N"	18	34	20	18	34	20	15	25	20	4	15	12
"Š"	7	40	17	7	40	17	7	40	16	4	12	7

Table 5.1 Minimum, maximum, and median values of the optimum AR order according to the FPE, AIC, CAT, and BIC for 5 vowels, 2 nasals, and 1 unvoiced fricative.

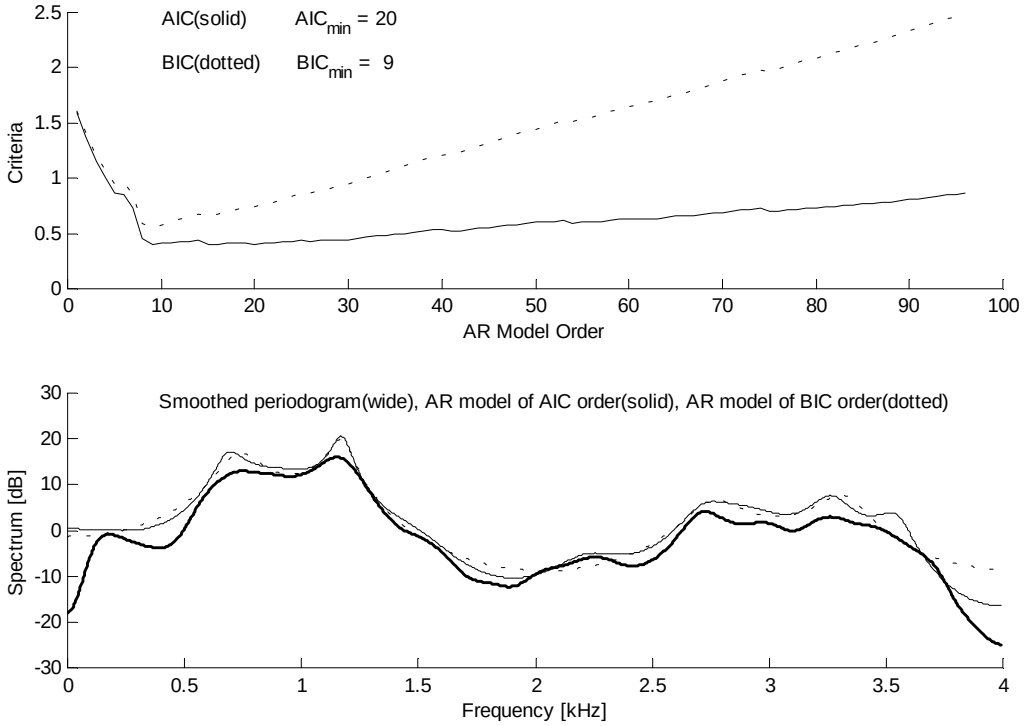

Figure 5.2 Upper: Dependence of AIC and BIC on AR order for a frame of the vowel "A". Lower: AR spectra of the optimum AIC order 20 and BIC order 9 together with the smoothed periodogram (RMS(AIC) = 4 dB, RMS(BIC) = 4.05 dB).

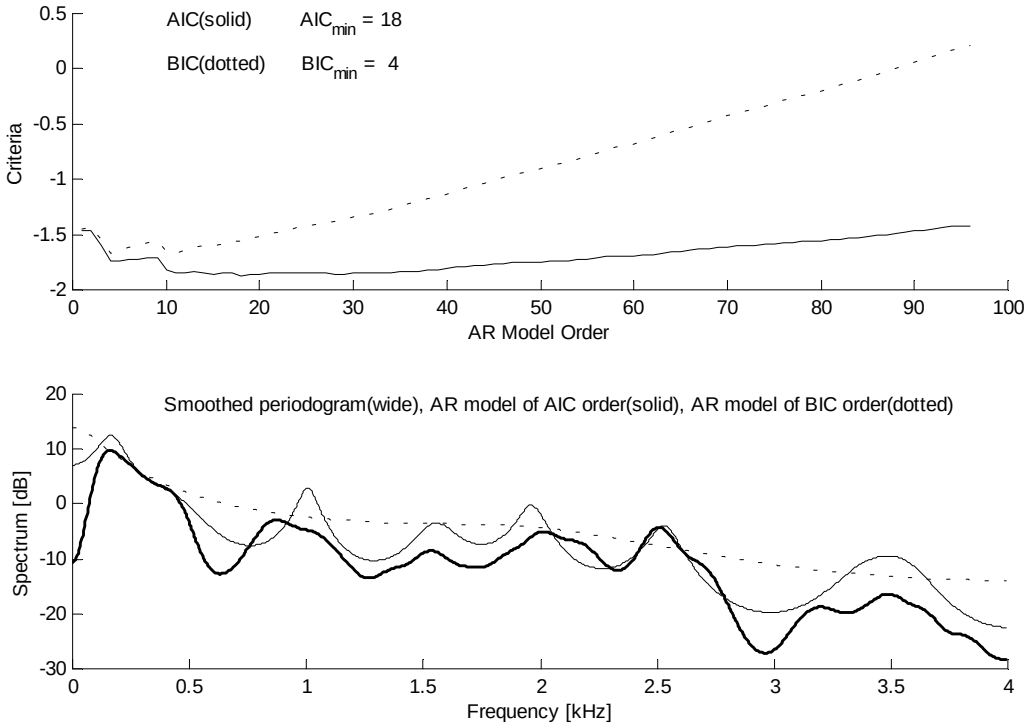

Figure 5.3 Upper: Dependence of AIC and BIC on AR order for a frame of the nasal "M". Lower: AR spectra of the optimum AIC order 18 and BIC order 4 together with the smoothed periodogram (RMS(AIC) = 4.81 dB, RMS(BIC) = 7.47 dB).

sound	RMS(FPE) [dB]	RMS(AIC) [dB]	RMS(CAT) [dB]	RMS(BIC) [dB]	RMS(8) [dB]
"A"	3.96	3.96	3.98	4.03	3.73
"E"	3.07	3.07	3.07	3.23	2.95
"I"	3.94	3.93	3.98	4.92	5.70
"O"	3.66	3.65	3.67	3.86	3.86
"U"	4.29	4.29	4.29	4.63	4.63
"M"	4.09	4.07	4.16	4.87	5.73
"N"	3.97	3.95	4.03	5.10	6.01
"Š"	2.56	2.53	2.60	3.67	4.28

Table 5.2 Mean values of the RMS log spectral measure. RMS(8) means the RMS log spectral measure between the spectrum of the 8^{th} order AR model and the smoothed periodogram. RMS with indices FPE, AIC, CAT, and BIC means the RMS log spectral measure between the spectrum of the optimum order AR model according to respective criteria and the smoothed periodogram.

Results of the FPE, AIC, CAT, and BIC minimization should give optimum orders for the AR modelling. However, comparison according to the RMS log spectral measure has not approved it in general. The median optimum order according to the criteria is at least 8 for voiced sounds. It agrees with the theoretical order for modelling of basic resonances of speech sounds. The BIC criterion exhibits worse results than the FPE, AIC, and CAT criteria, which give higher optimum orders. The FPE, AIC, and CAT criteria give excellent results for the measured frames of the sounds "I", "M", "N", and "Š", where the mean spectral measure of the optimum order is better than that of the 8^{th} order. The BIC criterion also gives good results for these sounds, though the mean spectral measure is higher what might be caused by lower BIC orders. On the other hand, the measured frames of the vowels "A" and "E" give bad results, where the optimum order model mean spectral measure is higher than the 8^{th} order model mean spectral measure. Neither the high order according to the criteria improves the RMS log spectral measure in general. It can be seen in Figure 5.4, where the model of the optimum FPE order 20 gives even worse spectral measure than the 8^{th} order model. Experiments have shown that in general the use of the FPE, AIC, CAT, and BIC criteria is not justified for choice of the AR model order of speech signals. Conventionally used model order of twice the bandwidth of the signal in kilohertz is always a good compromise.

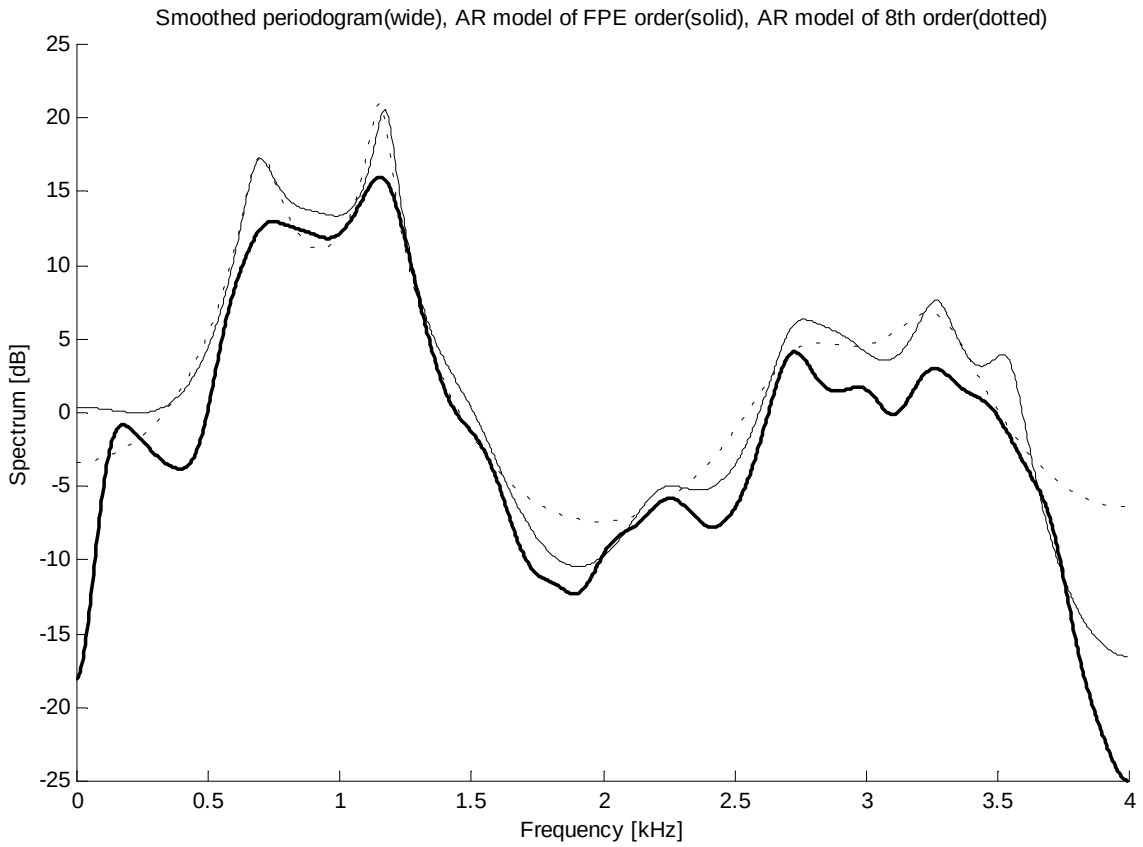

Figure 5.4 Comparison of the AR spectra for a frame of the vowel "A" with the FPE order 20, and the conventional 8th order, and the smoothed periodogram (RMS(FPE) = 4 dB, RMS(8) = 3.81 dB).

5.1.2 Cepstral Model

The cepstrum $\{c_n\}$ of a signal (sometimes referred to as the real cepstrum) is defined as the inverse Fourier transform of the logarithm of the magnitude of the Fourier transform [29], i.e.

$$c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln\left|S\left(e^{j\omega}\right)\right| \cdot \exp(jn\omega) \cdot d\omega \,. \tag{5.15}$$

Then, the logarithmic speech spectrum can be described by the cepstral coefficients $\{c_n\}$ using the expression

$$\ln\left|S\left(e^{j\omega}\right)\right| = \sum_{n=-\infty}^{\infty} c_n \cdot \exp(-jn\omega), \tag{5.16}$$

where ω denotes the normalized angular frequency.

Using N_F-point FFT it may be rewritten as

$$\ln|S_k| = \sum_{n=0}^{N_F-1} c_n \cdot \exp\left(-jn\frac{2\pi}{N_F}k\right), \qquad k = 0, 1,..., N_F-1. \tag{5.17}$$

A nonrealizable minimum-phase digital filter corresponding to the cepstral speech model, whose logarithmic magnitude frequency response approximates the function (5.16), is defined by the transfer function [30]

$$\widetilde{S}(z) = \exp(c_0) \cdot \exp\left(2\sum_{n=1}^{N_F/2-1} c_n z^{-n} + c_{N_F/2} z^{-N_F/2}\right), \tag{5.18}$$

and the transfer function of the filter modelling the vocal tract drawn in Figure 5.1 is given by truncation of the real cepstrum to N_C cepstral coefficients as follows

$$P(z) = \exp(c_0) \cdot \exp\left(2\sum_{n=1}^{N_C-1} c_n z^{-n}\right). \tag{5.19}$$

In other words, the frequency response of this filter determined by N_C cepstral coefficients is realized as a product of exponential functions. Each of the $N_C - 1$ exponential elements of the product included in the second exponential function of (5.19) can be approximated by Padé approximation of the continued fraction expansion of the exponential function [30], or Maclaurin approximation [117]. The former is performed as a cascade of IIR filters, the latter is performed as a cascade of FIR filters. A detailed block diagram of the cepstral speech model is in Figure 5.5 [108]. For unvoiced speech the excitation is formed by the noise generator. For voiced speech it is formed by the impulse generator plus high-pass filtered noise according to the value of the spectral flatness measure S_F. The impulse generator produces pulses with the shape of the impulse response of the Hilbert transformer in the pitch period intervals. The first cepstral coefficient c_0 is comprised in the gain of the filter and the next $N_C - 1$ coefficients are implemented in a cascade of $N_C - 1$ digital filters with the frequency response given by the chosen approximation. The output of the cascade is a synthesized speech signal.

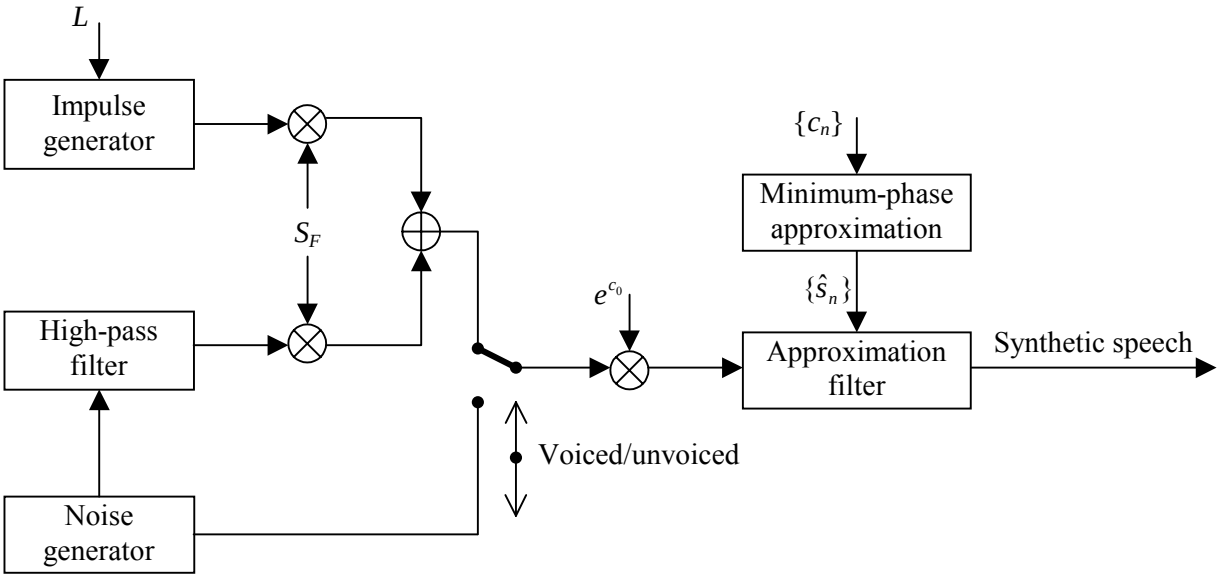

Figure 5.5 Block diagram of the cepstral speech model.

5.1.2.1 Cepstral Model Parameters Determination

The cepstral coefficients are determined using the definition (5.15). Using N_F-point inverse FFT it may be rewritten as

$$c_n = \frac{1}{N_F}\sum_{k=0}^{N_F-1}\ln|S_k|\cdot\exp\left(jn\frac{2\pi}{N_F}k\right), \tag{5.20}$$

where S_k are the FFT spectral values of the speech signal $s(n)$ weighted by the normalized Hamming window and zero-padded to N_F points. The window must be normalized in such a way that the energy of the original signal weighted by the Hamming window and a rectangular window with unitary magnitude are equal. The Hamming window $w(n)$ without normalization is specified by following equation [29]

$$w(n) = 0.54 - 0.46\cdot\cos\left(\frac{2\pi\cdot n}{N-1}\right), \qquad 0 \le n \le N-1, \tag{5.21}$$

where N is the length of the analysis frame.

The window $w_n(n)$ after the aforementioned normalization must satisfy the relation

$$\sum_{n=0}^{N-1} w_n^2(n) = N, \qquad 0 \le n \le N-1. \tag{5.22}$$

Then the normalized window can be derived as follows

$$w_n(n) = \frac{w(n)}{\sqrt{\frac{\sum_{n=0}^{N-1} w^2(n)}{N}}}, \qquad 0 \le n \le N-1. \tag{5.23}$$

The spectral flatness measure S_F, according to which high frequency noise is mixed in voiced frames, is determined by [118]

$$S_F = \frac{\exp(c_0)}{\frac{2}{N_F}\sum_{k=0}^{N_F/2-1} \ln|S_k|^2}. \tag{5.24}$$

5.1.2.2 Cepstral Model Order Selection

For the cepstral speech model there exist no way similar to AR model order selection described in Section 5.1.1.2. It was found out by simulation that the minimum number of 26 cepstral coefficients (a cascade of 25 approximation filters) is necessary for sufficient log spectrum approximation at 8-kHz sampling [119], and 51 cepstral coefficients (50 approximation filters in a cascade) are necessary for sufficient log spectrum approximation at 16-kHz sampling [120]. Both the values stem from the suggestion that the approximation error should be maximally about 1 dB. The real order of the filter cascade depends on the order of composite IIR and FIR filters approximating each of $N_C - 1$ exponential functions.

5.2 Harmonic Model

The principle of the harmonic speech model is shown in Figure 5.6. It is performed as a sum of harmonically related sine waves with frequencies given by pitch harmonics, and amplitudes and phases given by sampling the frequency response of the vocal tract model at these frequencies.

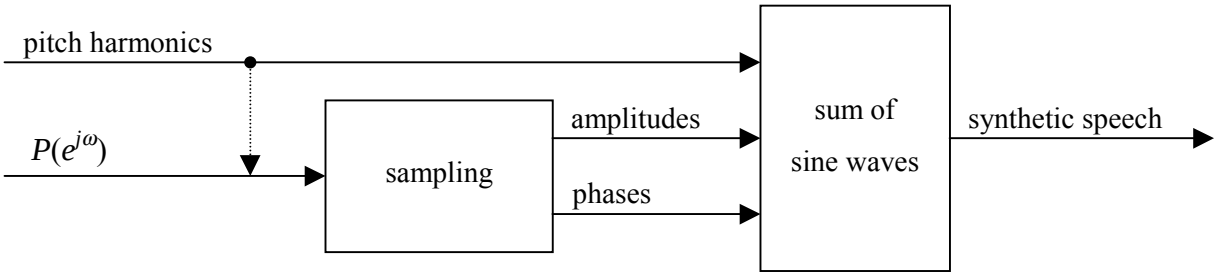

Figure 5.6 Principle of the harmonic speech model.

The number of the sine waves to be summed depends on the pitch period, and their amplitudes depend also on the fact whether the number of samples of the pitch period L is even or odd. Let us illustrate it for $L = 16$ and $L = 17$ corresponding to $F_0 = 500$ Hz and $F_0 = 470$ Hz at $f_s = 8$ kHz using a synthetic signal with the constant magnitude frequency response of the vocal tract model. The first case can be seen in Figure 5.7. Here, L is even and the number of composite sine waves is $L/2$. The amplitudes, obtained from spectral sampling at pitch harmonics, must be multiplied by 2 for first $L/2-1$ harmonics. The amplitude of the last harmonic ($L/2$) must not be multiplied by 2, because it coincides with $f_s/2$. The second case can be seen in Figure 5.8. In this case L is odd and the number of composite sine waves is the integral part of the half the pitch period in samples $[L/2]$. All the amplitudes, obtained from spectral sampling at pitch harmonics, must be multiplied by 2. It can be generalized in the following way:

1st step: number of pitch harmonics = $[L/2]$,

2nd step: sampling the frequency response of the vocal tract model at pitch harmonics

=> amplitudes $\{amplitude_m\}$,

3rd step: for the first $([L/2]-1)$ harmonics: $\{A_m\} = \{2 \cdot amplitude_m\}$,

4th step: $L \text{ modulo } 2 = 0 \quad => \quad \{A_{[L/2]}\} = \{amplitude_{[L/2]}\}$,

$L \text{ modulo } 2 \neq 0 \quad => \quad \{A_{[L/2]}\} = \{2 \cdot amplitude_{[L/2]}\}$. (5.25)

For voiced speech, using a minimum-phase assumption not only for the vocal tract but also for the glottal pulse contribution, the logarithm of the magnitude frequency response and the phase frequency response form a Hilbert transform pair. For unvoiced speech, the phases are randomized instead of using a random noise source known from the source-filter model. Such a phase randomization models a noise-like character of unvoiced speech while preserving its magnitude spectral shape.

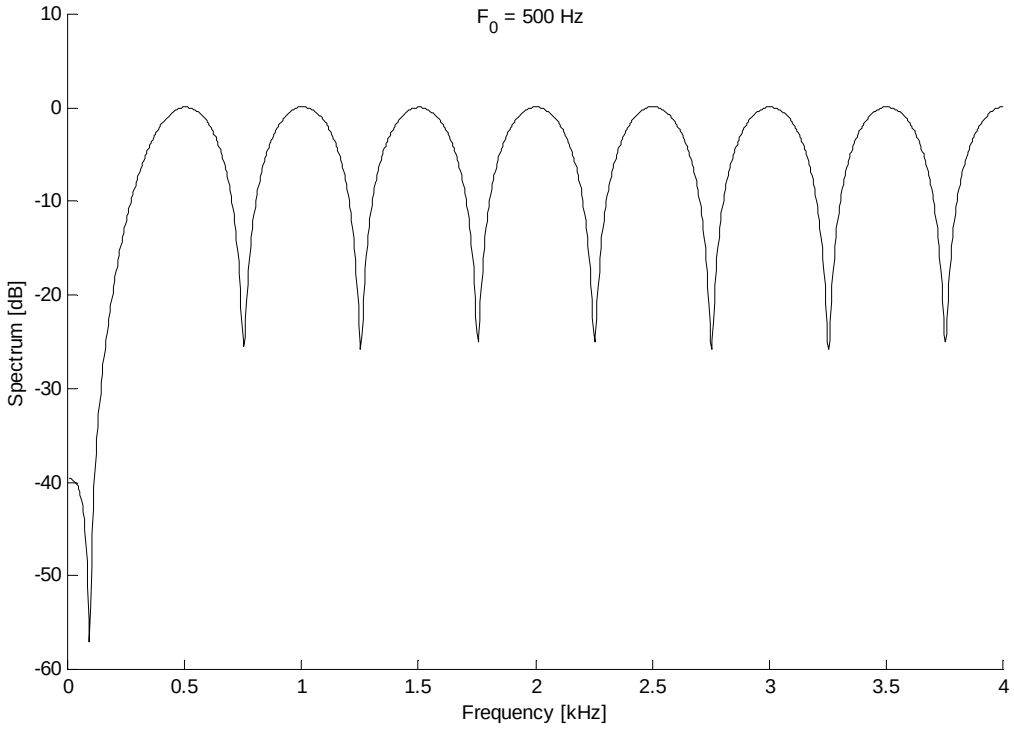

Figure 5.7 Spectrum of a synthetic signal with constant frequency response of the vocal tract model for $L = 16$, $f_s = 8$ kHz.

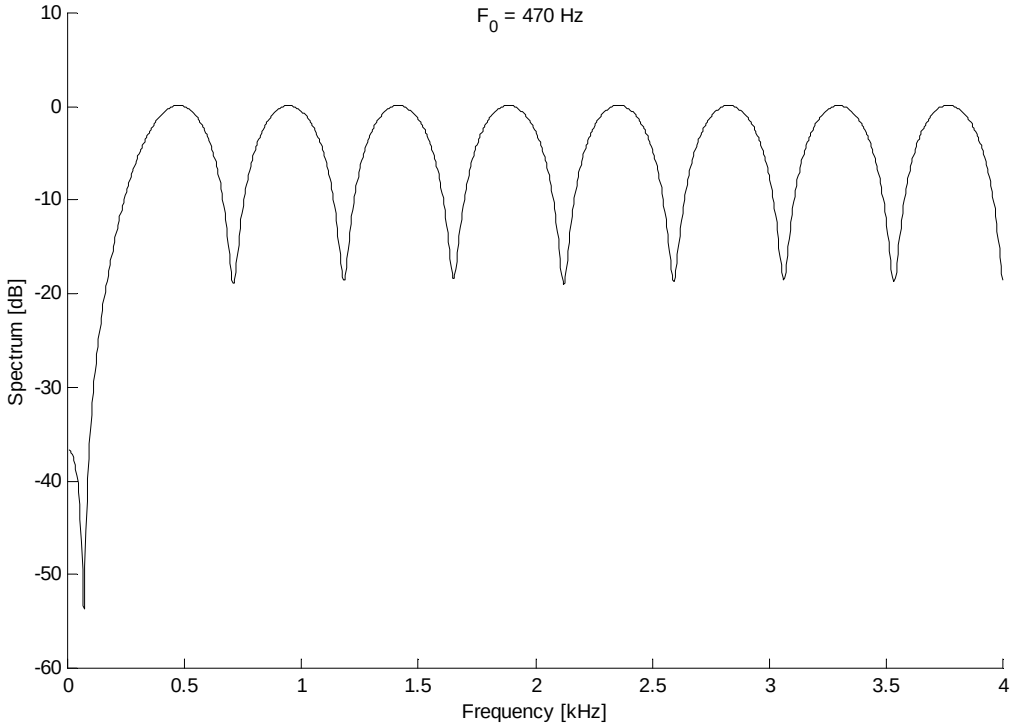

Figure 5.8 Spectrum of a synthetic signal with constant frequency response of the vocal tract model for $L = 17$, $f_s = 8$ kHz.

Before computing the spectrum, the analysis frame is weighted by the normalized Hamming window. The Hamming window is chosen as a good compromise between the side-lobe attenuation and the main-lobe bandwidth. The highest side-lobe level of the Hamming window is -43 dB, and the 6-dB bandwidth is 1.81 bins; where a bin is the frequency resolution ω_s / N [121], where $\omega_s = 2\pi f_s$. Then, to resolve two adjacent pitch harmonics, the following relation must be fulfilled:

$$1.81 \cdot \frac{f_s}{N} \leq F_0 . \tag{5.26}$$

After substituting for the pitch period in samples

$$L = \frac{f_s}{F_0}, \tag{5.27}$$

the length of the Hamming window should be

$$N \geq 1.81 \cdot L . \tag{5.28}$$

In [37], [39], [40] the window length is made at least 2.5 times the average pitch period to maintain the resolution properties. The same condition is taken over for the analysis frame duration throughout this thesis.

For the harmonic speech model the normalization of the window is different than the normalization for the source-filter model given by (5.23). The peak values of the periodogram must correspond to the amplitudes of the harmonic model. The peak signal gain of a window is defined by [121]

$$W(0) = \sum_{n=0}^{N-1} w(n). \tag{5.29}$$

As the peak values of the periodogram must be retained after spectral sampling, the normalized Hamming window for harmonic speech modelling must be of the form

$$w_n(n) = \frac{w(n)}{\sum_{n=0}^{N-1} w(n)}, \qquad 0 \leq n \leq N-1, \tag{5.30}$$

where $w(n)$ is defined by (5.21).

Synthesis of consecutive pitch-synchronous speech frames can be performed by simple concatenation or by OLA used e.g. in [40], [53]-[56], [64], [65], [78], [84]-[86], mentioned in Section 3.2.1. An asymmetric Hanning window is chosen for the OLA synthesis in this thesis. For every pair of consecutive pitch-synchronous frames the pitch period L_1 of the first frame is used for determination of the pitch harmonics $\{f_m\}$. The AR or cepstral parameters of the first and the second frame of the pair ($\mathbf{T}_1$, $\mathbf{T}_2$) are averaged. The harmonic parameters are determined according to Section 5.2.1.2 or 5.2.2.2. Speech is synthesized as a sum of sine waves during two consecutive pitch-synchronous frames of the length L_1 and L_2, as follows

$$s_y(l) = \sum_{m=1}^{[L_1/2]} A_m \cos(2\pi f_m l + \varphi_m), \qquad 0 \le l \le L_1 + L_2. \tag{5.31}$$

Then, this pair of frames is weighted by an asymmetric Hanning window with its left and right parts corresponding to the pitch periods of the first and the second frame. A symmetric Hanning window is defined by [29]

$$w(n) = \frac{1}{2}\left[1 - \cos\left(\frac{2\pi \cdot n}{N-1}\right)\right], \qquad 0 \le n \le N-1. \tag{5.32}$$

For the consecutive pitch periods L_1 and L_2, the asymmetric Hanning window may be written as

$$\begin{aligned} w(n) &= \frac{1}{2}\left[1 - \cos\left(\frac{\pi \cdot n}{L_1}\right)\right], && 0 \le n \le L_1, \\ w(n) &= \frac{1}{2}\left[1 + \cos\left(\frac{\pi \cdot (n - L_1)}{L_2}\right)\right], && L_1 + 1 \le n \le L_1 + L_2, \end{aligned} \tag{5.33}$$

so that the left part of the current asymmetric window has the same length as the right part of the previous window, and the right part of the current window has the same length as the left part of the next window, and the overlapped asymmetric windows are complementary. For the final synthesis the weighted overlapped consecutive pairs of pitch-synchronous frames are added to avoid discontinuities at the frame boundaries. The idea of the asymmetric Hanning window was inspired by [65], where this term had been mentioned only in the context of the noise component calculation by the harmonic component subtraction from the original signal.

In next sections the concatenated synthesis will be compared with the OLA for various parametrizations of the harmonic model.

5.2.1 AR Parametrization of the Harmonic Model

The harmonic model with AR parametrization (HAP) uses the description (5.1) to code the frequency response of the vocal tract model determining the amplitudes and phases of the composite sine waves.

5.2.1.1 AR Parameters Determination of the HAP

The HAP may use the same method of AR parameters determination as the AR model, however, much better results are given by the method with prior spectral envelope determination [78]. It will be discussed in detail In Section 5.2.1.4.

5.2.1.2 Harmonic Parameters Determination of the HAP

To determine the magnitude frequency response, first, the relation (5.1) is rewritten by

$$P\left(e^{j\omega}\right)=\frac{G}{B\left(e^{j\omega}\right)}, \tag{5.34}$$

where

$$B\left(e^{j\omega}\right)=1+\sum_{n=1}^{N_A} a_n \exp\left(-jn\omega\right). \tag{5.35}$$

Using N_F−point FFT, the relation (5.35) may be rewritten as

$$B_k = \sum_{n=0}^{N_F-1} b_n \cdot \exp\left(-jn\frac{2\pi}{N_F}k\right), \tag{5.36}$$

where

$$b_n = \begin{cases} 1, & n=0, \\ a_n, & 1 \le n \le N_A, \\ 0, & N_A < n \le N_F-1. \end{cases} \tag{5.37}$$

Then the AR model magnitude frequency response is given by

$$\left|P_k\right| = \frac{G}{\left|B_k\right|}. \tag{5.38}$$

Using a definition of the Hilbert transform [29] the relation between the magnitude frequency response and the minimum-phase frequency response may be rewritten in the following convolution form

$$\varphi^{\min}(k) = -\ln|P_k| * h(k), \tag{5.39}$$

where $h(k)$ is the impulse response of a 90-degree phase shifter corresponding to its frequency response

$$H(e^{j\omega}) = \begin{cases} -j, & 0 \le \omega < \pi, \\ j, & -\pi \le \omega < 0. \end{cases} \tag{5.40}$$

Using the N_F-point FFT, the relation (5.40) may be rewritten as

$$H_k = \begin{cases} 1, & k = 0,\ N_F/2, \\ 2, & k = 1, 2, \ldots, N_F/2 - 1, \\ 0, & k = N_F/2 + 1, \ldots, N_F - 1. \end{cases} \tag{5.41}$$

The minimum-phase frequency response may be computed using the FFT-s of the first and the second element of the convolution (5.39), and the inverse FFT of their product. Sampling the magnitude and minimum-phase frequency responses ($|P_k|$ and $\varphi^{\min}(k)$) at the pitch harmonics $\{f_m\}$ using the algorithm (5.25) gives the amplitudes $\{A_m\}$ and phases $\{\varphi_m\}$ of a composite synthetic speech signal

$$s_y(l) = \sum_{m=1}^{[L/2]} A_m \cos(2\pi f_m l + \varphi_m), \qquad 0 \le l \le L-1. \tag{5.42}$$

5.2.1.3 Number of Parameters for the HAP

In Section 5.1.1.2 it has been stated that the minimum AR model order of preemphasized speech should be twice the bandwidth of the signal in kilohertz. For the harmonic speech model no preemphasis is performed prior to the analysis and no postemphasis is performed prior to the synthesis. As the preemphasis is done by passing the speech signal through a single-zero filter given by (5.12), the AR model order of speech without preemphasis must be at least one order higher than that of speech with preemphasis. For the sampling frequency of f_s [kHz] it means the minimum AR model order of f_s+1. The real number of the HAP parameters is given by the pitch period using the algorithm (5.25).

5.2.1.4 AR Parameters Determination with Prior Spectral Envelope

Experiments with speech analysis and synthesis were performed on a male voice with the mean pitch frequency of about 110 Hz, recorded in an unechoic room using a 12-bit A/D converter with the sampling frequency of 8 kHz, and a magnetodynamic microphone. The analysis was performed in the frame intervals of 12 ms with the frame length of 24 ms, i.e. in 24-ms overlapping frames. The same conditions were used for application of the methods described in Sections 5.2.2.4, 5.2.2.5, and 5.2.4. In Figure 5.9, we can see spectra of a 24-ms voiced frame. Three orders of the AR model were compared: $N_A = 9$ (as it is the minimum necessary order as described in Section 5.2.1.3), $N_A = 25$ (as it corresponds to the same number of parameters as the cepstral model, see Sections 5.1.2.2 and 5.2.2.3), and $N_A = 17$ (as it is in the middle between these two values). Figure 5.9 shows comparison of the original speech spectrum and the AR spectrum of these orders using the standard autocorrelation method and a proposed new method.

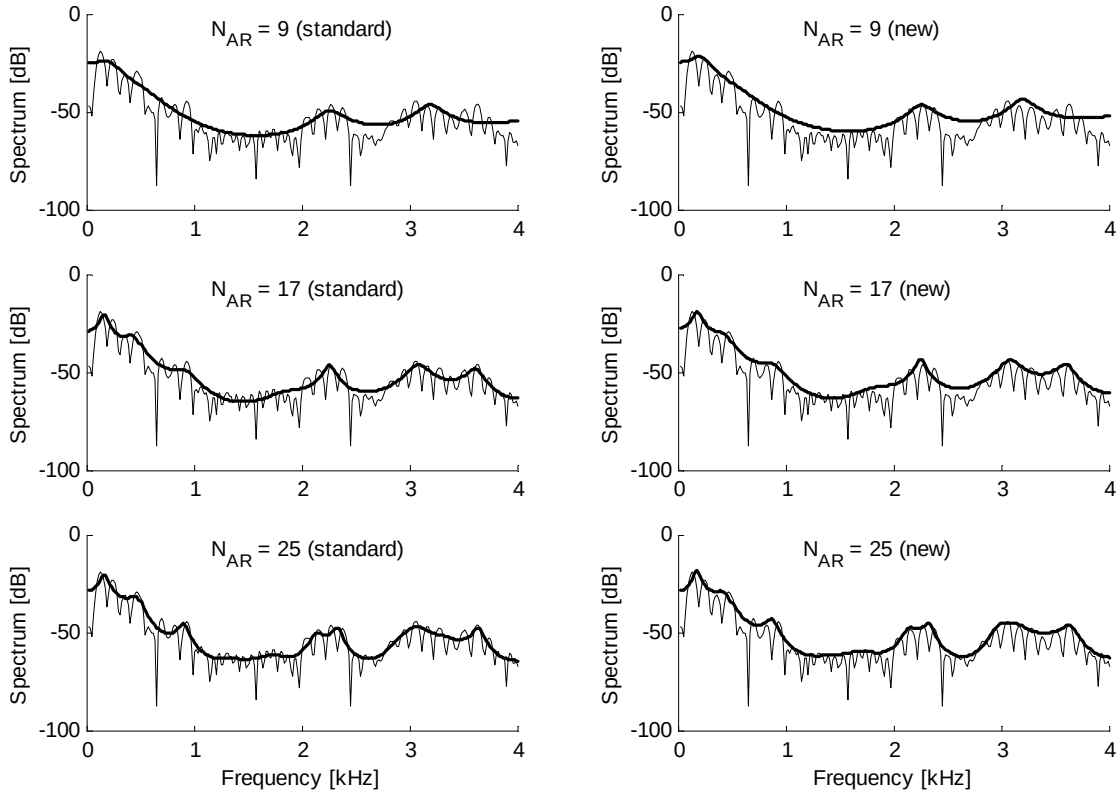

Figure 5.9 The original speech spectra (solid) together with the AR spectra of the order N_A (wide) for the standard autocorrelation method (left) and the proposed new method (right) for a 24-ms frame of a vowel "I" spoken by the male voice.

Let the standard method is denoted by HAP1, and the new method is denoted by HAP2. The block diagram of speech analysis using HAP2 is depicted in Figure 5.10. Standard autocorrelation method (HAP1) is performed with the dashed block omitted. In HAP2 the AR parameters are computed from the time-domain signal corresponding to the spectral envelope instead of the original speech signal. First, the staircase log spectral envelope is determined using steps of a pitch-frequency width. In each of the intervals of a pitch width the local maxima are found by detection of the slope change from positive to negative. The mean value of their amplitudes is chosen as the amplitude of the step. If no local maximum is found in the interval using this algorithm, the mean value of the interval boundary amplitudes is chosen as the amplitude of the step. The idea of the staircase envelope was inspired by the "piecewise constant interpolation of the sine wave amplitude measurements" mentioned in [39], where it had been used to compute what they had called "cepstral envelope". However, in this thesis, a new algorithm is proposed and the resulting staircase envelope is smoothed using the weighed moving average having the shape of the Blackman window defined by (5.13) and normalized in such a way that

$$\sum_{n=0}^{M-1} w(n) = 1 . \tag{5.43}$$

The length of the window has been determined experimentally to be one and a half of the pitch frequency. Then, the inverse Fourier transform of the smoothed spectral envelope was treated as a real speech signal. Thus, the AR parameters were computed by the autocorrelation method from the time-domain signal corresponding to the speech spectral envelope of the original speech signal.

To consider mixed voicing of many sounds detected as voiced by the pitch detector with binary voicing decision, a maximum voiced frequency f_{max} determines the degree of voicing. For totally unvoiced frames f_{max} is set to zero. For voiced frames f_{max} is computed from the magnitude spectrum comparing the frequency distances between the pitch harmonics and the spectral local maxima. If there is no spectral peak in the predefined vicinity of the pitch harmonic for two consecutive pitch harmonics, the last pitch harmonic before these two harmonics is chosen as the maximum voiced frequency. The predefined value of the distance between the pitch harmonic and the spectral peak was determined experimentally as a portion (0.4) of the pitch frequency.

The proposed algorithm for the staircase envelope determination may be written as follows:

S = log spectrum from 0 to f_s
For the first interval of $F_0/2$ duration
    $\mathbf{S}_x = \mathbf{S}$ from 0 to $F_0/2$
    Approximate gradient **G** of $\mathbf{S}_x$
    Find frequencies **F** of change **G** from positive to negative
    If **F** not found
        T = mean of $\mathbf{S}_x$ at 0 and $F_0/2$
    Else
        T = mean of **F**
    End
    Step of T amplitude from 0 to $F_0/2$
End
For next $L-1$ intervals of F_0 durations
    $\mathbf{S}_x = \mathbf{S}$ from $(2l-1)F_0/2$ to $(2l+1)F_0/2$
    Approximate gradient **G** of $\mathbf{S}_x$
    Find frequencies **F** of change **G** from positive to negative
    If **F** not found
        T = mean of $\mathbf{S}_x$ at $(2l-1)F_0/2$ and $(2l+1)F_0/2$
    Else
        T = mean of **F**
    End
    Step of T amplitude from $(2l-1)F_0/2$ to $(2l+1)F_0/2$
End
For the last interval of $F_0/2$ duration
    $\mathbf{S}_x = \mathbf{S}$ from $(2L-1)F_0/2$ to f_s
    Approximate gradient **G** of $\mathbf{S}_x$
    Find frequencies **F** of change **G** from positive to negative
    If **F** not found
        T = mean of $\mathbf{S}_x$ at $(2L-1)F_0/2$ and f_s
    Else
        T = mean of **F**
    End
    Step of T amplitude from $(2L-1)F_0/2$ to f_s
End (5.44)

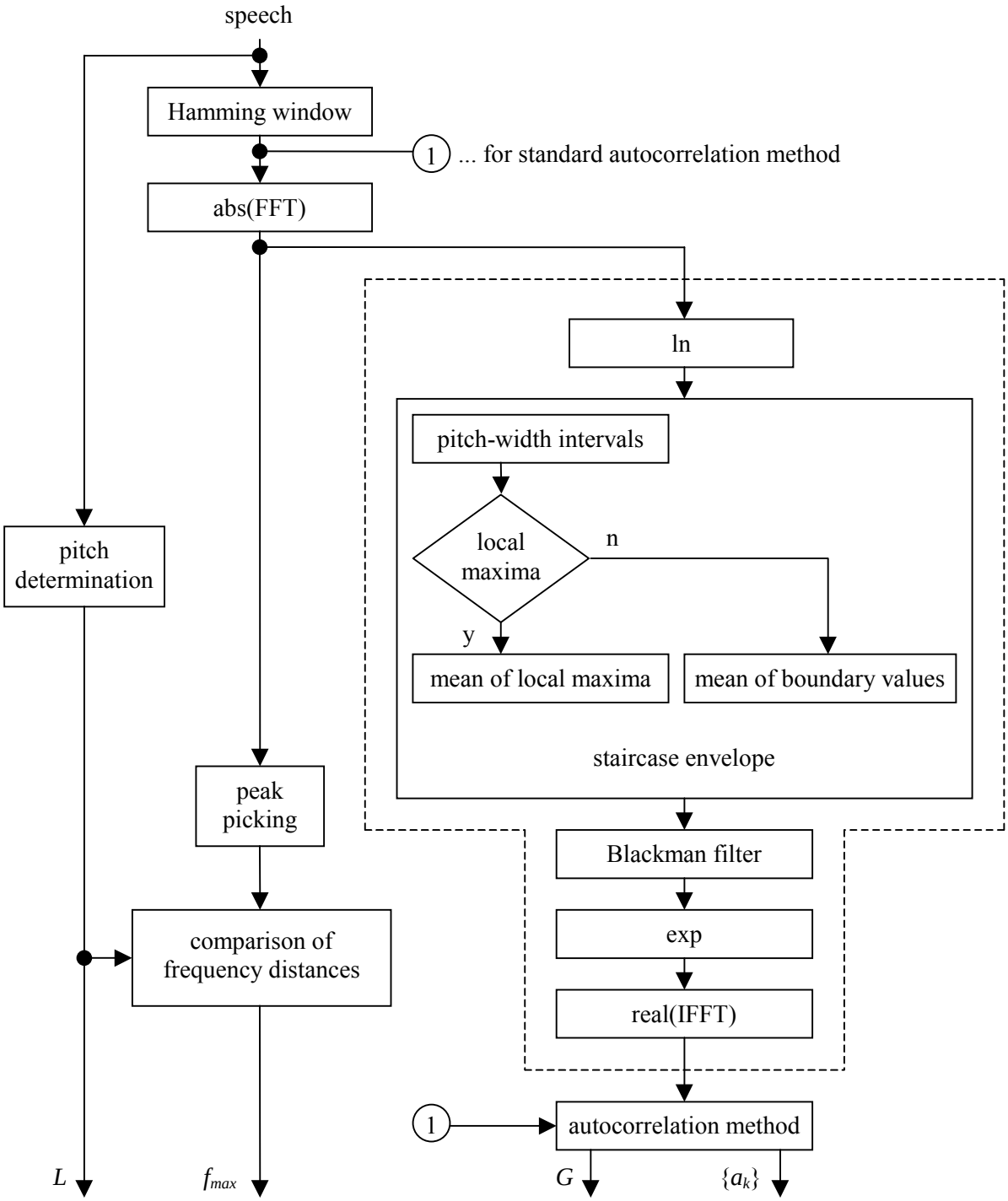

Figure 5.10 Analysis of one equidistant speech frame with determination of the maximum voiced frequency for the HAP model.

5.2.1.5 Speech Synthesis Using the HAP

The block diagram of speech synthesis as a sum of sine waves coded by AR parameters is shown in Figure 5.11. The output parameters of the analysis (the pitch period L, the gain G, the coefficients $\{a_k\}$, and the maximum voiced frequency f_{max}) serve as the input parameters for the speech synthesis. The vocal tract transfer function block is performed using the procedure described in Section 5.2.1.2. The phases $\{\varphi_m\}$ for unvoiced frames are randomized in the interval $< -\pi, \pi >$. The phases at frequencies lower than f_{max} are minimum phases $\{\varphi_m^{\min}\}$ and the phases at frequencies higher than f_{max} are randomized in the same way as the phases of unvoiced frames.

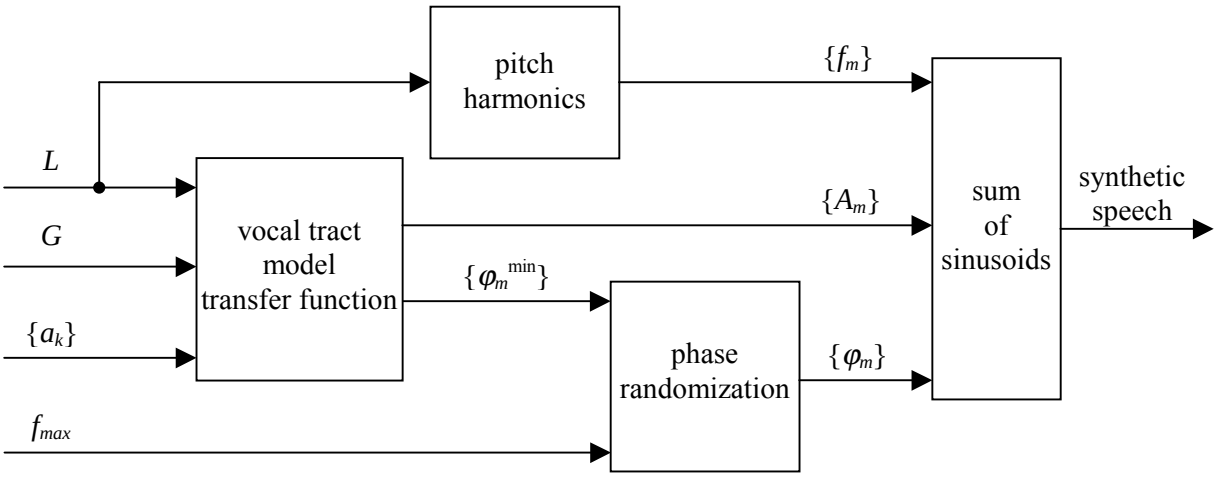

Figure 5.11 Synthesis of one pitch-synchronous speech frame using the HAP model.

5.2.1.6 Quantitative Comparison of Several Approaches to the HAP

The RMS logarithmic spectral measure was used to compare the smoothed spectra of original and resynthesized speech. The speech material consisted of about 450 stationary parts of 5 vowels, 2 nasals, and 1 unvoiced fricative. The RMS values were computed for the spectra of the speech frames weighed by a 24-ms Hamming window normalized by (5.23) and zero padded to 2048-point FFT. The computational complexity of the methods was determined with the help of the MATLAB function *flops* (floating point operation count) and has been referred to one sample of the processed speech signal of a 0.9-ms word "AFÉRA" having both voiced and unvoiced frames. The same conditions were used for application of other methods described in Sections 5.2.2.4 and 5.2.2.5.

Statistical values for the 9^{th}, 17^{th}, and 25^{th} order AR using standard autocorrelation method (HAP1) with concatenation of pitch-synchronous frames are shown in Tables 5.3 to 5.5. Results with the new method described in Section 5.2.1.4 (HAP2) using concatenated synthesis are shown Tables 5.6 to 5.8. Comparison of the three orders for all the voiced sounds, i.e. 5 vowels and 2 nasals, is represented in Figure 5.12. It can be seen that the mean value of the RMS log spectral measure slightly decreases with increasing AR order for both the analysis methods. However, the mean values for the method with prior spectral envelope (HAP2) are about 0.6 dB lower than that for the standard autocorrelation method (HAP1). The difference between the standard method and the new method is more evident for higher AR orders (9^{th} order: 0.54 dB, 17^{th} order: 0.63 dB, 25^{th} order: 0.64 dB). For the HAP1 the standard deviation gives almost the same values; for the HAP2 it is slightly higher.

Results for the synthesis performed by OLA of pairs of frames (5.31) weighted by the asymmetric Hanning window (5.33) are given in Tables 5.9 to 5.11 for the HAP1, and in Tables 5.12 to 5.14 for the HAP2. Averaged results for voiced frames are shown in Figure 5.13. Here we can see that for the HAP1, the mean value of the RMS log spectral measure and its standard deviation exhibit the same trend as that for the HAP1 with frame concatenation (compare HAP1 in Figures 5.12 and 5.13). However, the mean value is about 0.1 dB lower, and the standard deviation is about 0.2 dB lower for the OLA synthesis. Comparing the HAP1 and HAP2 with OLA in Figure 5.13 the same trend is seen as for concatenated synthesis: the mean RMS log spectral measure is about 0.6 dB lower than for the standard method. The difference between the standard method and the new method is even more evident for higher AR orders (9^{th} order: 0.52 dB, 17^{th} order: 0.66 dB, 25^{th} order: 0.76 dB). Comparing HAP2 in Figures 5.12 and 5.13, it can be concluded that OLA gives better results. The difference between the mean RMS log spectral measure for concatenation and OLA is not so evident as for the standard autocorrelation method (9^{th} order: 0.09 dB, 17^{th} order: 0.15 dB, 25^{th} order: 0.2 dB). However, the difference between the standard deviation for concatenation and OLA is more important (9^{th} order: 0.35 dB, 17^{th} order: 0.37 dB, 25^{th} order: 0.45 dB).

sound/ number of frames	RMS log spectral measure [dB]			
	minimum	maximum	mean	standard deviation
A/81	2.32	5.42	3.80	0.79
E/60	2.17	6.10	3.86	0.92
I/60	1.91	7.00	4.13	1.05
O/69	1.91	6.66	3.94	0.91
U/60	2.35	7.13	3.87	1.07
M/44	2.33	7.38	3.98	0.89
N/69	3.20	6.89	4.55	0.88
S/10	2.64	6.68	4.70	1.21

Table 5.3 RMS log spectral measure between the original and concatenated synthetic speech for the HAP with the 9th order AR using the standard autocorrelation method (HAP1).

sound/ number of frames	RMS log spectral measure [dB]			
	minimum	maximum	mean	standard deviation
A/81	1.97	5.87	3.56	0.86
E/60	1.73	8.42	3.68	1.04
I/60	1.63	7.14	3.67	1.06
O/69	2.35	6.19	3.64	0.79
U/60	2.49	8.10	4.02	1.14
M/44	2.01	4.89	3.37	0.65
N/69	2.12	7.59	3.68	0.87
S/10	2.97	7.42	4.67	1.27

Table 5.4 RMS log spectral measure between the original and concatenated synthetic speech for the HAP with the 17th order AR using the standard autocorrelation method (HAP1).

sound/ number of frames	RMS log spectral measure [dB]			
	minimum	maximum	mean	standard deviation
A/81	2.05	6.10	3.50	0.83
E/60	1.74	5.17	3.50	0.81
I/60	1.84	6.07	3.56	0.98
O/69	2.05	6.07	3.58	0.70
U/60	2.64	10.69	3.94	1.40
M/44	1.85	4.97	3.21	0.69
N/69	1.47	8.11	3.44	0.92
S/10	3.49	5.82	4.50	0.70

Table 5.5 RMS log spectral measure between the original and concatenated synthetic speech for the HAP with the 25th order AR using the standard autocorrelation method (HAP1).

sound/ number of frames	RMS log spectral measure [dB]			
	minimum	maximum	mean	standard deviation
A/81	1.58	4.51	2.54	0.65
E/60	1.12	6.08	2.65	0.92
I/60	1.45	7.23	3.33	1.34
O/69	1.75	7.40	3.40	1.02
U/60	1.80	8.41	3.80	1.40
M/44	2.54	11.21	3.93	1.37
N/69	3.04	7.87	4.72	1.19
S/10	3.22	9.25	5.10	1.67

Table 5.6 RMS log spectral measure between the original and concatenated synthetic speech for the HAP with the 9^{th} order AR using the inverse Fourier transform of the spectral envelope (HAP2).

sound/ number of frames	RMS log spectral measure [dB]			
	minimum	maximum	mean	standard deviation
A/81	1.35	4.31	2.37	0.53
E/60	1.23	5.08	2.46	0.72
I/60	1.25	5.95	2.92	1.00
O/69	1.49	6.97	3.04	1.02
U/60	1.61	9.63	3.88	1.66
M/44	1.94	4.71	3.10	0.70
N/69	1.78	7.70	3.43	1.10
S/10	3.10	6.92	4.89	1.40

Table 5.7 RMS log spectral measure between the original and concatenated synthetic speech for the HAP with the 17^{th} order AR using the inverse Fourier transform of the spectral envelope (HAP2).

sound/ number of frames	RMS log spectral measure [dB]			
	minimum	maximum	mean	standard deviation
A/81	1.13	3.86	2.29	0.53
E/60	1.14	5.47	2.40	0.80
I/60	1.18	6.40	2.84	0.97
O/69	1.45	6.80	2.87	0.94
U/60	1.24	11.05	3.84	1.86
M/44	1.78	4.90	2.85	0.69
N/69	1.67	8.80	3.15	1.12
S/10	3.46	6.57	4.73	0.90

Table 5.8 RMS log spectral measure between the original and concatenated synthetic speech for the HAP with the 25^{th} order AR using the inverse Fourier transform of the spectral envelope (HAP2).

sound/ number of frames	RMS log spectral measure [dB]			
	minimum	maximum	mean	standard deviation
A/81	2.47	5.45	3.79	0.77
E/60	2.15	5.95	3.86	0.84
I/60	2.17	5.52	3.87	0.71
O/69	2.83	5.24	3.89	0.60
U/60	2.39	5.86	3.59	0.67
M/44	1.74	5.47	3.84	0.70
N/69	3.15	6.30	4.50	0.72
S/10	4.41	7.01	5.45	0.89

Table 5.9 RMS log spectral measure between the original and OLA synthetic speech for the HAP with the 9th order AR using the standard autocorrelation method (HAP1).

sound/ number of frames	RMS log spectral measure [dB]			
	minimum	maximum	mean	standard deviation
A/81	2.02	5.95	3.58	0.79
E/60	1.96	5.47	3.63	0.83
I/60	2.04	5.38	3.51	0.81
O/69	2.16	4.78	3.59	0.59
U/60	2.35	5.15	3.43	0.60
M/44	1.71	5.09	3.42	0.66
N/69	2.34	5.05	3.65	0.67
S/10	3.71	7.24	5.35	1.12

Table 5.10 RMS log spectral measure between the original and OLA synthetic speech for the HAP with the 17th order AR using the standard autocorrelation method (HAP1).

sound/ number of frames	RMS log spectral measure [dB]			
	minimum	maximum	mean	standard deviation
A/81	2.13	6.41	3.58	0.82
E/60	1.57	5.61	3.62	0.84
I/60	2.00	5.44	3.38	0.83
O/69	1.84	4.85	3.52	0.63
U/60	2.13	4.89	3.37	0.60
M/44	1.52	5.13	3.26	0.69
N/69	1.93	6.19	3.40	0.77
S/10	3.74	6.77	5.34	1.05

Table 5.11 RMS log spectral measure between the original and OLA synthetic speech for the HAP with the 25th order AR using the standard autocorrelation method (HAP1).

sound/ number of frames	RMS log spectral measure [dB]			
	minimum	maximum	mean	standard deviation
A/81	1.55	5.04	2.67	0.52
E/60	1.31	4.30	2.71	0.56
I/60	1.51	5.80	3.20	0.88
O/69	1.80	5.10	3.31	0.73
U/60	1.82	6.86	3.50	0.96
M/44	2.54	5.68	3.68	0.71
N/69	3.06	6.75	4.65	0.84
S/10	3.75	6.14	4.87	0.82

Table 5.12 RMS log spectral measure between the original and OLA synthetic speech for the HAP with the 9th order AR using the inverse Fourier transform of the spectral envelope (HAP2).

sound/ number of frames	RMS log spectral measure [dB]			
	minimum	maximum	mean	standard deviation
A/81	1.48	4.65	2.50	0.52
E/60	1.23	3.50	2.49	0.50
I/60	1.46	5.71	2.70	0.78
O/69	1.82	4.77	2.87	0.68
U/60	1.62	6.07	3.19	0.86
M/44	2.15	4.73	3.06	0.54
N/69	1.96	5.73	3.36	0.79
S/10	3.73	6.78	5.00	1.10

Table 5.13 RMS log spectral measure between the original and OLA synthetic speech for the HAP with the 17th order AR using the inverse Fourier transform of the spectral envelope (HAP2).

sound/ number of frames	RMS log spectral measure [dB]			
	minimum	maximum	mean	standard deviation
A/81	1.27	4.02	2.36	0.50
E/60	1.22	3.45	2.38	0.48
I/60	1.40	4.71	2.54	0.65
O/69	1.56	4.58	2.63	0.63
U/60	1.50	5.88	3.09	0.90
M/44	2.02	4.42	2.81	0.51
N/69	1.75	5.89	3.01	0.78
S/10	2.89	7.15	4.72	1.27

Table 5.14 RMS log spectral measure between the original and OLA synthetic speech for the HAP with the 25th order AR using the inverse Fourier transform of the spectral envelope (HAP2).

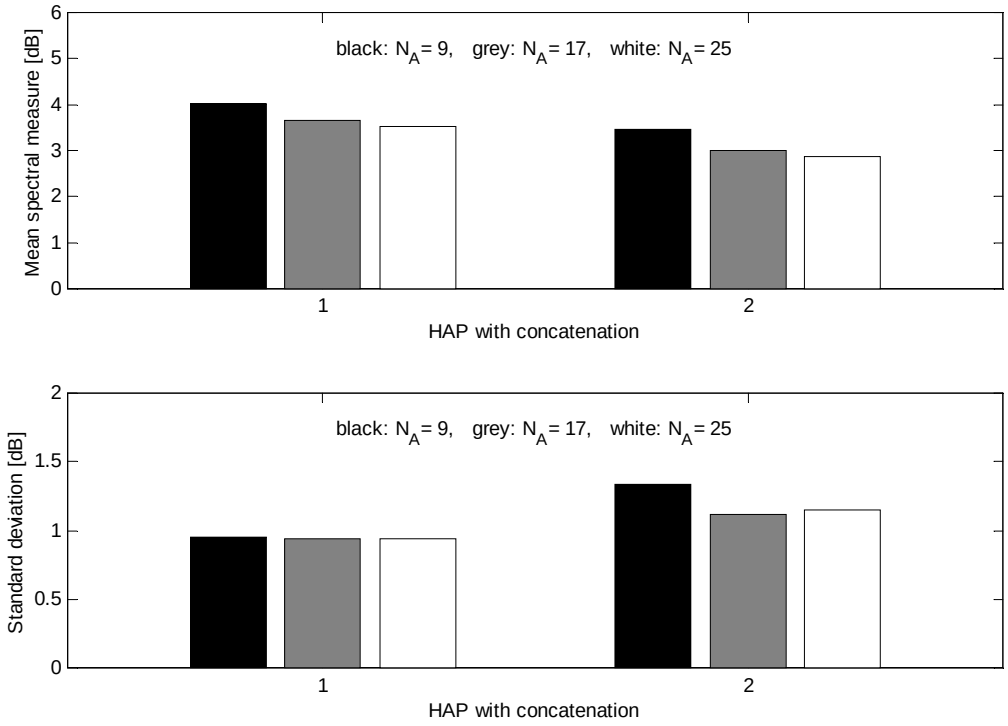

Figure 5.12 RMS log spectral measure between the original and concatenated synthetic voiced speech for the HAP with the 9th, 17th, and 25th order AR using standard autocorrelation method (HAP1), and using the inverse Fourier transform of the spectral envelope (HAP2).

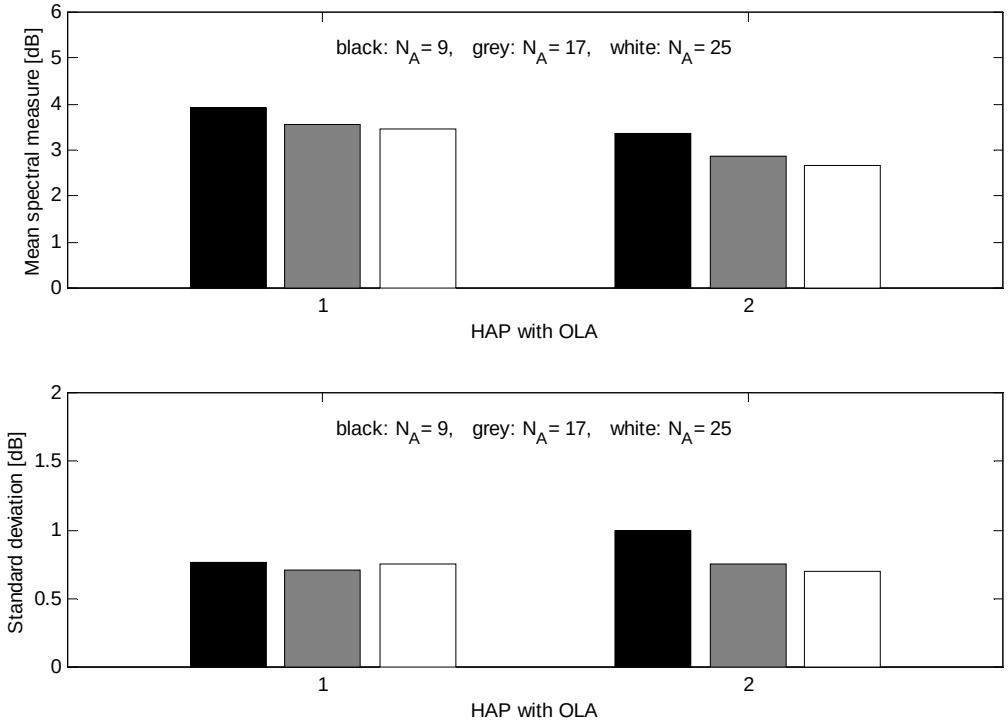

Figure 5.13 RMS log spectral measure between the original and OLA synthetic voiced speech for the HAP with the 9th, 17th, and 25th order AR using standard autocorrelation method (HAP1), and using the inverse Fourier transform of the spectral envelope (HAP2).

Computational complexity for individual blocks of the HAP1 analysis and the concatenated synthesis is shown in Tables 5.15 to 5.17. The overall results are in Figure 5.14. The computational complexity of the same HAP1 method with the OLA synthesis is given in Tables 5.18 to 5.20, summarized in Figure 5.15. Computational complexity for individual blocks of the HAP2 analysis and the concatenated synthesis is shown in Tables 5.21 to 5.23. The overall results are in Figure 5.16. The computational complexity of the same HAP2 method with the OLA synthesis is given in Tables 5.24 to 5.26, summarized in Figure 5.17.

Inspecting Tables 5.15 to 5.26, the blocks A0 and A1 give always the same computational complexity. However, the computational complexity of the block A0 depends on the duration of the processed signal because the normalized window (5.30) is computed once for the whole signal. Interesting might be comparison of the block A2 computational complexity for the HAP1 (Tables 5.15 to 5.20), and the HAP2 (Tables 5.21 to 5.26). Its much lower value for the HAP2 is due to using the magnitude spectrum for AR parameters determination. The same magnitude spectrum may be used for the maximum voiced frequency determination lowering thus its computational complexity.

Let us compare the computational complexity in Figures 5.14 to 5.17 with the RMS log spectral measure in Figures 5.12 and 5.13. We can find approximately reverse proportion between the mean RMS log spectral measure and the total computational complexity. Although the HAP1 with concatenation gives the lowest computational complexity (see Figure 5.14), it is not very useful because of the high mean RMS log spectral measure (see Figure 5.12). The computational complexity of the HAP1 with OLA is somewhat higher (see Figure 5.15), but the mean RMS log spectral measure is still rather high (see Figure 5.13). The lowest values of the RMS log spectral measure as well as the lowest standard deviation are manifested by the HAP2 with OLA, however at the expense of the highest computational complexity.

The HAP2 of the 25^{th} order AR with both the synthesis methods will be of interest also further in this work for comparison with the harmonic model with cepstral parametrization using the same number of parameters.

block	corresponding operations	complexity [flops/sample]
A0	segmentation, windowing	2.3
A1	pitch detection	448.3
A2	maximum voiced frequency determination (V/UV)	221/0
A3	AR parameters determination	62.4
Σ A	total analysis with AR parametrization (V/UV)	734.1/513.1
S1	AR to harmonic parameters transformation (V/UV)	920.1/266.3
S2	harmonic synthesis	204.5
Σ S	total harmonic synthesis with AR parametrization (V/UV)	1124.6/470.9

Table 5.15 Computational complexity for the HAP with the 9th order AR using the standard autocorrelation method (HAP1) with concatenated synthesis.

block	corresponding operations	complexity [flops/sample]
A0	segmentation, windowing	2.3
A1	pitch detection	448.3
A2	maximum voiced frequency determination (V/UV)	221/0
A3	AR parameters determination	115.5
Σ A	total analysis with AR parametrization (V/UV)	787.2/566.2
S1	AR to harmonic parameters transformation (V/UV)	920.1/266.3
S2	harmonic synthesis	204.5
Σ S	total harmonic synthesis with AR parametrization (V/UV)	1124.6/470.9

Table 5.16 Computational complexity for the HAP with the 17th order AR using the standard autocorrelation method (HAP1) with concatenated synthesis.

block	corresponding operations	complexity [flops/sample]
A0	segmentation, windowing	2.3
A1	pitch detection	448.3
A2	maximum voiced frequency determination (V/UV)	221/0
A3	AR parameters determination	171.2
Σ A	total analysis with AR parametrization (V/UV)	842.9/621.9
S1	AR to harmonic parameters transformation (V/UV)	920.1/266.3
S2	harmonic synthesis	204.5
Σ S	total harmonic synthesis with AR parametrization (V/UV)	1124.6/470.9

Table 5.17 Computational complexity for the HAP with the 25th order AR using the standard autocorrelation method (HAP1) with concatenated synthesis.

block	corresponding operations	complexity [flops/sample]
A0	segmentation, windowing	2.3
A1	pitch detection	448.3
A2	maximum voiced frequency determination (V/UV)	221/0
A3	AR parameters determination	62.4
Σ A	total analysis with AR parametrization (V/UV)	734.1/513.1
S1	AR to harmonic parameters transformation (V/UV)	922.3/268.6
S2	harmonic synthesis	540.2
Σ S	total harmonic synthesis with AR parametrization (V/UV)	1462.5/808.8

Table 5.18 Computational complexity for the HAP with the 9th order AR using the standard autocorrelation method (HAP1) with OLA synthesis.

block	corresponding operations	complexity [flops/sample]
A0	segmentation, windowing	2.3
A1	pitch detection	448.3
A2	maximum voiced frequency determination (V/UV)	221/0
A3	AR parameters determination	115.5
Σ A	total analysis with AR parametrization (V/UV)	787.2/566.2
S1	AR to harmonic parameters transformation (V/UV)	922.6/268.8
S2	harmonic synthesis	540.2
Σ S	total harmonic synthesis with AR parametrization (V/UV)	1462.8/809

Table 5.19 Computational complexity for the HAP with the 17th order AR using the standard autocorrelation method (HAP1) with OLA synthesis.

block	corresponding operations	complexity [flops/sample]
A0	segmentation, windowing	2.3
A1	pitch detection	448.3
A2	maximum voiced frequency determination (V/UV)	221/0
A3	AR parameters determination	171.2
Σ A	total analysis with AR parametrization (V/UV)	842.9/621.9
S1	AR to harmonic parameters transformation (V/UV)	922.8/269.1
S2	harmonic synthesis	540.2
Σ S	total harmonic synthesis with AR parametrization (V/UV)	1463/809.2

Table 5.20 Computational complexity for the HAP with the 25th order AR using the standard autocorrelation method (HAP1) with OLA synthesis.

block	corresponding operations	complexity [flops/sample]
A0	segmentation, windowing	2.3
A1	pitch detection	448.3
A2	maximum voiced frequency determination (V/UV)	50.3/0
A3	AR parameters determination	689
Σ A	total analysis with AR parametrization (V/UV)	1189.9/1139.6
S1	AR to harmonic parameters transformation (V/UV)	920.1/266.3
S2	harmonic synthesis	204.5
Σ S	total harmonic synthesis with AR parametrization (V/UV)	1124.6/470.9

Table 5.21 Computational complexity for the HAP with 9th order AR using the inverse Fourier transform of the spectral envelope (HAP2) with concatenated synthesis.

block	corresponding operations	complexity [flops/sample]
A0	segmentation, windowing	2.3
A1	pitch detection	448.3
A2	maximum voiced frequency determination (V/UV)	50.3/0
A3	AR parameters determination	822
Σ A	total analysis with AR parametrization (V/UV)	1323/1272.7
S1	AR to harmonic parameters transformation (V/UV)	920.1/266.3
S2	harmonic synthesis	204.5
Σ S	total harmonic synthesis with AR parametrization (V/UV)	1124.6/470.9

Table 5.22 Computational complexity for the HAP with 17th order AR using the inverse Fourier transform of the spectral envelope (HAP2) with concatenated synthesis.

block	corresponding operations	complexity [flops/sample]
A0	segmentation, windowing	2.3
A1	pitch detection	448.3
A2	maximum voiced frequency determination (V/UV)	50.3/0
A3	AR parameters determination	957.7
Σ A	total analysis with AR parametrization (V/UV)	1458.7/1408.4
S1	AR to harmonic parameters transformation (V/UV)	920.1/266.3
S2	harmonic synthesis	204.5
Σ S	total harmonic synthesis with AR parametrization (V/UV)	1124.6/470.9

Table 5.23 Computational complexity for the HAP with 25th order AR using the inverse Fourier transform of the spectral envelope (HAP2) with concatenated synthesis.

block	corresponding operations	complexity [flops/sample]
A0	segmentation, windowing	2.3
A1	pitch detection	448.3
A2	maximum voiced frequency determination (V/UV)	50.3/0
A3	AR parameters determination	691.6
Σ A	total analysis with AR parametrization (V/UV)	1192.6/1142.2
S1	AR to harmonic parameters transformation (V/UV)	922.3/268.6
S2	harmonic synthesis	540.2
Σ S	total harmonic synthesis with AR parametrization (V/UV)	1462.5/808.8

Table 5.24 Computational complexity for the HAP with 9th order AR using the inverse Fourier transform of the spectral envelope (HAP2) with OLA synthesis.

block	corresponding operations	complexity [flops/sample]
A0	segmentation, windowing	2.3
A1	pitch detection	448.3
A2	maximum voiced frequency determination (V/UV)	50.3/0
A3	AR parameters determination	824.6
Σ A	total analysis with AR parametrization (V/UV)	1325.6/1275.3
S1	AR to harmonic parameters transformation (V/UV)	922.6/268.8
S2	harmonic synthesis	540.2
Σ S	total harmonic synthesis with AR parametrization (V/UV)	1462.8/809

Table 5.25 Computational complexity for the HAP with 17th order AR using the inverse Fourier transform of the spectral envelope (HAP2) with OLA synthesis.

block	corresponding operations	complexity [flops/sample]
A0	segmentation, windowing	8.3
A1	pitch detection	448.3
A2	maximum voiced frequency determination (V/UV)	50.3/0
A3	AR parameters determination	960.3
Σ A	total analysis with AR parametrization (V/UV)	1461.3/1411
S1	AR to harmonic parameters transformation (V/UV)	922.8/269.1
S2	harmonic synthesis	540.2
Σ S	total harmonic synthesis with AR parametrization (V/UV)	1463/809.2

Table 5.26 Computational complexity for the HAP with 25th order AR using the inverse Fourier transform of the spectral envelope (HAP2) with OLA synthesis.

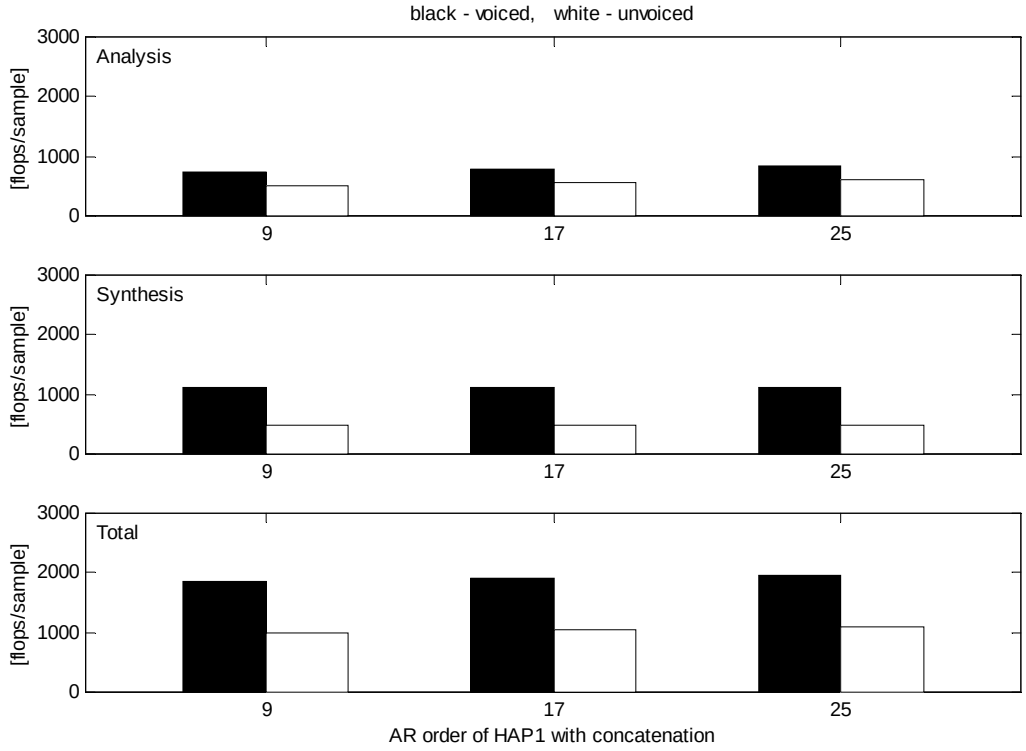

Figure 5.14 Computational complexity for the HAP with the 9th, 17th, and 25th order AR using the standard autocorrelation method (HAP1) with concatenated synthesis.

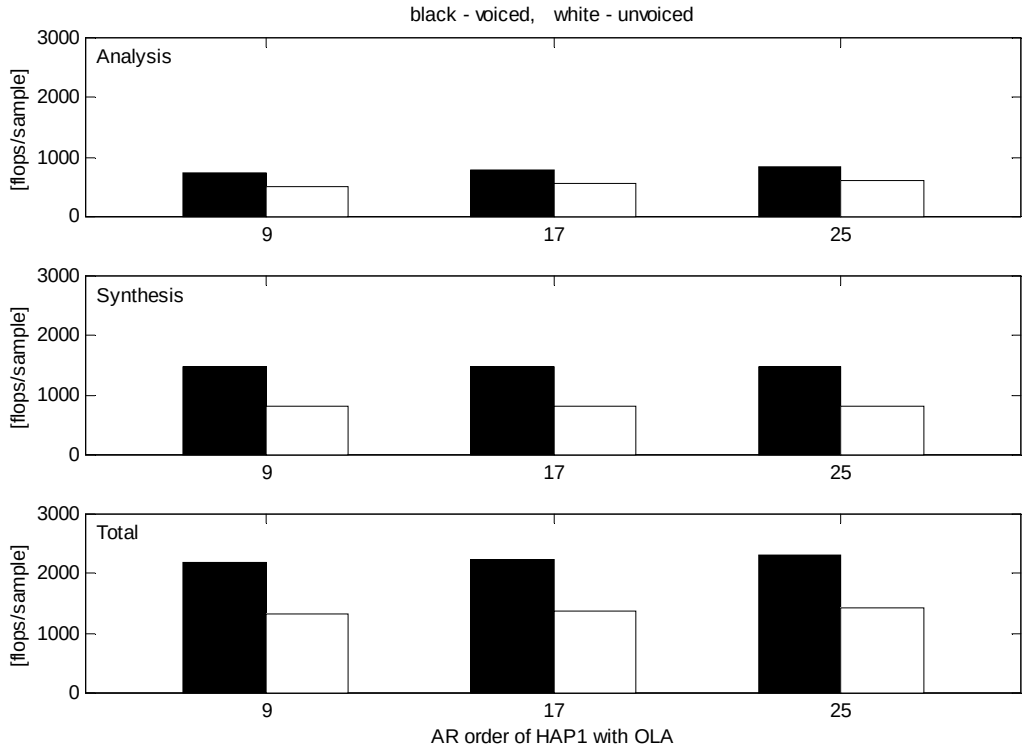

Figure 5.15 Computational complexity for the HAP with the 9th, 17th, and 25th order AR using the standard autocorrelation method (HAP1) with OLA synthesis.

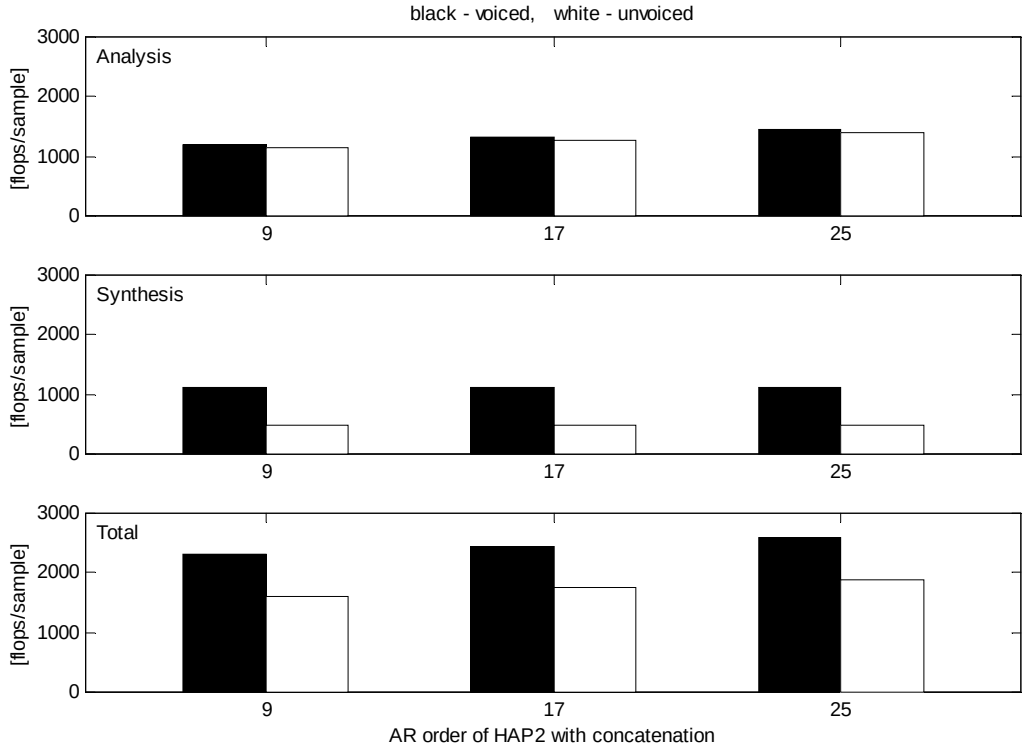

Figure 5.16 Computational complexity for the HAP with the 9th, 17th, and 25th order AR using the inverse Fourier transform of the spectral envelope (HAP2) with concatenated synthesis.

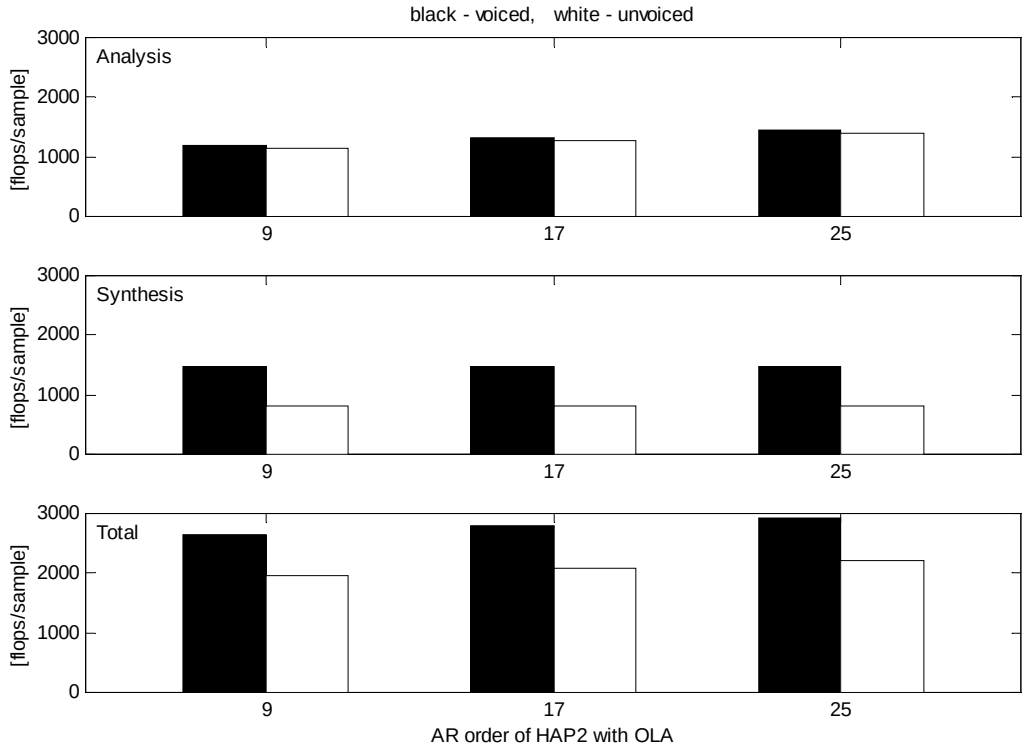

Figure 5.17 Computational complexity for the HAP with the 9th, 17th, and 25th order AR using the inverse Fourier transform of the spectral envelope (HAP2) with OLA synthesis.

5.2.1.7 An Experiment with Childish Voice Analysis and Synthesis

The advantage of the new AR parameters determination method for HAP is more evident for childish voice. Figure 5.18 presents a 9th order AR model frequency response determined by the standard autocorrelation method from a 10-ms speech frame of a childish voice sampled at 8 kHz. Its mean pitch frequency is about 300 Hz, and it was recorded in a domestic environment using a 16-bit SoundBlaster and a capacitor microphone. We can see that this model does not represent real spectral envelope for such a high pitch as can be found for example in a childish voice. The formant frequencies are biased toward pitch harmonics and formant bandwidth is underestimated. It is because the AR model frequency response shows tendency to follow the fine structure of the speech spectrum for high-pitch speakers. Problems would occur if this model were used in the TTS system, where prosodic modifications are necessary. It is evident that after sampling this frequency response at harmonics of modified pitch the original formant structure might be destroyed. The staircase envelope, and the smoothed staircase envelope is shown in Figure 5.19.

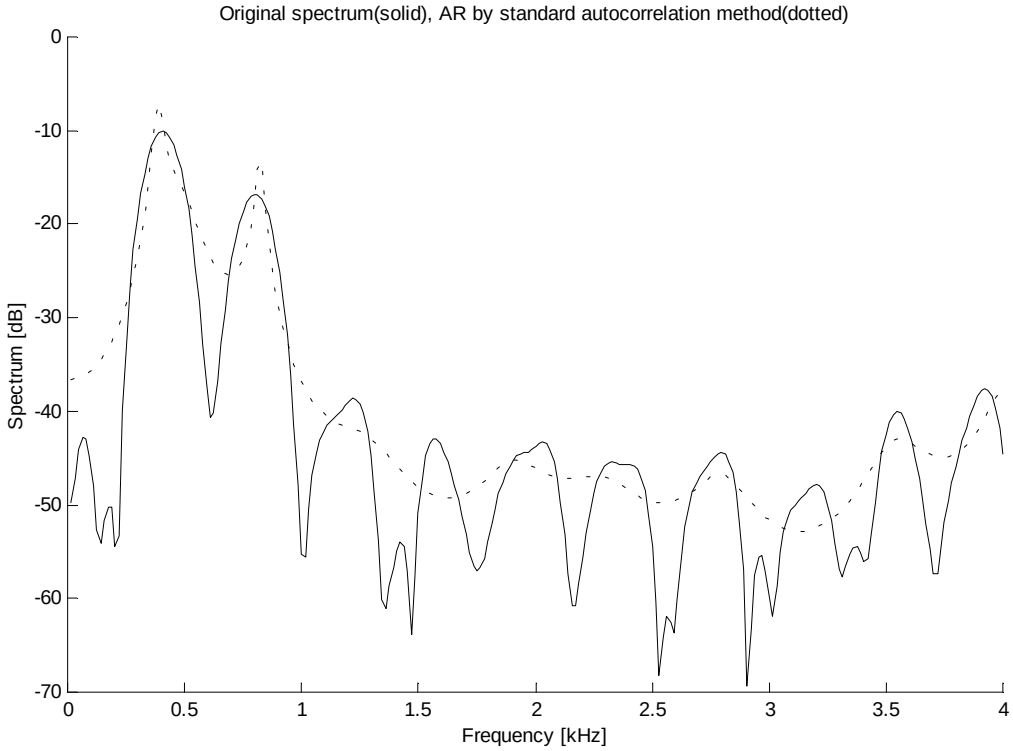

Figure 5.18 Original spectrum and AR magnitude frequency response determined by the standard autocorrelation method for a 10-ms frame of a vowel "I" spoken by the childish voice.

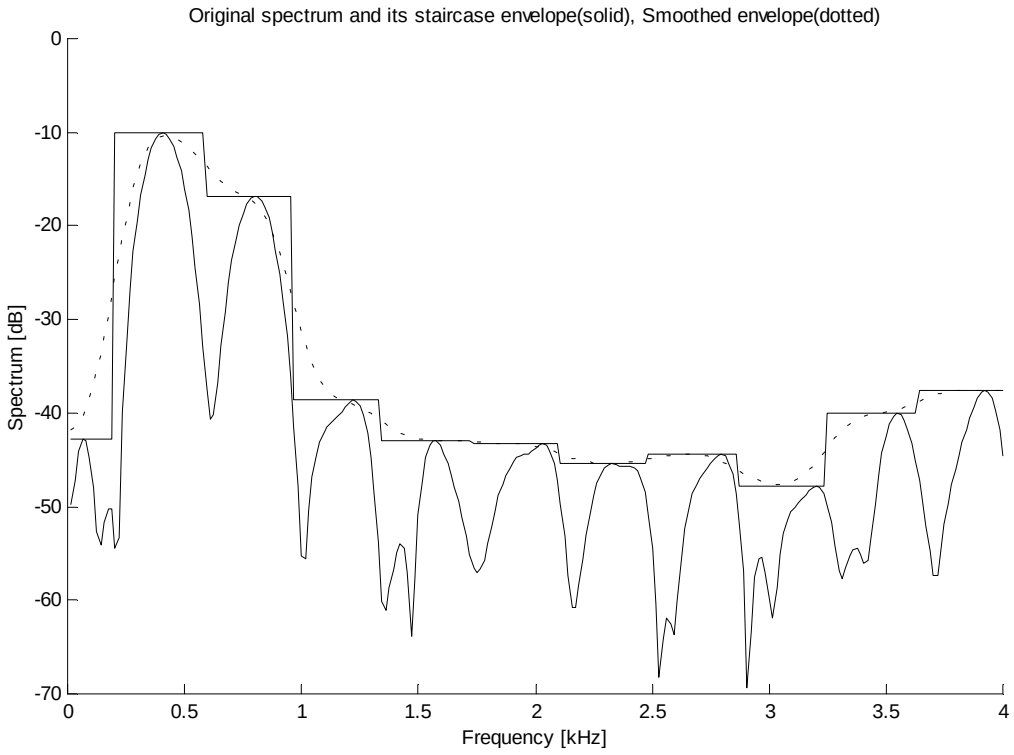

Figure 5.19 Original spectrum, staircase envelope, and smoothed staircase envelope for a 10-ms frame of a vowel "I" spoken by the childish voice.

The analysis of the childish voice was performed in the frame intervals of 5 ms with the frame length of 10 ms, i.e. in 10-ms overlapping frames. Here, no mixed voicing was used, so the block diagram of the analysis may be simplified as in Figure 5.20. Frames were either voiced or unvoiced, so the block diagram of the speech synthesis in a pitch-synchronous frame can be performed as in Figure 5.21. The AR magnitude frequency response of the vocal tract model corresponding to the new method is shown in Figure 5.22. This magnitude frequency response is sampled at frequencies $\{f_m\}$ to get the amplitudes $\{A_m\}$. For voiced frames ($L \neq 0$) the Hilbert transform of the logarithmic AR magnitude frequency response of the vocal tract model determines the phases $\{\varphi_m\}$. For unvoiced frames ($L = 0$) the phases $\{\varphi_m\}$ are randomized in the interval $< -\pi , \pi >$. OLA using (5.31) weighted by (5.33) was used for synthesis.

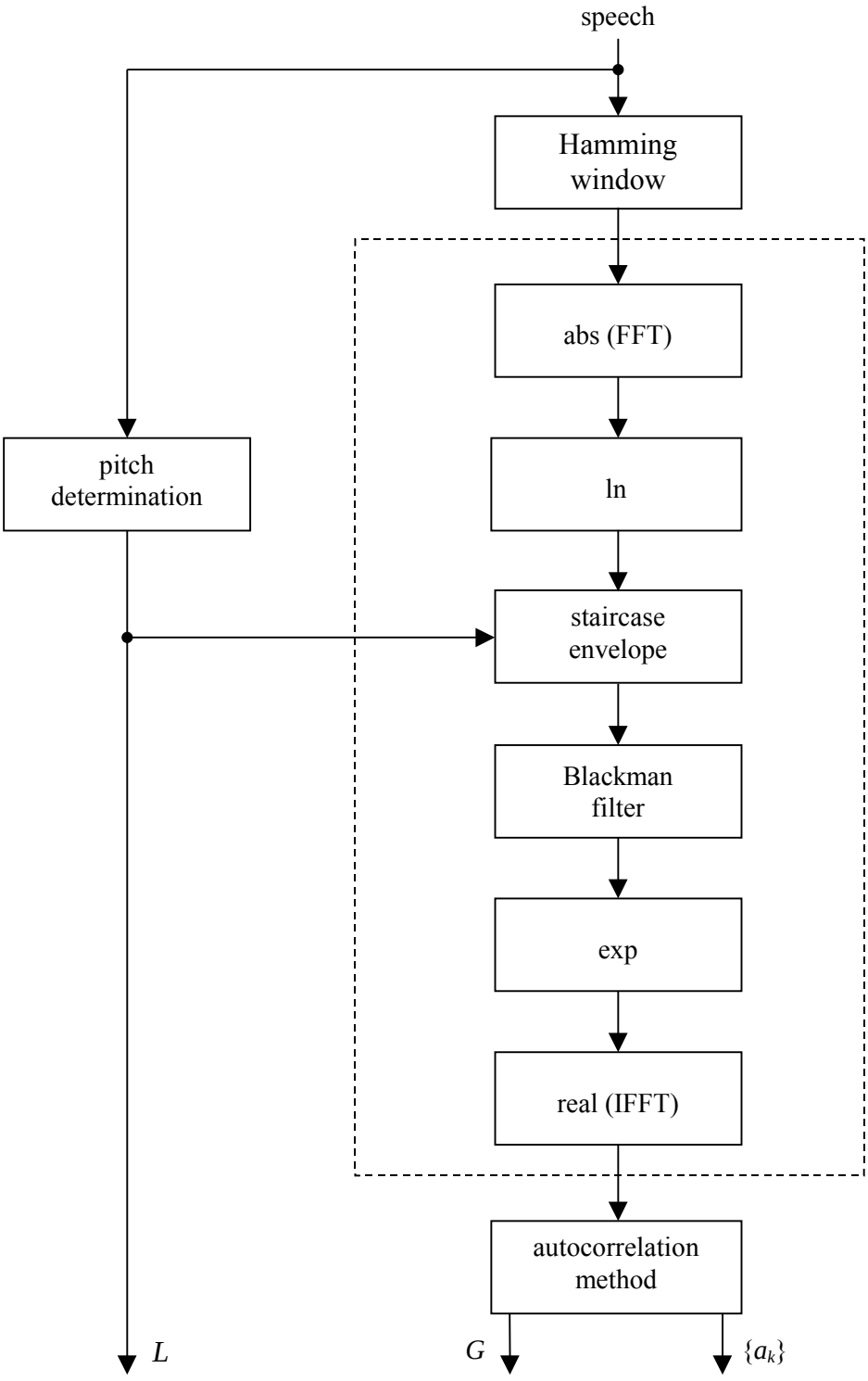

Figure 5.20 Analysis of one equidistant speech frame for the childish voice.

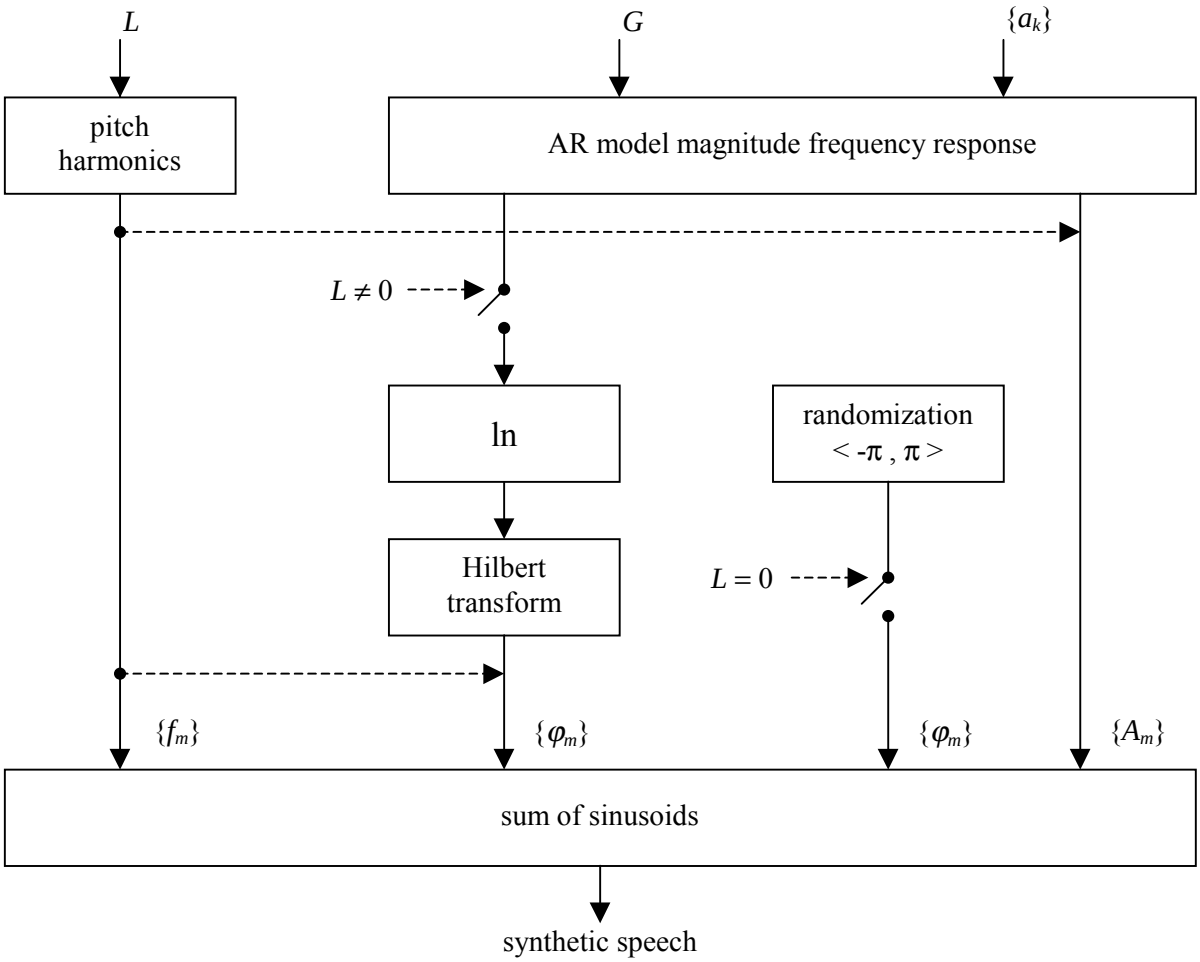

Figure 5.21 Synthesis of one pitch-synchronous speech frame for the childish voice.

The RMS log spectral measure was used to compare the smoothed spectra of original and resynthesized speech and both the methods were compared. The speech material consisted of 1687 stationary parts of vowels and nasals spoken by a childish voice with the mean pitch frequency of about 300 Hz. Comparison was made for 10-ms speech frames sampled at 8 kHz. The method using the spectral envelope computation described here was compared with the standard autocorrelation method of AR parameters determination from the windowed speech signal. The standard autocorrelation method was performed according to Figure 5.20 when leaving out the block drawn in the dashed rectangle. The results are shown in Table 5.27.

method	RMS log spectral measure [dB]			
	minimum	maximum	mean	standard deviation
AR1	0.29	13.12	4.46	1.66
AR2	0.53	9.45	3.18	1.28

Table 5.27 Statistical values of the RMS log spectral measure for the standard autocorrelation method (AR1), and the autocorrelation method using the inverse Fourier transform of the spectral envelope (AR2) for the childish voice.

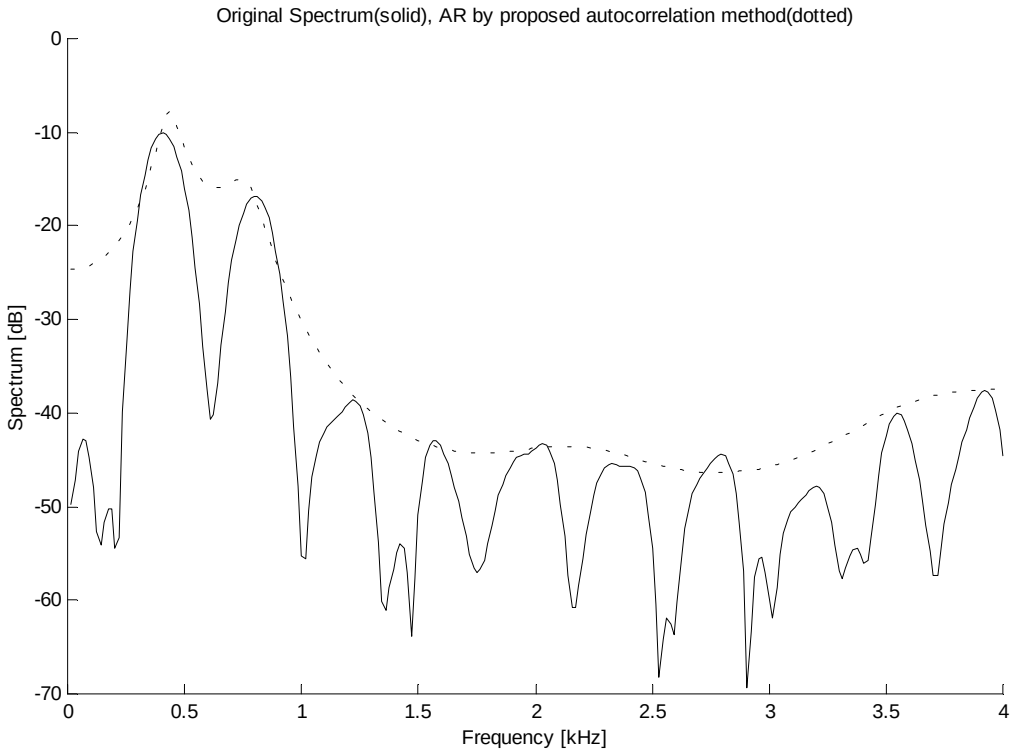

Figure 5.22 Original spectrum and AR magnitude frequency response determined by the autocorrelation method applied to the time-domain signal corresponding to the spectral envelope instead of the original speech signal for a 10-ms frame of a vowel "I" spoken by the childish voice.

Here, the standard method is denoted by AR1; the method with the inverse Fourier transform of the spectral envelope is denoted by AR2. The mean value of the RMS log spectral measure for AR2 is lower by 1.28 dB than that for AR1; the standard deviation of the RMS log spectral measure for AR2 is lower by 0.38 dB than that for AR1.

5.2.2 Cepstral Parametrization of the Harmonic Model

The harmonic model with cepstral parametrization (HCP) uses description similar to (5.19) to code the frequency response of the vocal tract model determining the amplitudes and phases of the composite sine waves. First, the logarithmic speech spectrum described by the real cepstrum $\{c_n\}$ using (5.16) is rewritten using the cosine expansion in the following form

$$\ln\left|S\left(e^{j\omega}\right)\right| = c_0 + 2\cdot\sum_{n=1}^{\infty} c_n \cdot \cos n\omega \,. \tag{5.45}$$

Then, the logarithmic frequency response of the vocal tract model can be given by truncation of the real cepstrum to N_C cepstral coefficients as follows

$$\ln\left|P\left(e^{j\omega}\right)\right| = c_0 + 2 \cdot \sum_{n=1}^{N_C-1} c_n \cdot \cos n\omega \,. \tag{5.46}$$

5.2.2.1 Cepstral Parameters Determination of the HCP

The HCP may use the cepstral parameters determination simply by the truncated cepstrum. The logarithmic spectrum computed from the truncated cepstrum should form the speech spectral envelope and its samples at pitch harmonics should model the spectral peaks. In Figure 5.23 we can see that it has the shape of the logarithmic envelope but it is vertically shifted towards lower amplitude values. Two solutions of this problem are compared in [122]. The first one is the cepstral coefficients determination with gain correction inspired by gain matching in the cepstral speech model [30], however, using different procedures. The method will be described in Section 5.2.2.4. The second solution uses the prior spectral envelope determination and the truncated cepstrum is computed from this spectral envelope instead of the original speech spectrum. The method will be described in Section 5.2.2.5. It was inspired by [39], [40], but different approach to the spectral envelope estimation is used here. The same spectral envelope determination was used in the harmonic speech model with AR parametrization [78] described in Section 5.2.1.4.

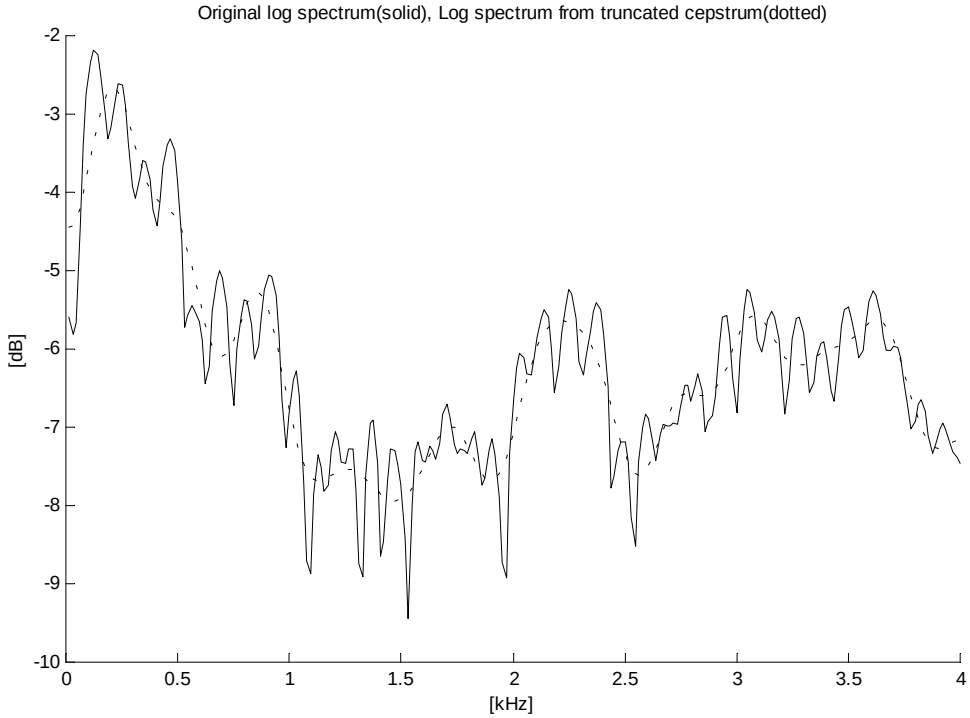

Figure 5.23 Speech spectrum from truncated cepstrum for a 24-ms frame of a vowel "I" spoken by the male voice.

5.2.2.2 Harmonic Parameters Determination of the HCP

If the magnitude frequency response of the vocal tract model is represented by N_C coefficients of the real cepstrum (5.46), then the amplitudes $\{A_m\}$ and phases $\{\varphi_m^{\min}\}$ of the minimum-phase spectrum of the original signal are given by

$$A_m = \exp\left(c_0 + 2\sum_{n=1}^{N_C-1} c_n \cos n\omega_m \right), \tag{5.47}$$

$$\varphi_m^{\min} = -2\sum_{n=1}^{N_C-1} c_n \sin n\omega_m , \tag{5.48}$$

where

$$\omega_m = \frac{2\pi \cdot m}{L}, \tag{5.49}$$

is the normalized angular frequency at the m-th pitch harmonic for

$$1 \le m \le [\, L/2 \,]. \tag{5.50}$$

According to the algorithm (5.25), for each pitch-synchronous speech frame the synthesized speech will be given by

$$s_y(l) = 2\sum_{m=1}^{[L/2]} A_m \cos(2\pi f_m l + \varphi_m), \qquad \text{for } L \text{ odd}, \tag{5.51}$$

$$s_y(l) = 2\sum_{m=1}^{L/2-1} A_m \cos(2\pi f_m l + \varphi_m) + A_{L/2}\cos(2\pi f_{L/2} l + \varphi_{L/2}), \quad \text{for } L \text{ even}, \tag{5.52}$$

where $0 \le l \le L-1$.

5.2.2.3 Number of Parameters for the HCP

As the cepstral speech model, described in Section 5.1.2, uses no preemphasis, the number of cepstral coefficients for HCP may be the same as it has been stated in Section 5.1.2.2, i.e. 26 cepstral coefficients for 8-kHz sampling and 51 cepstral coefficients for 16-kHz sampling. The real number of the HCP parameters is given by the pitch period as stated in the beginning of Section 5.2 and in the algorithm (5.25).

5.2.2.4 Cepstral Parameters Determination with Gain Correction

Gain correction of the vocal tract model transfer function in logarithmic scale means its vertical shift so that it represents the spectral envelope more properly than in Figure 5.23. The block diagram of the cepstral parameters determination using the gain correction (HCP1) is shown in Figure 5.24. The logarithmic spectrum computed using the N_F-point FFT of the speech frame is brought to the cepstral domain giving N_F cepstral coefficients. A truncated cepstrum of first N_C coefficients and a cepstrum of the residual signal are determined from this real cepstrum. The cepstrum of the residual signal is transformed back to the spectrum in logarithmic scale and then to the spectrum in linear scale. Peak picking is used to find all the local maxima of the spectrum of the residual signal. It means that all the frequencies at which the spectral slope changes from positive to negative are chosen. Amplitudes at these frequencies are averaged to get the correction gain. Its logarithm is summed with the first cepstral coefficient to get the modified first cepstral coefficient. The remaining (N_C-1) cepstral coefficients are left unchanged. Number of points of FFT N_F equals 512; number of cepstral coefficients N_C equals 26.

To illustrate more the described method we can see Figure 5.25. The upper figure shows the original logarithmic spectrum and the logarithmic spectrum computed from the truncated cepstrum. In the middle we can see the spectrum of the residual signal computed from the cepstrum of the residual signal. A dotted horizontal line represents the mean value of all the local maxima corresponding to the gain correction. The value of one corresponds to no gain correction representing the logarithmic spectrum from the truncated cepstrum of the upper figure. The lower figure represents the logarithmic spectrum from the truncated cepstrum with the first cepstral coefficient modified according to the gain correction. It really represents the speech spectral envelope better than that in upper figure.

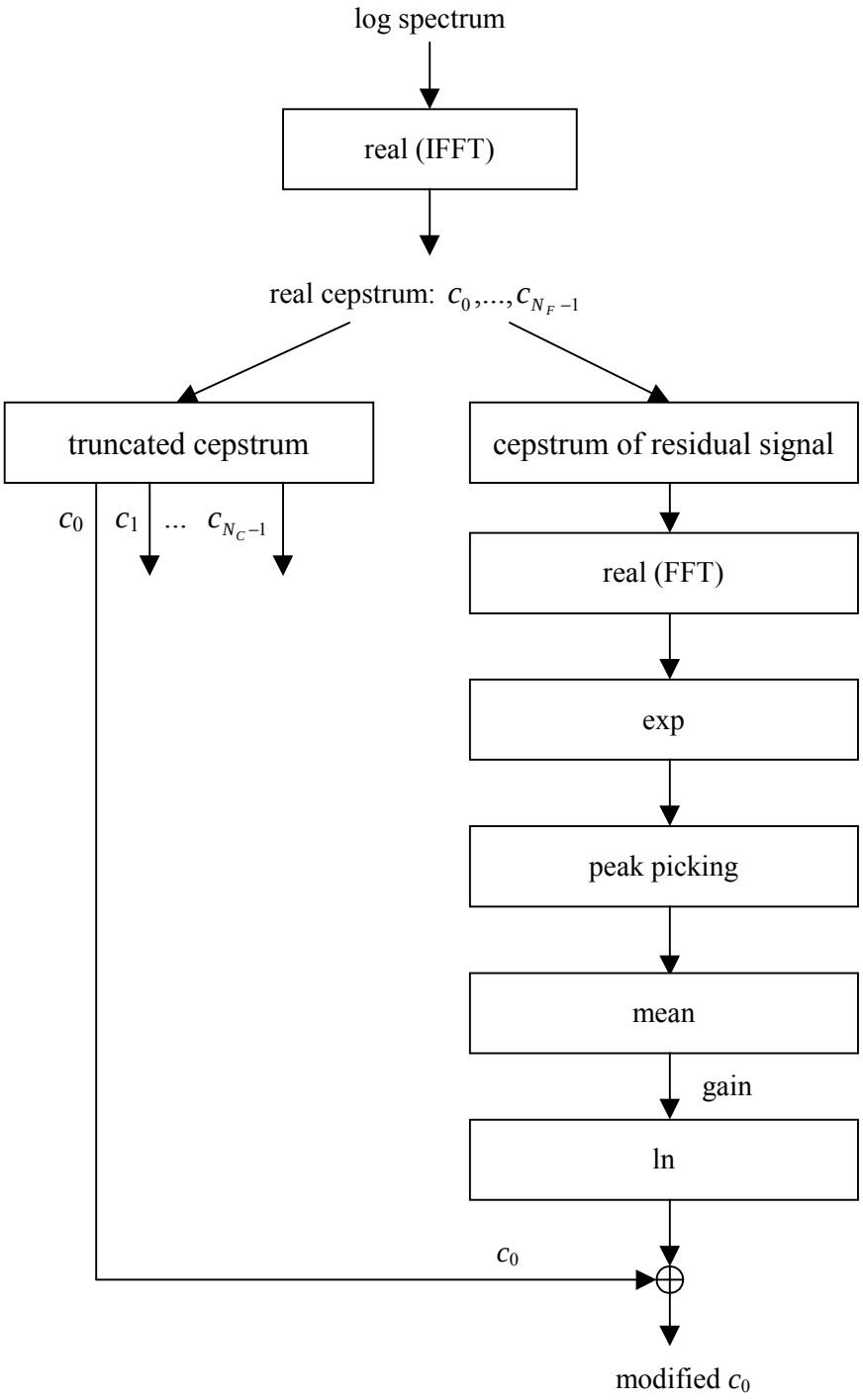

Figure 5.24 Cepstral coefficients determination with gain correction (HCP1).

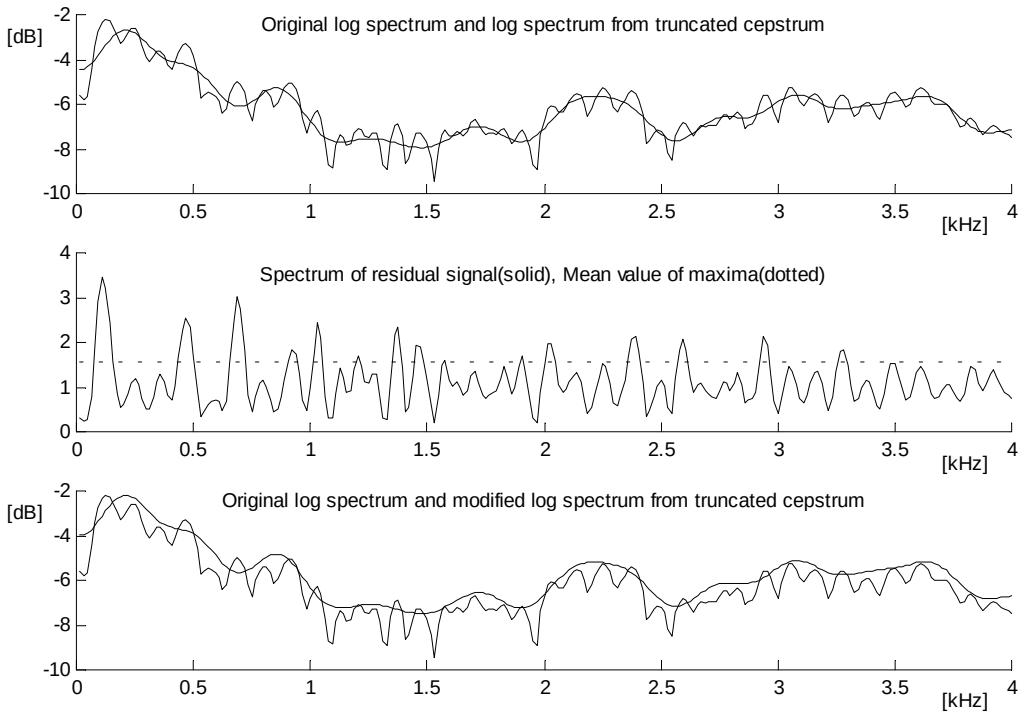

Figure 5.25 Spectrum from truncated cepstrum with gain correction (HCP1).

5.2.2.5 Cepstral Parameters Determination with Prior Spectral Envelope

The block diagram of the cepstral parameters determination using prior spectral envelope (HCP2) is shown in Figure 5.26. Similarly as in the method described in Section 5.2.2.4, the logarithmic spectrum is computed using the N_F-point FFT of the speech frame. Then its staircase envelope is computed and smoothed by Blackman filter in the same way as in Section 5.2.1.4 (see algorithm (5.44) and Figure 5.10). However, here the smoothed spectral envelope is used to compute a real cepstrum of N_F cepstral coefficients and a truncated cepstrum of N_C cepstral coefficients.

In illustration of this method Figure 5.27 depicts the logarithmic spectrum and its staircase envelope (upper figure). Its filtering using the Blackman window gives the smoothed spectral envelope in the middle figure. We can see that it really corresponds to the speech spectral envelope. After representing it by the truncated cepstrum we get the vocal tract model frequency response shown in the lower figure.

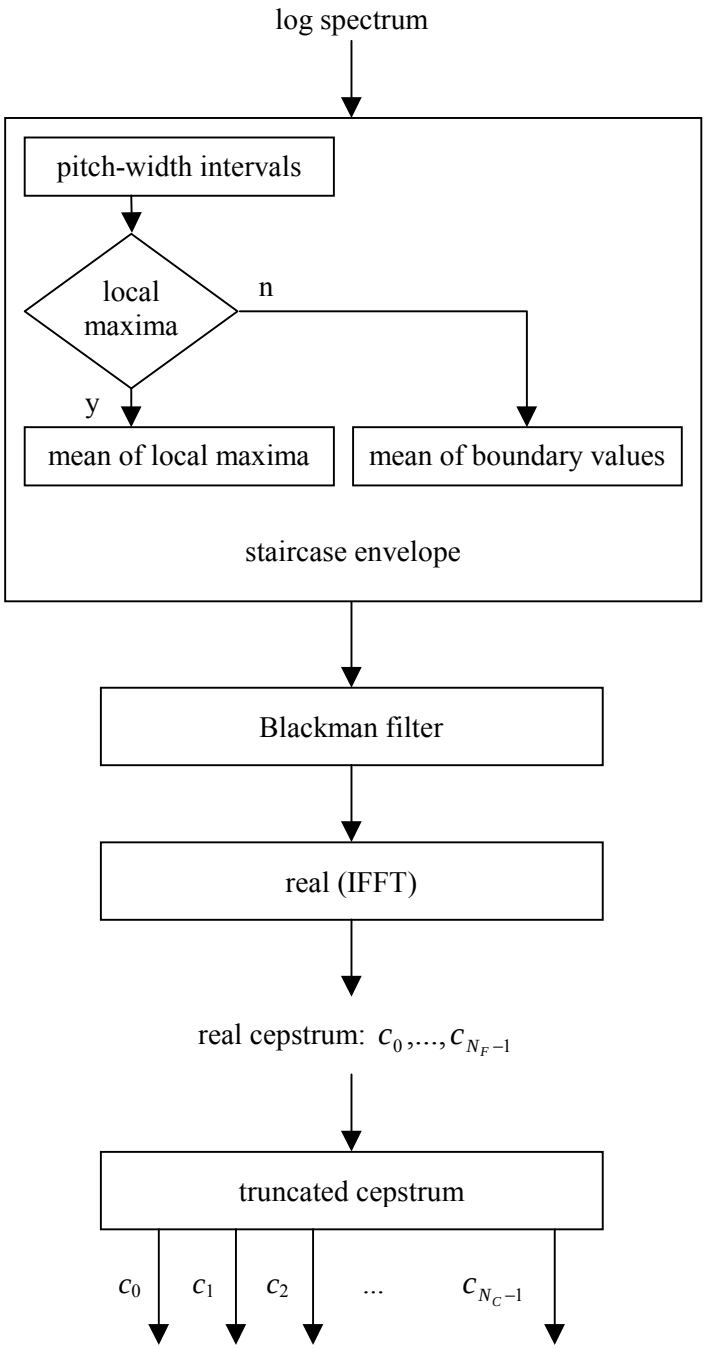

Figure 5.26 Cepstral coefficients determination from log spectral envelope (HCP2).

If we compare the lower figure of Figure 5.25 (HCP1) and the lower figure of Figure 5.27 (HCP2), it is evident that they are fairly similar.

The analysis of one pitch-synchronous frame for both HCP methods can be drawn in Figure 5.28 similarly as HAP in Figure 5.10. The block "cepstral coefficients determination" in Figure 5.28 corresponds either to the algorithm HCP1 (Figure 5.24) or the algorithm HCP2 (Figure 5.26).

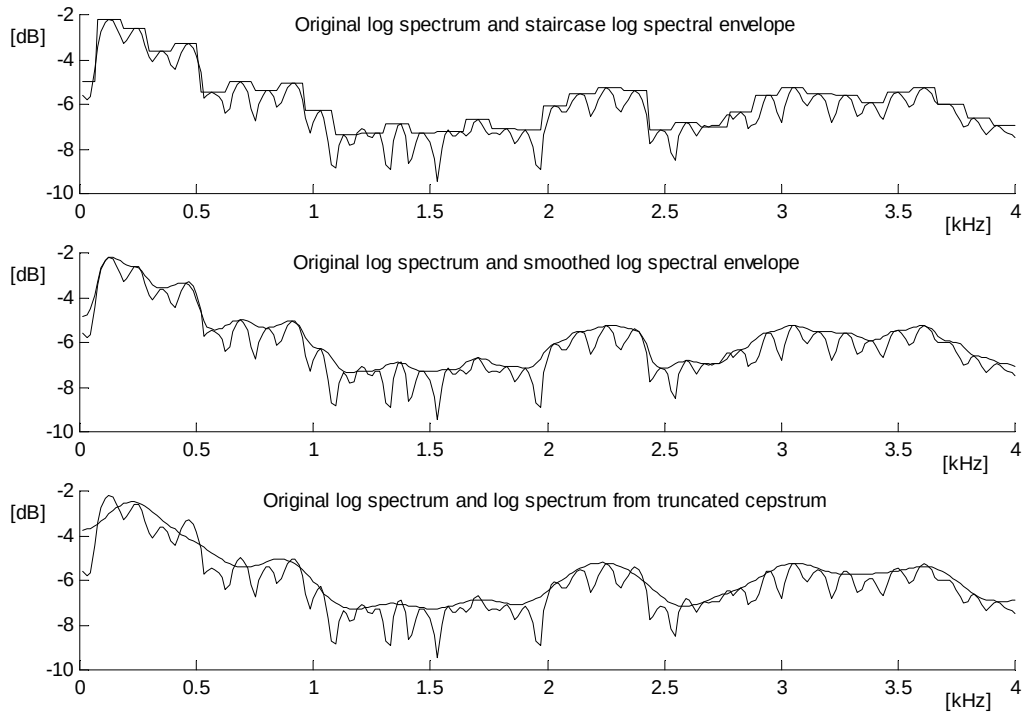

Figure 5.27 Spectrum from truncated cepstrum using prior spectral envelope (HCP2).

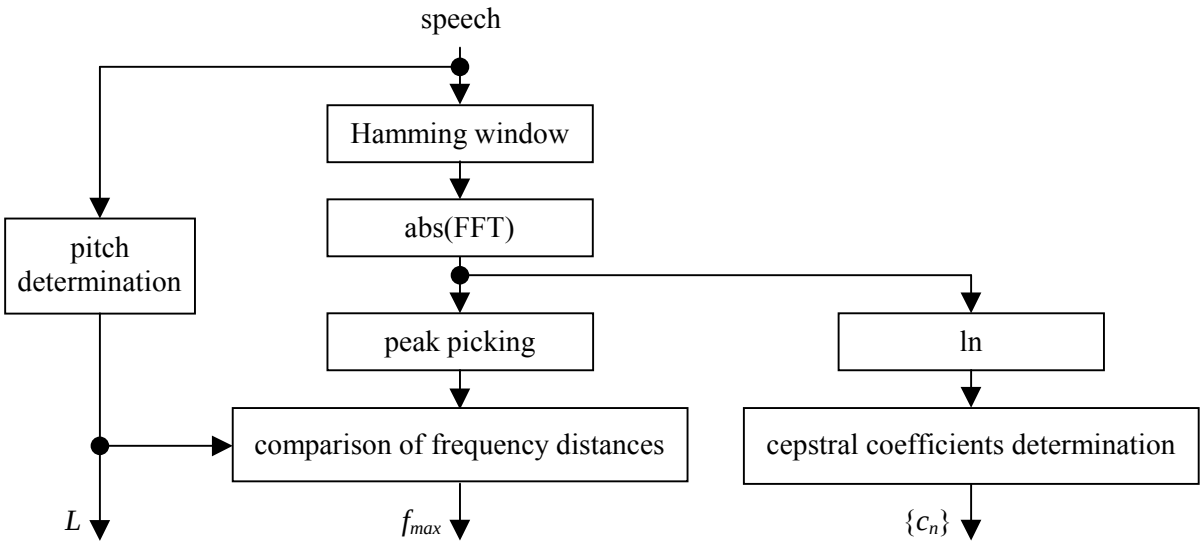

Figure 5.28 Analysis of one equidistant speech frame with determination of the maximum voiced frequency for the HCP model.

5.2.2.6 Speech Synthesis Using the HCP

The block diagram of the synthesis is shown in Figure 5.29. The phases at frequencies lower than f_{max} are computed from the cepstral coefficients using (5.48). Then, the phases at frequencies higher than f_{max} are randomized in the same way as the phases of unvoiced frames. Summing the sine waves with frequencies $\{f_m\}$, amplitudes $\{A_m\}$, and phases $\{\varphi_m\}$ gives the synthetic speech during one pitch-synchronous synthesis frame by (5.51) and (5.52).

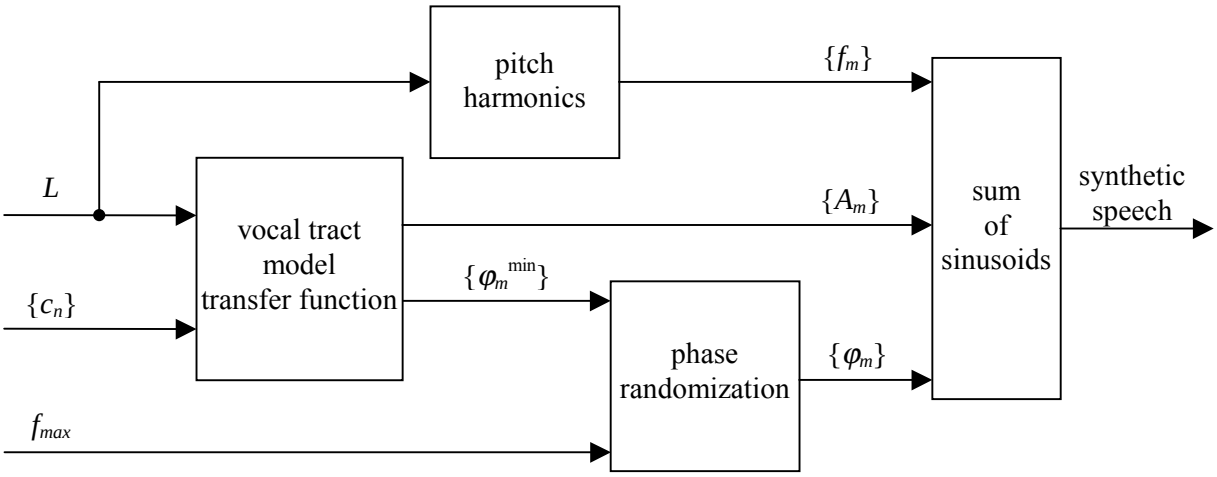

Figure 5.29 Synthesis of one pitch-synchronous speech frame using the HCP model.

5.2.2.7 Quantitative Comparison of Several Approaches to the HCP

Several approaches to the HCP were compared using the same conditions and in a similar way as the HAP in Section 5.2.1.6. Statistical values for the method with gain correction (HCP1) using concatenated synthesis are shown in Table 5.28. Results for the method with prior spectral envelope (HCP2) using concatenated synthesis are in Table 5.29. Results for the same analysis methods with OLA synthesis are given in Tables 5.30 and 5.31. Figure 5.30 shows comparison of all the mentioned combinations (HCP1+concatenation, HCP2+concatenation, HCP1+OLA, and HCP2+OLA) averaged for all the voiced sounds. We can see that the methods give almost the same mean RMS log spectral measure. The HCP2 gives slightly higher mean value then the HCP1 for both the synthesis methods. However, the standard deviation is the highest for the HCP1 with concatenation. The lowest standard deviation is observed in both the analysis methods with OLA synthesis, however, the HCP2 giving the mean value higher by 0.1 dB than the HCP1. It can be concluded that both the methods (the HCP1 of cepstral parameters determination with gain correction described in Section 5.2.2.4, and the HCP2 of cepstral parameters determination with prior spectral envelope described in

Section 5.2.2.5) can be regarded as almost identical in their results if OLA synthesis of pitch-synchronous frames is used.

sound/ number of frames	RMS log spectral measure [dB]			
	minimum	maximum	mean	standard deviation
A/81	0.89	3.68	2.33	0.59
E/60	1.31	4.72	2.34	0.61
I/60	1.28	4.85	2.57	0.71
O/69	1.33	7.04	2.81	1.06
U/60	1.75	8.01	3.44	1.34
M/44	1.72	7.85	2.98	1.03
N/69	1.65	5.60	2.87	0.86
S/10	3.39	6.09	4.81	0.96

Table 5.28 RMS log spectral measure between the original and concatenated synthetic speech for the HCP with 26 cepstral coefficients with gain correction (HCP1).

sound/ number of frames	RMS log spectral measure [dB]			
	minimum	maximum	mean	standard deviation
A/81	1.47	3.90	2.45	0.56
E/60	1.23	4.05	2.48	0.61
I/60	1.22	3.69	2.65	0.60
O/69	1.35	6.48	2.84	0.85
U/60	1.82	6.67	3.39	1.09
M/44	1.78	4.73	2.79	0.73
N/69	1.65	5.24	2.87	0.69
S/10	3.54	5.46	4.47	0.69

Table 5.29 RMS log spectral measure between the original and concatenated synthetic speech for the HCP with 26 cepstral coefficients with prior spectral envelope (HCP2).

sound/ number of frames	RMS log spectral measure [dB]			
	minimum	maximum	mean	standard deviation
A/81	1.17	3.76	2.50	0.48
E/60	1.50	3.51	2.46	0.49
I/60	1.12	4.62	2.56	0.68
O/69	1.55	5.73	2.68	0.66
U/60	1.63	6.19	3.12	0.96
M/44	1.89	4.50	2.75	0.56
N/69	1.82	4.79	2.84	0.61
S/10	3.81	5.39	4.62	0.52

Table 5.30 RMS log spectral measure between the original and OLA synthetic speech for the HCP with 26 cepstral coefficients with gain correction (HCP1).

sound/ number of frames	RMS log spectral measure [dB]			
	minimum	maximum	mean	standard deviation
A/81	1.78	4.07	2.67	0.50
E/60	1.59	3.81	2.58	0.52
I/60	1.56	5.60	2.72	0.74
O/69	1.71	4.54	2.73	0.55
U/60	1.90	5.82	3.21	0.84
M/44	1.59	4.38	2.83	0.59
N/69	1.54	4.79	2.89	0.64
S/10	4.02	6.90	5.23	0.88

Table 5.31 RMS log spectral measure between the original and OLA synthetic speech for the HCP with 26 cepstral coefficients with prior spectral envelope (HCP2).

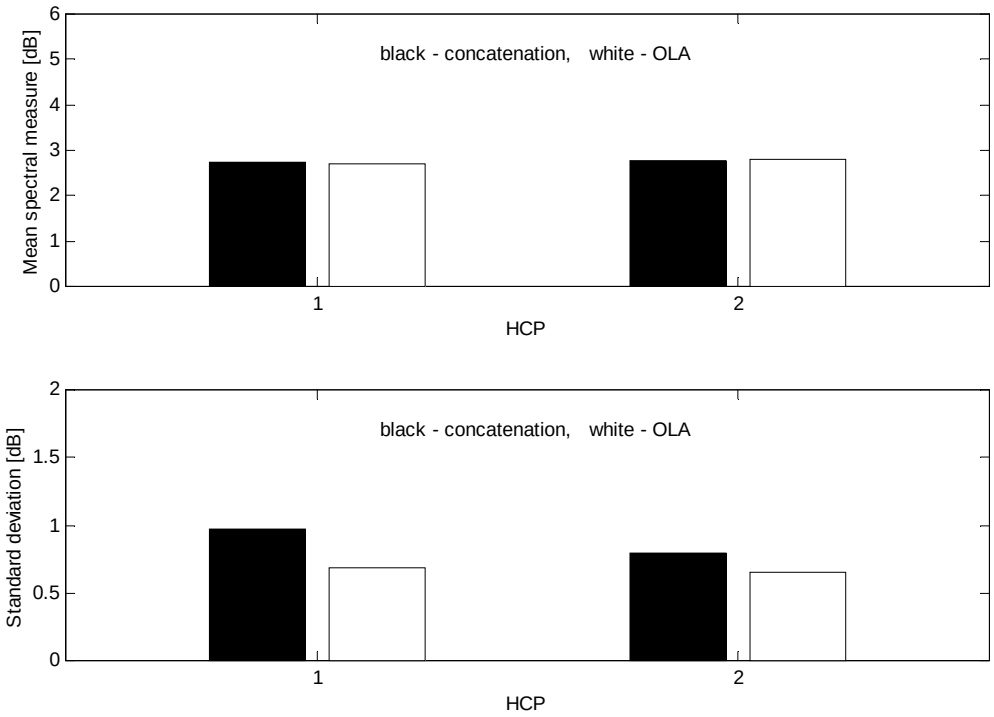

Figure 5.30 RMS log spectral measure between the original and synthetic voiced speech (concatenated and OLA) for the HCP with 26 cepstral coefficients with gain correction (HCP1), and the HCP with prior spectral envelope (HCP2).

Computational complexity for individual blocks of the HCP1 analysis and the concatenated synthesis is shown in Table 5.32. The computational complexity for individual blocks of the HCP2 analysis and the concatenated synthesis is shown in Table 5.33. The overall results for both the analysis methods are in Figure 5.31. In a similar way, the computational complexity with the OLA synthesis is given in Tables 5.34 and 5.35, summarized in Figure 5.32. The computational complexity of the HCP2 analysis is only slightly higher than that of the HCP1.

The synthesis using OLA has about twice the computational complexity of the synthesis using concatenation.

block	corresponding operations	complexity [flops/sample]
A0	segmentation, windowing	2.3
A1	pitch detection	448.3
A2	maximum voiced frequency determination (V/UV)	50.3/0
A3	cepstral parameters determination	523.4
Σ A	total analysis with cepstral parametrization (V/UV)	1024.4/974.1
S1	cepstral to harmonic parameters transformation (V/UV)	129.4/70.3
S2	harmonic synthesis	204.5
Σ S	total harmon. synthesis with cepstral parametrization (V/UV)	334/274.8

Table 5.32 Computational complexity for the HCP with 26 cepstral coefficients with gain correction (HCP1) with concatenated synthesis.

block	corresponding operations	complexity [flops/sample]
A0	segmentation, windowing	2.3
A1	pitch detection	448.3
A2	maximum voiced frequency determination (V/UV)	50.3/0
A3	cepstral parameters determination	523.8
Σ A	total analysis with cepstral parametrization (V/UV)	1024.8/974.5
S1	cepstral to harmonic parameters transformation (V/UV)	129.4/70.3
S2	harmonic synthesis	204.5
Σ S	total harmon. synthesis with cepstral parametrization (V/UV)	334/274.8

Table 5.33 Computational complexity for the HCP with 26 cepstral coefficients with prior spectral envelope (HCP2) with concatenated synthesis.

block	corresponding operations	complexity [flops/sample]
A0	segmentation, windowing	2.3
A1	pitch detection	448.3
A2	maximum voiced frequency determination (V/UV)	50.3/0
A3	cepstral parameters determination	523.4
Σ A	total analysis with cepstral parametrization (V/UV)	1024.4/974.1
S1	cepstral to harmonic parameters transformation (V/UV)	130.5/63.1
S2	harmonic synthesis	540.2
Σ S	total harmon. synthesis with cepstral parametrization (V/UV)	670.7/603.3

Table 5.34 Computational complexity for the HCP with 26 cepstral coefficients with gain correction (HCP1) with OLA synthesis.

block	corresponding operations	complexity [flops/sample]
A0	segmentation, windowing	2.3
A1	pitch detection	448.3
A2	maximum voiced frequency determination (V/UV)	50.3/0
A3	cepstral parameters determination	523.8
Σ A	total analysis with cepstral parametrization (V/UV)	1024.8/974.5
S1	cepstral to harmonic parameters transformation (V/UV)	130.5/63.1
S2	harmonic synthesis	540.2
Σ S	total harmon. synthesis with cepstral parametrization (V/UV)	670.7/603.3

Table 5.35 Computational complexity for the HCP with 26 cepstral coefficients with prior spectral envelope (HCP2) with OLA synthesis.

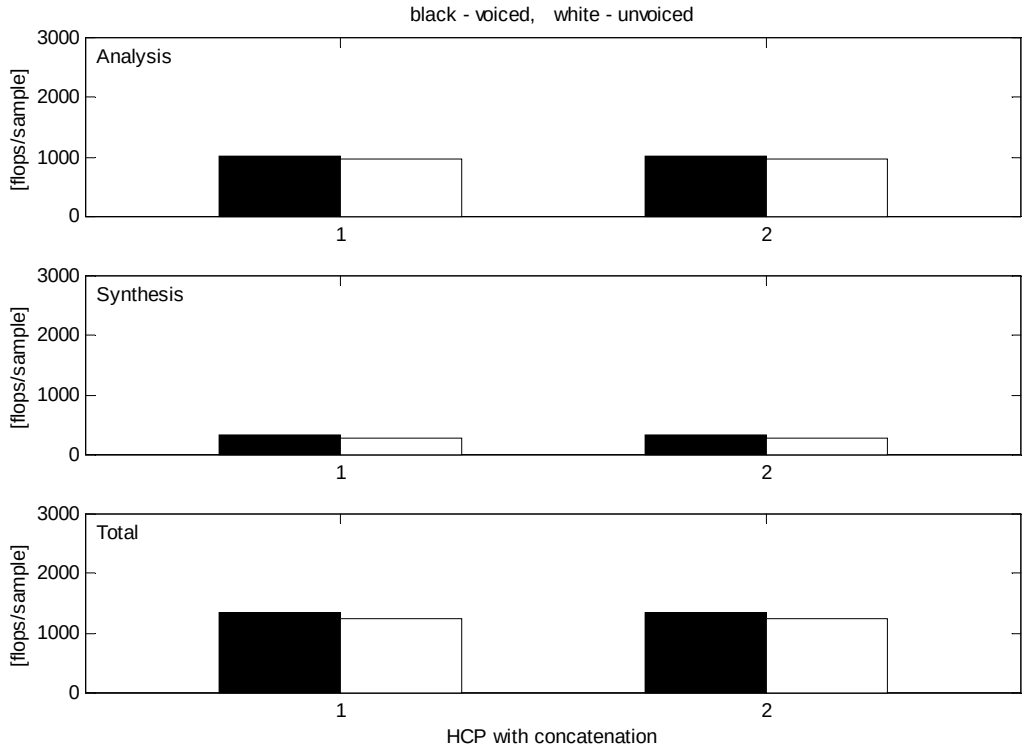

Figure 5.31 Computational complexity for the HCP with 26 cepstral coefficients with gain correction (HCP1), and the HCP with prior spectral envelope (HCP2) using concatenated synthesis.

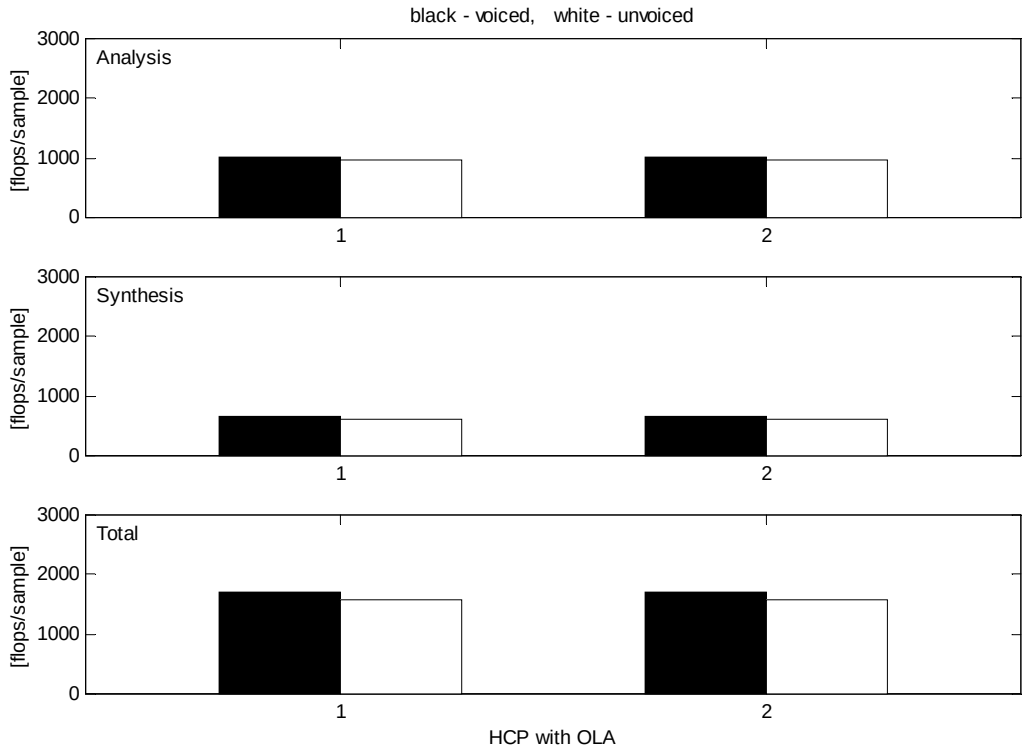

Figure 5.32 Computational complexity for the HCP with 26 cepstral coefficients with gain correction (HCP1), and the HCP with prior spectral envelope (HCP2) using OLA synthesis.

If we compare the computational complexity in Figures 5.31, and 5.32 with the RMS log spectral measure in Figure 5.30, the lowest mean RMS log spectral measure value is given by the HCP1+OLA with the computational complexity slightly lower than the HCP2+OLA. However, if we are interested mainly in low computational cost of synthesis, we may use the HCP with concatenation because of its lowest computational complexity. Although its RMS log spectral measure values are higher, it is still acceptable. In Section 5.2.4 this method will be compared with the source-filter cepstral model.

5.2.3 Comparison of AR and Cepstral Parametrization of the Harmonic Model

Let us compare HAP and HCP both using speech spectral envelope to compute 26 parameters, i.e. HAP2 with the 25^{th} order AR (Figures 5.12 and 5.13) and HCP2 with 26 cepstral coefficients (Figure 5.30). The results are resumed in Figure 5.33. Here we can see that the highest mean RMS log spectral measure as well as the highest standard deviation is given by the HAP2 with concatenation. The lowest mean value is observed for the HAP2 with OLA, while the standard deviation is almost the same for the HAP2 and the HCP2 with OLA. It

means that if the same spectral envelope is used to compute AR or cepstral parameters there is no advantage of HCP against HAP.

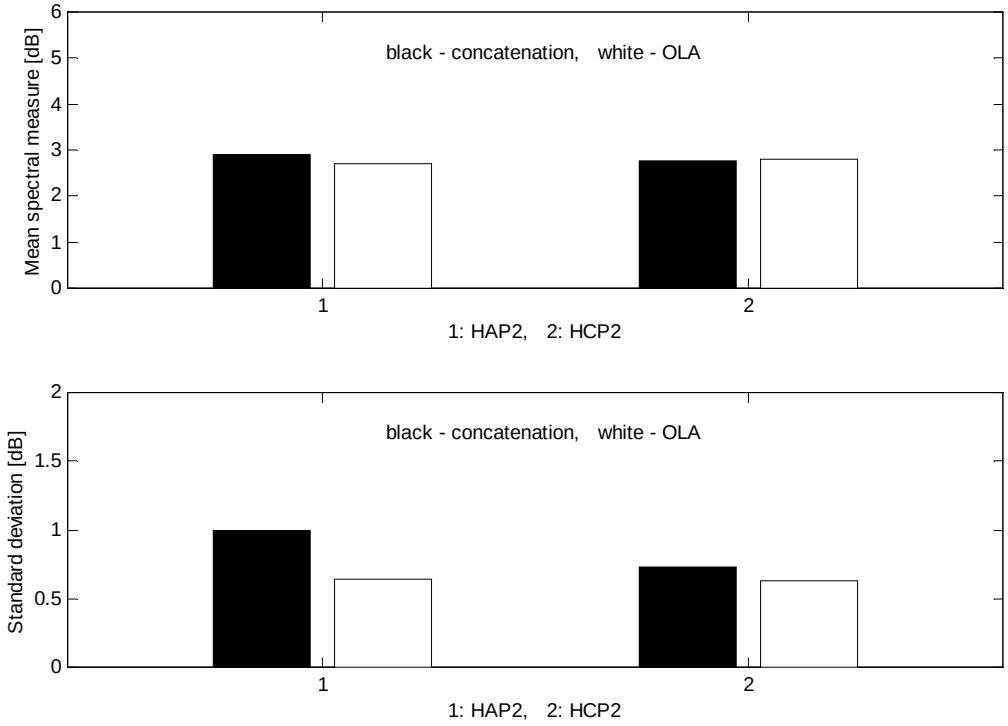

Figure 5.33 RMS log spectral measure between the original and synthetic speech (concatenated and OLA) for the HAP with the 25th order AR computed from the time signal corresponding to the spectral envelope (HAP2), and the HCP with 26 cepstral coefficients computed from the spectral envelope (HCP2).

Comparison of the speech spectra obtained from the AR and cepstral parameters is shown in Figures 5.34 and 5.35. Figure 5.34 shows the spectra computed from 26 parameters by the methods used in the models HAP2 and HCP2. We can see that sampling the spectrum corresponding to the HAP2 approximates the original spectral peaks more properly than sampling the spectrum corresponding to the HCP2. However, increasing the model order gives the spectra of the HAP2 and HCP2 models rather similar. It can be seen in Figure 5.35 for 42 parameters, i.e. 41st order AR model and 42 cepstral parameters.

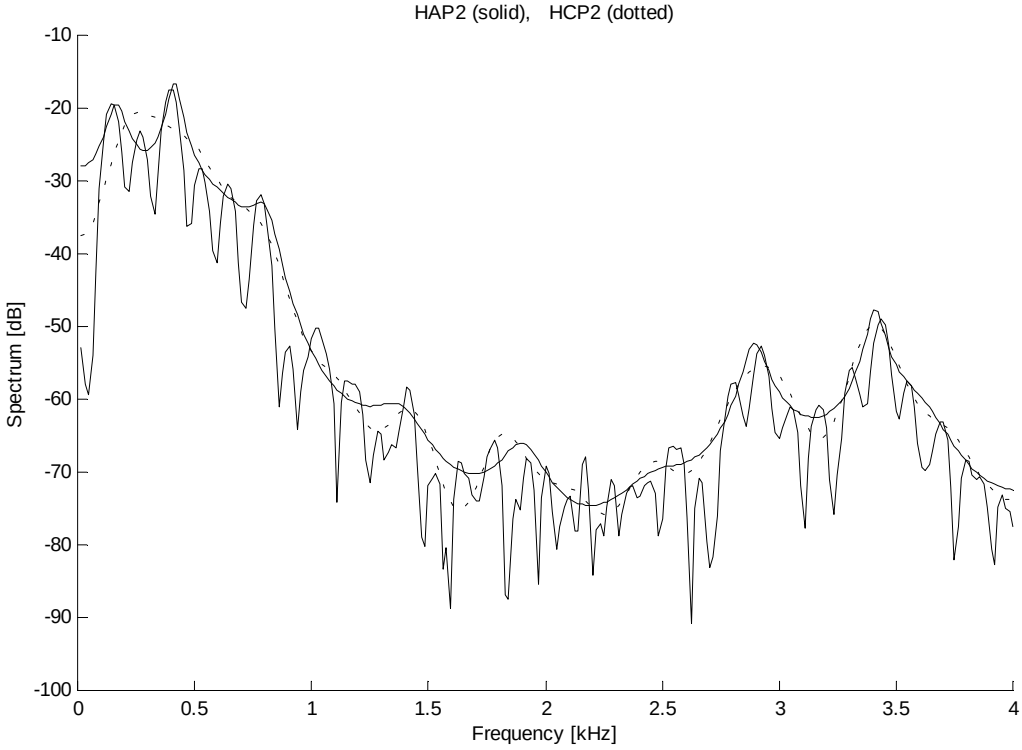

Figure 5.34 Comparison of the spectra obtained from 26 parameters of the AR and cepstral model with prior spectral envelope for a 24-ms frame of a vowel "U" spoken by the male voice.

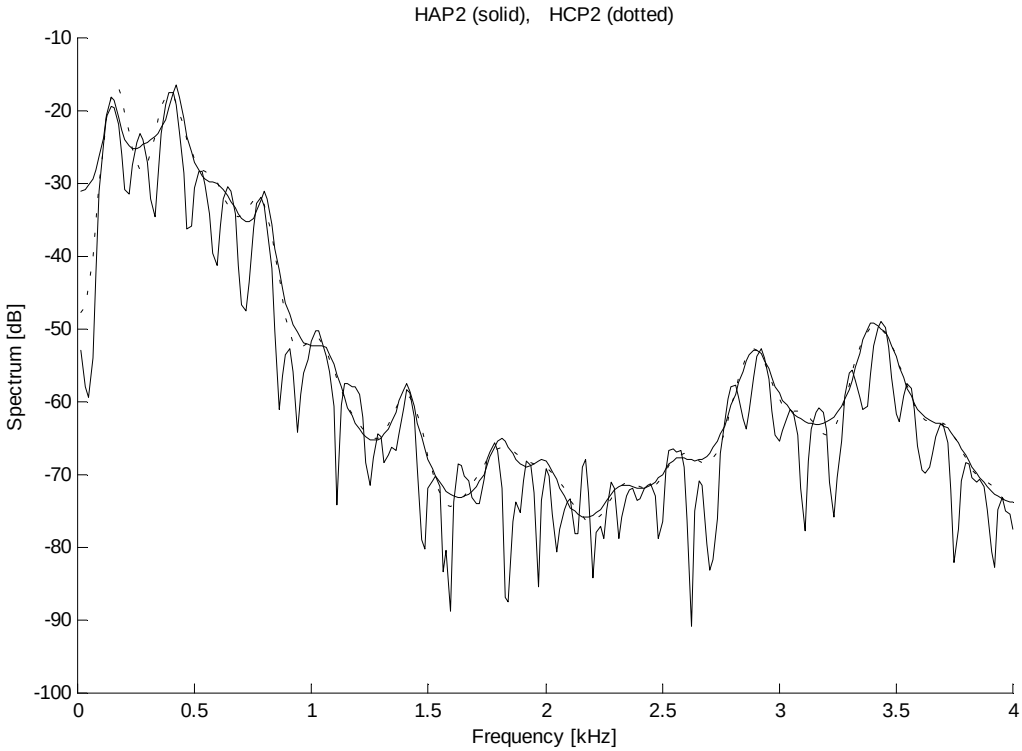

Figure 5.35 Comparison of the spectra obtained from 42 parameters of the AR and cepstral model with prior spectral envelope for a 24-ms frame of a vowel "U" spoken by the male voice.

Listening tests were performed in order to compare four combinations of the methods: HAP2+concatenation, HAP2+OLA, HCP2+concatenation, HCP2+OLA. Twenty words synthesized by these methods were grouped into pairs of the same word synthesized by two methods. The words of the pairs were grouped in the random order and listeners had to choose the word with better resemblance to the original. Scores given by the listeners are summarized in Table 5.36.

listener	HAP2	HCP2	concatenation	OLA
	concatenation vs. OLA	concatenation vs. OLA	HAP2 vs. HCP2	HAP2 vs. HCP2
AM1	2 : 18	1.5 : 18.5	6 : 14	9 : 11
AM2	2 : 18	3.5 : 16.5	2.5 : 17.5	10 : 10
AM3	5 : 15	9 : 11	5.5 : 14.5	13 : 7
JP1	9 : 11	10 : 10	9.5 : 10.5	12.5 : 7.5
JP2	6 : 14	11.5 : 8.5	10.5 : 9.5	14.5 : 5.5
JP3	8 : 12	9.5 : 10.5	9.5 : 10.5	12 : 8
JPu	12 : 8	9.5 : 10.5	11 : 9	11.5 : 8.5
PK	6.5 : 13.5	10.5 : 9.5	6.5 : 13.5	9 : 11
mean	6.07 : 13.93	7.93 : 12.07	7.36 : 12.64	11.36 : 8.64

Table 5.36 Preferences for the synthesis methods according to the listening tests of 20 words.

The listening tests were also evaluated for every word through all the listeners, and for every listener through all the words. Results are shown in Table 5.37. Here, new abbreviations were used:

AC = HAP2 + concatenation
AO = HAP2 + OLA
CC = HCP2 + concatenation
CO = HCP2 + OLA

The word “all” means that the listener regarded all four pairs of the same word as having the same quality. The question mark “?” means that there was some disagreement in the listener’s evaluation, e.g. AO and CO were regarded of the same quality, AC and CC were regarded of the same quality, AO was regarded as better than AC, but CC was regarded as better than CO. Table shows that the highest score is given to the HAP2 method with OLA. It is in accordance with the results of Figure 5.33.

word	the best method according to the listener								highest score
	AM1	AM2	AM3	JP1	JP2	JP3	JPu	PK	
aféra	AO	AO,CO	?	AC	AO	CO	AC	AC	AC,AO
fórum	CO	CC,CO	CC,CO	CC,CO	AO	AO	CC	CC	CC
gramo	CO	CO	AO,CO	AO	AO	AO	AO,CO	AO,CO	AO
kobra	AO,CO	AO,CO	?	?	?	CC	all	?	AO,CO
liga	AO	?	?	CO	CO	CO	?	?	CO
mařka	AO	AO,CO	AO	AC	CC	CC	AC	AC	AC,AO
naďa	CO	AO	AO,CO	AO	AO	AO	AC	AO	AO
náradiu	CO	AO,CO	?	?	?	?	?	?	CO
neguj	CO	CO	CC	AC,AO	?	?	AC,CO	CC,CO	CO
niekto	CO	AO,CO	?	AC,CC	AC	AC	AO,CO,CC	?	AC,CO
obor	AO	AO	AO	?	AO	AO	?	?	AO
očko	AO	AO	AO,CC	CC	AC	AC	?	?	AO
ovca	?	AO	AO	?	CC	CC	?	?	AO,CC
pauza	AO,CO	AO,CO,CC	AO,CO	CO	AO	AO	CC	CC	AO
racek	?	?	?	AC	AC	AC	?	AC.CC	AC
šifra	AC	CC	AC, CC	AO	AO	AO	?	AO	AO
spev	CO	AO	AO	AO,CC	CC	CC	all	AO,CO	AO
stĺp	AO,CO	CO	AO,CO	AC,CC	AC	AC	?	?	AC,CO
ufo	AO,CO	AO,CO	AO	AO	AC,AO	?	?	AO,CO	AO
ulietať	AO,CO	AO,CO	AO,CO,CC	CO	CO	CO	?	AO,CO,CC	CO
highest score	CO	AO	AO	AC,AO	AO	AO	AC	AO	AO

Table 5.37 Evaluation of the listening tests for every word through all the listeners, and for every listener through all the words.

The computational complexity of the same methods (HAP2, and HCP2) using 26 parameters (Figures 5.16, 5.17, 5.31, and 5.32) are resumed in Figures 5.36 and 5.37. The total computational complexity of the HAP2 is about twice that of the HCP2 for both the synthesis methods (concatenation, OLA). Although the HAP2 with OLA gives even slightly lower mean RMS log spectral measure than the HCP2 with OLA, it is not very useful because of its high computational complexity.

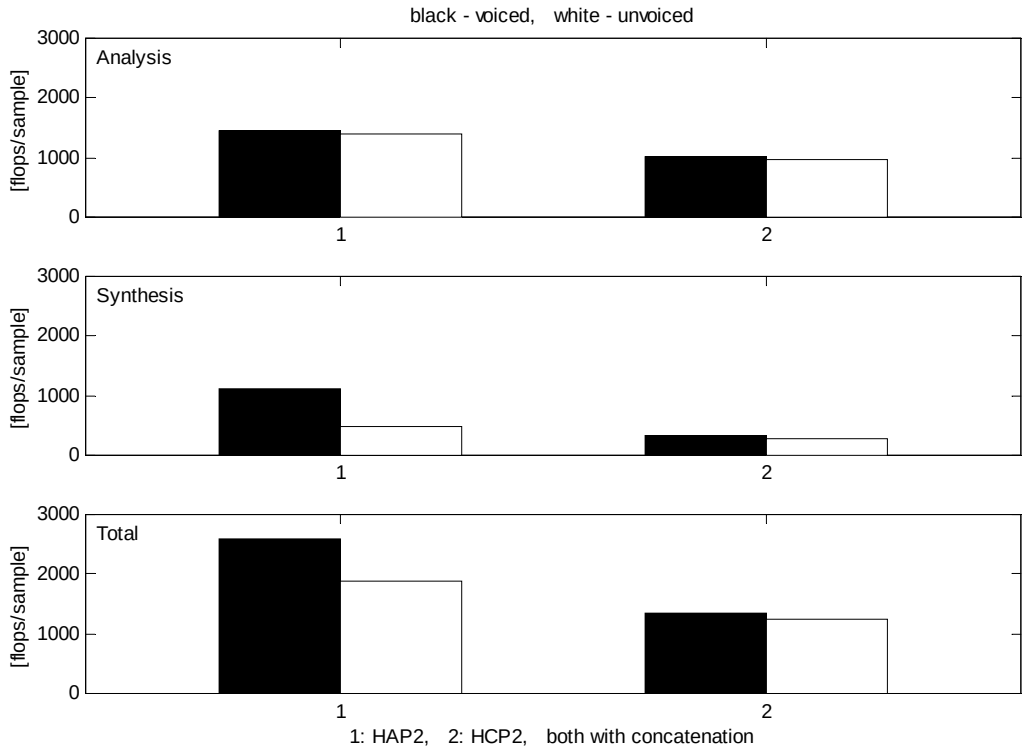

Figure 5.36 Computational complexity for the HAP with the 25th order AR computed from the time signal corresponding to the spectral envelope (HAP2), and the HCP with 26 cepstral coefficients computed from the spectral envelope (HCP2) using concatenated synthesis.

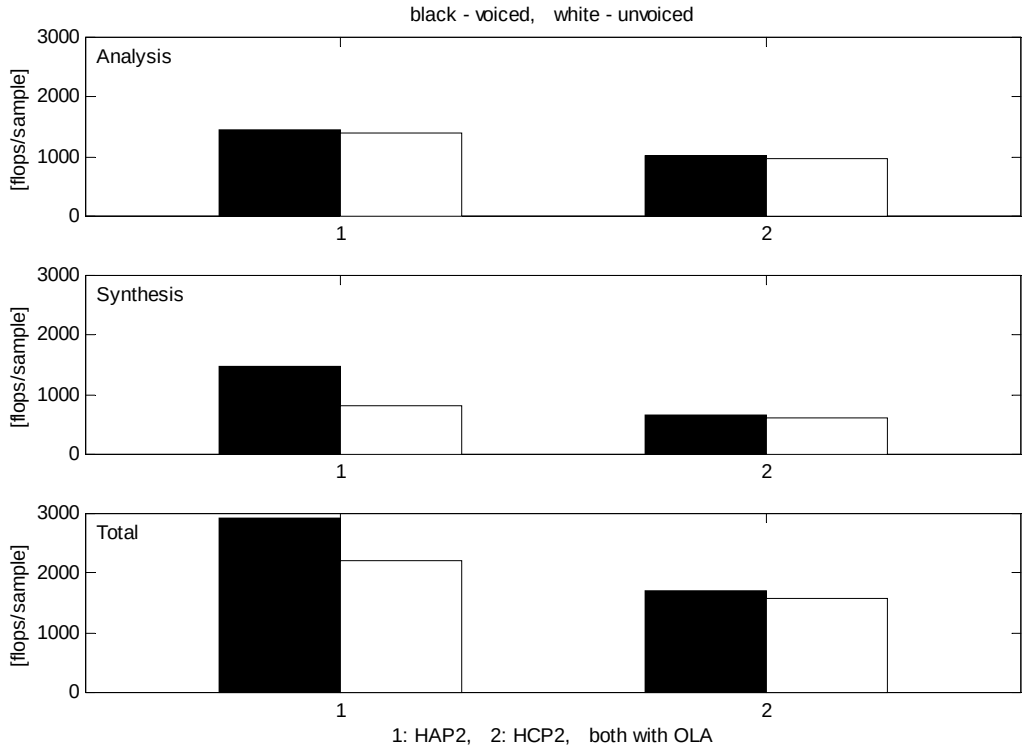

Figure 5.37 Computational complexity for the HAP with the 25th order AR computed from the time signal corresponding to the spectral envelope (HAP2), and the HCP with 26 cepstral coefficients computed from the spectral envelope (HCP2) using OLA synthesis.

5.2.4 Comparison of the Cepstral Model and the HCP

Here two approaches to speech modelling based on cepstral description will be compared: the cepstral model of speech synthesis described in Section 5.1.2, and the harmonic model with cepstral parametrization (HCP) described in Section 5.2.2. For the cepstral model the original speech spectrum is used to determine cepstral coefficients, and synthesis is performed by concatenation of pitch-synchronous frames. In order to make comparison with the same conditions the HCP1 method with concatenation is used here as a counterpart of the cepstral model. Both the compared models use 26 cepstral coefficients for 8-kHz sampling and 51 cepstral coefficients for 16-kHz sampling. Although the number of the cepstral coefficients is the same for both the methods, their values are different as a consequence of different normalization of the weighting window. In the cepstral model the Hamming window is normalized using (5.23), while in the harmonic model it is normalized using (5.30). In Figure 5.38 [108] we can see how to obtain parameters necessary for speech synthesis according to both the models. Apart from the cepstral coefficients, it is a pitch period L determined using a clipped autocorrelation function, and a parameter determining an extent of emphasizing noise at higher frequencies for voiced speech. For the cepstral model, it is a spectral flatness measure S_F (see (5.24) in Section 5.1.2.1), for the harmonic model, it is a maximum voiced frequency f_{max} (see Figure 5.28 and description at the end of Section 5.2.1.4). Output parameters of the cepstral analysis block are used as the input parameters of the synthesis blocks for both the models.

The cepstral synthesis is performed according to Figure 5.5 in Section 5.1.2. For voiced speech the excitation pulses with the shape of the impulse response of the Hilbert transformer are generated in the intervals of the pitch period. To preserve the same phase relations in the HCP the minimum phases $\{\varphi_m^{\min}\}$ are modified using all-pass phase correction. The phase response of the same Hilbert transformer is sampled at pitch harmonics and these phases are superimposed to the minimum phases. In this way Figure 5.29 is redrawn and the resulting block diagram of the HCP with all-pass phase modification is shown in Figure 5.39.

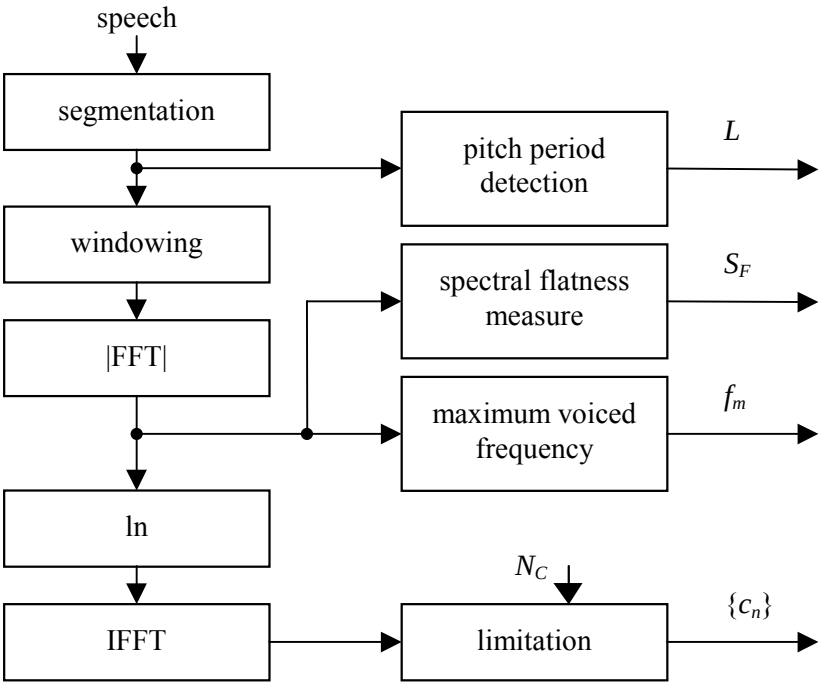

Figure 5.38 Block diagram of the cepstral analysis.

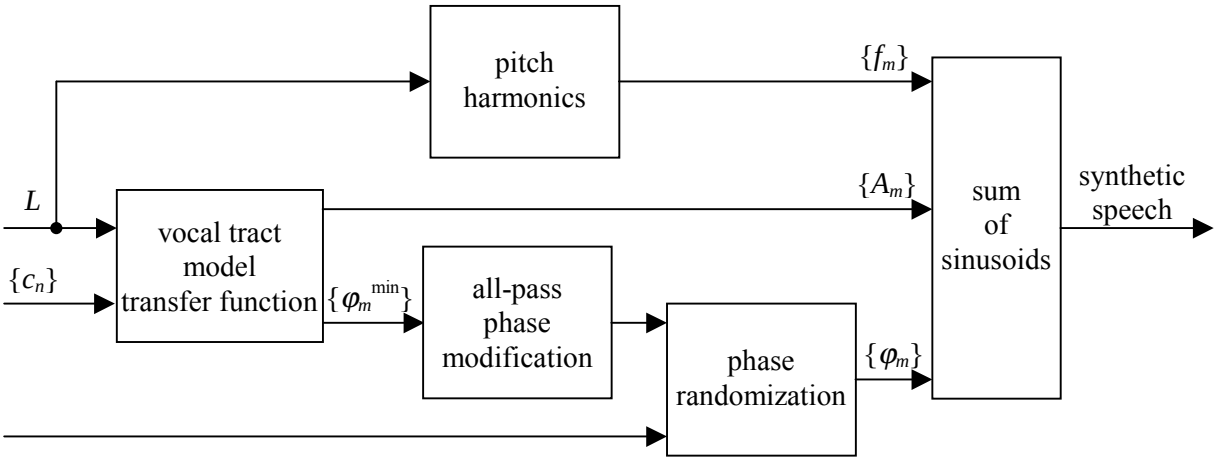

Figure 5.39 Block diagram of the HCP for one pitch-synchronous synthesis frame using excitation phase of the Hilbert transformer response.

The RMS log spectral measure was used to compare the smoothed spectra of original and resynthesized speech. The speech material consisted of the stationary parts of 5 vowels, 2 nasals, and 1 fricative spoken by the same male voice as in previous sections. Comparison was made for sampling frequencies of 8 kHz and 16 kHz. In upper part of Figure 5.40 we can see original and resynthesized signals for a 24-ms frame of the vowel "A" sampled at 16 kHz; wide line represents the original signal, solid line represents the cepstral resynthesis, dotted

line represents the HCP. The logarithmic speech spectra of these three signals are drawn in lower part of Figure 5.40. Table 5.38 shows mean values of the RMS log spectral measure for several frames of speech sounds. It is evident that for every speech sound and for both sampling frequencies, the RMS log spectral measure is higher for the cepstral synthesis than for the HCP.

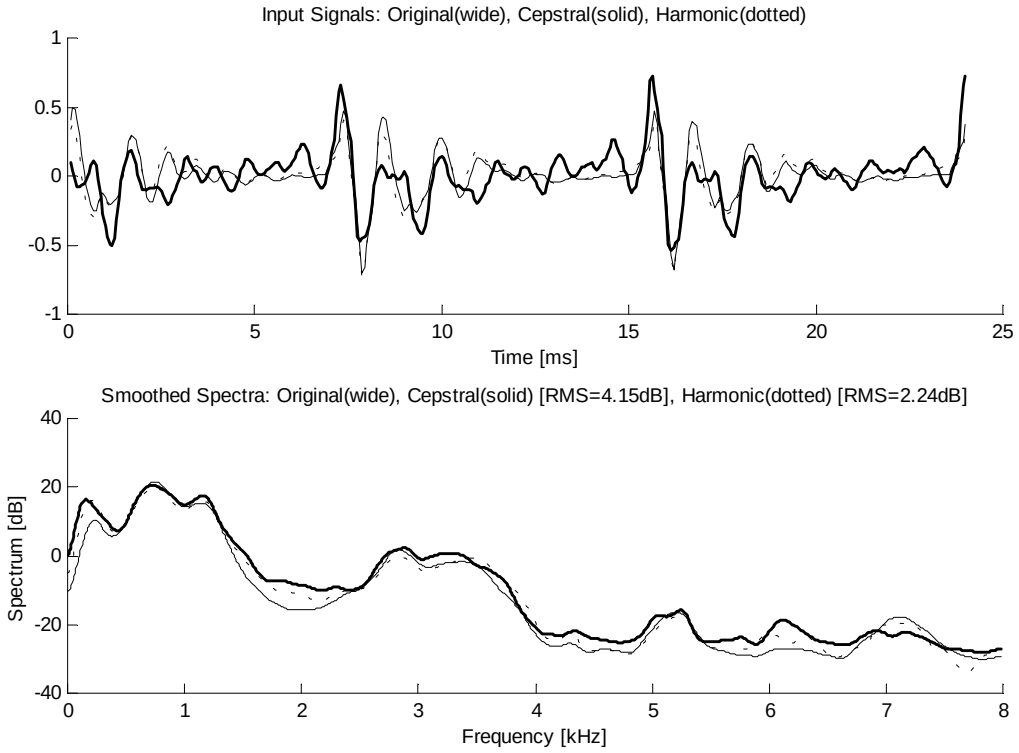

Figure 5.40 Input signal, smoothed spectra and RMS values for a 24-ms frame of a vowel ”A” spoken by the male voice.

The comparison of computational complexity was performed in the blocks shown in Figure 5.41 [107]:

a) cepstral speech analysis:
   - frame classification (voiced/unvoiced) and pitch-period detection
   - computing of coefficients for the cepstral as well as the harmonic model

b) speech synthesis :
   - cepstral synthesis of (voiced/unvoiced) frames
   - computing of parameters of the harmonic model from the cepstral coefficients
   - harmonic synthesis of (voiced/unvoiced) frames.

sound/ number of frames	mean RMS log spectral measure [dB]			
	f_s = 8 kHz		f_s = 16 kHz	
	cepstral synthesis	harmonic synthesis	cepstral synthesis	harmonic synthesis
A/4	3.65	2.49	4.74	2.05
E/4	3.91	2.27	4.44	2.20
I/4	4.51	3.02	4.44	3.18
O/4	4.44	2.78	4.41	3.01
U/4	3.54	2.96	4.00	2.30
M/8	5.39	3.38	6.57	4.03
N/9	4.08	2.64	4.43	2.45
S/9	4.03	3.91	4.16	3.97

Table 5.38 The mean RMS log spectral measure for the cepstral and HCP synthesis.

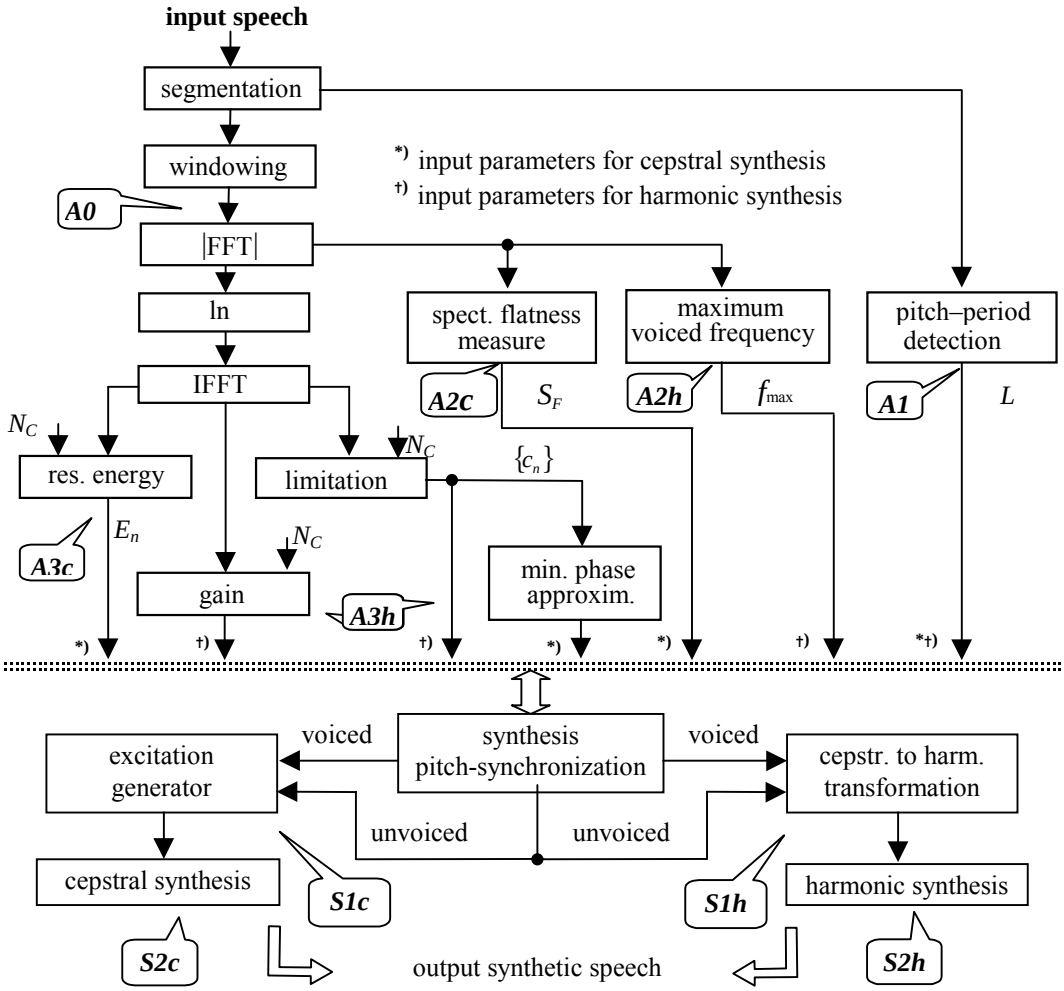

Figure 5.41 Block diagram for computational complexity of the cepstral model and the HCP.

Computational complexity as well as memory requirements were compared. The memory requirements are important for practical implementation in another programming languages (Assembler or C language for digital signal processors). The computational complexity has influence especially on real time applications (speech coders and decoders, or TTS systems). The input parameters for the cepstral and harmonic synthesis at 8 kHz and 16 kHz sampling frequencies are summarized in Table 5.39. Here the first cepstral coefficient c_0 is comprised in *En* or *G*. However, it can be realized in opposite way: *En* or *G* comprised in c_0, as was the case of the HCP in Section 5.2.2.4.

type of synthesis	input parameters necessary for one synthesis frame *)			
	f_s = 8 kHz	Σ	f_s = 16 kHz	Σ
cepstral	$1\times En, 25\times\{\hat{s}_n\}, 1\times S_F, 1\times L$	28	$1\times En, 50\times\{\hat{s}_n\}, 1\times S_F, 1\times L$	53
HCP	$1\times G, 25\times\{c_n\}, 1\times f_{\max}, 1\times L$	28	$1\times G, 50\times\{c_n\}, 1\times f_{\max}, 1\times L$	53

*) All the data are considered in the standard format Integer with the length of 2 bytes.

Table 5.39 Analysis to synthesis data transfer vector storage requirements for the cepstral model and the HCP.

block	corresponding operations	complexity [flops/sample]	
		f_s = 8 kHz	f_s = 16 kHz
A0	segmentation, windowing for cepstral model	34.1	34.1
A1	pitch-detection (V/UV)	452.3/452.6	477/477.1
A2c	spectral flatness measure	8.1	8
A2h	maximum voiced frequency (V/UV)	40.2/0	78.6/0
A3c	parameters determination for cepstral model	533.5	570.5
A3h	parameters determination for HCP	526.2	563
Σ Ac	analysis for cepstral model (V/UV)	1028/1028.2	1089.6/1089.7
Σ Ah	total analysis for harmonic model (V/UV)	1056.8/1016.9	1156.8/1078
S1c	excitation for cepstral model (V/UV)	8.9/3.4	8.5/3.2
S1h	ceps. to harm. parameters transformation (V/UV)	297.8/239.8	443.5/335.4
S2c	cepstral synthesis	320	518
S2h	harmonic synthesis (V/UV)	238.1/280.1	462.1/560.1
Σ Sc	total cepstral synthesis (V/UV)	328.9/323.4	526.5/521.2
Σ Sh	total harmonic synthesis (V/UV)	535.9/519.8	905.5/895.5

Table 5.40 Computational complexity at the points corresponding to Figure 5.41.

Quantitative comparison of computational complexity is shown in Table 5.40. Here, computational complexity was computed only for one speech frame and related to one sample. For that reason the block A0 has rather high computational complexity also for the HCP where the normalized Hamming window is computed only once per processed signal. Computational complexity of the block A2h was slightly lowered by a simplified assumption that the local maxima of the spectrum and the local maxima of the residual spectrum are the same what is not true in general. Computation of the phase response of the Hilbert transformer in the HCP increases the complexity of the block S1h. Results in Tables 5.39 and 5.40 show that the storage requirements of both the methods are identical, the computational complexity of the analysis is similar for both the methods, and the computational complexity of the harmonic synthesis is about 1.5-times higher than that of the cepstral synthesis.

At present, further improvements and lowering computational complexity of both the methods are having been worked at.

6 Conclusion

The thesis deals with evaluation of some parametric methods of speech processing with the emphasis on the sinusoidal model with harmonically related component sine waves. It presents improvements of known methods and new algorithms in order to achieve higher synthetic speech quality. Perhaps the major contribution of this work is a novel algorithm for speech spectrum envelope determination and its implementation in harmonic modelling with AR parametrization. Besides, a number of other contributions to different stages of speech analysis and synthesis using the harmonic model with AR as well as cepstral parametrization can be found in the thesis. The proposed methods are compared with respect to the spectral measure, the perceived synthetic speech quality, and the computational complexity.

6.1 Contributions of the Thesis

Almost in all the parts of Chapters 4 and 5, some original contribution can be found, except for Sections 5.1.1.1 a 5.1.2.

The most important scientific contributions of this thesis may be summarized as follows:

1. New algorithm of speech spectrum envelope determination using the staircase envelope and considering the spectral behaviour for voiced as well as unvoiced speech frames.

   The proposed method of spectral envelope determination can be used for AR as well as cepstral parametrization. This method outperforms other methods as it is shown to yield higher objective and subjective performance.

2. Use of the method determining AR parameters from the time-domain signal corresponding to the spectral envelope instead of the original speech signal.

   It has been shown that AR parametrization gives slightly better spectral measure than cepstral parametrization with the same number of parameters, however, at the expense of higher computational complexity. Results of the listening tests have justified use of the RMS log spectral measure for determining perceptual similarity of two speech signals.

3. Use of asymmetric Hanning window during synthesis with overlap-and-adding (OLA) pairs of consecutive speech frames.

Comparison of OLA method and simple pitch-synchronous frame concatenation has shown a considerable decrease in standard deviation of the spectral measure for all the analysis methods.

4. Original contributions are also in derivation of relations between pitch period precision in samples and pitch frequency precision in points of FFT, modification of pitch-synchronization method, use of information theoretic criteria for AR model order selection, AR model order determination without use of preemphasis, derivation of number of parameters for the harmonic model and their amplitudes depending on pitch, experiment with childish voice analysis and synthesis demonstrating considerable improvement of synthesis using the proposed method of spectral envelope determination. Contribution is also an approach to gain correction in cepstral parameters determination directly from the spectrum of original speech signal, and comparison of the harmonic model with the cepstral model, giving benefit of the harmonic model.

6.2 Future Research Directions

Although many efforts have been made to improve the speech analysis and synthesis methods in this work as well as in many other authors, there is still room for further improvement and enhancement. All the methods described in this work may be tested also for 16-kHz sampling. In all the models perceptual frequency scale can be used. Pitch and time-scale modification may be applied to a speech signal with further prosody transplantation. Voice conversion, i.e. modification of a speech signal of one speaker so that it sounds as if spoken by a different speaker, is also an interesting topic for further investigation. Interesting could also be comparison of OLA synthesis with asymmetric triangular or trapezoidal window instead of asymmetric Hanning window used in this work.

References

[1] Juang, B., H., Ed.: The Past, Present, and Future of Speech Processing. In: IEEE Signal Processing Magazine, vol. 15, no. 3., pp. 24-48, May 1998.

[2] Gersho, A.: Advances in Speech and Audio Compression. In: Proceedings of the IEEE, vol. 82, no. 6, pp. 900-918, June 1994.

[3] Vercoe B., L., Gardner, W., G., Scheirer, E., D.: Structured Audio: Creation, Transmission, and Rendering of Parametric Sound Representations. In: Proceedings of the IEEE, vol. 86, no. 5, pp. 922-939, May 1998.

[4] Spanias, A., S.: Speech Coding: A Tutorial Review. In: Proceedings of the IEEE, vol. 82, no. 10, pp. 1541-1582, October 1994.

[5] Kleijn, W., B., Paliwal, K., K.: An Introduction to Speech Coding. In: Kleijn, W. B., Paliwal, K., K., Eds.: Speech Coding and Synthesis, pp. 1-47, Elsevier Science, Amsterdam, 1995.

[6] Stevens, K.: Models of Speech Production. In: Crocker, M., J., Ed., Encyclopedia of Acoustics, pp. 1565-1578, John Wiley & Sons, Inc., 1997.

[7] Fant, G.: Acoustical Analysis of Speech. In: Crocker, M., J., Ed., Encyclopedia of Acoustics, pp. 1589-1598, John Wiley & Sons, Inc., 1997.

[8] Deller, J., R., Proakis, J., G., Hansen, J., H., L.: Discrete-Time Processing of Speech Signals, Macmillan Publishing Company, 1993.

[9] Makhoul, J.: Linear Prediction: A Tutorial Review. In: Proceedings of the IEEE, vol. 63, no. 4, pp. 561-580, April 1975.

[10] Kay, S., M., Marple, S., L.: Spectrum Analysis - A Modern Perspective. In: Proceedings of the IEEE, vol. 69, no. 11, pp. 1380-1419, November 1981.

[11] Markel, J., D., Gray, A., H.: A Linear Prediction Vocoder Simulation Based upon the Autocorrelation Method. In: IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-22, no. 2, pp. 124-134, April 1974.

[12] Burg, J., P.: Maximum Entropy Spectral Analysis. In: Proceedings of the 37th Meeting of the Society of Exploration Geophysicists, Oklahoma City, October 1967.

[13] Burg, J., P.: A New Analysis Technique for Time Series Data. In: NATO Advanced Study Institute on Signal Processing with Emphasis on Underwater Acoustics, Enschede, the Netherlands, August 1968.

[14] van den Bos, A.: Alternative Interpretation of Maximum Entropy Spectral Analysis. In: IEEE Transactions on Information Theory, vol. 17, no. 4, pp. 493-494, July 1971.

[15] Lacoss, R., T.: Data Adaptive Spectral Analysis Methods. In: Geophysics, vol. 36, no. 4, pp. 661-675, August 1971.

[16] Burg, J., P.: The Relationship between Maximum Entropy Spectra and Maximum Likelihood Spectra. In: Geophysics, vol. 37, no. 2, pp. 375-376, April 1972.

[17] Ables, J., G.: Maximum Entropy Spectral Analysis. In: Astron. Astrophys. Suppl. Series, vol.15, pp.383-393, 1974.

[18] Ulrych, T., J., Bishop, T., N.: Maximum Entropy Spectral Analysis and Autoregressive Decomposition. In: Rev. Geophysics, and Spac. Phys., vol. 13, pp. 183-200, February 1975.

[19] Madlová, A.: Maximum Entropy Spectrum Analysis, Preliminary PhD Report, Department of Radioelectronics, Faculty of Electrical Engineering and Information Technology, Slovak Technical University, Bratislava, Slovak Republic, October 1995.

[20] Madlová, A., Židek, F.: An Unconventional Spectrum Analysis Method. In: New Trends in Signal Processing III., pp. 11-14, Liptovský Mikuláš, Slovak Republic, May 1996.

[21] Makhoul, J.: Spectral Linear Prediction: Properties and Applications. In: IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-23, no.3, pp. 283-296, June 1975.

[22] Atal, B., S., Linear Prediction Analysis of Speech Signals. In: Schell, A. C. et al., Eds.: Programs for Digital Signal Processing, IEEE Acoustics, Speech, and Signal Processing Society, pp. 4.0-1 – 4.0-2, John Wiley & Sons, New York, 1979.

[23] Gray, A., H., Markel, J., D.: Linear Prediction Analysis Programs (AUTO-COVAR). In: Schell, A. C. et al., Eds.: Programs for Digital Signal Processing, IEEE Acoustics, Speech, and Signal Processing Society, pp. 4.1-1–4.1-7, John Wiley & Sons, New York, 1979.

[24] Gray, A., H., Wong, D., Y.: The Burg Algorithm for LPC Speech Analysis/Synthesis. In: IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-28, no. 6, pp. 609-615, December 1980.

[25] Vích, R.: Auswertung von Sprachanalyseverfahren im Spektralbereich. In: Digitale Sprach-verarbeitung – Prinzipien und Anwendungen, Bad Nauheim, pp. 43-50, October 1988.

[26] Makhoul, J.: Stable and Efficient Lattice Methods for Linear Prediction. In: IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-25, no. 5, pp. 423-428, October 1977.

[27] Viswanathan, R., Makhoul, J.: Efficient Lattice Methods for Linear Prediction. In: Schell, A. C. et al., Eds.: Programs for Digital Signal Processing, IEEE Acoustics, Speech, and Signal Processing Society, pp.4.2-1–4.1-14, John Wiley & Sons, New York, 1979.

[28] Strobach, P.: Pure Order Recursive Least-Squares Ladder Algorithms. In: IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-34, no. 4, pp. 880-897, August 1986.

[29] Oppenheim, A., V., Schafer, R., W.: Discrete-Time Signal Processing. Prentice Hall, 1989.

[30] Vích, R.: Cepstral Speech Model, Padé Approximation, Excitation and Gain Matching in Cepstral Speech Synthesis. In: Jan, J. et al., Eds.: EuroConference Biosignal '2000, Brno, Czech Republic, pp. 77-82, June 2000.

[31] Přibil, J.: Use of the Cepstral Model for Speech Synthesis, PhD Thesis, Praha, Czech Republic, November 1997 (in Czech).

[32] Imai, S., Kitamura, T., Takeya, H.: A Direct Approximation Technique of Log Magnitude Response for Digital Filters. In: IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-25, no. 2, pp. 127-133, April 1977.

[33] Imai, S.: Low Bit Rate Cepstral Vocoder Using the Log Magnitude Approximation Filter. In: IEEE International Conference on Acoustics, Speech, and Signal Processing '78, pp. 441-444, April 1978.

[34] Imai, S.: Cepstral Analysis Synthesis on the Mel Frequency Scale. In: IEEE International Conference on Acoustics, Speech, and Signal Processing '83, pp. 93-96, Boston, 1983.

[35] Tychtl, Z., Psutka, J.: Speech Production Based on the Mel-Frequency Cepstral Coefficients. In: Eurospeech '99, pp. 2335-2338, Budapest, Hungary, September 1999.

[36] Koishida, K., Hirabayashi, G., Tokuda, K., Kobayashi, T.: A 16 kb/s Wideband CELP-Based Speech Coder Using Mel-Generalized Cepstral Analysis, IEICE Transactions on Information and Systems, vol. E83-D, no.4, pp. 876-883, April 2000.

[37] McAulay, R., J., Quatieri, T., F.: Speech Analysis/Synthesis Based on a Sinusoidal Representation. In: IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-34, no. 4, pp. 744-754, August 1986.

[38] Quatieri, T., F., McAulay, R., J.: Speech Transformations Based on a Sinusoidal Representation. In: IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-34, no. 6, pp. 1449-1464, December 1986.

[39] McAulay, R., J., Quatieri, T., F.: Low-Rate Speech Coding Based on the Sinusoidal Model. In: Furui, S., Sondhi, M, M, Eds.: Advances in Speech Signal Processing, pp. 165-208, Marcel Dekker, New York, 1992.

[40] McAulay, R., J., Quatieri, T., F.: Sinusoidal Coding. In: Kleijn, W. B., Paliwal, K., K., Eds.: Speech Coding and Synthesis, pp. 121-173, Elsevier Science, Amsterdam, 1995.

[41] McAulay, R., J., Quatieri, T., F.: Pitch Estimation and Voicing Detection Based on a Sinusoidal Speech Model. In: IEEE International Conference on Acoustics, Speech, and Signal Processing '90, pp. 249-252, Albuquerque, 1990.

[42] Quatieri, T., F., McAulay, R., J.: Shape Invariant Time-Scale and Pitch Modification of Speech. In: IEEE Transactions on Signal Processing, vol. 40, no. 3, pp. 497-510, March 1992.

[43] Banga, E., R., García-Mateo, C.: Shape-Invariant Pitch-Synchronous Text-to-Speech Conversion. In: IEEE International Conference on Acoustics, Speech, and Signal Processing '95, pp. 656-659, Detroit, 1995.

[44] Banga, E., R., García-Mateo, C., Fernández-Salgado, X.: Shape-Invariant Prosodic Modification Algorithm for Concatenative Text-to-Speech Synthesis. In: Eurospeech '97, Rhodes, Greece, September 1997.

[45] Pollard, M., P., et al.: Voiced Speech Excitation Synthesis Using a Sinusoidal Model. In: Electronics Letters, vol. 34, no. 6, pp. 531-532, March 1998.

[46] O'Brien, D., Monaghan, A.: Shape Invariant Time-Scale Modification of Speech Using a Harmonic Model. In: IEEE International Conference on Acoustics, Speech, and Signal Processing '99, pp. 381-384, Phoenix, Arizona, March 1999.

[47] O'Brien, D., Monaghan, A., I., C.: Shape Invariant Pitch Modification of Speech Using a Harmonic Model. In: Eurospeech '99, pp. 1059-1062, Budapest, Hungary, September 1999.

[48] Trancoso, I., M., Marques, J., S., Ribeiro, C., M.: CELP and Sinusoidal Coders: Two Solutions for Speech Coding at 4.8-9.6 kbps. In: Speech Communication 9, pp.389-400, 1990.

[49] Nakhai, M., Marvasti, F.: A Hybrid Speech Coder Based on CELP and Sinusoidal Coding. In: IEICE Transactions on Information and Systems, vol. E83-D, no. 8, pp. 1190-1199, August 1999.

[50] Saito, S.: Speech Science and Technology. Ohmsha, Tokyo, 1992.

[51] Hashimoto, K. et al.: High Quality Synthetic Speech Generation Using Synchronized Oscillators. In: IEICE Transactions on Fundamentals of Electronics, Communications, and Computer Science, vol. E76-A, no. 11, pp. 1949-1945, November 1993.

[52] Spanias, A. et al.: Analysis/Synthesis of Speech Using the Short-Time Fourier Transform and a Time-Varying ARMA Process. In: IEICE Transactions on Fundamentals of Electronics, Communications, and Computer Science, vol. E76-A, no. 4, pp. 645-652, April 1993.

[53] Griffin, D., W., Lim, J., S.: Multiband Excitation Vocoder. In: IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-36, pp. 1223-1235, August 1988.

[54] Dutoit, H.: High Quality Text-to-Speech Synthesis: A Comparison of Four Candidate Algorithms. In: IEEE International Conference on Acoustics, Speech, and Signal Processing '94, pp. 565-568, Adelaide, Australia, 1994.

[55] Dutoit, T., Gosselin, B.: On the Use of a Hybrid Harmonic/Stochastic Model for TTS Synthesis by Concatenation. In: Speech Communication 19, pp. 119-143, 1996.

[56] Sercov, V., V., Petrovsky, A., A.: An Improved Speech Model with Allowance for Time-Varying Pitch Harmonic Amplitudes and Frequencies in Low Bit-Rate MBE Coders. In: Eurospeech '99, pp. 1479-1482, Budapest, Hungary, September 1999.

[57] Stylianou, Y., Dutoit, T., Schroeter, J.: Diphone Concatenation Using a Harmonic Plus Noise Model of Speech. . In: Eurospeech '97, pp. 613-616, Rhodes, Greece, September 1997.

[58] Stylianou, Y.: Concatenative Speech Synthesis Using a Harmonic Plus Noise Model. In: Third ESCA/COCOSDA Workshop on Speech Synthesis, pp. 261-266, Jenolan Caves, Blue Mountains, Australia, November 1998.

[59] Stylianou, Y.: Analysis of Voiced Speech Using Harmonic Models. In: ASA Meeting, Berlin, 1999.

[60] Stylianou, Y., Cappé, O., Moulines, E.: Continuous Probabilistic Transform for Voice Conversion. In: IEEE Transactions on Speech and Audio Processing, vol. 6, no. 2, pp. 131-142, March 1998.

[61] Stylianou, Y., Cappé, O.: A System for Voice Conversion Based on Probabilistic Classification and a Harmonic Plus Noise Model. In: IEEE International Conference on Acoustics, Speech, and Signal Processing '98, pp. 281-284, Seattle, Washington, May 1998.

[62] Stylianou, Y.: Removing Phase Mismatches in Concatenative Speech Synthesis. In: Third ESCA/COCOSDA Workshop on Speech Synthesis, pp. 267-272, Jenolan Caves, Blue Mountains, Australia, November 1998.

[63] Stylianou, Y.: Synchronization of Speech Frames Based on Phase Data with Application to Concatenative Speech Synthesis. In: Eurospeech '99, pp. 2343-2346, Budapest, Hungary, September 1999.

[64] Syrdal, A., et al.: TD-PSOLA Versus Harmonic Plus Noise Model in Diphone Based Speech Synthesis. In: IEEE International Conference on Acoustics, Speech, and Signal Processing '98, Seattle, Washington, 1998.

[65] Violaro, F., Böeffard, O.: A Hybrid Model for Text-to-Speech Synthesis. In: Transactions on Speech and Audio Processing, vol. 6, no. 5, pp. 426-434, September 1998.

[66] Schäfer-Vincent, K.: Pitch Period Detection and Chaining: Method and Evaluation. In: Phonetica, vol. 40, pp. 177-202, 1983.

[67] Psársky, M. (Madlová, A., advisor): Harmonic Speech Model, Diploma Thesis, KRE FEI STU Bratislava, Slovak Republic, December 1999 (in Slovak).

[68] Crespo, M., Á., R., et al.: On the Use of a Sinusoidal Model for Speech Synthesis in Text-to-Speech. In: van Santen, J., P., H., et al., Eds.: Progress in Speech Synthesis, pp. 57-70, Springer-Verlag, New York, 1997.

[69] Abe, T., Kobayashi, T., Imai, S.: Harmonics Tracking and Pitch Extraction Based on Instantaneous Frequency. In: IEEE International Conference on Acoustics, Speech, and Signal Processing '95, pp. 756-759, Detroit, 1995.

[70] Yang, G., Zanellato, G., Leich, H.: Band-Widened Harmonic Vocoder at 3 to 4 kbps. In: IEEE International Conference on Acoustics, Speech, and Signal Processing '95, pp. 504-507, Detroit, 1995.

[71] Ahmadi, S., Spanias, A., S.: A New Sinusoidal Phase Modelling Algorithm. In: IEEE International Conference on Acoustics, Speech, and Signal Processing '95, Detroit, 1995.

[72] Ahmadi, S., Spanias, A., S.: Low Rate Sinusoidal Coding of Speech Using an Improved Phase Matching Algorithm. In: IEEE Workshop on Speech Coding for Telecommunications, pp. 35-36, Pennsylvania, September 1997.

[73] Ahmadi, S., Spanias, A., S.: A New Phase Model for Sinusoidal Transform Coding of Speech. In: IEEE Transactions on Speech and Audio Processing, vol. 6, no. 5, pp. 495-501, September 1998.

[74] Cappé, O., Moulines, E.: Regularization Techniques for Discrete Cepstrum Estimation. In: IEEE Signal Processing Letters, vol. 3, no. 4, pp. 100-102, April 1996.

[75] Oudot, M., Cappé, O., Moulines, E.: Robust Estimation of the Spectral Envelope for "Harmonics + Noise" Models. In: IEEE Workshop on Speech Coding for Telecommunications, pp. 11-12, Pennsylvania, September 1997.

[76] Murthi, M., N., Rao, B., D.: All-Pole Modeling of Speech Based on the Minimum Variance Distortionless Response Spectrum. In: IEEE Transactions on Speech and Audio Processing, vol. 8, no. 3, pp. 221-239, May 2000.

[77] Jelinek, M., Adoul, J.-P.: Frequency-Domain Spectral Envelope Estimation for Low Rate Coding of Speech. In: IEEE International Conference on Acoustics, Speech, and Signal Processing '99, pp. 253-256, Phoenix, Arizona, March 1999.

[78] Madlová, A.: An Experiment with Childish Voice Analysis and Synthesis. In: Radioelektronika '2000, pp. III–112-115, Bratislava, Slovak Republic, September 2000.

[79] Přibil, J.: Comparison of Pitch Synchronous and Asynchronous Analysis in Cepstral Speech Modelling, In: Vích, R., ed.: Speech Processing, 9th Czech-German Workshop, pp. 35-36, Prague, Czech Republic, September 1999.

[80] Deisher, M., E., Spanias, A., S.: Speech Enhancement Using State-Based Estimation and Sinusoidal Modeling. In: Journal of Acoustical Society of America, vol. 102, no. 2, pp. 1141-1148, August 1997.

[81] Gardner, W., R., Rao, B., D.: Noncausal All-Pole Modeling of Voiced Speech. In: IEEE Transactions on Speech and Audio Processing, vol. 5, no. 1, pp. 1-10, January 1997.

[82] Chang, W.-W., Wang, D.-Y.: Techniques for Improving Sinusoidal Transform Vocoders. In: IEEE International Conference on Acoustics, Speech, and Signal Processing '98, pp. 525-528, Seattle, Washington, May 1998.

[83] Chang, W.-W., Wang, D.-Y.: Quality Enhancement of Sinusoidal Transform Vocoders. In: IEE Proceedings – Vision, Image, and Signal Processing, vol. 145, no. 6, pp. 379-383, December 1998.

[84] George, E., B., Smith, M., J., T.: Speech Analysis/Synthesis and Modification Using an Analysis-by-Synthesis/Overlap-Add Sinusoidal Model. In: IEEE Transactions on Speech and Audio Processing, vol. 5, no. 5, pp. 389-406, September 1997.

[85] Macon, M., W., Clements, M., A.: Sinusoidal Modelling and Modification of Unvoiced Speech. In: IEEE Transactions on Speech and Audio Processing, vol. 5, no. 6, pp. 557-560, November 1997.

[86] Macon, M., W., Clements, M., A.: An Enhanced ABS/OLA Sinusoidal Model for Waveform Synthesis in TTS. In: Eurospeech ’99, pp. 2327-2330, Budapest, Hungary, September 1999.

[87] Bailly, G., Bernard, E., Coisnon, P.: Sinusoidal Modelling. In: COST 258 Working Papers, 1998.

[88] Bailly, G.: A Parametric Harmonic + Noise Model. In: COST 258 Working Papers, 1999.

[89] Bailly, G.: A Parametric Harmonic + Noise Model. In: COST 258 Working Papers, 2000.

[90] Bailly, G.: Accurate Estimation of Sinusoidal Parameters in a Harmonic + Noise Model for Speech Synthesis. In: Eurospeech ’99, pp. 1051-1054, Budapest, Hungary, September 1999.

[91] Teague, K., A., Andrews, W., Walls, B.: Enhanced Modeling of Discrete Spectral Amplitudes. In: IEEE Workshop on Speech Coding for Telecommunications, pp. 13-14, Pennsylvania, September 1997.

[92] Teague, K., A., Andrews, W.: Enhanced Spectral Modeling for MBE Speech Coders. In: Thirty-first Asilomar Conference on Signals, Systems, and Computers, pp. 1071-1074, California, November 1998.

[93] Torres, S., Casajús-Quirós, F., J.: Phase Quantization by Pitch-Cycle Waveform Coding in Low Bit Rate Sinusoidal Coders. In: Eurospeech ’97, pp. 1303-1306, Rhodes, Greece, September 1997.

[94] Torres, S., Casajús-Quirós, F., J.: Vocal System Phase Coder for Sinusoidal Speech Coders. In: Electronics Letters, vol. 33, no. 20, pp. 1683-1685, September 1997.

[95] Li, C., Cuperman, V.: Enhanced Harmonic Coding of Speech with Frequency Domain Transition Modeling. In: IEEE International Conference on Acoustics, Speech, and Signal Processing ‘98, pp. 581-584, Seattle, Washington, May 1998.

[96] Shlomot, E., Cuperman, V., Gersho, A.: Combined Harmonic and Waveform Coding of Speech at Low Bit Rates. In: IEEE International Conference on Acoustics, Speech, and Signal Processing ‘98, pp. 585-588, Seattle, Washington, May 1998.

[97] Li, C., Gersho, A., Cuperman, V.: Analysis-by-Synthesis Low-Rate Multimode Harmonic Speech Coding. In: Eurospeech '99, pp. 1451-1454, Budapest, Hungary, September 1999.

[98] Li, C., Cuperman, V.: Analysis-by-Synthesis Multimode Harmonic Speech Coding at 4 kb/s. In: IEEE International Conference on Acoustics, Speech, and Signal Processing '2000, Istanbul, Turkey, June 2000.

[99] Li, C.: Analysis-by-Synthesis Multimode Harmonic Speech Coding at Low Bit Rate, PhD Thesis. University of California, Santa Barbara, USA, 2000.

[100] Etemoğlu, Ç., Ö., Cuperman, V., Gersho, A.: Speech Coding with an Analysis-by-Synthesis Sinusoidal Model. In: IEEE International Conference on Acoustics, Speech, and Signal Processing '2000, Istanbul, Turkey, June 2000.

[101] Jensen, J., Jensen, S., H., Hansen, E.: Exponential Sinusoidal Modeling of Transitional Speech Segments. In: IEEE International Conference on Acoustics, Speech, and Signal Processing '99, pp. 473-476, Phoenix, Arizona, March 1999.

[102] Fay, G. et al.: Polynomial Quasi-Harmonic Models for Speech Analysis and Synthesis. In: IEEE International Conference on Acoustics, Speech, and Signal Processing '98, pp. 865-868, Seattle, Washington, May 1998.

[103] Li, G., Qiu, L.: Speech Analysis and Synthesis Using Instantaneous Amplitudes. In: EUSIPCO '98, European Signal Processing Conference, Island of Rhodes, Greece, September 1998.

[104] Li, G. et al.: Signal Representation Based on Instantaneous Amplitude Models with Application to Speech Synthesis. In: IEEE Transactions on Speech and Audio Processing, vol. 8, no. 3, pp. 353-357, May 2000.

[105] Gray, A., H., Markel, J., D.: A Spectral-Flatness Measure for Studying the Autocorrelation Method of Linear Prediction of Speech Analysis. In: IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-22, no. 3, pp. 207-217, June 1974.

[106] Amodio, A., Feng, G.: A Wideband Speech Coder Based on Harmonic Coding at 16kbs. In: Eurospeech '99, pp. 1539-1542, Budapest, Hungary, September 1999.

[107] Přibil, J., Madlová, A.: Computational Complexity of Two Methods Based on Cepstral Parametrization of Speech Signal. In: New Trends in Signal Processing V., pp. 248-251, Liptovský Mikuláš, Slovak Republic, May 2000.

[108] Madlová, A., Přibil, J.: Comparison of Two Approaches to Speech Modelling Based on Cepstral Description. In: Jan, J., et al., Eds.: EuroConference Biosignal '2000, pp. 83-85, Brno, Czech Republic, June 2000.

[109] Haykin, S., Kesler, S.: Prediction-Error Filtering and Maximum-Entropy Spectral Estimation, In: Nonlinear Methods of Spectral Analysis (ed. Haykin, S.), Springer-Verlag, pp.9-72, 1983.

[110] Ulrych, T. J., Ooe, M.: Autoregressive and Mixed Autoregressive-Moving Average Models and Spectra, In: Nonlinear Methods of Spectral Analysis (ed. Haykin, S.), Springer-Verlag, pp.73-126, 1983.

[111] Weitkunat, R.: Digital Biosignal Processing, Elsevier, 1991.

[112] Lin, D., M., Wong, E., K.: A Survey of the Maximum Entropy Method and Parameter Spectral Estimation, In: Physics Reports, vol.193, no.2, pp. 41-135, September 1990.

[113] Mitra, S., K., Kaiser, J., F.: Handbook for Digital Signal Processing, John Wiley & Sons, 1993.

[114] Madlová, A.: Autoregressive Model Order Selection for Voiced Speech, Measurement '99, pp. 111-114, Smolenice, Slovak Republic, April 1999.

[115] Vích, R., Horák, P., Víchová, E.: Experimente mit der Synthese der Frauenstimme, In: DAGA 94, Teil C, pp.1317-1320, Dresden, 1994.

[116] Gray, A., Markel, J., D.: Distance Measures for Speech Processing. In: IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-24, no. 5, pp. 380-391, October 1976.

[117] Vích, R.: FIR Vocal Tract Model. In: Vích, R., ed.: Speech Processing, 10th Czech-German Workshop, pp. 21-24, Prague, Czech Republic, September 2000.

[118] Vích, R., Přibil, J.: Gemischte Anregung im cepstralen Sprachsynthesesystem. In: ITG-Fachbereicht Sprachkommunikation, pp. 105-107, Frankfurt am Main, September 1996.

[119] Přibil, J.: Comparison of Speech Spectral Features Using LPC Parameters and Cepstral Coefficients. In: 32nd Czech Acoustic Conference, pp.63-66, Prague, Czech Republic, September 1995.

[120] Přibil, J.: Comparison of Quality and Computational Complexity of Cepstral Speech Synthesis for Sampling Frequencies of 8 and 16 kHz. In: Applied Electronics, pp.140-144, Plzeň, Czech Republic, September 1999 (in Czech).

[121] Harris, F., J.: On the Use of Windows for Harmonic Analysis with the Discrete Fourier Transform. In: Proceedings of the IEEE, vol. 66, no. 1, pp. 51-84, January 1978.

[122] Madlová, A.: Harmonic Speech Model with Cepstral Parametrization. In: Vích, R., ed.: Speech Processing, 10th Czech-German Workshop, pp. 56-58, Prague, Czech Republic, September 2000.

[123] O'Brien, D., Monaghan, A., I., C.: Concatenative Synthesis Based on a Harmonic Model. In: IEEE Transactions on Speech and Audio Processing, vol. 9, no. 1, pp. 11-20, January 2001.

[124] Stylianou, Y.: Applying the Harmonic Plus Noise Model in Concatenative Speech Synthesis. In: IEEE Transactions on Speech and Audio Processing, vol. 9, no. 1, pp. 21-29, January 2001.