Lauri Juvela

Perceptual spectral matching utilizing mel-scale filterbanks for statistical parametric speech synthesis with glottal excitation vocoder

School of Electrical Engineering

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

Espoo 7.5.2015

Thesis supervisor:

Prof. Paavo Alku

Thesis advisor:

Tuomo Raitio, M.Sc. (Tech.)

Aalto University
School of Electrical
Engineering

AALTO UNIVERSITY
SCHOOL OF ELECTRICAL ENGINEERING

ABSTRACT OF THE
MASTER'S THESIS

Author: Lauri Juvela

Title: Perceptual spectral matching utilizing mel-scale filterbanks for statistical parametric speech synthesis with glottal excitation vocoder

Date: 7.5.2015 Language: English Number of pages: 7+62

Department of Signal Processing and Acoustics

Professorship: Speech communication technology Code: S-89

Supervisor: Prof. Paavo Alku

Advisor: Tuomo Raitio, M.Sc. (Tech.)

This thesis presents a novel perceptual spectral matching technique for parametric statistical speech synthesis with glottal vocoding. The proposed method utilizes a perceptual matching criterion based on mel-scale filterbanks.

The background section discusses the physiology and modelling of human speech production and perception, necessary for speech synthesis and perceptual spectral matching. Additionally, the working principles of statistical parametric speech synthesis and the baseline glottal source excited vocoder are described.

The proposed method is evaluated by comparing it to the baseline method first by an objective measure based on the mel-cepstral distance, and second by a subjective listening test. The novel method was found to give comparable performance to the baseline spectral matching method of the glottal vocoder.

Keywords: speech synthesis, perceptual spectral matching, glottal excitation vocoder, mel-spectrum, frequency warping

AALTO-YLIOPISTO SÄHKÖTEKNIIKAN KORKEAKOULU	DIPLOMITYÖN TIIVISTELMÄ

Tekijä: Lauri Juvela

Työn nimi: Perkeptuaalinen spektrisovitus glottisherätevokoodatussa tilastollisessa parametrisessa puhesynteesissä käyttäen mel-suodinpankkia

Päivämäärä: 7.5.2015 Kieli: Englanti Sivumäärä: 7+62

Signaalinkäsittelyn laitos ja akustiikan

Professuuri: Puhekommunikaatioteknologia Koodi: S-89

Valvoja: Prof. Paavo Alku

Ohjaaja: DI Tuomo Raitio

Tämä työ esittää uuden perkeptuaalisen spektrisovitustekniikan glottisvokoodattua tilastollista parametristä puhesynteesiä varten. Ehdotettu menetelmä käyttää mel-suodinpankkeihin perustuvaa perkeptuaalista sovituskriteeriä.

Työn taustaosuus käsittelee ihmisen puheentuoton ja havaitsemisen fysiologiaa ja mallintamista tilastollisen parametrisen puhesynteesin ja perkeptuaalisen spektrisovituksen näkökulmasta. Lisäksi kuvataan tilastollisen parametrisen puhesynteesin ja perusmuotoisen glottisherätevokooderin toimintaperiaatteet.

Uutta menetelmää arvioidaan vertaamalla sitä alkuperäiseen metodiin ensin käyttämällä mel-kepstrikertoimia käyttävää objektiivista etäisyysmittaa ja toiseksi käyttäen subjektiivisia kuuntelukokeita. Uuden metodin havaittiin olevan laadullisesti samalla tasolla alkuperäisen spektrisovitusmenetelmän kanssa.

Avainsanat: puhesynteesi, perkeptuaalinen spektrisovitus, glottisherätevokooderi, mel-spektri, taajuusvarppaus

Preface

First, I'd like to thank Paavo and Tuomo for their constant ideas, feedback and trust in my work. Second, I'd like to thank all the people at Aalto acoustics department for creating such a wonderful working environment. Last, but not the least, I'd like to thank Eeva for support.

Otaniemi, May 7, 2015

Lauri Juvela

Contents

Abbreviations

AC	Autocorrelation
AR	Autoregressive
CCR	Comparison category rating
DAP	Discrete all-pole
DCT	Discrete cosine transform
DFT	Discrete Fourier transform
EM	Expectation maximization
ERB	Equivalent rectangular bandwidth
ETSI	European Telecommunications Standards Institute
FFT	Fast Fourier transform
FIR	Finite impulse response
HMM	Hidden Markov model
HNR	Harmonic-to-noise ratio
HSMM	Hidden semi-Markov model
HTS	HMM-based Speech Synthesis System
IAIF	Iterative adaptive inverse filtering
IFFT	Inverse fast Fourer transform
IIR	Infinite impulse response
IPA	International phonetic alphabet
LP	Linear prediction, — predictive
LPC	Linear predictive coding
LSF	Line spectral frequency
MDL	Minimum description length
MFCC	Mel-frequency cepstral coefficient
ML	Maximum likelihood
MSE	Mean-squared error
NNLS	Non-negative least squares
STRAIGHT	Speech Transformation and Representation using Adaptive Interpolation of weiGHT spectrum
SPL	Sound pressure level
WIIR	Warped infinite impulse response

1 Introduction

Text-to-speech synthesis is currently under high scientific and commercial interest, and the methods are in rapid development – potential applications include personalised voice prostheses, human–machine speech interfaces, and automatic simultaneous interpretation.

Statistical parametric speech synthesis has gained popularity and been under great research interest in recent years. The advantages over the unit-selection type methods [18] include notably smaller memory footprint and greater flexibility in terms of voice adaptation to different speakers and speaking styles. [5, 61] Problem with statistical parametric speech synthesis has been the buzzy, robotic quality of the produced voice. Some improvement, especially in prosody, is expected from the ongoing transition in statistical modelling from hidden Markov models the to deep neural networks in speech synthesis. However, the parametric voice coder, or vocoder still plays an important role in improving the synthetic voice quality. [16,59]

The most widely used vocoder in statistical parametric speech synthesis is the STRAIGHT-system [58,60] that uses mel-generalized cepstral features to model the speech spectrum, and impulse-train type signal modified with aperiodicity parameters as excitation. The used parameters are well suited for statistical modelling, but the artificial excitation results in slightly robotic speech quality. Good results in improving the naturalness of the synthetic speech have been achieved by using a vocoder based on human speech production [43]. This glottal source vocoder uses a glottal pulse waveform estimated from natural speech as excitation.

In its basic form, the GlottHMM glottal source vocoder [43] uses a fixed excitation pulse, and a matching filter derived from a mean-square-error based spectral model to recreate the spectral characteristics of the target voice source. However, there is room for improvement in voice naturalness and overall quality. The method has been improved by using a more complex glottal excitation created with a glottal pulse library [45], or glottal pulses generated by speech parameter controlled deep neural network [44]. This thesis takes a different approach and attempts to improve the fixed excitation technique by replacing the MSE-based matching filter by a perceptually motivated matching filter.

Mel-scale filterbanks form the core spectral model in conventional mel-frequency cepstral coefficient features used in automatic speech recognition. Although the filterbank model has existed for a relatively long time, it is still used in modern speech recognition regardless of the advances in other aspects of speech recognition [16, 41]. Additionally, the mel-scale filterbank model has been utilized successfully in a generative context in artificial speech bandwidth extension [39]. Despite the simplicity of the filterbank, the model has the important perceptual properties of nonlinear frequency resolution and auditory filter approximation that are studied more closely later in this thesis. The success and simplicity of the mel-scale filterbank as a perceptual model in speech applications mark the model as a feasible criterion for perceptual spectral matching in speech synthesis.

This thesis presents a perceptual spectral matching scheme for speech synthesis using a mel-scale filterbank based matching criterion. The proposed method is based on the baseline GlottHMM vocoder [43], and for comparative evaluation purposes the matching scheme designed so that the parameter dimensionality matches that of the original system. Nonetheless, the presented spectral matching technique is easily adaptable to other configurations.

The thesis is organized as follows: Chapter 2 gives an overview of the physiology of human speech production and hearing, complemented with a general modelling viewpoint of speech production and sound perception. These models serve as foundation for the computational methods in the next section. Chapter 3 deals in speech parametrization techniques utilized in the parametric vocoder and the proposed spectral matching method. Chapter 4 first presents the general principle of statistical parametric speech synthesis with some detail on the hidden Markov models, and second, the speech synthesis system based on the GlottHMM vocoder with some alterations. Chapter 5 describes the proposed perceptual spectral matching method and its realisation as an all-pole matching filter. In Chapter 6, the proposed spectral matching method is evaluated in comparison to the baseline GlottHMM spectral matching, first by objective measures and statistical properties, and second by a subjective listening test. Finally, the results are discussed in Chapter 7 with some concluding remarks.

2 Speech production and perception

When one wishes to build a physiologically based speech synthesis system it is natural to first study the physiology of human speech production. On the other hand, knowledge on the human hearing and speech perception is equally needed, if one desires to apply psychoacoustics in designing the speech synthesizer. This chapter examines the physiological properties of speech production and perception, along with some high level modelling aspects.

2.1 Human speech production mechanism

2.1.1 Physiology of speech

A central component of the human speech is the so called voice source, or glottal excitation created in the larynx. The main speech function of the larynx is to modulate the airflow from the lungs by periodically closing the vocal folds. The opening between the vocal folds is called the glottis and the volume flow entering the vocal tract in voiced speech consists of periodic glottal pulses – thus the waveform directly related to the vocal fold movement is called the glottal volume velocity waveform. It should be noted that the movement of the vocal folds as a sound source is not analogous to the vibrating string or membrane as the vocal folds merely modulate the airflow [13, p. 265]. In order to create voiced speech, the glottal flow signal is altered by the resonances and turbulences in the vocal tract.

Other important component in the voice production system is the filter effect imposed on the glottal source by the vocal tract. Vocal tract refers to the supraglottal part of the voice production system, although some definitions include the entire voice production system. The vocal tract has varying resonant frequencies called formants, which play a major role in embedding linguistic information in speech sounds. Figure 1 shows a schematic of the voice production system. Air from the lungs enters the vocal tract through the larynx that in case of voiced speech modulates the airflow by closing and opening the glottis periodically. In the vocal tract, air passes through the pharyngeal and oral cavities, eventually exiting at the lips or nostrils. The cavities are separated at the uvula, and it is possible at this point that some of the airflow is diverted into the nasal cavity. The airflow to nasal cavity is controlled by the velum that opens while producing nasal sounds. [37]

Various types of speech sounds can be produced by the means of articulatory gestures, that is, by movements that alter the shape of the vocal tract. Some of the parts comprising the vocal tract are fixed in terms of articulation, as they do not permit voluntary movement: The nasal cavity has a large surface area to volume ratio and mucous walls, thus acting as a damped resonator. Also the teeth, the alveolar ridge and the hard palate are immovable and are fairly rigid. In contrast, the movable parts in the vocal tract are called the articulators. The most prominent articulator is the tongue, whose parts such as the tip and dorsum can be moved relatively independently. The rounding of lips affects vowel sounds especially. The larynx also has an articulatory function in addition to housing the vocal folds: the

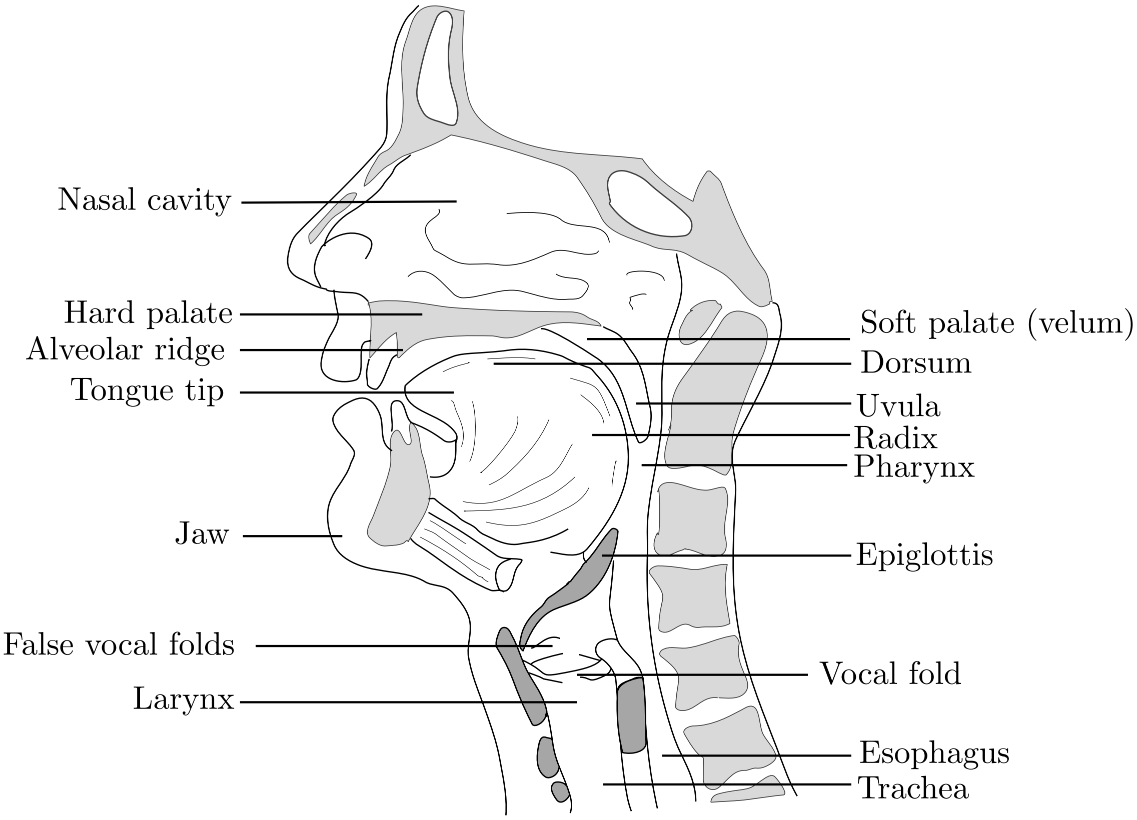


Figure 1: Schematic of the human voice producing apparatus. Adapted from [24]

larynx can be lowered to increase the pharyngeal cavity volume. Also the movement of the jaw participates in articaulation by altering the oral cavity volume. [37]

Articulatory phonetics give some insight on how the articulators apply linguistic information onto speech. Speech sounds are generally classified into two categories: vowels and consonants. Both are included in an articulation based classification system defined by the international phonetic alphabet (IPA). The vowels are always voiced and the vocal tract remains relatively open. Different vowels are characterised mainly by tongue position on the axes front-back and open-closed, and lip rounding.

The consonants are typically classified by phonation along with place and manner of articulation. Phonation indicates whether the sound is voiced or not. Places of articulation indicate the location of a constriction or other sound generating effect on the vocal tract. Manner of articulation tells which kind of articulatory gesture is used. Stop sounds such as /k/, /t/, /p/, and their voiced counterparts /g/, /d/, /b/ block the airflow completely at some point on the vocal tract. Fricatives such as /s/ and /f/ constrict the vocal tract so that the turbulent airflow at the constriction acts a sound source. Nasals allow airflow into the nasal cavity. Approximants such as the semivowels and the lateral approximant /l/ constrict the vocal tract partially but not enough to create turbulences. Other manners of articulation include flaps and trills, such as the alveolar trill /r/. Generally natural languages do not use all physically possible articulatory combinations, but tend to select some subset of the combinations. [37]

2.1.2 Source-filter model

Many modern speech processing methods rely on the source-filter model of speech production. The model states that phonation and articulation can be considered independent of each other. Phonation is represented by the glottal excitation and is modelled as the source, whereas the spectral effects of the vocal tract are modelled as a filter. In the model, the effects of each individual articulator are lumped into a single filter that can be estimated from the speech signal. [13]

In practice the source-filter relationship is realized by using digital filters. Typically the speech signal is analysed in short segments where the signal is assumed to be stationary within the segment. This enables the use of linear time-invariant filter models. In addition to glottal source and vocal tract filter, the spectro-temporal characteristics of speech are affected by lip radiation load. In the z-domain denote the glottal source as $G(z)$, the vocal tract filter with $V(z)$, and the lip radiation effect $L(z)$. According to the model, speech signal consists of three components in cascade: the glottal excitation, vocal tract filter and lip radiation load. In speech processing applications the latter two are often lumped into a single filter. When the signal is considered locally stationary, a linear z-domain relationship holds:

$$S(z) = G(z)V(z)L(z). \tag{1}$$

This type of model makes some simplifications in terms of the speech production physiology. For instance, in voiced fricatives the noise present in the waveform stems from the turbulences created by constrictions in the vocal tract, and not in the glottal source. The source-filter models considers the source in this type of signal as periodic glottal excitation with additional noise, while the actual noise source is not located in the glottis. Another physiological shortcoming is the modelling of the glottis as a simple volume velocity source, disregarding the effects of sub-glottal and and supra-glottal pressure differences. Nevertheless, the source-filter model has been proven quite effective by the multitude of applications based on the model.

2.1.3 Acoustical tube model

The classical acoustic model for the vocal tract filter has been a series of lossless cylindrical tubes with varying cross-sectional areas. These structures can indeed replicate the formant structures exhibited in speech when configured correctly. More detailed acoustical models have a branch in the tube system depicting nasal and oral tubes separately, resulting in zeros in the filter responses [37]. However, a non-branching tube is easier to model and has an interesting connection to all-pole digital filters: If the tube segments are of equal length it can be shown that the acoustic tube model is equivalent to the all-pole filter model given by linear prediction (see Section 3.2) [55] [42, p. 440-1].

The success of the linear predictive modelling of speech can be partly attributed to this approximative connection to voice production physiology. Additionally, the all-pole filters have some very useful technical properties briefly discussed in Section 3.2.

2.2 Hearing and sound perception

This section first describes the physiological processes that transform the acoustic signal received at the ear into a neural signal transmitted to the brain. Second, the section examines some psychoacoustical properties related to the perception of sound and presents models to recreate these properties. The aim is to illustrate how hearing physiology relates to the sound perception and, consequently, examine the perceptual modelling tools necessary for creating a perceptual spectral matching scheme.

2.2.1 Physiology of hearing

The human ear can be divided structurally into three sections: outer ear, middle ear and inner ear. The outer ear directs and transmits sound waves into the the middle ear, which acts as a mechanical transducer. The inner then ear transduces the vibrations transmitted from the middle ear into neural firings. An overview of the anatomy of the human ear is depicted in Figure 2.

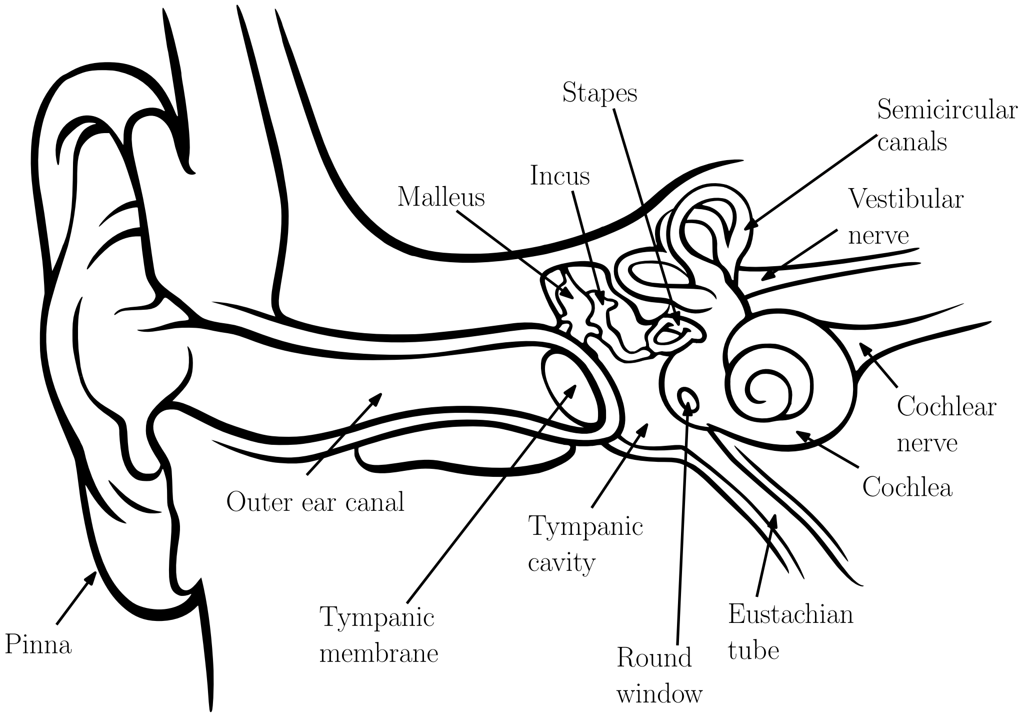


Figure 2: Schematic of the anatomy of human ear [9]. The drawing is not to scale, as the inner and middle ear are enlarged for clarity.

The function of the outer ea is to direct and amplify sound coming from the environment. The pinna focuses the incoming sound waves into the ear canal, applying direction-dependent spectral coloration to aid spatial hearing. Another spectral characteristic caused by the outer ear results from the resonance frequency of the ear canal. The average ear canal has a length of approximately 27 mm and can be modelled as tube, thus having its first resonance at 3 kHz. However, the ear canal resonances are not very sharp due to the non-rigid walls of the ear canal, resulting to approximately 15 dB amplification at 3–5 kHz region. [37, p. 110]

The middle ear is an air filled cavity connected to the outer ear at the tympanic membrane and to the inner ear at the oval window. In terms of hearing, the main functional part of the middle ear is the tranduction system comprised of the ossicular bones. Additionally, the Eustachian tube connects the middle ear cavity to the pharynx, enabling middle ear pressure equalization with the atmospheric pressure. The middle ear has two main functions: First, the ossicular bones, malleus, incus, and stapes, act as a mechanical impedance transducer between the ear canal and the fluid-filled inner ear, transforming the large displacement and small pressure signals in air into high pressure and small displacement signals in the cochlear fluid. The pressure variations in the ear canal excite the tympanic membrane which is connected to malleus. This motion is transmitted along the ossicular lever system to the oval window of the cochlea. The acoustic impedance of the inner ear is approximately 4000 times that of the air, and nearly no energy would be transmitted into the inner ear without the middle ear impedance matching. Most of this mismatch is overcome by the area-ratio of the tympanic membrane and the oval window, along with some mechanical transduction allowed by the ossicular lever system. [62] [37] The second function of the middle ear is to protect the inner ear from loud sounds via the auditory reflex. In the presence of loud sounds, the muscles connected to the ossicles contract, making the system more rigid and allowing less sound energy to be transmitted into the inner ear [33].

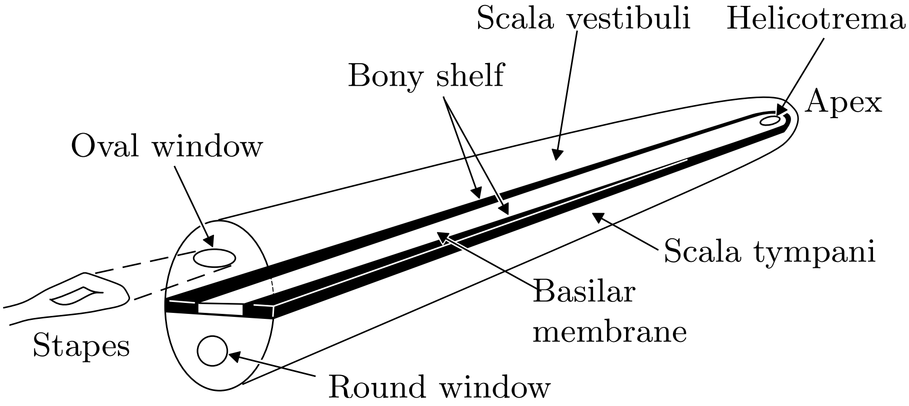


Figure 3: Linearised cochlea, adapted from [24]

In the inner ear, the functional part relevant for hearing is the cochlea, an organ of 32–35 mm length resembling a snail, coiled in approximately 2.5 turns. A linearised schematic of the cochlea where the coil has been unwound is depicted in Figure 3. The mechanical vibrations enter the cochlea via the oval window, which is connected to stapes in the middle ear. Motion of the oval window creates a travelling wave that reaches its maximum amplitude at a position depending on tje frequency so that the maxima near the base of the cochlea correspond to high frequencies and the maxima near the apex to low frequencies.

Neural transduction from pressure wave to neural impulses happens within the basilar membrane. The basilar membrane, stiffened by the bony shelf, lies within scala media and vibrates in conjunction with the fluid in scala tympani and scala vestibuli. The basilar membrane is stiffer at its base near the oval window, resonating more strongly with high frequencies, and less stiff near the apex of the cochlea, correspondingly resonating with low frequencies.

Locations on the basilar membrane can be related to a characteristic frequency. Frequency response for a specific location resembles a bandpass filter response centred at the characteristic frequency. Distance of the location from the stapes is roughly proportional to the logarithm of the characteristic frequency. This, with hair cells being somewhat evenly spaced on the basilar membrane, leads to a non-linear frequency resolution of the basilar membrane. In other words, a location with a higher characteristic frequency actuates neural fireings at a broader range of frequencies than a location with a lower characteristic frequency.

2.2.2 Critical bands

Critical band is an important auditory concept based on the idea that the auditory system processes sounds that are close in frequency together, and sounds that are far apart separately. The concept was originally introduced in [14], derived from experiments on tone masking by narrowband noise. Critical bands are closely related to auditory filters – the critical band mechanism can be seen as a band-pass filter whose frequency response approximates a tuning curve of an auditory neuron [37, p. 127]. In terms of modelling, particularly in speech parametrization, the most useful filter shapes can be derived from psychoacoustic tuning curves, which closely resemble their physiological counterparts. By assuming linearity, the auditory filters are obtained by inverting the psychoacoustic tuning curves [62, p. 68-71], [33, p. 79]. An example of these curves measured by the masked tone method is shown in Figure 4, exhibiting triangle-like shapes on the logarithmic scale.

Critical band filter shapes and bandwidths can be determined using various methods. One approach is to study the masking threshold of a test tone with narrow-band noise masker [33,34] to obtain the critical bandwidths. In the method, a test tone is kept at a fixed frequency and sound pressure level (SPL) while a narrow band masking noise is varied in center frequency. The frequency dependent masking thresholds are then determined by setting the masker SPL to a level where the tone is not perceived underneath the masker. Other, test tone free method for determining the critical bandwidths is by an experiment where the bandwidth of bandpass noise is varied while keeping the overall intensity constant. As long as the bandpass noise remains within a critical band, the perceived loudness also remains constant [24]. Many of these methods use rectangular band-pass filters in creating the test stimuli, although the critical band filters are more complex in shape, thus only obtaining an equivalent rectangular bandwidth estimates for the critical bands.

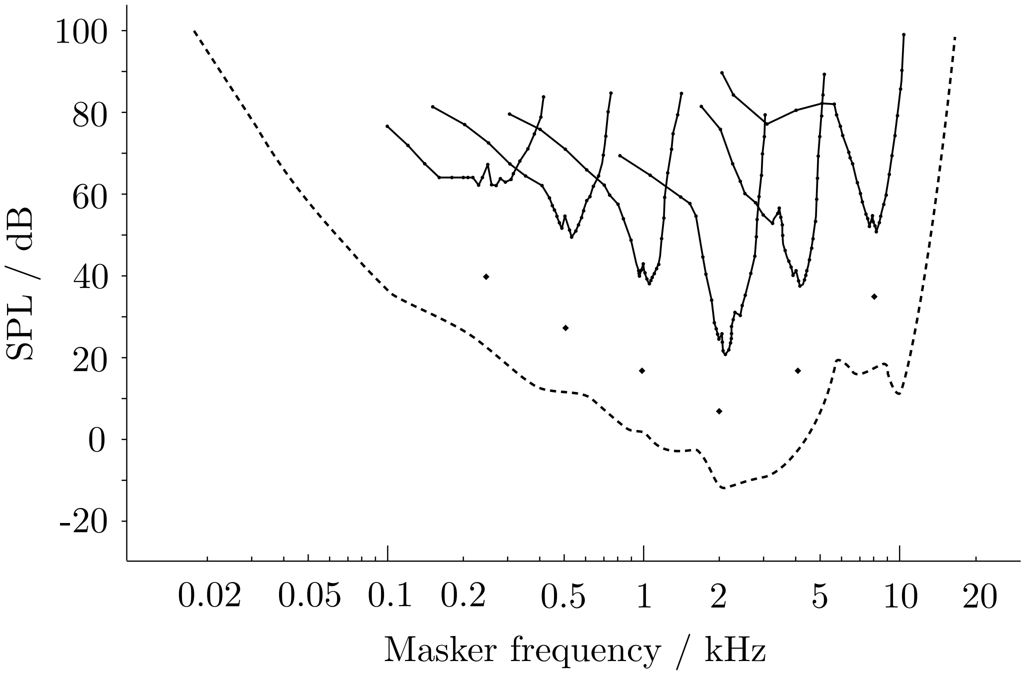


Figure 4: Psychoacoustic tuning curves measured with narrow-band variable frequency noise masker and fixed test tones [24].

Some formulatic expressions for auditory filter shapes can be found in [35].

It is somewhat unclear which part of the hearing system determines the exact shapes of the critical band auditory filters, but similar behaviour can be seen already in the excitation patterns of the basilar membrane. However, the neural and psychoacoustic tuning curves exhibit higher frequency selectivity, suggesting some kind of "second filter". The process is not well known, but direct modelling of psychoacoustic phenomena is sufficient in applying auditory knowledge to speech parametrization.

Another useful notion related to critical bands is the critical band intensity that is defined as an intensity integral over the critical band [62, p. 147]. The intensities can be used as an auditory representation of sound as they realize the non-linear frequency resolution and within-band masking effect of human hearing. Particularly, they are connected to the weighted summing within a filter utilized in, for example, the mel-scale filterbanks presented in Section 3.1.2.

2.2.3 Frequency resolution and pitch perception

An important perceptual property of human hearing is the non-linear frequency resolution of the human hearing, which can be modelled using various perceptually determined pitch scales. One such pitch scale can be derived from critical band measurements by finding a pitch scale where the critical bands are equally wide. This scale is called the Bark-scale and can be approximated functionally by [62]

$$z/\mathrm{Bark} = 13\arctan(0.76f/1000) + 3.5\arctan(f/7500)^2. \tag{2}$$

The resulting warping curve is illustrated in Figure 6. The functional expression is plotted with a solid line, complemented with experiment data from literature plotted with circles. Additionally, an all-pass warping approximation (see Section 3.1.1) is plotted with a dashed line.

Mel-scale (from melody) is an another perceptually determined scale that measures the relative pitch differences between tones or harmonic sounds. The scale is found experimentally by playing listeners two alternating tones, of which the first one is fixed. The listener then adjusts the second tone to a pitch which is half of the pitch of the first tone, i.e., an octave lower [49]. At low frequencies the pitch behaves nearly linearly as a function of frequency. The classical formula used for mapping from frequency to Mel scale is [31, 37]:

$$m = 2595 \log_{10}(1 + f/700). \tag{3}$$

The resulting warping curve is illustrated in Figure 5. Data points from listening experiments [4] are given for comparison. See Table A1 for the data point values.

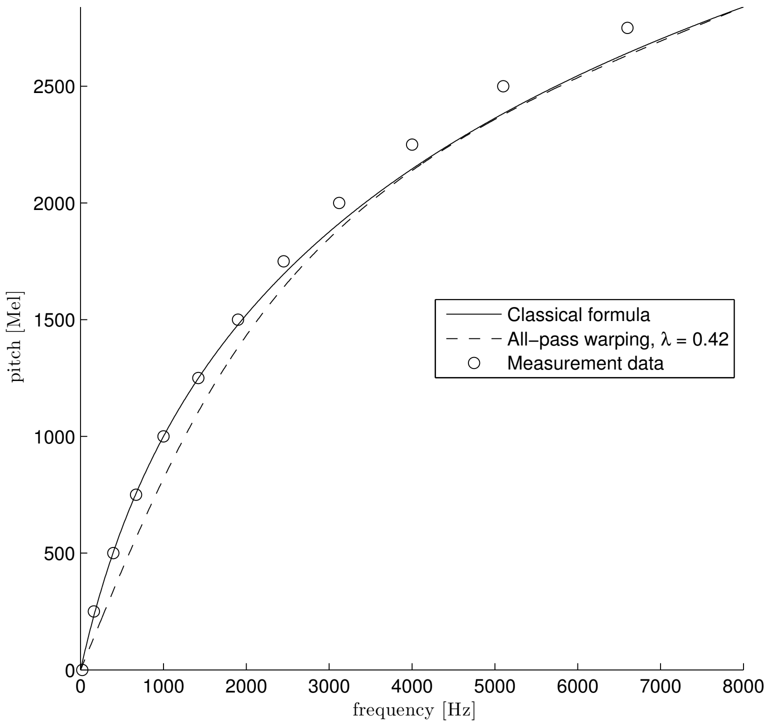


Figure 5: Mel–frequency warping curve at 16 kHz sample rate. The optimal warping coefficient depends on the sample rate. Measurement data points from [4] are listed in Table A1.

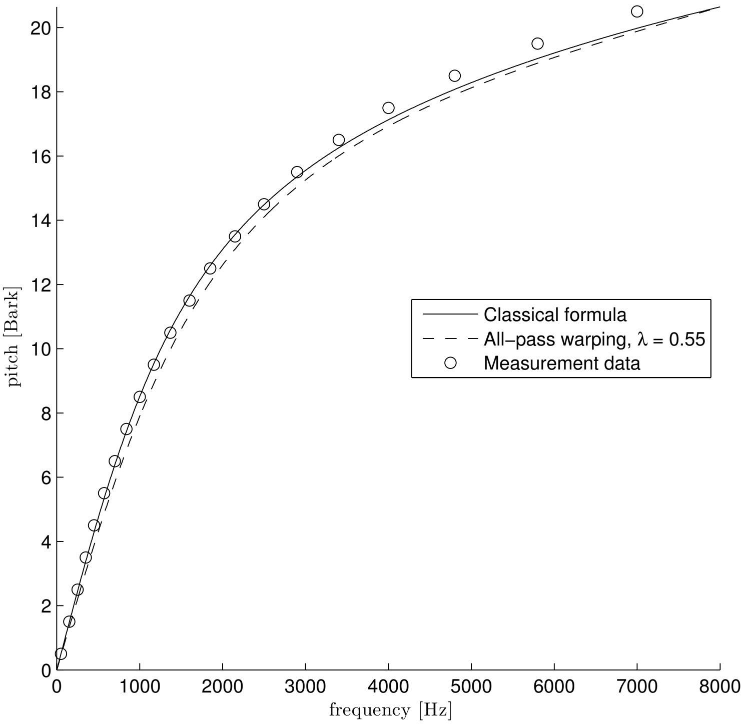


Figure 6: Bark–frequency warping curve at 16 kHz sample rate. Measurement data points from [62] are listed in Table A2.

3 Speech parametrization

3.1 Auditory models in speech parametrization

Auditory models are computational models that replicate some properties of human sound perception. Important aspects in auditory modelling include the nonlinear frequency resolution of hearing, the processing of spectral information on critical bands, and the sound pressure level and frequency dependence of perceived loudness. Nonlinear frequency resolution requires a model to implement some form of frequency warping. This property often comes embedded in filterbank design when using auditory filterbanks to model processing in critical bands. The following section focusses on the auditory modelling properties of the mel-scale filterbanks that are used as the base of the spectral matching scheme introduced in Chapter 5.

3.1.1 Frequency warping

Frequency warping is a general term for techniques for obtaining a frequency domain signal representation with a non-uniform frequency resolution. Usually frequency warping is used to model the nonlinear frequency resolution of human sound perception, and the warping function is chosen to approximate a pitch scale such as the mel-scale or the Bark-scale. Various methods can be used to realize the warping, yielding similar results.

Theoretically, any frequency warping function with certain limitations can be applied. First, the map should be monotonic and injective for the warped frequency to remain interpretable as frequency. Second, it is reasonable that zero frequency maps to zero. Third, the map should be antisymmetric with respect to the zero frequency, in order to map positive and negative frequencies similarly. Additionally, when dealing with sampled signals, the Nyquist frequency should map onto itself.

A simple method for calculating a warped magnitude spectrum from the fast Fourier transform (FFT) is to define a warped frequency axis and find the corresponding values by interpolating on the FFT values. This is proposed in the context of linear prediction and mel-scale warping by Makhoul and Cosell [31]. A detailed procedure using linear interpolation is presented by Yapanel et al. [57]. The attractiveness of the interpolation approach comes from its computational efficiency: both linear interpolation and the FFT are relatively fast to compute.

An arbitrary warping function can also be realized directly by using a warped discrete Fourier transform (DFT) matrix. The calculation of the warped spectrum $\tilde{\mathbf{X}}$ is analogous to the conventional DFT matrix: $\tilde{\mathbf{X}} = \tilde{\mathbf{F}}\mathbf{x}$, where $\tilde{\mathbf{X}} \in \mathbb{C}^N$ is the warped spectrum , $\tilde{\mathbf{F}} \in \mathbb{C}^{N \times N}$ is the warped DFT matrix, N is the discrete spectrum length, and $\mathbf{x}$ is a windowed signal padded to length N. $\tilde{\mathbf{F}}$ is constructed so that $\tilde{\mathbf{F}}_{k,n} = e^{-j\tilde{\omega}_k n}$, where $n = 0, \ldots, N-1$, $\tilde{\omega}_k = \frac{2\pi}{N}k$ and k are uniformly placed on the warped scale. The resulting warped spectrum is complex valued and conceptually directly analogous to the complex DFT spectrum. [26]

A specifically useful warping scheme can be constructed by using a simple conformal map. This scheme has one free parameter λ that can be adjusted to approximate mel and Bark scales with surprisingly high accuracy, as illustrated in Figures 5 and

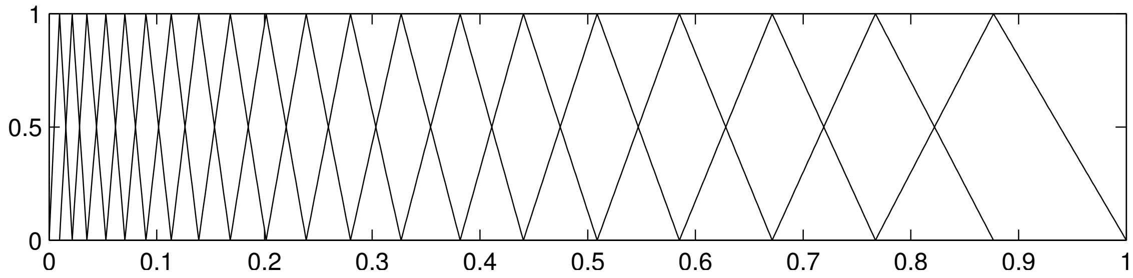


Figure 7: Triangular mel-scale filterbank on normalized frequency axis where 1 corresponds to the Nyquist frequency and filterbank amplitude is set to unity.

6. Additionally, the resulting formulation enables the design and use of warped filters, that have utility in audio signal processing generally. The map is derived by setting the zero and Nyquist frequencies as fixed points in a general first-order conformal map [47]

$$\frac{(\tilde{z}-\tilde{z}_1)(\tilde{z}_2-\tilde{z}_3)}{(\tilde{z}_2-\tilde{z}_1)(\tilde{z}-\tilde{z}_3)} = \frac{(z-z_1)(z_2-z_3)}{(z-z_1)(z-z_3)}, \tag{4}$$

where $z, \tilde{z} \in \mathbb{C}$. The map fulfils the necessary properties for frequency warping when the unit circle is set to map onto itself. Substitution of the fixed points at zero and Nyquist frequencies, $(z_1 = \tilde{z}_1 = 1)$ and $(z_2 = \tilde{z}_2 = -1)$ respectively, leads to a transformation of form

$$\tilde{z} = \frac{z-\lambda}{1-\lambda z}, \qquad \lambda = \frac{\tilde{z}_3 + z_3}{1 - z_3\tilde{z}_3}, \tag{5}$$

where λ is a free parameter determining the rate of warping. The above relation directly gives $\tilde{z}^{-1} = \frac{z^{-1}-\lambda}{1-\lambda z^{-1}}$. As z^{-1} denotes the unit delay, this type of warping corresponds to replacing the unit delay with a first-order all-pass filter. Practical use of this all-pass type warping with linear prediction is covered in Section 3.2.3.

3.1.2 Mel-scale filterbanks

The mel-scale filterbank is a set of filters with triangular frequency response, whose centre frequencies are evenly spaced on the mel-scale. This type of filter structure is widely used for auditory spectral modelling in speech recognition and speaker verification, specifically as a part of the mel-frequency cepstral coefficient calculation [17, 37, 41]. The filterbank output represents a smoothed spectrum with linear resolution on the mel-scale, and serves as a compact and easily computable auditory spectrum.

Conventionally the mel-scale filterbank is applied directly to either the FFT power or magnitude spectrum. Both types of spectra are used but the choice varies depending on the study, (see Section 3.1.3) for discussion. This thesis uses a definition based on the power spectrum, and the filterbank is constructed as follows: First, find the filter centers by determining the FFT indices that correspond to a linear spacing on the mel-scale, using the mel-to-frequency mapping of Eq. (3). Then set the filter value at each center index to one and connect it to the neighbouring centres with linear ramps to zero. This results in a set of pairwise overlapping triangular filters, as illustrated in Figure 7. Denote the i:th filter with $H_i(k)$. To calculate the filterbank output, each filter is applied to the positive frequencies of the FFT power spectrum $|X(k)|^2$, where k indexes the FFT from zero frequency to the Nyquist frequency. Then the i:th filter output is given by

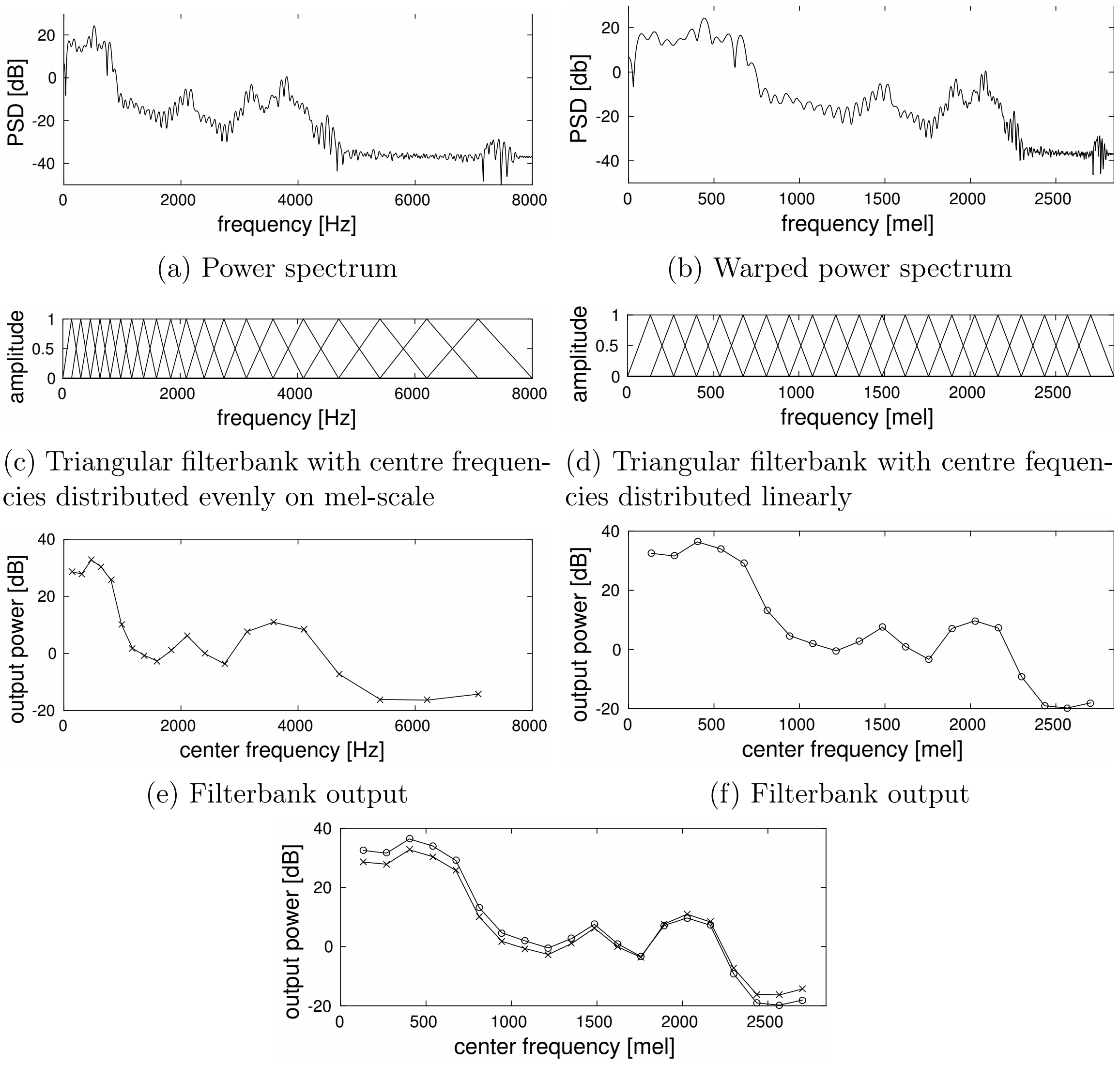
$$\mathcal{X}_i = \sum_{k=0}^{N-1} H_i(k)|X(k)|^2 \tag{6}$$

An alternative mel-filterbank formulation that has some interesting interpretative properties can be constructed by warping the full length FFT spectrum and applying a triangular filterbank with uniform filter bandwidths on it. A visual comparison of these two methods is presented in Figure 8. The conventional, or direct mel-filterbank calculation, is presented on the left column, and the warped calculation on the right column. Figures 8a and 8b show an example of a power spectrum and the corresponding warped power spectrum, respectively. The filterbanks applied on the power spectra are shown below in Figures 8c and 8d. The resulting outputs of the filterbanks are not exactly equivalent, but are nevertheless highly similar and based on the same high level concepts.

Regardless of its relative simplicity, the mel-filterbank structure has some properties that make it useful as an auditory spectrum model. First, the model output has a nonlinear frequency resolution that is set to approximate the auditory mel-scale. Second, the application of filterbanks leads to frequency band integration that is similar to the critical band model discussed in Section 2.2.2. With certain assumptions the filterbank output can be shown to be a *smoothed auditory spectrum sampled at the filterbank centre frequencies*. This type of filter bank summation is also conceptually related to auditory excitation pattern calculations, where a warped spectrum is convolved in frequency domain with a smoothing kernel function. This is covered in more detail in Appendix B.

A full length smoothed spectrum can be reconstructed from the mel-filterbank output by making some assumption of the spectral shape. We choose to work in the warped spectral domain, as the notion of the filterbank output as a sampled version of the smoothed spectrum becomes useful: the reconstruction can be seen simply as interpolation between the sampled points. Figure 9 presents an example power spectrum in the warped domain. The filterbank outputs representing samples of the smoothed spectrum are plotted with circles with the nearest-value and linear interpolation reconstructions. A matrix formulation for these reconstruction operations is presented in detail in Section 5.2.2. These can be combined with the matrix representation of the filterbank into a smoothing-operator matrix whose visualization can provide some insight into the nature of the smoothing.

Smoothing matrices with piecewise constant and piecewise linear reconstructions are shown in Figures 10a and 10b, respectively. Both are centred around the diagonal and thus show a resemblance to the unit matrix corresponding to a perfect reconstruction. However, the mel-filterbank reconstructions are not perfect due



(a) Power spectrum

(b) Warped power spectrum

(c) Triangular filterbank with centre frequencies distributed evenly on mel-scale

(d) Triangular filterbank with centre fequencies distributed linearly

(e) Filterbank output

(f) Filterbank output

(g) Both filterbank outputs on mel-scale

Figure 8: Comparison of direct (left column) and warped (right column) methods to calculate the mel-filterbank outputs of an /o/ vowel. The filterbank outputs are superimposed on the bottom picture to illustrate their similarity.

to the smoothing effect that can be seen as spreading around the diagonal. For comparison, a least-squares optimal reconstruction using the pseudoinverse of the filterbank matrix is presented in Figure 10c. The pseudoinverse smoother has the undesirable property of having negative values, thus making its use infeasible in this work. Similar negativity issues arise if higher order polynomial interpolations are used for reconstruction. Nevertheless, the pseudoinverse exhibits triangular main lobes similar in shape to the linear reconstruction, which suggests that the linear reconstruction approximates the pseudoinverse reasonably well. This is illustrated by the similarity of spreading patterns in Figures 10b and 10c.

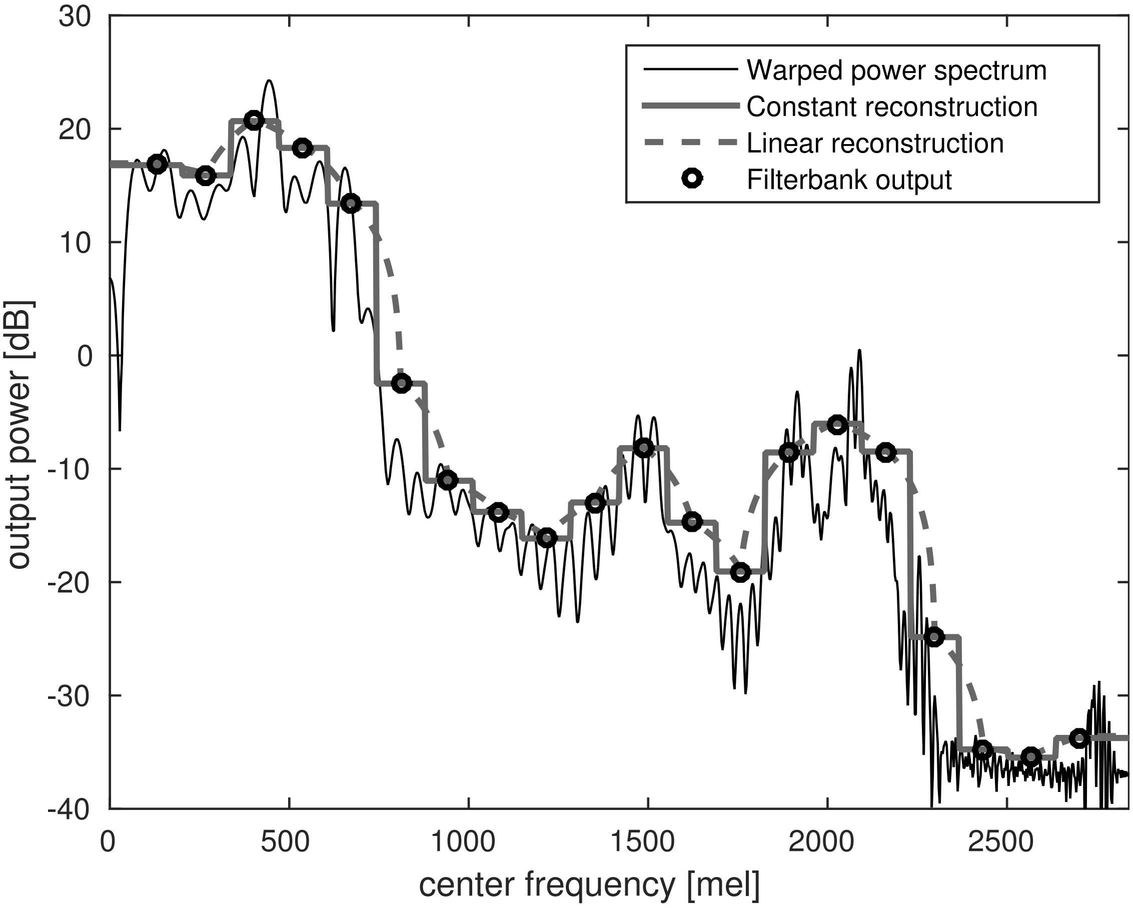
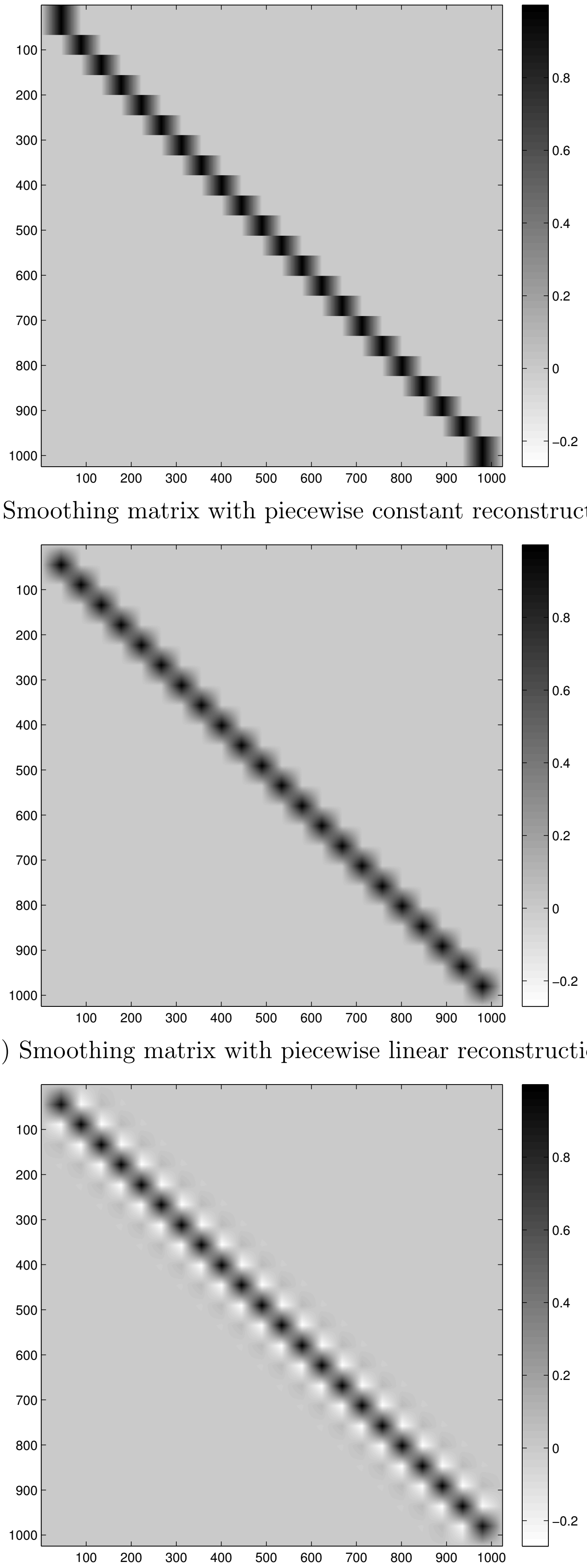


Figure 9: Mel filterbank as a spectral smoother – different spectral reconstructions in warped frequency domain. Linear reconstruction is piecewise linear in the power spectral domain and therefore piecewise logarithmic on the dB-scale

3.1.3 Mel-frequency cepstral coefficients

Mel-frequency cepstral coefficients (MFCC) [10] is a cepstrum domain representation of the mel-filterbank information. The MFCC are widely used in speech recognition and speaker verification due to their fast computation, compactness, relatively good representation of the auditory spectrum and good decorrelation properties. A mel-



(a) Smoothing matrix with piecewise constant reconstruction

(b) Smoothing matrix with piecewise linear reconstruction

(c) Smoothing matrix with filterbank matrix pseudoinverse

Figure 10: Triangular filterbank combined with various spectral reconstructions can be presented as spectral smoothing matrices.

scale filterbank is applied on the FFT spectrum and a discrete cosine transform (DCT) [1] is applied on the logarithm of the filterbank outputs.

There is some variation on whether the filterbank is applied on magnitude or power spectrum. Davis and Mermelstein [10] use the magnitude spectrum in the calculation, as does the European Telecommunications Standards Institute (ETSI) standard for MFCC in speech recognition [12]. In this case, the filterbank output can be interpreted as a smoothed magnitude, but the filterbank no longer preserves the total energy of the spectrum in the 2-norm sense. In contrast, Huang et al. present their version of MFCC with integration on the power spectrum [17]. With summation in power domain the filterbank output remains in the power domain, corresponding to the weighted total power on frequency band. Now, if the overlapping filters sum to unity, the total energy is preserved in the filterbank output. This property is lost when using filters with area normalization.

In conventional cepstrum, a Fourier transform is applied on the two-sided log-magnitude spectrum [42, p. 364]. The mel-cepstrum uses a similar idea, with the distinction that the FFT is replaced with type-II DCT. Since the mel-filterbank outputs $\mathcal{X}_j$ are calculated from, and correspond to, the positive frequencies of a symmetric power spectrum, the symmetry property can be extended to the mel-spectral representation. Then the application of the DCT on the one-sided mel-spectrum corresponds to application of the FFT on a two-sided mel-spectrum with even symmetry. The k:th mel-frequency cepstral coefficient is given by

$$c_k = \sum_{j=1}^{M} \log(\mathcal{X}_j) \cos\left(\frac{\pi k}{M}\left(j - \frac{1}{2}\right)\right), \tag{7}$$

where M is the size of the filterbank. The standard choices for speech recognition are $M = 23$ and $k = 0 \ldots 12$ [12]. Truncating the cepstrum in this manner provides dimensionality reduction which is beneficial in speech recognition, and smooths the spectral representation further. In addition to dimensionality reduction, the DCT is used to decorrelate the transform-domain representation. It has been shown that for signals with a specific type of covariance matrix, the DCT approximates the the Karhunen-Loève transform, which is the least-squares optimal linear decorrelator [1].

3.2 Linear predictive coding

Linear predictive coding (LPC) models the speech as an autoregressive process and estimates the optimal model coefficients using the MSE criterion. The technique has proven very useful in modelling the spectral characteristics of speech. Typically speech signal is analysed in short segments, and the underlying statistical process is assumed to be stationary within the segment. The classical formulation for linear prediction is given by Makhoul [30]. The following presentation is based on Rabiner and Schafer [42].

The AR-model for speech underlying linear prediction corresponds to a source-filter model, where the filter is of all-pole type. Figure 11 shows a schematic of this model. For voiced speech the source consists of an impulse train with impulse spacing

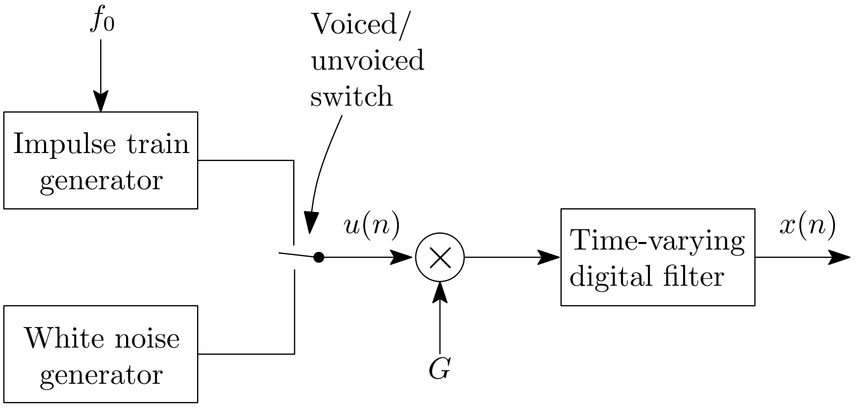


Figure 11: Simplified model for speech production, after [42]

matched to f_0. For unvoiced speech random noise is used for source. According to the model the system output $x(n)$ at time instant n depends only on the current input and past outputs

$$x(n) = \sum_{k=1}^{p} a_k x(n-k) + Gu(n), \tag{8}$$

where $u(n)$ is the input value of the filter at time n and G is a gain parameter that is assumed to be constant within the speech segment. To estimate the coefficients in this speech model we build a linear model that predicts signal value from its past values. Such linear predictor $\hat{x}(n)$ with prediction coefficients α_k is defined by

$$\hat{x}(n) = \sum_{k=1}^{p} \alpha_k x(n-k). \tag{9}$$

Prediction error is defined as the difference between the actual and predicted values. The error signal is also called the LPC residual:

$$e(n) = x(n) - \hat{x}(n) = x(n) - \sum_{k=1}^{p} \alpha_k x(n-k). \tag{10}$$

Prediction error filter, also called the analysis filter, is defined as

$$A(z) = 1 - \sum_{k=1}^{p} \alpha_k z^{-k}. \tag{11}$$

The analysis filter can be used in whitening the speech spectrum, a property which is seen from the relation

$$G \cdot U(z) = A(z)X(z). \tag{12}$$

3.2.1 Solving the optimal predictor

The optimal predictor is solved by finding the predictor coefficients that minimize the mean squared error σ^2

$$\sigma^2 = E\left[e_n(m)^2\right] = E\left[\left(x_n(m) - \sum_{k=1}^{p} \alpha_k x_n(m-k)\right)^2\right], \tag{13}$$

where $E[\cdot]$ is the expected value operator. The error function is quadratic and therefore has a single global minimum at

$$\frac{\partial \sigma^2}{\partial \alpha_i} = 0, \;\; i = 1, \ldots, p. \tag{14}$$

Minimization leads to the set of LPC normal equations for $i = 1, \ldots, p$:

$$E\left[x_n(m-i)x_n(m)\right] = \sum_{k=1}^{p} \alpha_k E\left[x_n(m-i)x_n(m-k)\right]. \tag{15}$$

The expected value operation is related to the autocorrelation method of linear prediction, and minimizes σ^2 on the infinite duration $-\infty < m < \infty$. Practically this is done by assuming $x_n(m)$ to be zero outside the analysis frame. The left hand side of (15) directly equals the autocorrelation $R_n(i)$. With an index change $m \rightarrow m + i$ in the summation, the right hand side can also be written in terms of autocorrelation

$$E\left[x_n(m-i)x_n(m-k)\right] = E\left[x_n(m+i-k)x_n(m)\right] = R_n(i-k). \tag{16}$$

Since autocorrelation is symmetric, $R_n(j) = R_n(-j)$, (15) becomes

$$\sum_{k=1}^{p} \alpha_k R_n(i-k) = R_n(i), \;\; i = 1, \ldots, p. \tag{17}$$

This can be written in matrix form as

$$\begin{bmatrix} R_n(0) & R_n(1) & R_n(2) & \cdots & R_n(p-1) \\ R_n(1) & R_n(0) & R_n(1) & \ddots & R_n(p-2) \\ R_n(2) & R_n(1) & R_n(0) & \ddots & R_n(p-3) \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ R_n(p-1) & R_n(p-2) & R_n(p-3) & \cdots & R_n(0) \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R_n(1) \\ R_n(2) \\ R_n(3) \\ \vdots \\ R_n(p) \end{bmatrix} \tag{18}$$

The resulting autocorrelation matrix has a Toeplitz structure, i.e., it is symmetric and has $R_n(i-1)$ on its i:th diagonal. Equation (17) can be solved efficiently by using the recursive Levinson–Durbin algorithm [11, 28].

The optimal predictor coefficients can also be solved directly in matrix form by treating linear prediction as a constrained optimization problem [6]. Consider the prediction error, or residual for the n:th sample $e_n = \mathbf{x}^\top \mathbf{a}$, where $\mathbf{x}$ is the signal and $\mathbf{a}$ contains the predictor coefficients. Linear prediction is then equivalent to minimizing the residual energy $\sigma^2 = E\left[e_n^2\right] = E\left[\mathbf{a}^\top \mathbf{x}\mathbf{x}^\top \mathbf{a}\right] = \mathbf{a}^\top E\left[\mathbf{x}\mathbf{x}^\top\right]\mathbf{a} = \mathbf{a}^\top \mathbf{R}\mathbf{a}$ where $\mathbf{R} \in \mathbb{R}^{p+1\times p+1}$ is an autocorrelation matrix. To make this minimization problem solvable, constrain $a_0 = 1$, or equivalently in vector notation $\mathbf{a}^\top \mathbf{v} = 1$ with $\mathbf{v} = \left[1, 0, 0, \cdots\right]^\top$. This constrained optimization problem can be solved with the Lagrange multiplier technique. The objective function becomes

$$L(\mathbf{a}, \lambda) = \mathbf{a}^\top \mathbf{R}\mathbf{a} + \lambda(\mathbf{a}^\top \mathbf{v} - 1). \tag{19}$$

Setting $0 = \frac{\partial L}{\partial \lambda}$ forces the constraint and $0 = \frac{\partial L}{\partial \mathbf{a}} = 2\mathbf{R}\mathbf{a} + \lambda \mathbf{v}$ gives the constrained solution. The Lagrange multiplier λ can be eliminated from the solution by left multiplication with $\mathbf{a}^\top$:

$$\begin{aligned} 2\mathbf{a}^\top \mathbf{R}\mathbf{a} &= -\lambda \mathbf{a}^\top \mathbf{v}, \\ \sigma^2 = \mathbf{a}^\top \mathbf{R}\mathbf{a} &= -\lambda/2. \end{aligned}$$

The optimal $\mathbf{a}$ is thus given by

$$\mathbf{R}\mathbf{a} = \sigma^2 \mathbf{v}, \tag{20}$$

where $\mathbf{R}_{i,j} = R(|i - j|)$ for $i, j = \{0, \ldots, p + 1\}$. The solution is related to the conventional formulation by $\mathbf{a}_{k+1} = -\alpha_k$ for $k = 0, \ldots, p$. A modified Levinson–Durbin recursion can be applied for efficient solution.

3.2.2 Line spectral pairs

The prediction error filter polynomial $A(z)$ can be decomposed to symmetric and antisymmetric parts $A(z) = \frac{1}{2}\left[P(z) + Q(z)\right]$, where

$$P(z) = A(z) + z^{-(p+1)}A(z^{-1}), \tag{21}$$

$$Q(z) = A(z) - z^{-(p+1)}A(z^{-1}). \tag{22}$$

This representation has the following useful properties [48]:

1. All zeros of $P(z)$ and $Q(z)$ lie on the unit circle;

2. Zeros of $P(z)$ and $Q(z)$ are interlaced if and only if $A(z)$ is minimum phase;

3. Zeros of $P(z)$ and $Q(z)$ are distinct.

By the first property, each root of $P(z)$ and $Q(z)$ is of form $e^{j\omega_k}$ and is thus determined by a single argument, the angular frequency ω_k. This is called the line spectral frequency (LSF) of the root. Since the frequency fully determines the root location it suffices to store only the frequency. Additionally, as $P(z)$ and $Q(z)$ have real coefficients, the roots form complex conjugate pairs on the unit circle, which facilitates the numerical evaluation of the roots. In practice, the evaluation can be done for example by utilizing Chebyshev polynomials [23].

LSF representation is robust in terms of quantization and interpolation: the minimum phase property of $A(z)$ is easily preserved in quantization by ensuring the above LSF properties are fulfilled [38]. Resulting synthesis filter stability is then guaranteed by the minimum phase property. Additionally, LSFs have been found to behave well as HMM speech synthesis features [15,43,45]. A theoretical review with proofs of LSF properties is given by Bäckström and Magi [7].

3.2.3 Warped linear prediction

The idea of warped linear prediction is to modify the minimization criterion for linear prediction in a manner that gives a higher resolution at low frequencies. An early suggestion for linear prediction method used with arbitrary warping function is given by Makhoul [31]. Specifically, a mel-warping as in Eq. (3) was used. This approach suffers from the necessity of unwarping the warped LPC spectrum for synthesis filter realization.

In modern context, warped linear prediction is primarily associated with warping achieved by replacing unit delays with first order all-pass elements, and the corresponding warped filter realizations, as presented by Strube [50]. The first-order all-pass element has unit magnitude response and a nonlinear phase response, and is therefore also called a diffuse delay element. Specifically, the diffuse delay element used for warping is given in the z-domain by

$$\tilde{z}^{-1} = D_1(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}} \ , \tag{23}$$

where λ is the warping parameter in the range $[-1, 1]$. Using $\lambda > 0$ bends the frequency axis so that the low frequencies obtain higher resolution than the high frequencies, which is desired for the approximation of auditory resolution. The frequency warping function is determined by the phase response of the mapping [36]. Thus the warped angular frequency $\tilde{\omega}$ is given by

$$\tilde{\omega} = \arctan\left(\frac{(1-\lambda^2)\sin(\omega)}{(1+\lambda^2)\cos(\omega) - 2\lambda}\right), \tag{24}$$

where ω is the non-warped angular frequency. With suitable choice of λ, this warping curve approximates mel or Bark scale warping quite closely, as illustrated in Figures 5 and 6.

Next we formulate the warped linear prediction similarly to the standard linear prediction presented previously. In order to work in the time domain, define a generalized delay operator

$$d_k[x(n)] = \underbrace{d_1(n) * d_1(n) * \cdots * d_1(n)}_{k\text{-fold convolution}} * x(n), \tag{25}$$

where d_1 is the impulse response of $D_1(z)$. The warped linear prediction error can now be written similarly to Eq. (13)

$$\tilde{\sigma}^2 = E\left[\tilde{e}^2\right] = E\left[\left(x(n) - \sum_{k=1}^{p} \tilde{\alpha}_k d_k[x(n)]\right)^2\right]. \tag{26}$$

Minimization leads to normal equations analogous to Eq. (15):

$$E\left[d_i[x(m)]d_0[x(m)]\right] = \sum_{k=1}^{p} \alpha_k E\left[d_i[x(m)]d_k[x(m)]\right]. \tag{27}$$

Since D_1 is a first order all-pass filter, it is straightforward to show that similar index change to Eq. (16) can be applied. This modifies the warped normal equations for $i = 1, \ldots, p$ to

$$\sum_{k=1}^{p} \tilde{\alpha}_i E\left[d_{i-k}[x(n)]d_0[x(n)]\right] = E\left[d_i[x(n)]d_k[x(n)]\right] \tag{28}$$

$$\sum_{k=1}^{p} \tilde{\alpha}_i \tilde{R}(i-k) = \tilde{R}(i), \tag{29}$$

where $\tilde{R}$ is the warped autocorrelation. Equation (29) leads to the same matrix equation form with Eq. (17), and it can be solved likewise with the Levinson–Durbin algorithm.

The resulting warped linear predictor has some equivalent properties to the non-warped version. The linear prediction analysis filter is defined by replacing the unit delays with dispersive delays $\tilde{z}$:

$$\tilde{A}(z) = 1 - \sum_{k=1}^{p} \tilde{\alpha}_k \tilde{z}^k. \tag{30}$$

The warped linear prediction polynomial has the minimum phase properties of the non-warped version, thus the LSF and other (e.g. reflection coefficient) representations of LPC information can be used directly. The warped filter coefficients can also be converted into conventional filter coefficients and vice versa [50]. However, numerical stability issues arise when realizing warped synthesis filters with non-warped filter structures, so the warped LPC information is usually kept in the warped domain.

Synthesis filter implementation with warped infinite impulse response (WIIR) structures is non-trivial because a delay free loop is present in a straightforward setting. A modification for preventing this is presented already by Strube [50]. Realizable warped filters and their applications are studied more closely in Härmä et al. [20, 21, 25], and the warped filters used in this work are based on the related toolbox by the authors.

The choice of warping coefficient depends on the desired warping curve and the sample rate. Choice of $\lambda = 0$ simplifies the first order all-pass into a unit delay and does not warp the frequency axis at all, whereas values closer to $\lambda = 1$ lead to more extreme warping. Typically λ is chosen so that the warping curve approximates some pitch scale. Some optimal values for mel and Bark scale approximation with various sampling frequencies are presented in Table 1. Techniques for deriving optimal warping parameters are presented by for example Smith and Abel [47].

Table 1: Optimal warping coefficient values for different sampling rates [54].

Sampling frequency	8 kHz	10 kHz	12 kHz	16 kHz	20 kHz	22.05 kHz
mel scale	0.31	0.35	0.37	0.42	0.44	0.45
Bark scale	0.42	0.47	0.50	0.55	–	–

4 Statistical parametric speech synthesis

In statistical parametric speech synthesis the speech signal is coded into a parametric representation which enables the use of statistical modelling for first learning a model relating speech parameter sequences with corresponding word sequences, and second using this model to generate a speech parameter sequence from an arbitrary word sequence. A good overview of hidden Markov model (HMM) based statistical parametric speech synthesis is given in [61].

In speech synthesis, we are interested mainly in two types of statistical inference problems: first how to train the statistical model so that it best describes the relationship between given acoustical and text data, and second how to use the model to generate acoustical features given some text sequence. This chapter gives a short general overview of the methodology and describes in some detail how HMMs can be used for statistical modelling in speech synthesis.

The model training problem can be stated as follows: given the model structure $\mathbf{\Phi}$ and the set of text sequences $\mathscr{W}$ corresponding to the set of acoustic observation sequences $\boldsymbol{O}$, find the model parameters $\hat{\mathbf{\Phi}}$ that maximize the likelihood of $\boldsymbol{O}$

$$\hat{\mathbf{\Phi}} = \arg\max_{\mathbf{\Phi}}\{p(\boldsymbol{O}|\mathscr{W}, \mathbf{\Phi})\}. \tag{31}$$

Conversely, the acoustic observation sequence generation problem is: given a word sequence w and the model parameters $\hat{\mathbf{\Phi}}$, find the output sequence $\hat{\boldsymbol{o}}$ that maximizes the likelihood of generating the sequence $\boldsymbol{o}$

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}}\{p(\boldsymbol{o}|w, \hat{\mathbf{\Phi}})\}. \tag{32}$$

In principle any generative statistical modelling method can be applied in achieving these tasks. The HMMs have been the most prominent method and are also used in this work. However, recently deep neural networks have given some promising results when used for the generative modeling task instead of HMMs [59].

4.1 Hidden Markov models

Hidden Markov model (HMM) is a tool suitable for modelling data with a time-structure, such as speech. The HMMs were initially published by Baum et. al in the late 1960s to early 1970s. They gained popularity in the speech processing community in the 1980s and have gained specific success in speech recognition and synthesis. The following presentation is based on [40] and [17, p. 380-383]

4.1.1 Model structure

The HMM is a finite state machine that involves of two types of random processes: First, the underlying Markov process which holds the information of the current state of the system and probabilities of transitions between the states. Second, the random process that generates observations at state transitions with probabilities depending on the current state. An example HMM typical for speech synthesis is presented in Fig. 12. The system state is hidden, but the most likely state or state sequence can be inferred from the observations. In terms of modelling, the system state represents a physical state of the system, whereas the observed value represents the output signal of the system.

The underlying mathematical model in HMMs is the first-order Markov process, where the transition probabilities to the next state depends only on the current state. If the process has N possible states, they define a state space $S = \{s_1, s_2, \ldots, s_N\}$. Denote the system state at time index t with q_t. In a Markov process the probability of getting to a state q_t depends only on the previous state of the system:

$$p(q_t \mid q_{t-1}, \ldots, q_0) = p(q_t \mid q_{t-1}) \tag{33}$$

This is called the Markov property. When the assumption holds, a state transition probability matrix $\mathbf{A} = \{a_{ij}\}$ can be used to describe the system. Here $a_{ij} = P(q_t = s_j \mid q_{t-1} = s_i)$ is the transition probability from state i to state j. To fully describe a Markov process, also the initial state probabilities $\boldsymbol{\pi} = \{\pi_i\}$ are needed, where π_i is the probability $P(q_0 = s_i)$.

In HMMs the state of the system cannot be observed directly. Instead, a state transition emits an observation with some probability depending on the current state of the system. If the output quantity is discrete, or easily discretised by quantization, the possible outputs are represented by an observation alphabet $V = \{v_1, v_2, \ldots, v_M\}$. The probability of outputting a certain observation O_t at time index t depends on the current state of the system, but is independent of previous observations.

- The state space $S = \{s_1, s_2, \ldots, s_N\}$. For the state of the system q_t at time index t it holds $q_t \in S$

- Output observation alphabet $V = \{v_1, v_2, \ldots, v_M\}$. The outputs can be vector valued

- Transition probability matrix $\mathbf{A} = \{a_{ij}\}$, where $a_{ij} = p(q_{t+1} = j | q_t = i)$ is the state transition probability from state i to state j . If no transitions to previous states are allowed, the HMM is of left-to-right type, and the transition matrix is upper triangular.

- Output probability matrix $\mathbf{B} = \{b_i(k)\}$, where $b_i(k) = p(O_t = v_k \mid q_t = s_i)$ is the probability that the observation O_t emitted by the state $q_t = s_i$ equals v_k. Note that the observed outputs are assumed independent: $p(O_t | O_1^{t-1}, q_1^t) = p(O_t | q_t)$

- Initial state probabilities $\boldsymbol{\pi} = \{\pi_i\}$, where π_i is the probability of $q_0 = i$

The HMM model $\mathbf{\Phi}$ is then the composite of these

$$\mathbf{\Phi} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi}). \tag{34}$$

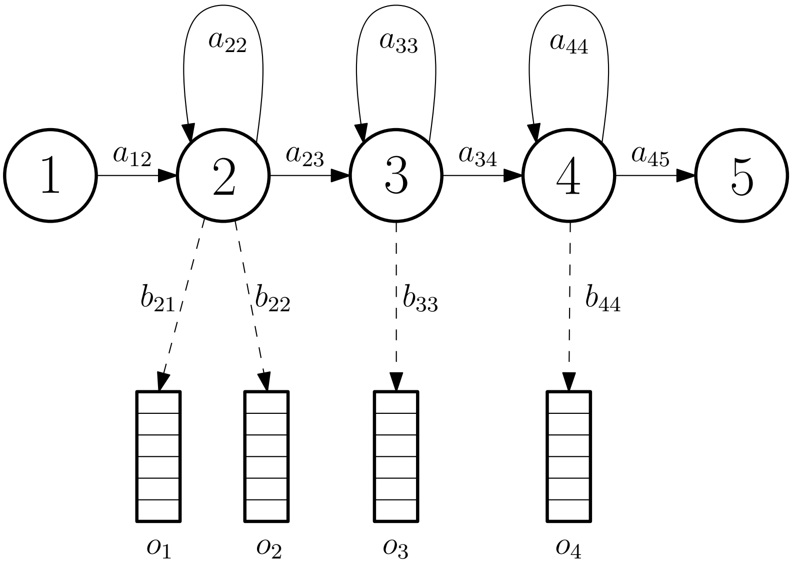


Figure 12: An example of a left-to-right HMM structure, after [24]. The speech synthesis system in this work uses this type of HMMs.

4.1.2 Learning problem

The first important problem with the HMM-based speech synthesis is the learning problem, or how to find the model parameters that best describe the training data. The problem is formulated as: given an observation sequence $\boldsymbol{O} = \{O_1, O_2, \ldots, O_T\}$ and the model $\boldsymbol{\Phi} = (\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{\pi})$, how to adjust model parameters to maximize $p(\boldsymbol{O} \mid \boldsymbol{\Phi})$?

There exists no solution in closed form, but the standard approaches for solving the learning problem include the Baum–Welch iteration, the expectation maximization algorithm (EM), and gradient descent type optimization algorithms. These are covered in detail by Rabiner [40], with focus on speech recognition. The same model structure can be used generatively for speech synthesis.

4.1.3 Parameter generation problem

The parameter generation algorithm for HMM-based statistical parametric speech synthesis was first presented by Tokuda et al. [53]. The following presentation is based on an overview given in [61]. The parameter generation problem of Eq. (32) can be simplified by approximations when using HMMs with Gaussian mixture probability densities. Start from the maximum likelihood estimate

$$\hat{\boldsymbol{o}} = \arg\max_{\boldsymbol{o}} \{p(\boldsymbol{o}|w, \hat{\boldsymbol{\Phi}})\}$$

Partition the probability space on every possible state sequence

$$= \arg\max_{\boldsymbol{o}} \{\sum_{\boldsymbol{q}} p(\boldsymbol{o}, \boldsymbol{q}|w, \hat{\boldsymbol{\Phi}})\} \tag{35}$$

Approximate by using the ML state sequence with given w and $\hat{\boldsymbol{\Phi}}$

$$\approx \arg\max_{\boldsymbol{o}} \max_{\boldsymbol{q}} \{p(\boldsymbol{o}, \boldsymbol{q} | w, \hat{\boldsymbol{\Phi}})\} \tag{36}$$

Rewrite joint probability as conditional

$$\propto \arg\max_{\boldsymbol{o}} \max_{\boldsymbol{q}} \{p(\boldsymbol{q} | w, \hat{\boldsymbol{\Phi}}) \cdot p(\boldsymbol{o} | \boldsymbol{q}, \hat{\boldsymbol{\Phi}})\} \tag{37}$$

Insert the ML state sequence estimate $\hat{\boldsymbol{q}}$

$$\approx \arg\max_{\boldsymbol{o}} \{p(\boldsymbol{o} | \hat{\boldsymbol{q}}, \hat{\boldsymbol{\Phi}})\} \tag{38}$$

$$= \arg\max_{\boldsymbol{o}} \{\mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_{\hat{\boldsymbol{q}}}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{q}}})\}, \tag{39}$$

where $\mathcal{N}(\boldsymbol{o}; \boldsymbol{\mu}_{\hat{\boldsymbol{q}}}, \boldsymbol{\Sigma}_{\hat{\boldsymbol{q}}})$ is the multivariate Gaussian distribution related to $\hat{\boldsymbol{q}}$ with a mean sequence $\boldsymbol{\mu}_{\hat{\boldsymbol{q}}}$ and a covariance sequence $\boldsymbol{\Sigma}_{\hat{\boldsymbol{q}}}$, and

$$\hat{\boldsymbol{q}} = \arg\max_{\boldsymbol{o}} \{P(\boldsymbol{q} | w, \hat{\boldsymbol{\Phi}})\} \tag{40}$$

maximizes its state transition probabilities.

In short, the algorithm first finds the state sequence which is most likely associated with the given text sequence, and then uses the Gaussian models of the state sequence to generate an appropriate sequence of speech parameters. Typically, the process involves generating the delta and delta-delta features along with the static features. This is realized by setting suitable constraints in the generation algorithm [53].

4.2 Speech synthesis system

The speech synthesis system used in this work is based on the single library pulse GlottHMM, proposed by Raitio et al. in [43], with the harmonic-to-noise ratio analysed as in [45]. The new contribution of this work is a perceptual spectral matching scheme that is used instead of the original MSE-based glottal source matching.

In this section, we review how the speech parametrization and synthesis is done by the GlottHMM vocoder and also look at the statistical modelling part of the system. The focus is on the vocoder functions, as the HMM modelling framework remains virtually the same and changes are made only on the vocoder. Here the vocoder is presented in analysis-synthesis manner due to the analysis-by-synthesis nature of the novel spectral matching scheme. The spectral matching system is later elaborated on in Chapter 5.

An overview of the speech synthesis system is shown in Figure 13. In the training part, context dependent HMMs are trained using features extracted from speech signal. These features are given by the analysis part of the vocoder that extracts a parametric representation from a speech waveform. The speech parameters are then aligned to context dependent phonetic labels corresponding to the text. After this the parameters and their labels are used to train context dependent HMMs for use in synthesis part.

Synthesis part of the system takes text as input and processes it into a context dependent phonetic labels. The labels are then used to select the most suitable

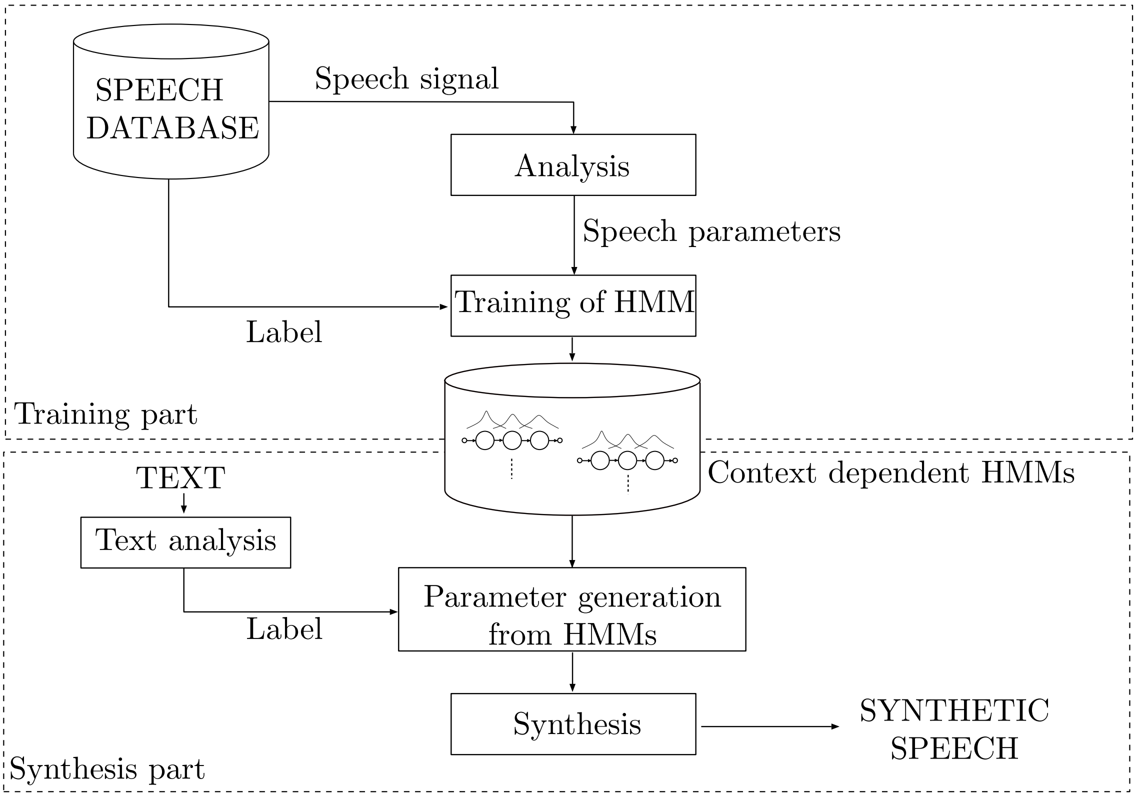


Figure 13: Speech synthesis system overview, after [61]

HMMs from the trained model. This results in a sequence of context dependent single phone HMMs that are catenated to form a sentence HMM which generates a speech parameter sequence on sentence level. Finally, the synthesis part of the vocoder recreates a synthetic speech waveform from the generated parameters.

4.3 GlottHMM glottal source vocoder

The core concept in the vocoder is physiologically motivated estimation of the vocal tract filter, and consequently the glottal volume velocity waveform used in synthesis. This approach results in a more natural excitation signal compared to traditional impulse train excited systems. In practice the vocal tract filter estimation is performed by using glottal inverse filtering techniques.

The speech production model used in the vocoder is based on the source filter model (see Section 2.1.2). Speech spectrum $S(z)$ is regarded to consist of three independent main components: glottal excitation source $G(z)$, vocal tract filter $V(z)$, and the lip radiation effect $L(z)$. The relationship is described by Eq. 1:

$$S(z) = G(z)V(z)L(z).$$

Here the z-domain notation is used for convenience, but it is good to note the importance of the time domain waveform of the glottal excitation. The frequency domain notation is more natural for the lip radiation $L(z)$ which is modelled with a fixed differentiator filter, and $V(z)$ which is an all-pole filter. It should also be noted that the speech model differs from the one used for conventional LPC analysis in Section 3.2. There are two distinctions: first, the glottal excitation signal is not spectrally white, but has a varying spectral tilt that depends on, for example, the speaker and phonation style. Second, the lip radiation effect is now modelled separately and is not lumped into the LPC filter.

Generally, glottal inverse filtering is a technique used for solving the glottal excitation in the speech model

$$G(z) = \frac{S(z)}{V(z)L(z)}. \tag{41}$$

The parameters used in the vocoder aside from the vocal tract filter are mostly describing the glottal excitation. The power of the system comes largely from the naturalness of the excitation and therefore it is sensible to allocate many parameters into describing it.

4.3.1 Iterative Adaptive Inverse Filtering

The glottal inverse filtering method used in the vocoder is the Iterative Adaptive Inverse Filtering (IAIF) [43], originally published in [3]. The inverse filtering is not performed pitch-synchronously in this work, although this is possible with the GlottHMM vocoder. In order to function properly, pitch-synchronous estimation requires accurate glottal closure instant estimation, which is a non-trivial task in itself. This work aims to keep the vocoder setup as simple as possible to effectively evaluate the suggested modification.

At the core of the vocoder lies the IAIF algorithm that estimates an all-pole model for the vocal tract filter using only speech signal as input. A block diagram for the algorithm is presented in Figure (14). Orders of the all-pole models for the vocal tract filter and glottal source spectra are determined by parameters p and q, respectively. In the algorithm, the lip radiation effect is modelled as signal differentiation, and the role of the integration blocks is to cancel the effect.

First the input signal $s(n)$ is highpass filtered with a linear phase finite impulse response (FIR) filter with its cut-off frequency set below a predetermined f_0 lower limit. This step eliminates any possible low frequency hum present in the signal. After this, a first order LP model $H_{\mathrm{g1}}(z)$ is fitted to obtain a preliminary estimate on the contribution of the glottal flow and lip radiation effect on the signal spectrum. This signal is then inverse filtered with $H_{\mathrm{g1}}(z)$ and subjected to p:th order LP analysis to obtain the first estimate for the vocal tract filter $H_{\mathrm{vt1}}(z)$. The iteration then proceeds to re-estimating the glottal flow by inverse filtering the signal with $H_{\mathrm{vt1}}(z)$ and cancelling the lip radiation effect by integration. The refined glottal source q:th order LP-spectrum $H_{\mathrm{g2}}(z)$ is similarly used to estimate the refined vocal tract model, obtaining $H_{\mathrm{vt2}}(z)$. Finally $H_{\mathrm{vt2}}(z)$ is applied to the signal along with integration to obtain an estimate of the glottal flow waveform $g(n)$.

The method has proved to be relatively well suited for vocoding in statistical parametric speech synthesis by the success of the GlottHMM vocoder. However,

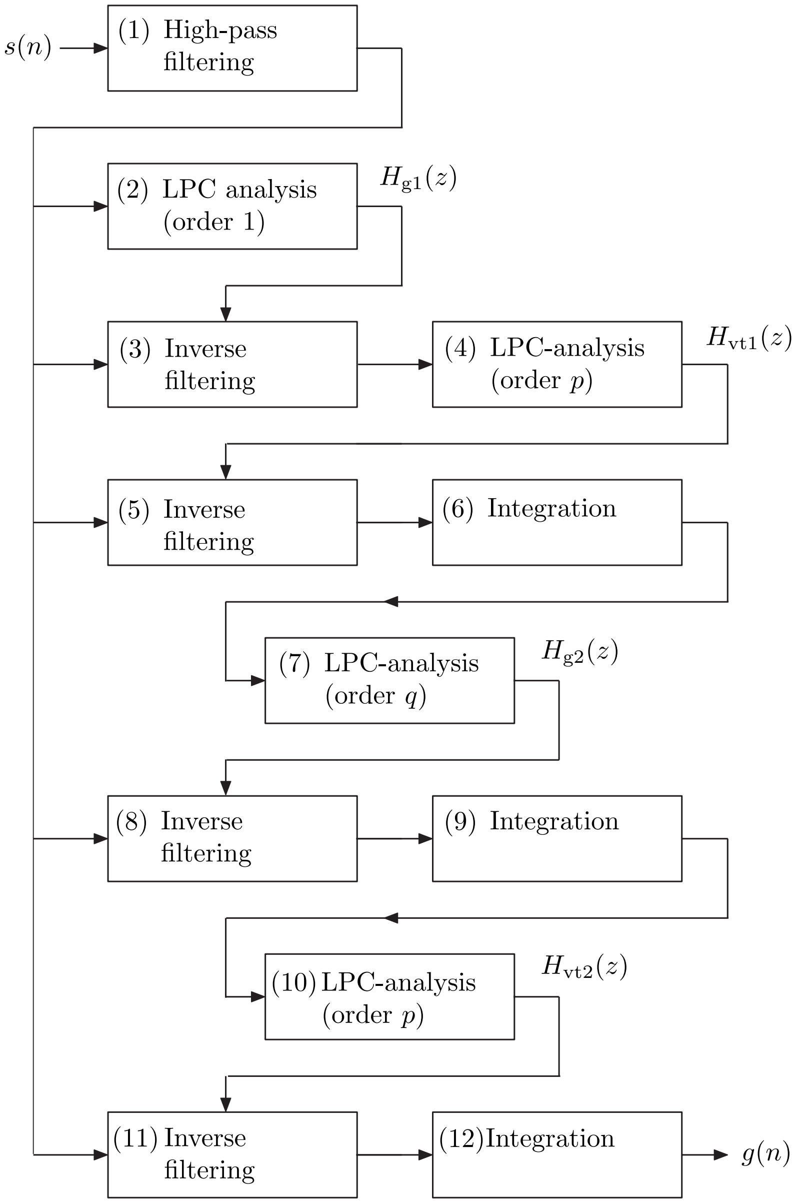


Figure 14: IAIF block diagram

there are some limitations in the method, some of which are specific to the method and some that are intrinsic to all-pole modelling of the vocal tract. One main concern is the lack of an explicit convergence criterion in the method that may cause additional variation between frames in the estimated vocal tract filter. General limitations of source-filter based all-pole vocal tract model include inaccuracy in modelling branching vocal tract, i.e., nasals, and disregarding the possible interaction between the glottal source and vocal tract.

Table 2: Vocoder features and their dimensions

Feature	Dimension
Energy	1
F0	1
Harmonic to noise ratio HNR	5
Source filter LSF **or** Matching filter LSF	10
Vocal tract filter LSF	30

4.3.2 Analysis

In the vocoder, signal is analysed framewise by sliding a 25 ms window at 5 ms intervals. Table 2 lists the vocoder features and their dimensions per frame. An estimate for the vocal tract filter H_{vt} is obtained by the IAIF method as described above, and stored in the LSF representation. The other parameters are used to characterise the glottal excitation.

Fundamental frequency (f_0) is estimated from $g(t)$ by the autocorrelation method. f_0 estimation from autocorrelation is based on the fact that with quasi-periodic signals, the autocorrelation function peaks at lags corresponding to the multiples of the fundamental period. A longer analysis window of 45 ms is used in f_0 estimation to allow several periods in a frame even on low fundamental frequencies. See Rabiner and Schafer [42, p. 150-58] for details on f_0 estimation from autocorrelation.

Harmonic-to-noise ratio (HNR) measures bandwise noise power relative to the power of the harmonic components in voiced speech. Three types of noise sources are captured in HNR. First and mainly, there is the phonation induced additive noise present in the glottal excitation. Second, HNR sees the noise-like aperiodicity effects resulting from slight variation in the duration of adjacent excitation pulses due to naturally varying f_0. Additionally, in voiced consonants, some noise is created by the vocal tract constriction turbulences. This vocal-tract-related noise is not modelled by the all-pole vocal tract filter, and is instead caught in the HNR parameter, making HNR not strictly an excitation parameter.

HNR is calculated on bands obtained by partitioning the frequency axis uniformly on the ERB scale. [45]. The harmonic peaks are detected adaptively by using the estimated f_0 value as an initial guess for the first peak location. The first peak is selected as the local maximum of the FFT-spectrum in the neighbourhood of f_0. The process then continues to search the second harmonic in the neighbourhood of twice f_0. This type of estimation continues for the higher harmonics, while refining the estimate for f_0. After the harmonic locations are detected, the local noise floor level is taken to be the FFT spectrum value at the midpoint between two harmonic peaks. Finally, bandwise HNR is calculated as the average power ratio between the harmonics and the noise floor within the band.

In baseline GlottHMM the LSF-representation of the glottal source LPC-spectrum is stored as a feature and used in synthesis. In the perceptual version, these source

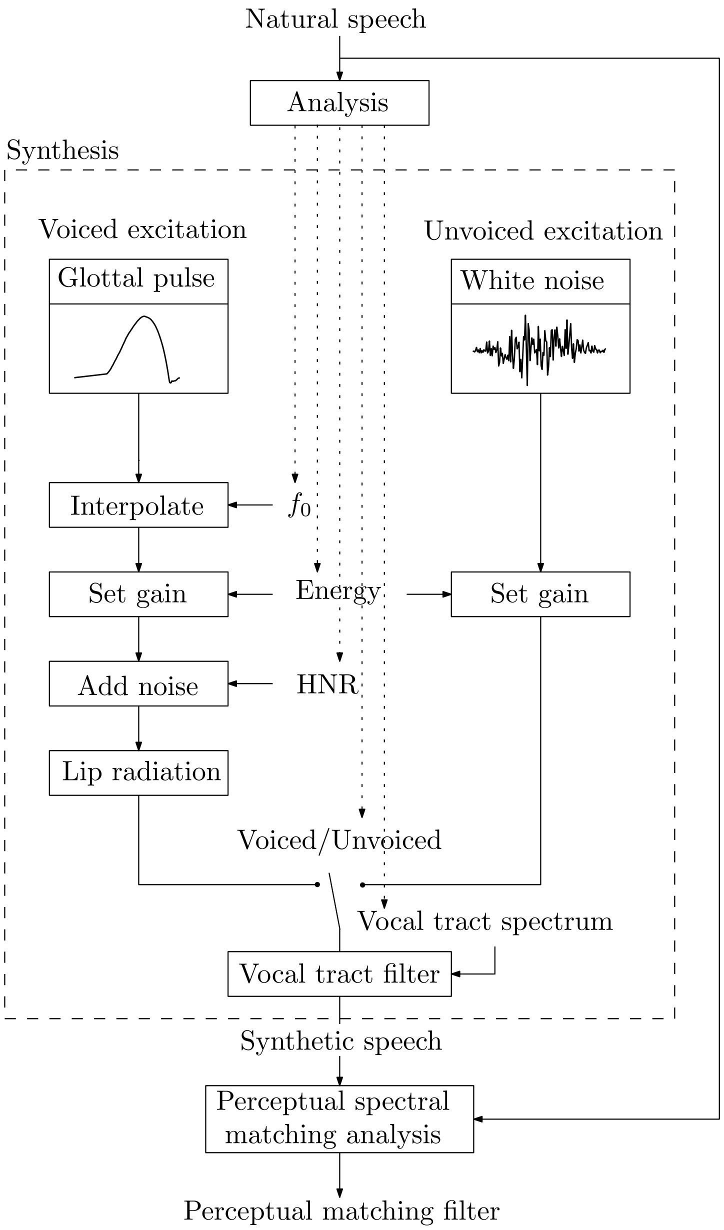


Figure 15: Synthesis procedure and estimation of the perceptual matching filter via analysis-by-synthesis

LSF features are replaced a by perceptual matching filter LSFs. The matching filter for each frame is evaluated by minimizing the mel-spectral distance between the analysed speech and synthetic speech generated from analysed features. Figure 15 shows a general overview of the analysis-by-synthesis scheme. Waveform synthesis from features in elaborated in the next section.

4.3.3 Synthesis

The synthesis part of the system is shown as a white-box block in Figure 15. Voiced excitation is created from a natural glottal flow pulse. First the pulse is interpolated to match the desired f_0 and scaled to suitable energy level. After this noise is added bandwise to match the HNR on each band. Lip radiation effect is applied by a fixed differentiator. For unvoiced excitation, white noise is scaled to match the frame energy, and the other excitation parameters are disregarded. The excitation is fed through the vocal tract filter to create synthetic speech. Finally, the perceptual matching filter is applied to the synthesized speech signal to compensate for the perceptual difference between the synthetic and desired source spectra.

4.3.4 HMM training and parameter generation

The HMM training and parameter generation uses the same system structure as [43]. State durations are modelled separately, making the self-transition probability dependent on the time spent in the state. This makes the system actually use hidden semi-Markov models (HSMM) [29]. The acoustic parameters other than f_0 are modelled as continuous variables in separate streams, and a multi-space distribution (MSD) [52] is used for f_0. This is due to that f_0 is also used as indicator in the voiced–unvoiced decision: in voiced regions f_0 is modelled as a continuous numeric variable and in unvoiced regions as symbol strings.

Five-state left-to-right MSD-HSMMs similar to the one in Figure 12 are used in the system. State output probability density functions are single Gaussians with diagonal covariances and the state durations are modelled with single Gaussians with scalar variance. In training, monophone models are first created and estimated with the EM algorithm. These are then converted to context-dependent models and re-estimated. Finally, a decision-tree based clustering using the minimum description length (MDL) [46] criterion is applied to tie similar states together in order to reduce the total model complexity.

In parameter generation, input text is first translated into context dependent phonetic labels. Using the labels, context dependent tied-state HMM models are selected and concatenated into a sentence HMM. Duration of the sentence is governed by the state-duration models embedded in the HMMs. Finally, the parameter sequence is generated from the sentence HMM by a generation algorithm [51] that also recreates the global variance properties estimated from the training data.

5 Perceptual spectral matching

Perceptual spectral matching for glottal excitation vocoder seeks to address two problems in the baseline vocoder: First, psychoacoustic knowledge is not explicitly utilized in spectral matching in the original GlottHMM method. Second, the strong spectral tilt in the glottal source filter results in skewed feature distributions which may cause difficulties in the HMM training [2]. The approach of training a matching filter instead of glottal source spectrum expectedly gives a flatter spectrum model that results in more even distribution of the LSF frequencies. This section presents a novel spectral matching method utilizing the mel-scale filterbank structure. The method presentation is preceded by an overview of constrained least-squares fitting method that is required by the method.

5.1 Non-negative least squares

Non-negative least squares (NNLS) is a variant of the standard least squares fitting technique where the regression coefficients are restricted to non-negative values. In this work the technique is used in optimizing an equation whose solution corresponds to a magnitude spectrum, and thereby non-negativity is required. Let matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ and vector $\mathbf{b} \in \mathbb{R}^m$. Then the NNLS problem is then defined by

$$\min_{\mathbf{x} \geqslant 0} f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2. \tag{42}$$

The problem is convex and therefore the Karush-Kuhn-Tucker conditions are sufficient for optimality. Let the objective function be $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$. Then its gradient is $\nabla f(\mathbf{x}) = \mathbf{A}^\top(\mathbf{A}\mathbf{x} - \mathbf{b})$ and the optimality conditions at the saddle point $\mathbf{x}^*$ are [8]

$$\begin{aligned} \mathbf{x}^* &\geqslant 0, \\ \nabla f(\mathbf{x}^*) &\geqslant 0, \\ \nabla f(\mathbf{x}^*)^\top \mathbf{x}^* &= 0. \end{aligned} \tag{43}$$

The classical method for solving the NNLS problem is presented by Lawson and Hanson [27], and this algorithm is used also in this work. Pseudocode for the method is listed in Algorithm 1. The algorithm is an active set method that utilizes the knowledge that usually only a subset of the constraints is active. A constraint is considered active if the related regression coefficient of the unconstrained regression violates the constraint. Using this, the regression variables can be divided into the active set $\mathcal{R}$ and the passive set $\mathcal{P}$. Starting with all variables in $\mathcal{R}$, the method iteratively moves variables into $\mathcal{P}$ until the true active set is found. When the true active set is known, the constrained solution is simply the least squares solution on $\mathcal{P}$ expanded with the coefficients related to $\mathcal{R}$ set to zero. It can be shown first that there exists a unique solution to the constrained problem, and second that the algorithm converges smoothly to the solution and terminates in a finite amount of operations [27].

The algorithm presented here is suitable for relatively small problems. For this work, the algorithm is sufficient as size of the matrix $\mathbf{A}$ will be the number of mel-scale filters squared. On large scale problems, the iterated calculation of a matrix pseudoinverse for the least squares fit becomes inhibitively expensive. Other algorithms based on, for example, the gradient method or the Newton-Raphson method are available, (see [8] for a review on the methods for solving the NNLS problem.)

Algorithm 1 NNLS

```
Input: $\mathbf{A} \in \mathbb{R}^{m \times n}$, $\mathbf{b} \in \mathbb{R}^m$
Output: $\mathbf{x}^* = \arg\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ such that $\mathbf{x}^* \geqslant \mathbf{0}$
$\mathcal{P} \leftarrow \varnothing, \mathcal{R} \leftarrow \{1, \ldots, n\}$
$\mathbf{x} = \mathbf{0}$
$\mathbf{w} \leftarrow \mathbf{A}^\top(\mathbf{b} - \mathbf{A}\mathbf{x})$                                  ▻ Negative gradient
while $\mathcal{R} \neq \varnothing$ and $\max_{i \in \mathcal{R}}(w_i) > tol$ do
    $j \leftarrow \arg\max_{i \in \mathcal{R}}(w_i)$                          ▻ Dimension of steepest descent
    Include $j$ in $\mathcal{P}$ and remove it from $\mathcal{R}$
    $\mathbf{s}^{\mathcal{P}} \leftarrow \left[(\mathbf{A}^{\mathcal{P}})^\top \mathbf{A}^{\mathcal{P}}\right]^{-1} (\mathbf{A}^{\mathcal{P}})^\top \mathbf{b}$          ▻ Least squares solution in $\mathcal{P}$
    while $\min(\mathbf{s}^{\mathcal{P}}) \leqslant 0$ do
        $\alpha \leftarrow \min_{i \in \mathcal{P}}(x_i/(x_i - s_i))$
        $\mathbf{x} \leftarrow \mathbf{x} + \alpha(\mathbf{s} - \mathbf{x})$
        Move to $\mathcal{R}$ all indexes $j$ in $\mathcal{P}$ such that $\mathbf{x}_j = 0$
        $\mathbf{s}^{\mathcal{P}} \leftarrow \left[(\mathbf{A}^{\mathcal{P}})^\top \mathbf{A}^{\mathcal{P}}\right]^{-1} (\mathbf{A}^{\mathcal{P}})^\top \mathbf{b}$
        $\mathbf{s}^{\mathcal{R}} \leftarrow \mathbf{0}$
    $\mathbf{x} \leftarrow \mathbf{s}$
    $\mathbf{w} \leftarrow \mathbf{A}^\top(\mathbf{b} - \mathbf{A}\mathbf{x})$
```

5.2 Method overview

According to the source-filter model, a natural reference speech signal $s(t)$ can be decomposed to the convolution of a natural (derivative) glottal excitation $g(t)$ with a vocal tract filter $h_{\mathrm{vt}}(t)$. The vocal tract filter is estimated with glottal inverse filtering as presented in Section 4.3.1.

$$s(t) = g(t) * h_{vt}(t). \tag{44}$$

The vocal tract filter $h_{\mathrm{vt}}(t)$ is used to create a synthetic signal $\hat{s}(t)$ from a known excitation $c(t)$. This excitation is a pulse train formed by interpolating and concatenating a fixed waveform natural glottal pulse. To recreate the aperiodicity properties of the natural speech segment, bandpass noise is added to the synthetic excitation according to the HNR estimated from the natural signal. Additionally, a matching filter $h_{\mathrm{m}}(t)$ is applied on the synthetic signal in order to match the perceptual spectra of the signals. The synthetic signal $\hat{s}(t)$ is given by the convolution of the excitation with the vocal tract and matching filters.

$$\hat{s}(t) = c(t) * h_{\mathrm{vt}}(t) * h_{\mathrm{m}}(t). \tag{45}$$

The goal of perceptual spectral matching is to find a filter h_{m} that matches the mel-spectrum of a synthetic speech segment to the corresponding mel-spectrum of the target natural speech segment. More specifically, the matching filter is estimated by minimizing the distance between the mel-scale filterbank outputs of the two signals.

5.2.1 Piecewise constant matching spectrum

In the mel-spectrum calculation, the dimensionality of the spectral representation is reduced in the filterbank energy summation, and therefore there exists no unique inverse mapping back to the full-length spectrum. In order to create an inverse transformation from mel-filterbank output domain to spectrum, we first assume the matching filter magnitude spectrum $H_{\mathrm{m}}(z)$ to be piecewise constant on bands related to each filter.

Let $\mathbf{T} \in \mathbb{R}^{M \times N}$ be a matrix containing the triangular filterbank used in the mel-spectral coefficient calculation such that the rows correspond to the M individual filters. Then the mel-spectral coefficients are obtained by left multiplication on the warped FFT power spectrum of length N, namely $\mathbf{s} = \mathbf{T}|\tilde{\mathcal{F}}\{s(t)\}|^2$ for the natural speech segment and $\hat{\mathbf{s}} = \mathbf{T}|\tilde{\mathcal{F}}\{\hat{s}(t)\}|^2$ for the synthetic segment.

For scalar notation for individual filters, denote $\mathbf{T}_{i,k} = T_i(k)$, that is, T_i corresponds to the i:th filter with $i = 0, \ldots, M-1$. Let X_i be the constant matching filter power spectrum value at the region where the i:th filter has the highest value. This construction partitions the spectrum indices contributing to the i:th filter output into three sets: center region K_i^{c} where the i:th filter has the highest value, lower region K_i^{l} where the $i-1$:th filter is has the highest value and higher region K_i^{h} where the $i+1$:th filter has the highest value. Define the regions as:

$$\begin{aligned} K_i^{\mathrm{c}} &= \{k \mid T_i(k) > T_{i-1}(k) \wedge T_i(k) > T_{i+1}(k)\} \\ K_i^{\mathrm{l}} &= \{k \mid T_{i-1}(k) \geqslant T_i(k) \wedge T_i(k) \neq 0\} \\ K_i^{\mathrm{h}} &= \{k \mid T_{i+1}(k) \geqslant T_i(k) \wedge T_i(k) \neq 0\} \end{aligned} \tag{46}$$

Relationships between the filters and the regions of constant magnitude are illustrated in Figure 16. The values of the matching filter spectrum $H_{\mathrm{m}}(k)$ at each $k = 0, \ldots, N-1$ are constant within triangular filter bands and depend only on the filter index i

$$H_{\mathrm{m}}(k) = X_i \text{ when } k \in K_i^{\mathrm{c}}. \tag{47}$$

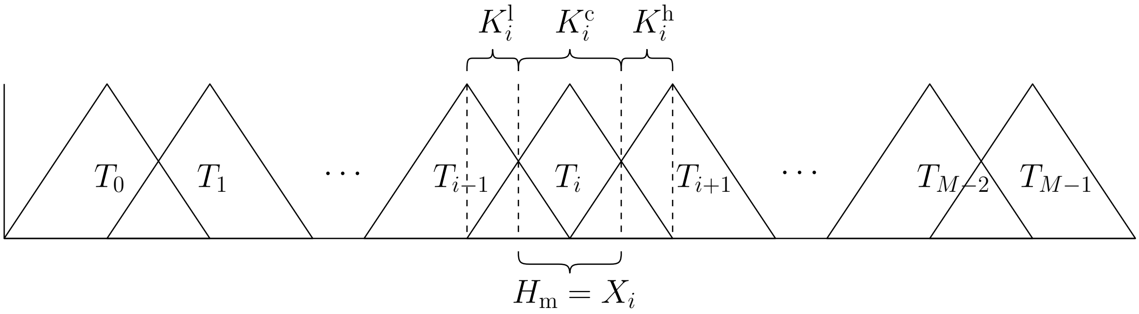


Figure 16: Piecewise constant definition of the matching spectrum H_{m}: X_i is the unknown constant equal to H_{m} at the region where the triangular filter T_i has the highest value. Spectrum regions contributing to the i:th filter output are denoted with K_i^{c}, K_i^{l} and K_i^{h}.

Now we can construct the filter bank output for the synthetic signal segment. Denote the warped FFT power spectrum of the unmatched synthetic signal as $F(k) = |\tilde{\mathcal{F}}\{c(t) * h_{\mathrm{vt}}(t)\}(k)|^2$. The i:th mel-spectral coefficient of the matched frame is then given by

$$\begin{aligned}\hat{S}_i = &\sum_{k \in K_i^{\mathrm{c}}} T_i(k) \cdot F(k) \cdot X_i \\ &+ \sum_{k \in K_i^{\mathrm{l}}} T_i(k) \cdot F(k) \cdot X_{i-1} \\ &+ \sum_{k \in K_i^{\mathrm{h}}} T_i(k) \cdot F(k) \cdot X_{i+1} \ .\end{aligned} \tag{48}$$

Note that X_i is independent of k for all i and write

$$\begin{aligned}G_i^{\mathrm{c}} &= \sum_{k \in K_i^{c}} T_i(k) \cdot F(k) \\ G_i^{\mathrm{l}} &= \sum_{k \in K_i^{l}} T_i(k) \cdot F(k) \\ G_i^{\mathrm{h}} &= \sum_{k \in K_i^{h}} T_i(k) \cdot F(k) \ .\end{aligned} \tag{49}$$

Now the i:th filter output can be expressed as

$$\hat{S}_i = G_i^{\mathrm{l}} X_{i-1} + G_i^{\mathrm{c}} X_i + G_i^{\mathrm{h}} X_{i+1} \ . \tag{50}$$

This results in a tridiagonal matrix representation where all the matrix elements are non-negative

$$\underbrace{\begin{bmatrix} G_0^{\mathrm{c}} & G_0^{\mathrm{h}} & 0 & \cdots & & 0 \\ G_1^{\mathrm{l}} & G_1^{\mathrm{c}} & G_1^{\mathrm{h}} & & & \vdots \\ 0 & G_2^{\mathrm{l}} & G_2^{\mathrm{c}} & G_2^{\mathrm{h}} & & \\ \vdots & & \ddots & \ddots & \ddots & 0 \\ & & & G_{K-2}^{\mathrm{l}} & G_{M-2}^{\mathrm{c}} & G_{M-2}^{\mathrm{h}} \\ 0 & \cdots & & 0 & G_{M-1}^{\mathrm{l}} & G_{M-1}^{\mathrm{c}} \end{bmatrix}}_{=\mathbf{G}} \underbrace{\begin{bmatrix} X_0 \\ X_1 \\ X_2 \\ \vdots \\ X_{M-2} \\ X_{M-1} \end{bmatrix}}_{=\mathbf{x}} = \underbrace{\begin{bmatrix} \hat{S}_0 \\ \hat{S}_1 \\ \hat{S}_2 \\ \vdots \\ \hat{S}_{M-2} \\ \hat{S}_{M-1} \end{bmatrix}}_{=\hat{\mathbf{s}}} \tag{51}$$

For an exact match we would set $\hat{\mathbf{s}} = \mathbf{Gx} = \mathbf{s}$ and solve $\mathbf{x}$ directly. However, $\mathbf{x}$ corresponds to a power spectrum and therefore all its elements must be positive, which is not guaranteed by a direct calculation. A suitable solution can be obtained by minimizing $|\hat{\mathbf{s}} - \mathbf{s}|^2$ with an element-wise non-negativity constraint $\mathbf{x} \geqslant 0$. The minimization problem is then stated by

$$\mathbf{x}_{\mathrm{opt}} = \underset{\mathbf{x} \geqslant 0}{\arg\min}\{|\mathbf{Gx} - \mathbf{s}|^2\}. \tag{52}$$

The constrained problem can be solved with the NNLS-algorithm (see Section 5.1) that converges to a unique solution when the pseudoinverse of $\mathbf{G}$ is well defined [27].

5.2.2 Matrix formulation

To give a direct matrix formulation for the problem, we define an expansion matrix $\mathbf{E}$ that expands the unknown M-vector $\mathbf{x}$ into a full length matching spectrum vector $\mathbf{H}_{\mathrm{m}}$, such that $\mathbf{H}_{\mathrm{m}} = \mathbf{E}^{\top}\mathbf{x}$. In the piecewise constant case

$$\mathbf{E}(i,k) = \begin{cases} 1, \text{when } k \in K_i^{\mathrm{c}} \\ 0, \text{otherwise} \end{cases} \tag{53}$$

Denote the spectrum vector of the unmatched synthetic signal by $\mathbf{F} = |\tilde{\mathcal{F}}\{c(t) * h_{\mathrm{vt}}(t)\}|^2$. Then the matched synthetic spectrum is given by $|\tilde{\mathcal{F}}\{\hat{s}\}|^2 = |\tilde{\mathcal{F}}\{c(t) * h_{\mathrm{vt}}(t) * h_{\mathrm{m}}(t)\}|^2 = \mathbf{F} \circ \mathbf{H}_{\mathrm{m}}$, where $\circ$ is the element-wise vector product. To form the matrix equation, rewrite element-wise vector multiplication as matrix-to-vector multiplication with a diagonal matrix that has the left vector elements on its diagonal. Denote this as $\mathbf{D} = \mathrm{diag}(\mathbf{F})$. Now the mel-filterbank output vector for $\hat{s}$ is given by

$$\begin{aligned} \hat{\mathbf{s}} &= \mathbf{T}(\mathbf{F} \circ \mathbf{H}_m) \\ &= \mathbf{TDH}_m \\ &= \underbrace{\mathbf{TDE}^{\top}}_{=\mathbf{G}}\mathbf{x}\ . \end{aligned} \tag{54}$$

When $\mathbf{E}$ is defined as in Eq. (53) the resulting $\mathbf{G}$ is equivalent to the one in Eq. (51). The direct matrix formulation of Eq. (54) facilitates the use of other types of shape assumptions for the matching spectrum. The values of $\mathbf{x}$ can be interpreted as sampled point values of $\mathbf{H}_\mathrm{m}$, and therefore $\mathbf{E}$ can be chosen to be some polynomial interpolation on the points defined in $\mathbf{x}$. Conveniently, linear interpolation, i.e., assuming $\mathbf{H}_\mathrm{m}$ is piecewise linear is given by the choice $\mathbf{E} = \mathbf{T}$. This interpretation is illustrated in Figure 17. The non negative solution is obtained as before with NNLS.

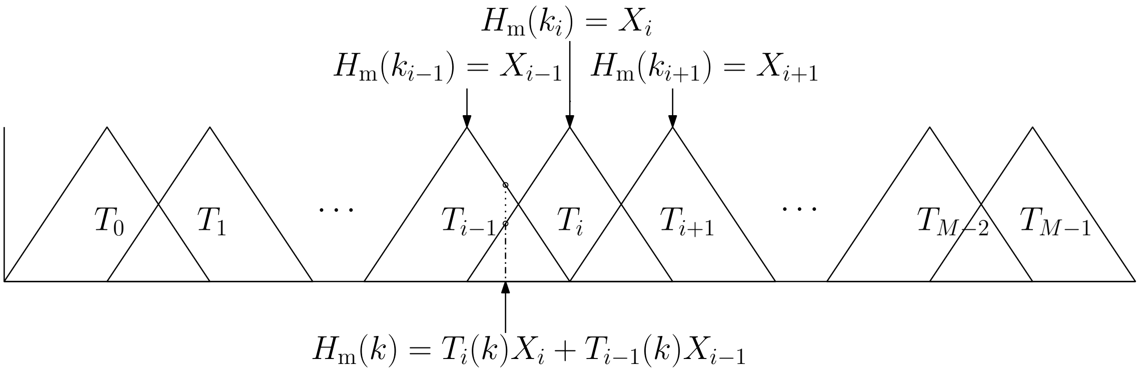

$H_\mathrm{m}(k_i) = X_i$
$H_\mathrm{m}(k_{i-1}) = X_{i-1}$
$H_\mathrm{m}(k_{i+1}) = X_{i+1}$
T_0
T_1
$\cdots$
T_{i-1}
T_i
T_{i+1}
$\cdots$
T_{M-2}
T_{M-1}
$H_\mathrm{m}(k) = T_i(k)X_i + T_{i-1}(k)X_{i-1}$

Figure 17: Piecewise linear definition of the matching spectrum $H_\mathrm{m}(k)$, where k indexes the spectrum: X_j represent the matching spectrum value sampled at the filter centers for $j = 0, \ldots, M-1$. The triangular filters $T_j(k)$ define a linear interpolation between the center points.

5.2.3 Warped all-pass filter realization

It is worth noting that the method presented above is well defined regardless of the exact filter shape. The shape results from how the frequency warping is realized in the mel-spectrum calculation. Traditionally warping is incorporated directly into the filter bank and the filters operate on the non-warped magnitude spectrum. This results in increasing filter base width as a function of frequency, and consequently decreasing frequency resolution.

In contrast, this work uses uniform width filters on a warped magnitude spectrum. There are two reasons for this: first, the obtained matching spectrum has a constant resolution on the warped frequency scale, which is beneficial for the interpolation used in the method. Second, working in the all-pass warped frequency domain enables convenient use of warped filters for realizing of the matching filter.

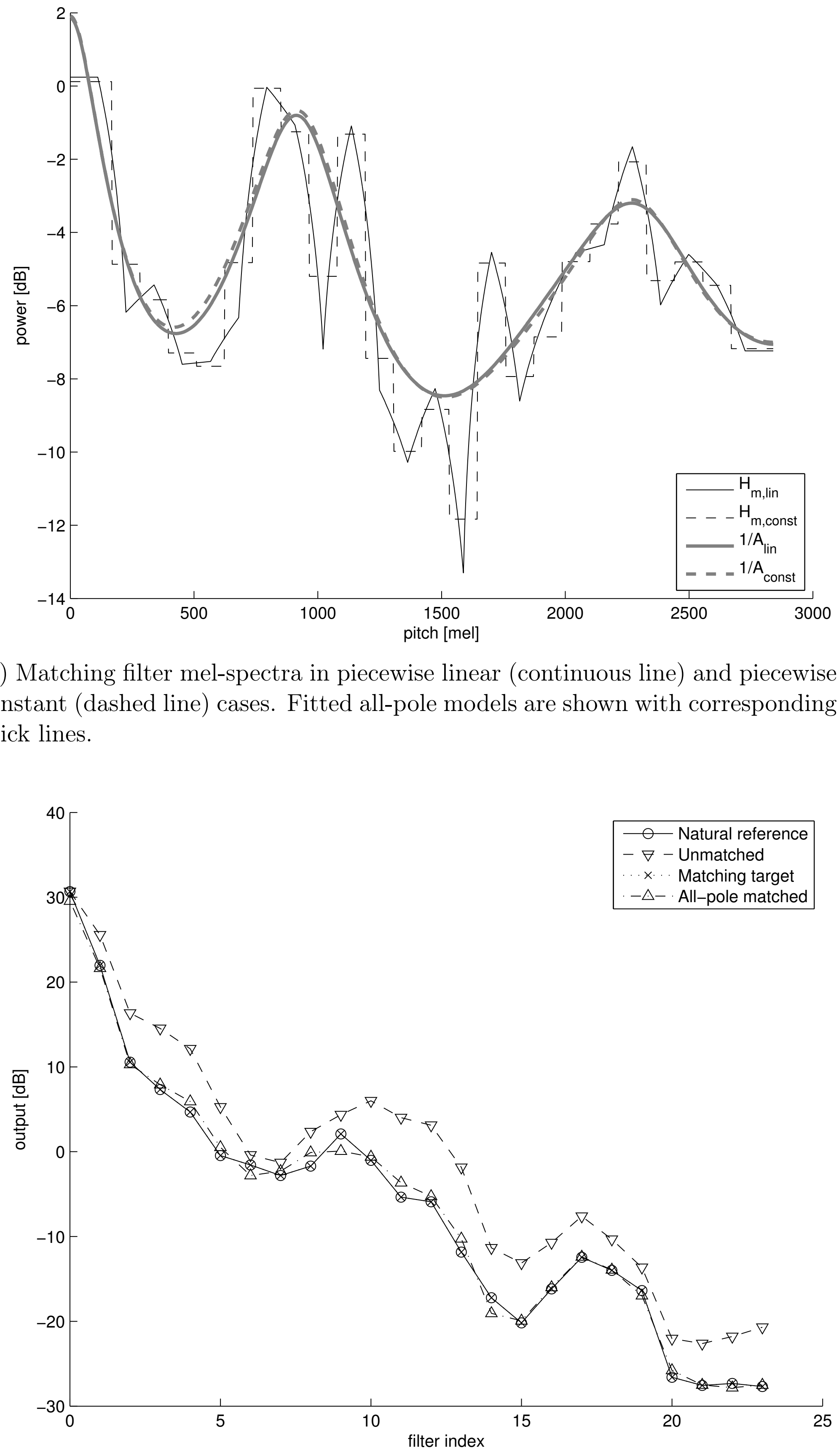
The magnitude spectrum previously obtained for the matching filter is fitted onto an all-pole filter with the following procedure:

Algorithm 2 Power spectrum to LPC

```
Input: One-sided power spectrum $\mathbf{x} \in \mathbb{R}^N$
Output: LPC filter coefficients $\mathbf{a}$
1: $\mathbf{x} \leftarrow \begin{bmatrix} \mathbf{x}_{0:N-1} \\ \mathbf{x}_{N-1:-1:0} \end{bmatrix}$    ▷ Extend to two sided spectrum
2: $\mathbf{r} \leftarrow \mathrm{IFFT}(\mathbf{x})$    ▷ Get autocorrelation from spectrum
3: $\mathbf{r} \leftarrow \mathbf{r}_{N-1:N+p-1}$    ▷ Select $\mathbf{r}$ at lags $0, \ldots, p$
4: $\mathbf{R}_{i,j} \leftarrow \mathbf{r}_{|i-j+1|}$    ▷ Form autocorrelation matrix
5: $\mathbf{a} \leftarrow \mathbf{R}^{-1}\mathbf{r}$    ▷ Solve Yule-Walker equations
```

The linear predictive model gives a filter that can be used as an all-pole synthesis filter. When working with a warped magnitude spectrum, the resulting filter can also be considered warped. The filter is effectively a warped LP synthesis filter and can be handled with the methods described in [20]. For HMM synthesis the filter is converted into its LSF representation, similarly to the other all-pole filters used in the synthesis system.

An example of piecewise constant and linear matching filters is presented in Figure 18. Figure 18a shows matching filter mel-spectra in the piecewise linear (continuous line) and piecewise constant (dashed line) cases. Note that the logarithmic scale bends the piecewise linear spectrum to appear piecewise logarithmic. Fitted all-pole models are shown with corresponding thick lines. Note that the all-pole spectra are relatively close to each other, regardless of the assumed shape of the matching spectrum. Figure 18b shows the outputs of mel-spectral filterbanks for the signals involved. Here only the piecewise linear matching spectrum is applied to the unmatched speech signal. The piecewise matching spectrum represents the "matching target"—the optimal matched spectrum, and "all-pole matched" gives the filter realization of the matching filter. Filterbank output of the natural reference signal corresponds closely with that of the synthetic signal matched with the optimal matching filter.



(a) Matching filter mel-spectra in piecewise linear (continuous line) and piecewise constant (dashed line) cases. Fitted all-pole models are shown with corresponding thick lines.

(b) Filterbank outputs for various mel spectra. The filters used in matching are represented by the solid lines in Fig. 18a Matching target is obtained by multiplying the piecewise defined matching spectrum with the unmatched synthetic spectrum.

Figure 18

6 Evaluation

6.1 Objective measures

6.1.1 MFCC distance to natural reference

To compare the method performances with an objective measure, a test based on MFCC distances was performed. The MFCC distance is commonly used as a dissimilarity measure in speech recognition [41].

Three GlottHMM-based systems were used to create analysis-synthesis representations of a natural reference signal. The analysis-synthesis scheme allows the comparison to the reference on a frame-to-frame basis. The three systems were the proposed mel-spectral matching (Mel), baseline GlottHMM MSE-based spectral matching (MSE), and GlottHMM without any spectral matching (nomatch). A male and a female speaker were used as references, and between-frame MFCC distances to reference were calculated for each system in non-silent frames based on an energy threshold. For both speakers 50 utterances were used, yielding a different number of frames due to differences in utterance content an rate of speech.

The results were analysed by comparing the distances between methods pairwise so that the distances in a certain frame are paired. The default approach is to use the paired t-test to test for the equality of means, or equivalently that the mean difference of the paired distances is zero. However, the t-test requires normality, and a Klomogorov-Smirnoff normality test [32] strongly rejects ($p < 0.001$) the normality of distance differences in all combinations of methods and speakers. Therefore, a non-parametric Wilcoxon signed rank test [56] for equality of medians with 2-sided alternative hypothesis was used.

Test results for male and female voices are listed in Tables 4 and 3, respectively. The first column in the tables lists the method pair under comparison. The second column gives the median difference of the distances, negative value indicating that the first method in the pair comparison performed better than the second in MFCC distance sense. The third column lists the decision on the null-hypothesis of equal medians, rejection indicating that the difference between the median distances is statistically significant, while the corresponding p-value is given in the fourth column.

Table 3: MFCC distance test for female voice, Number of frames = 4243.

	median difference	H_0: equal medians	p
Mel vs MSE	0.0243	reject $p < .001$	2.3409e-47
Mel vs nomatch	-0.2457	reject $p < .001$	0
MSE vs nomatch	-0.3124	reject $p < .001$	0

The test shows that in all pair comparisons the null hypothesis can be rejected at $p < .001$ significance level, meaning that the differences are statistically significant. Both spectral matching methods show higher performance when compared to the

Table 4: MFCC distance test for male voice, Number of frames = 9032.

	median difference	H_0: equal medians	p
Mel vs MSE	-0.0121	reject $p < .001$	3.8911e-11
Mel vs nomatch	-0.5296	reject $p < .001$	0
MSE vs nomatch	-0.5284	reject $p < .001$	0

no-spectral-matching case for both the male and female voices. However, the median difference for Mel and MSE matching methods is relatively small when compared to the no-match comparisons. Furthermore, the Mel matching method performs better for the male speaker and worse for the female in comparison to MSE. Based on the data, it is not justified to claim that either of the spectral matching methods is better, but only that they both clearly improve on using no spectral matching at all.

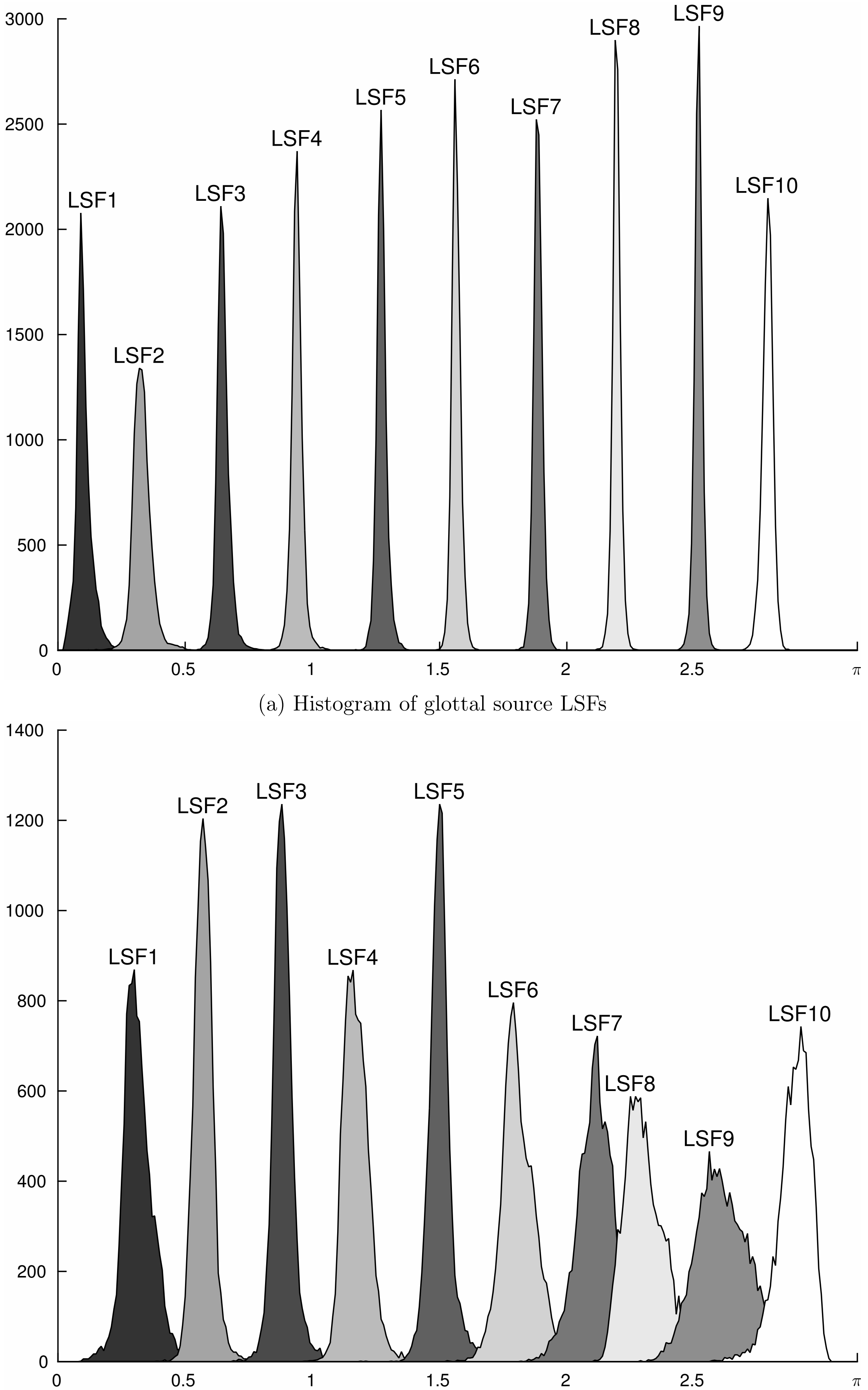
6.1.2 Statistical properties

Study of vocoder parameter statistics is justified because of the distribution assumptions made in the HMM-part of statistical parametric speech synthesis. By assumption, the parameter distributions are Gaussian mixtures with diagonal covariances, with a single Gaussian modelling a single parameter. The correlation of vocoder parameter statistics with analysis-synthesis has been studied in [2].

As the methods under comparison differ only with respect to the source or matching filter, it suffices to study the statistics of these. The LSF parameters correspond to frequencies between zero and the Nyquist frequency and thus lie numerically between 0 and π. It should be noted that the source LSFs of the baseline method are on a linear frequency scale, whereas the mel-matching filter LSFs are on a warped scale.

Figure 19 shows histograms of estimated LSFs for a male speaker. Statistics related to Gaussianity are listed on Table 5 for the source filter LSFs and matching filter LSFs. In addition to conventional skewness and kurtosis statistics, a negentropy [19] value is given. Negentropy estimates distribution distance from a Gaussian with same mean and variance, with negentropy values closer to zero indicating higher Gaussianity.

From the histograms and kurtosis values it is evident that in both cases the distributions are super-Gaussian. However, the parameters of the baseline method exhibit somewhat higher kurtosis, having sharper peaks in the distributions. Neither method's parameters seem markedly skewed by the histograms, and this is supported by the skewness values being relatively close to zero. The negentropy values for matching filter LSFs tend to be lower than for source filter LSFs, suggesting higher normality for the matching filter parameters. This may indicate that the mel-matching would be better suited for HMM synthesis in terms of parameter normality.



(a) Histogram of glottal source LSFs

(b) Histogram of mel-matching filter LSFs

Figure 19

Table 5: Statistics describing the degree of Gaussianity in distributions of the source LSFs and mel-matching filter LSFs. Values indicating higher normality are bolded.

	skewness		kurtosis		negentropy	
	source	mel-match	source	mel-match	source	mel-match
LSF1	1.1942	**0.3292**	6.5947	**3.9721**	0.1878	**0.0318**
LSF2	0.6489	**0.5879**	6.2710	**5.8542**	0.1562	**0.0901**
LSF3	0.8629	**0.8321**	**6.3486**	6.7976	**0.1993**	0.3460
LSF4	**0.4666**	0.6932	6.0749	**4.5876**	0.3528	**0.0657**
LSF5	**0.4904**	0.5762	**5.0238**	5.6270	**0.2578**	0.2926
LSF6	0.3089	**0.2561**	4.3045	**3.3609**	0.0415	**0.0285**
LSF7	**0.0881**	-0.2033	4.0432	**3.5243**	**0.0246**	0.0517
LSF8	**0.0151**	0.7048	4.7062	**3.5412**	0.1755	**0.0069**
LSF9	-0.1491	**0.0678**	4.6887	**2.5998**	0.1025	**0.0137**
LSF10	**-0.3270**	-0.8677	**3.7195**	4.3180	**0.0334**	0.0446
Mean	0.3599	**0.2976**	5.1775	**4.4183**	0.1531	**0.0972**

6.2 Subjective listening test

6.2.1 Synthetic voices

Synthetic voices were trained using speech data from two Finnish speakers, one male and one female. The training dataset for the male speaker consists of 599 utterances, totalling 53 minutes and correspondingly 513 utterances, totalling 33 minutes for the female speaker.

HMM voices were trained for both speakers using the baseline GlottHMM and the proposed modified GlottHMM with mel-spectral matching. Identical acoustic and linguistic data was used in training of the voices. Furthermore, the method parameters differ only in their matching filter, which in the baseline vocoder is calculated at the synthesis end from the HMM-generated vocal source LSFs, whereas the proposed method directly gives a matching filter from the statistical model. Parameter stream structures and dimensions were chosen to be identical, as presented in Table 2. For the listening test, 30 unseen test utterances for both speakers were generated with both methods.

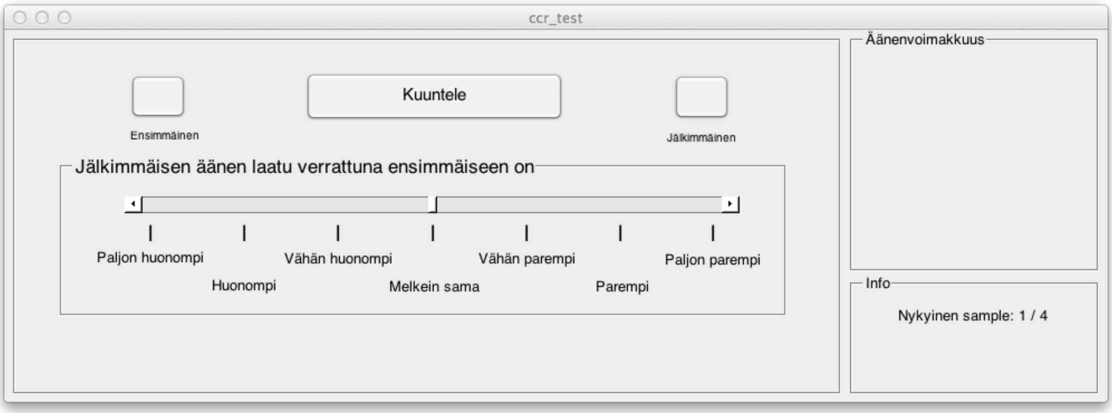
6.2.2 Test setup

The subjective quality of the speech synthesis systems was evaluated with a category comparison rating (CCR) test based on [22]. The test took place in a quiet listening booth and Sennheiser HD-650 headphones were used. In the test, the listener was presented with a pair of sound samples played successively, and asked to evaluate the quality of the latter sample compared to the first on a scale from 3 to -3, with description attached to the numbers according to Table 6. A sample pair consists of two utterances with the same linguistic content, and the method ordering is chosen at random. Also the ordering of utterances is randomized for each listener individually.

Table 6: CCR test rating scale.

3	Much better
2	Better
1	Slightly better
0	No difference
-1	Slightly worse
-2	Worse
-3	Much worse

Listeners were able to use the scale continuously and listen to each sample pair as many times they liked. The listening test interface (in Finnish) is depicted in Figure 20. Pressing the button labelled "Kuuntele" plays the sample pair and the indicators labelled "Ensimmäinen" and "Jälkimmäinen" would turn green while playing the first and second samples, respectively.



Nykyinen sample: 1 / 4

Figure 20: User interface of the CCR listening test.

6.2.3 Results

Results from the listening test are processed as follows: looking at each sample pair determine which method was played first and assign the listener score to it

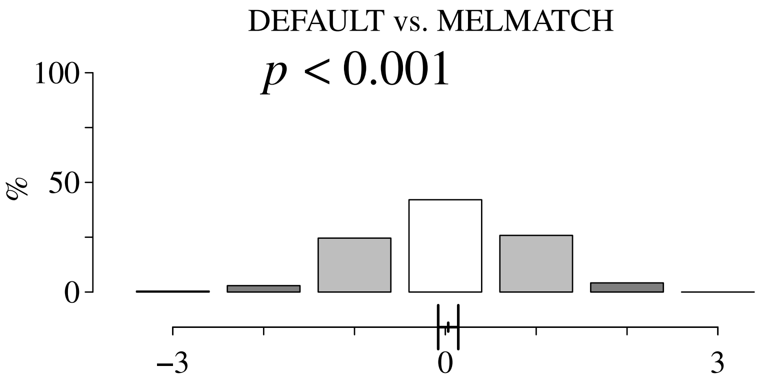


Figure 21: CCR test score distribution for a male speaker

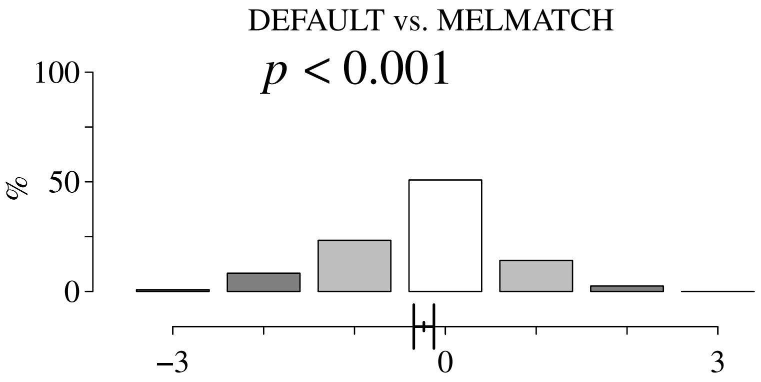


Figure 22: CCR test score distribution for a female speaker

accordingly and a score with an opposing sign to the other method. To obtain a mean opinion score, average of these ratings is taken. Score distributions for all listeners are presented in Figure 21 for the male voices and in Figure 22 for the female voices. Evaluation scores for individual listeners are included in Appendix C. The scores are reordered so that the baseline GlottHMM labeled "DEFAULT" is treated as the first sample and the mel matching method labeled "MELMATCH" as the second. Then a negative mean value of the scores indicates that the first method has better subjective quality. The mean value is plotted on a scale from -3 to 3 with 95% confidence intervals. Additionally, the p-value of a t-test for zero mean is presented.

For the male speaker, the difference is very close to zero with zero lying within the 95% confidence interval. Nevertheless, the t-test indicates that the mel-matching method gives slightly better quality, as the null hypothesis of equality of means is rejected. For the female speaker, the result is opposite and somewhat clearer: the baseline method gives slightly better quality.

7 Discussion and conclusion

The chapters 2–4 of this thesis provide the background needed, first in human speech production and perception (Chapter 2), second in speech parametrization and the involved auditory models (Chapter 3), and finally the used statistical parametric speech synthesis system (Chapter 4). The novel contribution in the thesis begins from Chapter 5 onwards. This discussion chapter focuses on the choices made in developing the new method and on the evaluation of the method.

7.1 Discussion

The main contribution of this thesis is a novel perceptual spectral matching method based on mel-scale filterbank distance criterion in Chapter 5. The method involves assuming the matching spectrum to be defined in piecewise constant or linear manner. This piecewise definition makes the matching problem well-determined and provides some regularity to the solution. Additionally, the solution corresponds directly to a power spectrum and therefore a positivity constraint must be included. The NNLS algorithm (Section 5.1) provides a suitable method for solving the problem in a least-squares sense with the solution constrained to non-negative values.

Alternatively, the assumption on the spectral shape could be relaxed. The problem would remain under-determined but could be solved in least-squares sense with additional constraints to the solution, for example, by modifying the cost function to include a penalty for great differences between subsequent values in the solution. This was not done, first for the added complexity and second because of the connection of piecewise definitions to filterbank output interpretation and the filterbank matrix pseudoinverse.

In practice, the shape assumption on the matching spectrum defines a reconstruction from the filterbank outputs to a full-length spectrum. The spectral representation of the mel-scale filterbank is studied in detail in Section 3.1.2, based mostly on an alternative version of the triangular filterbank applied on the warped spectrum. The resulting spectral representation can be conveniently interpreted as a sampling of the smoothed warped spectrum at the centre frequencies of the filterbank. Then the piecewise constant of linear spectral shape assumptions correspond to nearest-value or linear interpolation on the sampled values, respectively. Additionally, the piecewise linear reconstruction resembles the least-squares optimal reconstruction by the filterbank matrix pseudoinverse.

In this work, all-pass type warping was chosen for the mel-scale filterbank due to its several benefits: First, the all-pass warping curve approximates the mel-scale well in and allows the use of uniformly shaped an spaced filters in the warped frequency domain. Second, the filterbank output has a natural and precise interpretation in this domain as show in Appendix B. Finally, the all-pass warped domain allows the direct use of warped filters using standard filter design techniques. Potential drawbacks of the all-pass warping include the increased computational complexity compared to the standard mel-scale filterbank and non-warped filters. Furthermore, this alternative form of the mel-filterbank has not been widely used or studied previously.

Objective measures (Section 6.1.1) based on the MFCC distance between analysis-synthesis and natural reference shows that the proposed method works approximately equally well as the baseline method. Comparison with using no spectral matching shows that both spectral matching methods improve the MFCC similarity clearly. Differences between the spectral matching methods, though small, were statistically significant with the proposed method being slightly better than the baseline with male voice and vice versa for female voice. Subjective listening experiment (Section 6.2) for overall quality comparison with HMM-generated voices gives similar results: for a male voice the methods are equally good, and a female voice the proposed method performs slightly worse.

7.2 Conclusion

This thesis presents a novel perceptual spectral matching technique that uses a mel-scale filterbank output distance as its matching criterion. In addition, the auditory modelling properties of an all-pass type warping variant of the mel-filterbank were studied in detail. The devised matching filter was realized with a warped all-pole filter and incorporated into the GlottHMM speech synthesis system. After this, the novel method was evaluated in comparison to the baseline method by an objective measure test based on the MFCC-distance, and a subjective listening test. The evaluation shows that the proposed method gives approximately the same MFCC-similarity to natural reference and subjective quality as the baseline method.

The similar performance of the spectral matching methods stems mainly from the fact that both the mel-filterbank approach and the baseline linear predictive method operate on the high-level spectral envelope of the signal, ignoring spectral fine structure. Indeed, it is well known that LP modelling does well in capturing the spectral envelope of speech, even without any perceptual criterion. Additionally, the mel-matching method currently has the caveat of sometimes forming resonances up to 10 dB, typically on high frequency bands, if the library pulse has low energy on that particular band. This may result in audible artefacts and particularly affects the female voice when using a library pulse from a male speaker. The problem is not present in LP based spectral matching, as the MSE-based error criterion results in a spectral model that attaches itself to spectral peaks and largely ignores the low energy areas, resulting in no resonances in the corresponding (inverted) all-pole filter. One possible solution to alleviate this problem is weighting the optimization in Chapter 5 so that the error on the lower mel-bands is given more significance.

Future work with perceptual spectral matching involves use of more sophisticated filterbank models and different configurations with the speech synthesis system. For example gamma-tone filterbanks can relatively easily be adopted into the spectral matching scheme. In fact, any filter structure defined in the magnitude spectral domain is a suitable basis for matching. In terms of the speech synthesis stream configurations, an interesting option would be to train the MFCC of the natural speech along with the LSF-vocal tract information, and apply a mel-matching filter based on the MFCC entirely at the synthesis end of the system.

References

[1] N. Ahmed, T. Natarajan, and K. Rao, "Discrete Cosine Transform," *Computers, IEEE Transactions on*, vol. C-23, no. 1, pp. 90–93, Jan 1974.

[2] M. Airaksinen, "Analysis/synthesis comparison of vocoders utilized in statistical parametric speech synthesis," Master's thesis, Aalto University, 2012.

[3] P. Alku, "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, no. 2–3, pp. 109 – 118, 1992, Eurospeech '91.

[4] L. Beranek, *Acoustic measurements.* J. Wiley, 1949.

[5] A. W. Black, H. Zen, and K. Tokuda, "Statistical parametric speech synthesis," in *Proc. of ICASSP*, vol. 4, 2007, pp. IV–1229.

[6] T. Bäckström and P. Alku, "On the stability of constrained linear predictive models," in *Proc. of ICASSP*, vol. 6, Apr 2003, pp. VI–285–8 vol.6.

[7] T. Bäckström and C. Magi, "Properties of line spectrum pair polynomials—a review," *Signal processing*, vol. 86, no. 11, pp. 3286–3298, 2006.

[8] D. Chen and R. J. Plemmons, "Nonnegativity constraints in numerical analysis," in *in A. Bultheel and R. Cools (Eds.), Symposium on the Birth of Numerical Analysis, World Scientific.* Press, 2009.

[9] L. Chittka and A. Brockmann, "Perception space—the final frontier," PLoS Biol 3(4): e137. , note = Creative Commons Attribution 2.5 Generic, Apr 2005.

[10] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[11] J. Durbin, "The fitting of time-series models," *Revue de l'Institut International de Statistique*, pp. 233–244, 1960.

[12] European Telecommunications Standards Institute, *Standard ETSI ES 201 108 V1.1.3*, Sep 2003.

[13] G. Fant, *Acoustic Theory of Speech Production*, ser. D A C S R Series. Mouton De Gruyter, 1970.

[14] H. Fletcher, "Auditory patterns," *Rev. Mod. Phys.*, vol. 12, pp. 47–65, Jan 1940.

[15] Z. Heiga, T. Tomoki, and K. Tokuda, "The Nitech-NAIST HMM-based speech synthesis system for the Blizzard Challenge 2006," *Information and systems, IEICE transactions on*, vol. 91, no. 6, pp. 1764–1773, 2008.

[16] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, Nov 2012.

[17] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing: A Guide to Theory, Algorithm, and System Development.* Prentice Hall PTR, 2001.

[18] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. of ICASSP*, vol. 1, 1996, pp. 373–376.

[19] A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural networks*, vol. 13, no. 4, pp. 411–430, 2000.

[20] A. Härmä, M. Karjalainen, L. Savioja, V. Välimäki, U. K. Laine, and J. Huopaniemi, "Frequency-warped signal processing for audio applications," in *the 108th Audio Engineering Society (AES) Convention, preprint no. 5171*, Paris, France, 2000, p. 42 p.

[21] A. Härmä and U. K. Laine, "A comparison of warped and conventional linear predictive coding," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 5, pp. 579–588, Jul 2001.

[22] "Methods for Subjective Determination of Transmission Quality," ITU-T SG12, Geneva, Switzerland, Recommendation P.800, Aug. 1996.

[23] P. Kabal and R. P. Ramachandran, "The computation of line spectral frequencies using Chebyshev polynomials," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 34, no. 6, pp. 1419–1426, Dec 1986.

[24] M. Karjalainen, *Kommunikaatioakustiikka.* Helsinki University of Technology, 1999, in Finnish.

[25] M. Karjalainen, A. Härmä, and U. K. Laine, "Realizable warped IIR filters and their properties," in *Proc. of ICASSP*, vol. 3, Apr 1997, pp. 2205–2208 vol.3.

[26] T. Kinnunen, M. J. Alam, P. Matejka, P. Kenny, J. Cernocký, and D. D. O'Shaughnessy, "Frequency warping and robust speaker verification: a comparison of alternative mel-scale representations." in *Proc. of Interspeech*, 2013, pp. 3122–3126.

[27] C. Lawson and R. Hanson, *Solving Least Squares Problems.* Prentice-Hall, 1974.

[28] N. Levinson, "The Wiener RMS (root mean square) error criterion in filter design and prediction," *J. Math. Phys.*, vol. 25, no. 4, pp. 261–278, 1947.

[29] S. E. Levinson, "Continuously variable duration hidden markov models for automatic speech recognition," *Computer Speech & Language*, vol. 1, no. 1, pp. 29–45, 1986.

[30] J. Makhoul, "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, pp. 561–580, Apr 1975.

[31] J. Makhoul and L. Cosell, "LPCW: An LPC vocoder with linear predictive spectral warping," in *Proc. of ICASSP*, vol. 1, Apr 1976, pp. 466–469.

[32] F. J. Massey, "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[33] B. Moore, *An introduction to the psychology of hearing*, 2nd ed. Academic Press, 1982.

[34] B. C. J. Moore and B. R. Glasberg, "Auditory filter shapes derived in simultaneous and forward masking," *The Journal of the Acoustical Society of America*, vol. 70, no. 4, pp. 1003–1014, 1981.

[35] ——, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *The Journal of the Acoustical Society of America*, vol. 74, no. 3, pp. 750–753, 1983.

[36] A. Oppenheim, D. Johnson, and K. Steiglitz, "Computation of spectra with unequal resolution using the fast Fourier transform," *Proceedings of the IEEE*, vol. 59, no. 2, pp. 299–301, Feb 1971.

[37] D. O'Shaughnessy, *Speech communication: human and machine*, 2nd ed. IEEE press, 2000.

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

[39] H. Pulakka and P. Alku, "Bandwidth extension of telephone speech using a neural network and a filter bank implementation for highband mel spectrum," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 7, pp. 2170–2183, Sept 2011.

[40] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, Feb 1989.

[41] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition.* Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[42] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals.* Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1978.

[43] T. Raitio, A. Suni, J. Yamagishi, H. Pulakka, J. Nurminen, M. Vainio, and P. Alku, "HMM-based speech synthesis utilizing glottal inverse filtering," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 1, pp. 153–165, Jan 2011.

[44] T. Raitio, A. Suni, L. Juvela, M. Vainio, and P. Alku, "Deep neural network based trainable voice source model for synthesis of speech with varying vocal effort," in *Proc. of Interspeech*, Singapore, Sept 2014, pp. 1969–1973.

[45] T. Raitio, A. Suni, H. Pulakka, M. Vainio, and P. Alku, "Utilizing glottal source pulse library for generating improved excitation signal for HMM-based speech synthesis," in *Proc. of ICASSP*, Prague, Czech Republic, May 2011, pp. 4564–4567.

[46] K. Shinoda and T. Watanabe, "Acoustic modeling based on the MDL criterion for speech recognition," in *Proc. EuroSpeech*, no. 1, 1997, pp. 99–102.

[47] J. O. Smith and J. S. Abel, "Bark and ERB bilinear transforms." *Speech and Audio Processing, IEEE Transactions on*, vol. 7, no. 6, pp. 697–708, 1999.

[48] F. Soong and B.-H. Juang, "Line spectrum pair (LSP) and speech data compression," in *Proc. of ICASSP*, vol. 9, Mar 1984, pp. 37–40.

[49] S. S. Stevens, J. Volkmann, and E. B. Newman, "A scale for the measurement of the psychological magnitude pitch," *The Journal of the Acoustical Society of America*, vol. 8, no. 3, pp. 185–190, 1937.

[50] H. W. Strube, "Linear prediction on a warped frequency scale," *The Journal of the Acoustical Society of America*, vol. 68, no. 4, pp. 1071–1076, 1980.

[51] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *Information and Systems, IEICE Transactions on*, vol. 90, no. 5, pp. 816–824, 2007.

[52] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Hidden Markov models based on multi-space probability distribution for pitch pattern modeling," in *Proc. of ICASSP*, vol. 1, Mar 1999, pp. 229–232 vol.1.

[53] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proc. of ICASSP*, vol. 3, 2000, pp. 1315–1318 vol.3.

[54] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, "Mel-generalized cepstral analysis - a unified approach to speech spectral estimation," in *Proc. of ICSLP*, 1994, pp. 1043–1046.

[55] H. Wakita, "Direct estimation of the vocal tract shape by inverse filtering of acoustic speech waveforms," *Audio and Electroacoustics, IEEE Transactions on*, vol. 21, no. 5, pp. 417–427, 1973.

[56] F. Wilcoxon, “Individual comparisons by ranking methods,” *Biometrics bulletin*, pp. 80–83, 1945.

[57] U. H. Yapanel and J. H. L. Hansen, “A new perceptually motivated MVDR-based acoustic front-end (PMVDR) for robust automatic speech recognition,” *Speech Communication*, vol. 50, no. 2, pp. 142–152, Feb. 2008.

[58] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. W. Black, and K. Tokuda, “The HMM-based speech synthesis system version 2.0,” in *Proc. 6th ISCA Workshop on Speech Synthesis (SSW6-2007)*, 2007, pp. 294–299.

[59] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in *Proc. of ICASSP*, May 2013, pp. 7962–7966.

[60] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, “Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005,” *Information and systems, IEICE transactions on*, vol. 90, no. 1, pp. 325–333, 2007.

[61] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” *Speech Communication*, vol. 51, no. 11, pp. 1039 – 1064, 2009.

[62] E. Zwicker and H. Fastl, *Psychoacoustics: Facts and Models*, ser. Springer series in information sciences. Springer, 1990.

A Pitch scale values from literature

Table A1: Mel-scale values from listening experiment [4, p. 524]

frequency [mel]	frequency [Hz]
0	20
250	160
500	394
750	670
1000	1000
1250	1420
1500	1900
1750	2450
2000	3120
2250	4000
2500	5100
2750	6600
3000	9000
3250	14000

Table A2: Bark-scale values from listening experiment [62, p. 142]

frequency [Bark]	frequency [Hz]
0.5	50
1.5	150
2.5	250
3.5	350
4.5	450
5.5	570
6.5	700
7.5	840
8.5	1000
9.5	1170
10.5	1370
11.5	1600
12.5	1850
13.5	2150
14.5	2500
15.5	2900
16.5	3400
17.5	4000
18.5	4800
19.5	5800
20.5	7000

B Filterbank application as convolutive smoothing

When working in the warped spectral domain, the triangular filters in the mel-filterbank become uniform. In this case it can be shown that the filterbank output equals the warped spectrum convolved with the triangular filter shape and sampled at filterbank centre frequencies.

Consider the mel-spectrum variant where the filterbank of equal width triangles is applied on a warped spectrum. Let this triangle shape be the spreading function in an auditory spectrum, as illustrated in Figure B1. Let m be the frequency index of the M:th filter center. Then the filter H_M is a shifted version of the prototype filter H_0, i.e. $H_M(k) = H_0(k-m)$. Convolution of a spectrum X with the smoothing function H_0 is given by

$$\begin{aligned}
(X * H_0)(m) &= \sum_{k=-\infty}^{\infty} X(m-k)H_0(k) && H_0(k) = 0, |k| > K \\
&= \sum_{k=-K}^{K} X(m-k)H_0(k) && \text{index change } k \to -k \\
&= \sum_{k=K}^{-K} X(m+k)H_0(-k) && H_0(k) \text{ is symmetric} \\
&= \sum_{k=K}^{-K} X(m+k)H_0(k) && \text{reverse summation order} \\
&= \sum_{k=-K}^{K} X(m+k)H_0(k) && H_M(k) = H_0(k-m) \Leftrightarrow H_0(k) = H_M(k+m) \\
&= \sum_{k=-K}^{K} X(m+k)H_M(m+k) && \text{index change } k \to k-m \\
&= \sum_{k=m-K}^{m+K} X(k)H_M(k) && \text{add zeros: } H_M(k) = 0, |k-m| > K \\
&= \sum_{k=0}^{N} X(k)H_M(k)
\end{aligned}$$

which is the output of the M:th filter.

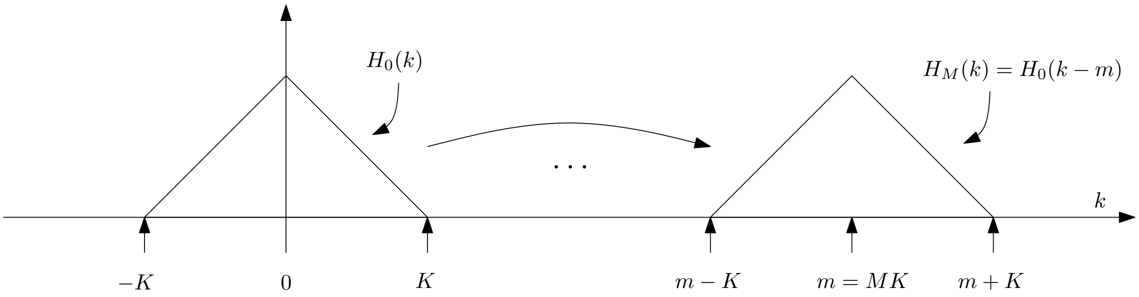


Figure B1: Prototype filter of the triangular filterbank shifted to create the M:th filter

C Individual ratings in listening test

Individual listener ratings for the CCR listening test

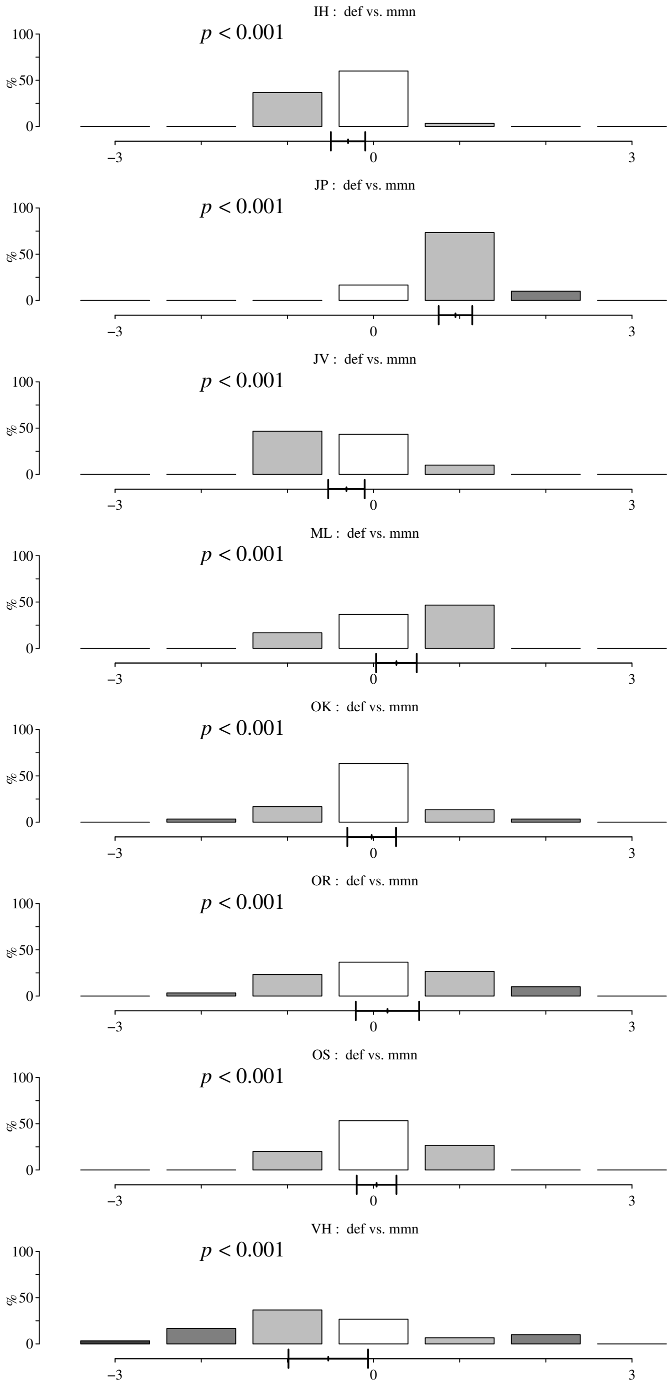


Figure C1: CCR test individual listener score distributions for the male speaker

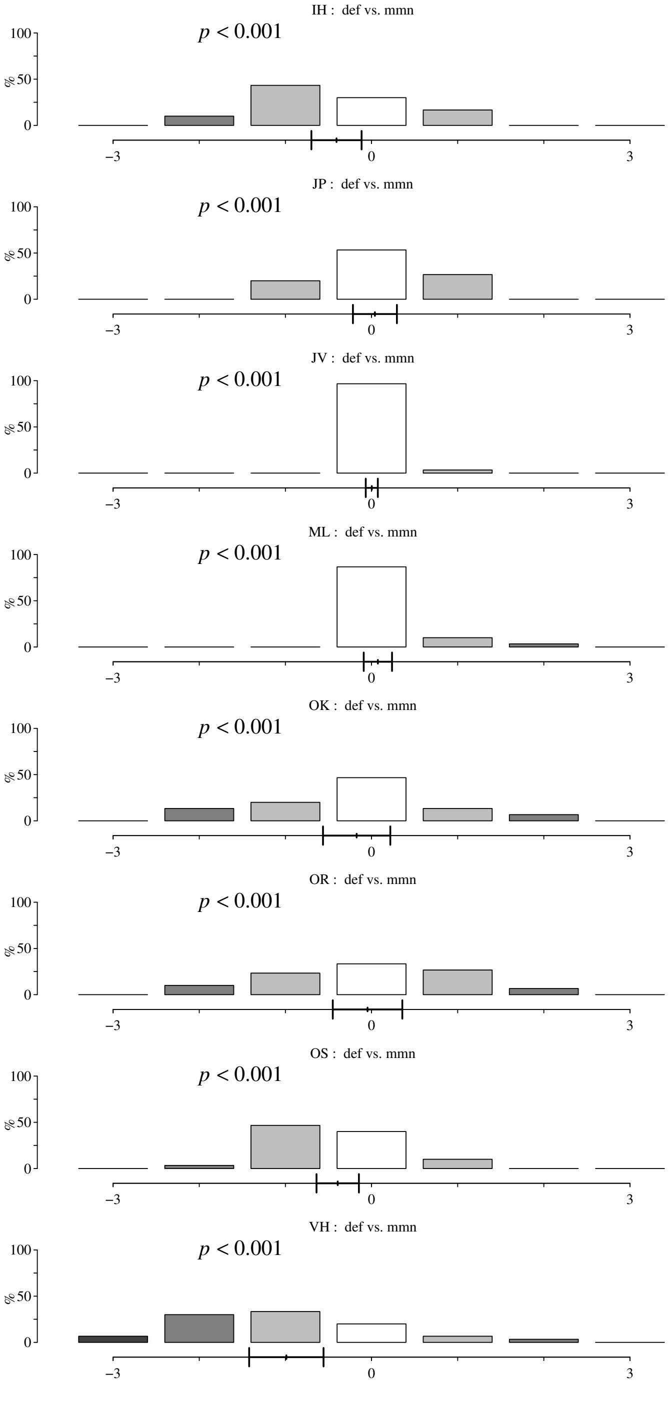


Figure C2: CCR test individual listener score distributions for the female speaker