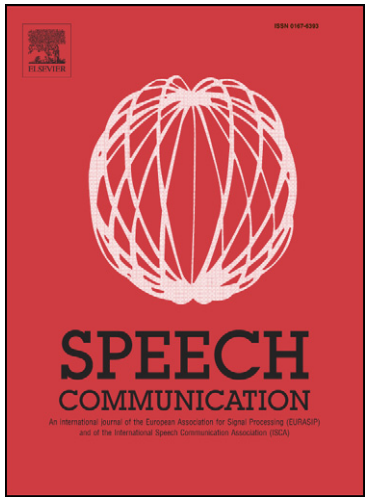
Signal Adaptive Spectral Envelope Estimation for Robust Speech Recognition

Matthias Wölfel



Signal Adaptive Spectral Envelope Estimation for Robust Speech Recognition

Matthias Wölfel

PII: S0167-6393(09)00025-9
DOI: 10.1016/j.specom.2009.02.006
Reference: SPECOM 1787

To appear in: *Speech Communication*

Received Date: 29 May 2008
Revised Date: 26 January 2009
Accepted Date: 24 February 2009

Please cite this article as: Wölfel, M., Signal Adaptive Spectral Envelope Estimation for Robust Speech Recognition, *Speech Communication* (2009), doi: 10.1016/j.specom.2009.02.006

Signal Adaptive Spectral Envelope Estimation for Robust Speech Recognition

Matthias Wölfel

Institut für Theoretische Informatik, Universität Karslruhe (TH), Am Fasanengarten 5, 76131 Karlsruhe, Germany

Abstract

This paper describes a novel spectral envelope estimation technique which adapts to the characteristics of the observed signal. This is possible via the introduction of a second bilinear transformation into warped *minimum variance distortionless response* (MVDR) spectral envelope estimation. As opposed to the first bilinear transformation, however, which is applied in the time domain, the second bilinear transformation must be applied in the frequency domain. This extension enables the resolution of the spectral envelope estimate to be steered to lower or higher frequencies, while keeping the overall resolution of the estimate and the frequency axis fixed. When embedded in the feature extraction process of an automatic speech recognition system, it provides for the emphasis of the characteristics of speech features that are relevant for robust classification, while simultaneously suppressing characteristics that are irrelevant for classification. The change in resolution may be steered, for each observation window, by the normalized first autocorrelation coefficient.

To evaluate the proposed adaptive spectral envelope technique, dubbed *warped-twice MVDR*, we use two objective functions: class separability and word error rate. Our test set consists of development and evaluation data as provided by NIST for the Rich Transcription 2005 Spring Meeting Recognition Evaluation. For both measures, we observed consistent improvements for several speaker-to-microphone distances. In average, over all distances, the proposed front–end reduces the word error rate by 4% relative compared to the widely used mel-frequency cepstral coefficients as well as perceptual linear prediction.

Key words: Adaptive Feature Extraction, Spectral Estimation, Minimum Variance Distortionless Response, Automatic Speech Recognition, Bilinear Transformation, Time vs. Frequency Domain

1. Introduction

Acoustic modeling in *automatic speech recognition* (ASR) requires that a windowed speech waveform is reduced to a set of representative features which preserves the information needed to determine the phonetic class while being invariant to other factors. Those factors might include speaker differences such as fundamental frequency, accent, emotional state or speaking rate, as well as distortions due to ambient noise, the channel or reverberation. In the traditional feature extraction process of ASR systems, this is achieved through successive feature transformations (e.g. a spectral envelope and/or filterbank followed by cepstral transformation, cepstral normalization and linear discriminant analysis) whereby all phoneme types are treated equivalently.

Different phonemes, however, have different properties such as voicing where the excitation is due to quasi-periodic opening of the vocal cord or classification relevant frequency regions [1–3]. While low frequencies are more relevant for vowels, high frequencies are more relevant for fricatives. It is thus a natural extension to the traditional feature extraction approach to vary the spectral resolution for each observation window according to some characteristics of the observed signal. To improve phoneme classification, the spectral resolution may be adapted such that characteristics relevant for classification are emphasized while classification irrelevant characteristics are attenuated.

To achieve these objectives, we have proposed to extend the warped *minimum variance distortionless response* (MVDR) through a second bilinear transformation [4]. This spectral envelope estimate has two free parameters to control spectral resolution: the model order, which changes the number of linear prediction coefficients, and the warp factor. While the model order allows the overall spectral resolution to be changed, the warp factor enables the spectral resolution to be steered to lower or higher frequency regions without changing the frequency axis. Note that this is in contrast to the previously proposed warped MVDR [5,6],

wherein the warp factor has an influence on both the spectral resolution and the frequency axis.

A note about the differences between the present publication and [4] is perhaps now in order. The present publication presents important background information which had been discarded in the conference publication [4] because of space limitations including: A comparison of well known and not so well known ASR front-ends on close and distant recordings in terms of word error rate and class separability. The present publication also includes a detailed analysis and discussion of phoneme confusability. In addition it fosters understanding by highlighting the differences between warping in the time and frequency domain and investigating of the values of the steering function in relation to single phonemes and phoneme classes.

The balance of this paper is organized as follows. A brief review of spectral envelope estimation techniques with a focus on MVDR is given in Section 2. The bilinear transformation is reviewed in Section 3 where its properties in the time and frequency domains are discussed. Section 4 introduces a novel adaptive spectral estimation technique, dubbed warped-twice MVDR, and a fast implementation thereof. A possible steering function, to emphasize phoneme relevant spectral regions, is discussed in Section 5. The proposed signal-adaptive feature extraction scheme is evaluated in Section 6. Our conclusions are presented in the final section of this paper.

2. MVDR Spectral Envelope

In the feature extraction stage of speech recognition systems, particular characteristics of the spectral estimate are required. To name a few: provide a particular spectral resolution, be robust to noise, and model the frequency response function of the vocal tract during voiced speech. To satisfy these requirements, both *non-parametric* and *parametric* methods have been proposed. Non-parametric methods are based on periodograms, such as power spectra, while parametric methods such as linear prediction estimate a small number of parameters from the data. Table 1 summarizes the characteristics of different spectral estimation methods. Two widely used methods in ASR are mel-scale power spectrum [7] and warped or perceptual linear prediction [8].

In order to overcome the problems associated with (warped or perceptual) linear prediction, namely overestimation of spectral power at the harmonics of voiced speech, Murti and Rao [9,10] proposed the use of *minimum variance distortionless response* (MVDR), which is also known as Capon's method [11] or the maximum-likelihood method [12], for all-pole modeling of speech in 1997. They demonstrated that MVDR spectral envelopes cope well with the aforementioned problem. Some years later, in 2001, MVDR was applied to speech recognition by Dharanipragada and Rao [13]. To account for the frequency resolution of the human auditory system, we have introduced *warped MVDR* [5,6]. It extends the MVDR approach by *warping* the frequency axis with a bilinear transformation in the time domain.

Table 1
Properties of spectral estimation methods.
PS = power spectrum; LP = linear prediction;
MVDR = minimum variance distortionless response
* no particular name is given in the work by Nakatoh *et al.*

Spectral Estimate	Properties		
	detail	resolution	sensitive to pitch
PS	exact	linear, static	very high
mel-scale PS [15]	smooth	mel, static	high
LP [16,17]	approx.	linear, static	medium
perceptual LP [8]	approx.	mel, static	medium
warped LP [18,19]	approx.	mel, static	medium
warped-twice LP* [20]	approx.	mel, adaptive	medium
MVDR [9,10,13]	approx.	linear, static	low
warped MVDR [6]	approx.	mel, static	low
perceptual MVDR [14]	approx.	mel, static	low
warped-twice MVDR	approx.	mel, adaptive	low

In this section, we briefly review MVDR spectral estimation. A detailed discussion of speech spectral estimation by MVDR can be found in [10], with focus on speech recognition and warped MVDR in [6], and with focus on robust feature extraction for recognition in [14].

MVDR methodology

MVDR spectral estimation can be posed as a problem in filterbank design, wherein the final filterbank is subject to the *distortionless constraint* [21]:

The signal at the frequency of interest ω_{foi} must pass undistorted with unity gain.

$$H_{\text{foi}}(e^{j\omega_{\text{foi}}}) = \sum_{k=0}^{M} h_{\text{foi}}(k) e^{-jk\omega_{\text{foi}}} = 1, \tag{1}$$

where the impulse response $h_{\text{foi}}(k)$ of the distortionless finite impulse responce filter of order M is specifically designed to minimize the output power. Defining the *fixed frequency vector*

$$\mathbf{v}(e^{j\omega}) = [1, e^{+j\omega}, \ldots, e^{+jM\omega}]^T \tag{2}$$

allows the constraint to be rewritten in vector form as

$$\mathbf{v}^H(e^{j\omega_{\text{foi}}}) \cdot \mathbf{h}_{\text{foi}} = 1, \tag{3}$$

where $(\bullet)^H$ represents the Hermitian transpose operator and

$$\mathbf{h}_{\text{foi}} = [h_{\text{foi}}(0), h_{\text{foi}}(1), \ldots, h_{\text{foi}}(M)]^T \tag{4}$$

is the distortionless filter.

Upon defining the autocorrelation sequence

$$R[n] = \sum_{m=0}^{L-n} x[m]x[m-n] \tag{5}$$

of the input signal x of length L as well as the $(M+1) \times (M+1)$ Toeplitz autocorrelation matrix $\mathbf{R}$ whose $(l,k)^{th}$ element is given by

$$\mathbf{R}_{l,k} = R[l-k], \tag{6}$$

it is readily shown that $\mathbf{h}_{\text{foi}}$ can be obtained by solving the constrained minimization problem:

$$\min_{\mathbf{h}_{\text{foi}}} \mathbf{h}_{\text{foi}}^T \mathbf{R}\, \mathbf{h}_{\text{foi}} \quad \text{subject to} \quad \mathbf{v}^H(e^{j\omega_{\text{foi}}})\mathbf{h}_{\text{foi}} = 1. \tag{7}$$

The solution to this problem is given by [21]:

$$\mathbf{h}_{\text{foi}} = \frac{\mathbf{R}^{-1}\,\mathbf{v}(e^{j\omega_{\text{foi}}})}{\mathbf{v}^H(e^{j\omega_{\text{foi}}})\,\mathbf{R}^{-1}\,\mathbf{v}(e^{j\omega_{\text{foi}}})}. \tag{8}$$

This implies that $\mathbf{h}_{\text{foi}}$ is the impulse response of the distortionless filter for the frequency ω_{foi}. The MVDR envelope of the power spectrum of the signal $P(e^{j\omega})$ at frequency ω_{foi} is then obtained as the output of the optimized constrained filter:

$$S_{\text{MVDR}}(e^{j\omega_{\text{foi}}}) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \left|\mathbf{H}_{\text{foi}}(e^{j\omega})\right|^2 P(e^{j\omega})d\omega. \tag{9}$$

Although MVDR spectral estimation was posed as a distortionless filter design for a given frequency ω_{foi}, the MVDR spectrum can be represented in parametric form for all frequencies [21]

$$S_{\text{MVDR}}(e^{j\omega}) = \frac{1}{\mathbf{v}^H(e^{j\omega})\mathbf{R}^{-1}\mathbf{v}(e^{j\omega})}. \tag{10}$$

Fast computation of the MVDR envelope

Assuming that the $(M+1)\times(M+1)$ Hermitian Toeplitz correlation matrix $\mathbf{R}$ is positive definite and thus invertible, Musicus [12] derived a fast algorithm to calculate the MVDR spectrum from a set of *linear prediction coefficients* (LPCs). The steps (i until iii) of Musicus' algorithm [12] are:

(i) *Computation of the LPCs $a_{0\cdots M}^{(M)}$ of order M including the prediction error variance ϵ_M*

(ii) *Correlation of the LPCs*

$$\mu_k = \begin{cases} \dfrac{1}{\epsilon_M}\displaystyle\sum_{m=0}^{M-k}(M+1-k-2m)a_m^{(M)}a_{m+k}^{*(M)} & , k \geq 0 \\ \mu_{-k}^* & , k < 0 \end{cases} \tag{11}$$

(iii) *Computation of the MVDR envelope*

$$S_{\text{MVDR}}(e^{j\omega}) = \frac{1}{\sum_{m=-M}^{M}\mu_m e^{-j\omega m}}. \tag{12}$$

(iv) *Scaling of the MVDR envelope*
In order to improve robustness to additive noise it has been argued in [6] to adjust the highest spectral peak of the MVDR envelope to match to the highest spectral peak of the power spectrum to get the so called *scaled* envelope.

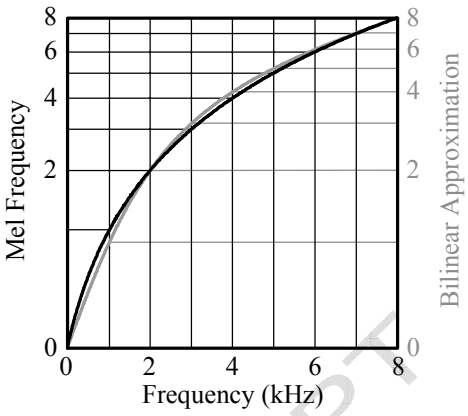


Fig. 1. Mel-frequency (scale shown along left edge) can be approximated by a bilinear transformation (scale shown along right edge) demonstrated for a sampling rate of 16 kHz, $\alpha_{\text{mel}} = 0.4595$.

3. Warping — Time vs. Frequency Domain

In the speech recognition community it is well known that features based on a non-linear frequency mapping improve the recognition accuracy over features on a linear frequency scale [7]. Transforming the linear frequency axis ω to a non-linear frequency axis $\tilde{\omega}$ is called *frequency warping*. One way to achieve frequency warping is to apply a non-linear scaled filterbank, such as a mel-filterbank, to the linear frequency representation. An alternative possibility is to use a conformal mapping such as a first order all-pass filter, also known as a *bilinear transformation* [22,23], which preserves the unit circle. The bilinear transformation is defined in the z-domain as

$$\tilde{z}^{-1} = \frac{z^{-1}-\alpha}{1-\alpha\cdot z^{-1}} \;\forall -1 < \alpha < +1, \tag{13}$$

where α is the *warp factor*. The relationship between $\tilde{\omega}$ and ω is non-linear as indicated by the phase function of the all-pass filter [19]

$$\arg\left(e^{-j\tilde{\omega}}\right) = \tilde{\omega} = \omega + 2\arctan\left(\frac{\alpha\sin\omega}{1-\alpha\cos\omega}\right). \tag{14}$$

The mel-scale, which along with the Bark scale is one of the most popular non-linear frequency mappings, was proposed by Stevens *et al.* in 1937 [15]. It models the non-linear pitch perception characteristics of the human ear and is widely applied in audio feature extraction. A good approximation of the mel-scale by the bilinear transformation is possible, if the warp factor is set accordingly. The optimal warp factor depends on the sampling frequency and can be found by different optimization methods [24]. Fig. 1 compares the mel-scale with the bilinear transformation for a sampling frequency of 16 kHz.

Frequency warping by bilinear transformation can either be applied in the *time domain* or in the *frequency domain*. In both cases, the frequency axis is non-linearly scaled; however, the effect on the spectral resolution differs for the two domains. This effect can be explained as follows:

– *Warping in the time domain* modifies the values in the autocorrelation matrix and therefore, in the case of linear

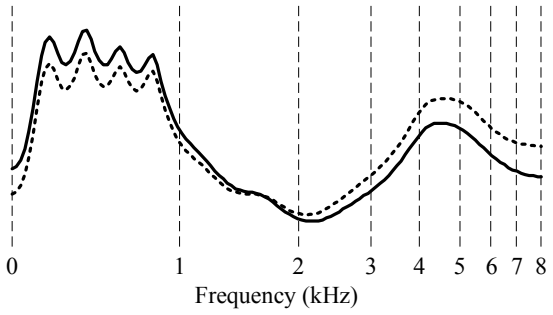


Fig. 3. The plot of two warped-twice MVDR spectral envelopes demonstrates the effect of spectral tilt. While the spectral tilt is not compensated for the dashed line, it is compensated for the solid line. It is clear to see that high frequencies are emphasized if no compensation is applied.

prediction, more linear prediction coefficients are used, for $\alpha > 0$, to describe lower frequencies and less coefficients to describe higher frequencies.

– *Warping in the frequency domain* does not change the spectral resolution as the transformation is applied after spectral analysis. As indicated by Nocerino et al. [25], a general warping transformation in the same domain, such as the bilinear transformation, is equivalent to a matrix multiplication

$$f_{\text{warp}}[n] = \mathbf{L}(\alpha) f[n],$$

where the matrix $\mathbf{L}(\alpha)$ depends on the warp factor. It follows that the values $f_{\text{warp}}[n]$ on the warped scale are a linear interpolation of the values $f[n]$ on the linear scale. In the case of linear prediction or MVDR, the prediction coefficients are not altered as they are calculated before the bilinear transformation is applied.

Fig. 2 demonstrates the effect of warping applied either in the time or in the frequency domain on the spectral envelope and compares the warped spectral envelopes with the unwarped spectral envelope.

For clarity we briefly investigate the change of spectral resolution, for the most interesting case, where the bilinear transformation is applied in the time domain with warp factor $\alpha > 0$. In this case we observe that spectral resolution decreases as frequency increases. In comparison to the resolution provided by the linear frequency scale, $\alpha = 0$, the warped frequency resolution increases for low frequencies up to the *turning point frequency* [26]

$$f_{\text{tp}}(\alpha) = \pm \frac{f_{\text{s}}}{2\pi} \arccos(\alpha), \tag{15}$$

where f_{s} represents the sampling frequency. At the turning point frequency, the spectral resolution is not affected. Above the turning point frequency, the frequency resolution decreases in comparison to the resolution provided by the linear frequency scale. For $\alpha < 0$, spectral resolution increases as frequency increases.

As observed by Strube [18], prediction error minimization of the predictors $\tilde{a}_m$ in the warped domain is equivalent to the minimization of the output power of the warped inverse filter

$$\tilde{A}(z) = 1 + \sum_{m=1}^{M} \tilde{a}_m \tilde{z}^{-m}(z) \tag{16}$$

in the linear domain, where each unit delay element z^{-1} is replaced by a bilinear transformation $\tilde{z}^{-1}$. The prediction error is therefore given by

$$E(e^{j\omega}) = |\tilde{A}(e^{j\omega})|^2 P(e^{j\omega}), \tag{17}$$

where $P(e^{j\omega})$ is the power spectrum of the signal. The total prediction error power can be expressed as

$$\sigma^2 = \int_{-\pi}^{\pi} E(e^{j\tilde{\omega}}) d\tilde{\omega} = \int_{-\pi}^{\pi} E(e^{j\omega}) W^2(e^{j\omega}) d\omega \tag{18}$$

with

$$W(z) = \frac{\sqrt{1-\alpha^2}}{1-\alpha z^{-1}}. \tag{19}$$

The minimization of the prediction error σ^2, however, does *not* lead to minimization of the power, but minimization of the power of the error signal filtered by the weighting filter $W(z)$, which is apparent from the presence of this factor in (18). Thus, the bilinear transformation introduces an unwanted spectral tilt. To compensate for this negative effect, we apply the inverted weighting function

$$\left| \tilde{W}(\tilde{z}) \cdot \tilde{W}(\tilde{z}^{-1}) \right|^{-1} = \frac{\left|1 + \alpha \cdot \tilde{z}^{-1}\right|^2}{1-\alpha^2}. \tag{20}$$

The effect of the spectral tilt of the bilinear transformation and the remedy by (20) are depicted in Fig. 3.

4. Warped-Twice MVDR Spectral Envelope

The use of two bilinear transformations, one in time domain and the other in frequency domain, introduces two additional free parameters into the MVDR approach [4]. The first free parameter, the model order, is already determined by the underlying linear prediction model. Due to the application of two bilinear transformations which apply two warping stages into MVDR spectral estimation, the proposed approach is dubbed *warped-twice MVDR*. While the model order varies the overall spectral resolution of the estimate, which becomes apparent by comparing the different envelopes for model order 30, 60 and 90 in Fig. 4a, the warp factors bend the frequency axis as already seen in Section 3. Bending the frequency axis can be used to apply the mel-scale or, when done on a speaker-dependent basis, to implement *vocal tract length normalization* (VTLN), although the latter is not used in the experiments described in Section 6, as piece-wise linear warping leads to better results [27].

As already mentioned in Section 1, our aim is to change the spectral resolution while keeping the frequency axis fixed. This becomes possible by compensating for the unwanted bending of the frequency axis, introduced by the first warping stage in the time domain, by a second warping stage in the frequency domain. An example is given in Fig. 4b.

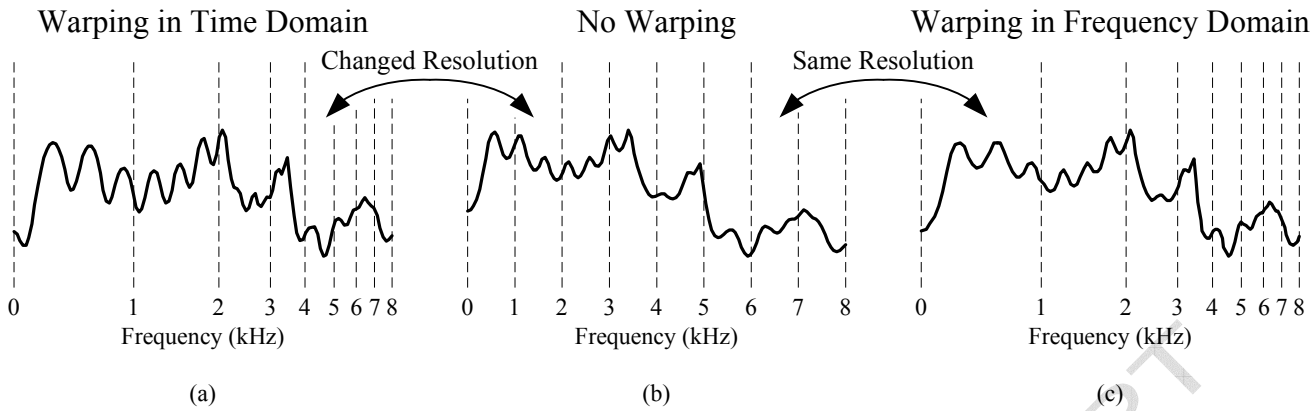


Fig. 2. Warping in (a) time domain, (b) no warping and (c) warping in frequency domain. While warping in the time domain is changing the spectral resolution and frequency axis, warping in frequency domain does not alter the spectral resolution but still changes the frequency axis.

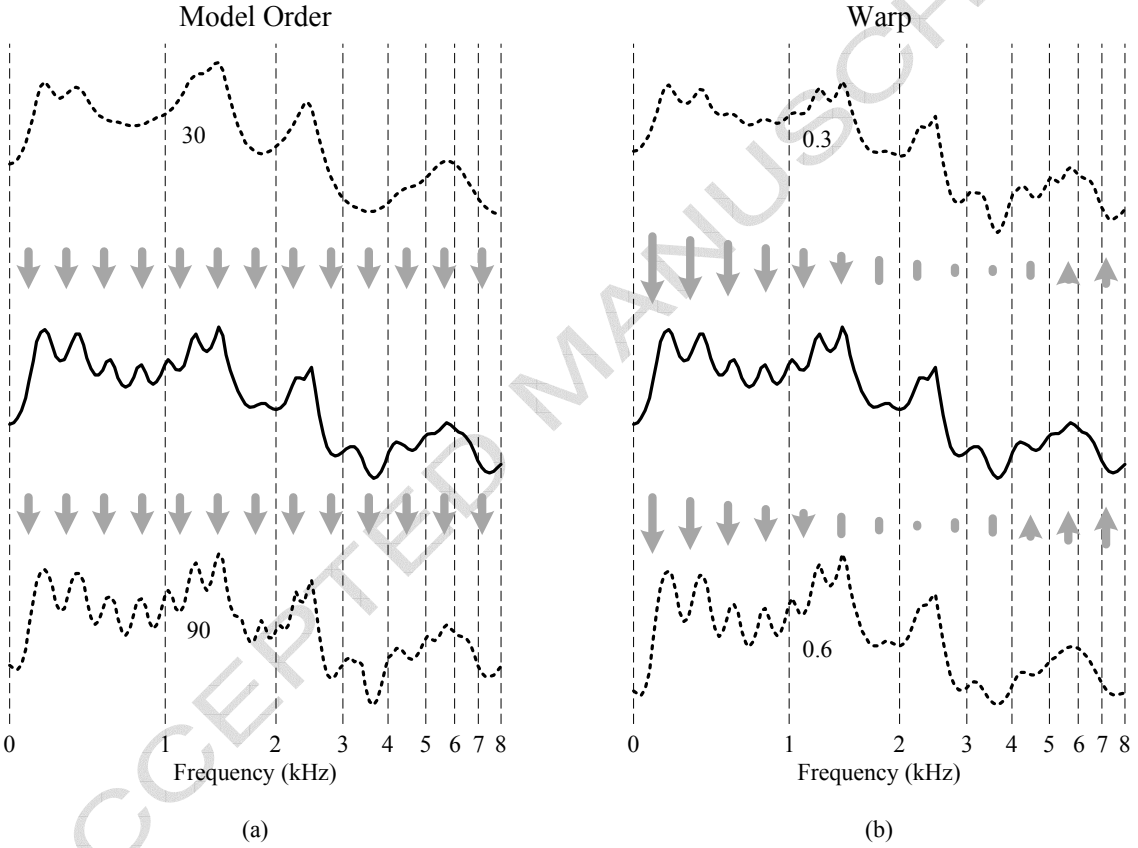


Fig. 4. The solid lines show warped-twice MVDR spectral envelopes with model order 60, $\alpha = 0.4595$ and $\alpha_{\text{mel}} = 0.4595$ which, except for the spectral tilt, are identical to a warped MVDR spectral envelope. Its counterparts with lower and higher (a) model order and (b) warp factor α are given by dashed lines. The arrows point in the direction of higher resolution. While the model order changes the overall spectral resolution at all frequencies, the warp factor moves spectral resolution to lower or higher frequencies. At the turning point frequency, the resolution is not affected and the direction of the arrows changes.

Fast computation of the warped-twice MVDR envelope

A fast computation of the warped-twice MVDR envelope of model order M is possible by extending Musicus' algorithm. A flowchart diagram of the individual processing steps is given in Fig. 5.

(i) *Computation of the warped autocorrelation coefficients* $\tilde{R}[0] \cdots \tilde{R}[M+1]$

To compute warped autocorrelation coefficients, the linear frequency axis ω has to be transformed to a warped frequency axis $\tilde{\omega}$ by replacing the unit delay element z^{-1} with a bilinear transformation (13). This leads to the warped autocorrelation coefficients [28,19]

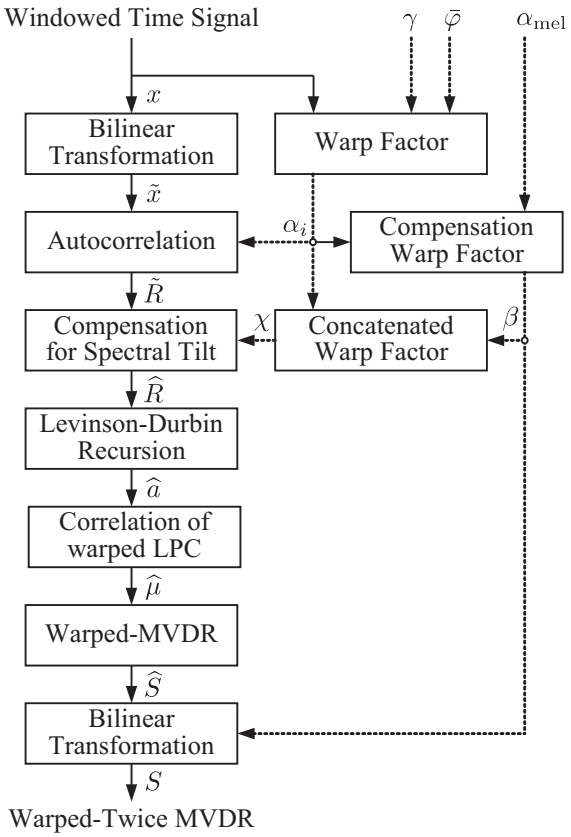


Fig. 5. Overview of warped-twice minimum variance distortionless response. Symbols are defined as in the text.

$$\tilde{R}[n] = \sum_{m=0}^{L-n-1} x[m]y_n[m] \tag{21}$$

where $y_n[m]$ is the sequence of length L given by

$$y_n[m] = \alpha \cdot (y_n[m-1] - y_{n-1}[m]) - y_{n-1}[m-1] \tag{22}$$

and initialized with $y_0[m] = x[m]$.

Note that we need to calculate $M+1$ warped autocorrelation coefficients (the additional coefficient is used in the compensation step).

(ii) *Calculation of the compensation warp factor*
To fit the final frequency axis to the mel-scale, we need to compensate for the first warping stage with value α in a second warping stage with the warp factor

$$\beta = \frac{\alpha - \alpha_{\text{mel}}}{1 - \alpha \cdot \alpha_{\text{mel}}}. \tag{23}$$

(iii) *Compensation for the spectral tilt*
To compensate for the distortion introduced by the concatenated bilinear transformations with warp factors α and β, we first concatenate the cascade of warping stages into a single warping stage with the warp factor

$$\chi = \frac{\alpha + \beta}{1 + \alpha \cdot \beta}. \tag{24}$$

A derivation of (24) is provided in [29]. To get a flat transfer function, we now apply the inverted weighting function

$$\left| \tilde{W}(\tilde{z}) \cdot \tilde{W}(\tilde{z}^{-1}) \right|^{-1} \tag{25}$$

to the warped autocorrelation coefficients, which can be realized as a second order finite impulse response filter:

$$\widehat{R}[m] = \frac{1 + \chi^2 + \chi \cdot \tilde{R}[m-1] + \chi \cdot \tilde{R}[m+1]}{1 - \chi^2}. \tag{26}$$

(iv) *Computation of the warped LPCs $\widehat{a}_{0\cdots M}^{(M)}$ including the warped prediction error variance $\widehat{\epsilon}_M$*
The warped LPCs can now be estimated using the Levinson-Durbin recursion [30], by replacing the linear autocorrelation coefficients R with their warped and spectral tilt compensated counterparts $\widehat{R}$.

(v) *Correlation of the warped LPCs*
The MVDR parameters $\widehat{\mu}_{-k}$ can be related to the LPC by

$$\widehat{\mu}_k = \begin{cases} \dfrac{1}{\widehat{\epsilon}_M} \displaystyle\sum_{m=0}^{M-k} (M+1-k-2m) \widehat{a}_m^{(M)} \widehat{a}_{m+k}^{*(M)} & , k \geq 0 \\ \widehat{\mu}_{-k}^* & , k < 0 \end{cases} \tag{27}$$

(vi) *Computation of the warped-twice MVDR envelope*
The spectral estimate can now be obtained by

$$S_{\text{W2MVDR}}(e^{j\omega}) = \frac{1}{\sum_{m=-M}^{M} \widehat{\mu}_m \frac{e^{j\omega} - \beta}{1 - \beta \cdot e^{j\omega}}}. \tag{28}$$

Note that the spectrum (28), if β is set appropriately, is already resembling the non-linear frequency axis as discussed in Section 3. In those cases it is necessary to either:

(a) eliminate the non-linear spaced triangular filterbank as for example used in the extraction of mel-frequency cepstral coefficients or perceptual linear prediction coefficients, or

(b) replace the non-linear spaced triangular filterbank by a filterbank of uniform half-overlapping triangular filters in order to provide feature reduction and additional spectral smoothing.

(vii) *Scaling of the warped-twice MVDR envelope*
To provide more robustness we match the warped-twice MVDR envelope to the highest spectral peak of the power spectrum.

Implementation Issues

Frequency warping including linear or non-linear VTLN can be realized using filterbanks. Carefully adjusted, those filterbanks can simulate the bilinear transformation in the frequency domain. In the case of warped-twice MVDR spectral estimation those filterbanks can be adjusted for each individual frame according to the compensation warp factor β and the VTLN parameter. In practice it is sufficient to

use a limited number of pre-calculated filterbanks; in this way, warped-twice MVDR spectral estimation can be implemented with only a very small overhead when compared to warped MVDR spectral estimation.

5. Steering Function

To support automatic speech recognition, the free parameters of the warped-twice MVDR envelope have to be adapted in such a way that classification relevant characteristics are emphasized while less relevant information is suppressed. Nakatoh *et al.* [20] proposed a method for steering the spectral resolution to lower or higher frequencies whereby for every frame i, the first two autocorrelation coefficients were used to define the *steering function*

$$\varphi_i = \frac{R_i[1]}{R_i[0]}. \tag{29}$$

The zero autocorrelation coefficient $R[0]$ represents the average power while the first autocorrelation coefficient $R[1]$ represents the correlation of a signal. Thus φ has a high value for voiced signals and a low value for unvoiced signals. Fig. 6 gives the different values of the normalized first autocorrelation coefficient φ averaged over all samples for each individual phoneme. A clear separation between the fricatives and non-fricatives can be observed. *Fricatives* are consonants produced by forcing air through a narrow channel made by placing two articulators close together. The *sibilants* are a particular subset of fricatives made by directing a jet of air through a narrow channel in the vocal tract towards the sharp edge of the teeth. Sibilants are louder than their non-sibilant counterparts, and most of their acoustic energy occurs at higher frequencies than by non-sibilant fricatives. A detailed discussion about the properties of different phoneme classes can be found in [1].

To adjust for the sensitivity to the steering function the factor γ is introduced, and the subtraction of the bias $\bar{\varphi} = \frac{1}{I}\sum_i \varphi_i$ (i.e., the mean over all values I in the training set) keeps the average of α close to α_{mel}. This leads to

$$\alpha_i = \gamma \cdot (\varphi_i - \bar{\varphi}) + \alpha_{\text{mel}}. \tag{30}$$

The last equation is a slight modification of the original formulation proposed by Nakatoh *et al.* As preliminary experiments have revealed that the word accuracy is not very sensitive to γ, we kept γ fixed at 0.1; values around 0.1 might lead to slightly, however, not significantly different results. The influence of γ has been, in more detail, investigated in [20].

6. Evaluation

To evaluate the proposed warped-twice MVDR spectral estimation and steering function against traditional front-ends such as *perceptual linear prediction* (PLP) [8], *mel frequency cepstral coefficients* (MFCC) [7] and more recently proposed front-ends based on warped-twice LP or warped MVDR spectral envelopes, we used NIST's development and evaluation data of the Rich Transcription 2005 Spring Meeting Recognition Evaluation [31]. The data has been chosen as a test environment as it contains challenging acoustic environments on both close and distant speech recordings. The development data, sampled at 16 kHz, consists of 5 seminars with approximately 130 minutes of speech. The evaluation data, also sampled at 16 kHz, consists of 16 seminars with approximately 180 minutes of speech. The data was collected under the *Computers in the Human Interaction Loop* (CHIL) project [32] and contains spontaneous, native and non-native speech.

We have used the *Janus Recognition Toolkit* (JRTk). To train acoustic models only relatively little supervised in-domain speech data is available. Therefore, we decided to train the acoustic models on close talking channels of meeting corpora and the *Translanguage English Database* (TED) corpus [33], summing up to a total of approximately 100 hours of acoustic training material. After split and merge training the acoustic model consisted of approximately 3,500 context-dependent codebooks with up to 64 diagonal covariance Gaussians each, summing up to a total of 180,000 Gaussians.

Each front-end provided features every 10 ms (first and second pass) or 8 ms (third pass). Spectral estimates have been obtained by the Fourier transformation (MFCC), PLP, warped MVDR, warped-twice LP and warped-twice MVDR spectral estimation. While the Fourier transformation is followed by a mel-filterbank, warped MVDR, warped-twice LP and warped-twice MVDR are followed by a linear filterbank. The 30 (13 or 20 in the case of PLP) spectral features have been truncated to 13 or 20 cepstral coefficients after cosine transformation. After mean and variance normalization, the cepstral features were stacked (seven adjacent left and right frames providing either 195 or 300 dimensions) and truncated to the final feature vector dimension of 42 by a multiplication with the optimal feature space matrix (the linear discriminant analysis matrix multiplied with the global semi-tied covariance transformation matrix [34]).

To train a four-gram language model, we used corpora consisting of broadcast news, proceedings of conferences such as ICSLP, Eurospeech, ICASSP, ACL and ASRU and talks in TED. The vocabulary contains approximately 23,000 words, the perplexity is 120 with an out-of-vocabulary rate of 0.25%.

We compare the different front-ends on class separability and *word error rate* (WER).

6.1. *Class Separability*

Class separability is a classical concept in pattern recognition, usually expressed using a scatter matrix. We can define

– the *within-class scatter matrix* ($\mathbf{S}_w$)

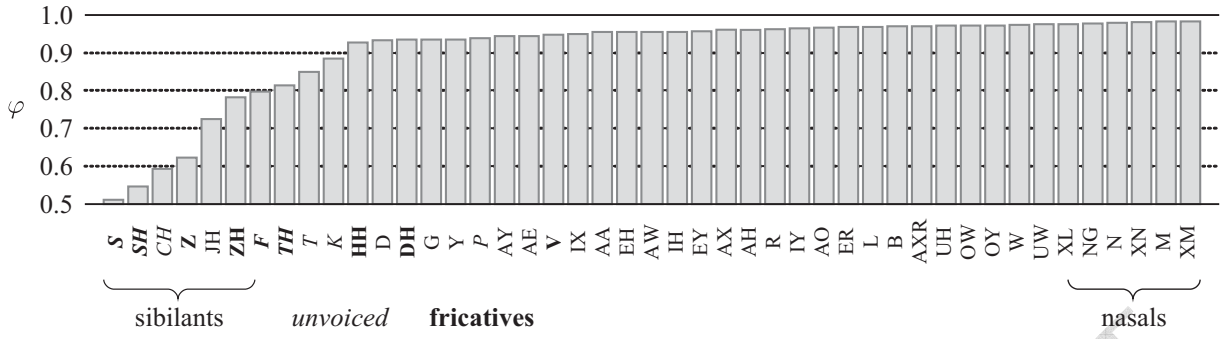


Fig. 6. Values of the normalized first autocorrelation coefficient by phonemes. Different phone classes group either for small values, e.g. sibilants, unvoiced (italic) and fricatives (bold) or for high values, e.g. nasals.

$$\mathbf{S}_w = \sum_{c=1}^{C} \left[\sum_{n=1}^{N_c} (\mathbf{x}_{cn} - \boldsymbol{\mu}_c)(\mathbf{x}_{cn} - \boldsymbol{\mu}_c)^T \right], \tag{31}$$

– the *between-class scatter matrix* ($\mathbf{S}_b$)

$$\mathbf{S}_b = \sum_{c=1}^{C} N_c (\boldsymbol{\mu}_c - \boldsymbol{\mu})(\boldsymbol{\mu}_c - \boldsymbol{\mu})^T \tag{32}$$

– and the *total scatter matrix* ($\mathbf{S}_t$)

$$\mathbf{S}_t = \mathbf{S}_w + \mathbf{S}_b = \sum_{c=1}^{C} \left[\sum_{n=1}^{N_c} (\mathbf{x}_{cn} - \boldsymbol{\mu})(\mathbf{x}_{cn} - \boldsymbol{\mu})^T \right], \tag{33}$$

where N_c denotes the number of samples in class c, $\boldsymbol{\mu}_c$ is the mean vector for the cth class, and $\boldsymbol{\mu}$ is the global mean vector over all classes C.

We would like to derive feature vectors such that all vectors belonging to the same class (e.g. phoneme) are close together in feature space and well separated from the feature vectors of other classes (e.g. all other phonemes). This property can be expressed using the scatter matrices; a small within-class scatter and a large between-class scatter stand for large class separability. Therefore, an approximate measure of class separability can be expressed by [35]

$$D_d = \mathrm{trace}_d \left\{ {\mathbf{S}_w}^{-1} \mathbf{S}_b \right\}, \tag{34}$$

where trace_d is defined as the sum of the first d eigenvalues λ_i of ${\mathbf{S}_w}^{-1} \cdot \mathbf{S}_b$ (a d-dimensional subspace) and hence the sum of the variances in the principal directions.

Comparing the class separability of different spectral estimation methods in Table 2 we first note that a higher number of cepstral coefficients always results in a higher class separability. Comparing the class separability, for 20 cepstral coefficients, on different front-ends we observe that class separability increases from PLP, warped-twice LP, warped MVDR, power spectrum to warped-twice MVDR. The class separability is significantly lower for PLP and significantly higher for warped-twice MVDR, while warped-twice LP, warped MVDR and power spectrum have nearly the same value.

On close talking microphone recordings in Table 3, we observe that warped-twice MVDR provides features with the highest separability on the development as well as the evaluation set. Averaging development and evaluation set the warped-twice MVDR is followed by warped MVDR, warped-twice MVDR, power spectrum and PLP. On distant microphone recordings, where the distance between speakers and microphones varies between approximately one and three meters, the power spectrum has the highest class separability on the development set. On the evaluation set, warped-twice MVDR performs equally well as warped MVDR, see Table 4. Averaging development and evaluation set on the distant data the power spectrum provides the highest class separability followed by warped-twice MVDR, warped-twice LP, warped MVDR and PLP.

6.2. *Word error rates*

The WERs of our speech recognition experiments for different spectral estimation techniques and recognition passes are shown for close talking microphone recordings in Table 3 and for distant microphone recordings in Table 4. The first pass is unadapted while the second and third pass are adapted on the hypothesis of the previous pass using *maximum likelihood linear regression* (MLLR) [36], constrained MLLR (CMLLR) [37] and VTLN [38].

Comparing the WERs of different spectral estimation methods in Table 2 we observe that a higher number of cepstral coefficients does not always result in a lower WER. Power spectra, warped and warped-twice MVDR envelopes tend to better performance with 20 cepstral coefficients while PLP performs better with 13 cepstral coefficients. The following discussion always refers to the lower WER. In average warped-twice MVDR provides the lowest WER followed by warped-twice LP and warped MVDR which perform equally well. PLP has a lower WER on the first and second pass which equals on the third compared to the power spectrum. PLP provides the lowest feature resolution which seems to be an advantage on the first pass, however, after model adaptation the lower feature resolution seems to be a disadvantage.

Investigating the WER on close microphone recordings, Table 3, we observe that the warped-twice MVDR front-end provides the best recognition performance, followed by PLP and warped-twice LP which are equally off. Warped MVDR ranks before the power spectrum which had the

Table 3
Class separability and word error rates for different front-end types and settings on close microphone recordings

Spectrum	Model	Cepstra	Class Separability			Word Error Rate %					
Test Set			Train	Develop	Eval	Develop			Eval		
Pass						1	2	3	1	2	3
power spectrum	–	13	11.007	16.470	16.088	36.1	30.3	28.0	35.3	29.7	27.7
power spectrum	–	20	11.620	17.929	16.299	36.0	29.7	27.7	37.2	31.3	28.4
PLP	13	13	10.699	17.110	15.152	34.7	29.3	27.2	34.2	29.6	27.1
PLP	20	20	11.029	18.059	16.068	34.7	29.5	27.7	34.9	30.3	27.9
warped MVDR	60	13	10.768	16.813	16.261	35.0	30.0	28.2	35.5	29.9	27.6
warped MVDR	60	20	11.337	18.022	16.614	34.5	29.1	27.3	35.3	29.6	27.3
warped-twice LP	20	13	10.772	17.038	16.254	35.3	30.5	28.5	36.2	29.8	27.1
warped-twice LP	20	20	11.333	17.864	16.436	34.4	29.5	27.4	37.1	29.4	26.8
warped-twice MVDR	60	13	10.893	17.673	16.456	34.5	29.5	27.5	34.1	29.2	27.0
warped-twice MVDR	60	20	11.473	18.510	16.818	34.1	28.8	26.8	35.4	29.0	26.3

Table 4
Class separability and word error rates for different front-end types and settings on distant microphone recordings

Spectrum	Model	Cepstra	Class Separability			Word Error Rate %					
Test Set			Train	Develop	Eval	Develop			Eval		
Pass						1	2	3	1	2	3
power spectrum	–	13	11.007	14.786	13.470	61.9	52.0	51.1	60.8	54.2	51.1
power spectrum	–	20	11.620	15.806	13.944	59.8	50.4	48.9	61.0	55.0	51.7
PLP	13	13	10.699	15.121	12.917	60.7	51.8	50.5	59.9	53.4	51.8
PLP	20	20	11.029	15.399	12.975	59.8	52.1	50.2	59.6	54.4	52.7
warped MVDR	60	13	10.768	13.836	13.885	62.9	53.7	52.0	60.7	52.8	50.7
warped MVDR	60	20	11.337	14.487	14.161	60.9	51.2	49.7	59.6	51.7	49.5
warped-twice LP	20	13	10.772	14.524	13.393	62.8	53.8	52.1	61.1	54.5	50.9
warped-twice LP	20	20	11.333	15.119	13.803	58.9	50.8	49.3	59.9	53.0	50.2
warped-twice MVDR	60	13	10.893	14.895	13.901	63.1	53.6	51.6	60.7	52.7	49.3
warped-twice MVDR	60	20	11.473	15.380	14.116	60.3	51.1	49.8	59.9	50.4	47.9

lowest recognition performance.

On distant microphone recordings, Table 3, the warped-twice MVDR front-end shows robust performance and has, in average, the lowest WER. On the development set, however, the power spectrum has the lowest WER. In average the warped-twice MVDR is followed by warped MVDR, then warped-twice LP, thereafter the power spectrum due to a weak performance on the evaluation set and PLP on the last place.

The reduced improvements of the warped-twice MVDR in comparison to the warped MVDR on distant recordings can be explained by the fact that, in comparison to close talking microphone recordings, the range of the values φ_i over all i is reduced. Therefore, the effect of spectral resolution steering is attenuated and consequently warped-twice MVDR envelopes behave more similarly to warped MVDR envelopes.

6.3. *Phoneme Confusability*

We investigate the confusability between phonemes by calculating the minimum distances, on the final features, between different phoneme pairs. In order to account for the range of variability of the sample points in both phoneme classes Ω_p and Ω_q, expressed by the covariance matrices $\mathbf{\Sigma}_p$ and $\mathbf{\Sigma}_q$, we extend the well known Mahalanobis distance by a second covariance matrix

$$D_{p,q} = \sqrt{(\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)^T \left(\mathbf{\Sigma}_p + \mathbf{\Sigma}_q\right)^{-1} (\boldsymbol{\mu}_p - \boldsymbol{\mu}_q)}.$$

Here $\boldsymbol{\mu}_p$ denotes the sample mean of phoneme class Ω_p and $\boldsymbol{\mu}_q$ denotes the sample mean of phoneme class Ω_q respectively.

As the comparison of the confusion matrix itself would be impractical, we limit our investigations on the comparison of the *distance* between the *nearest* phoneme to a given *phoneme* for different spectral estimation techniques

Table 5
Nearest phoneme distance for different phonemes (ordered by φ) and spectral estimation methods.

phoneme	S	SH	CH	Z	JH	ZH	F	TH	T	K	⋯	OW	OY	W	UW	XL	NG	N	XN	M	XM
φ	0.51	0.55	0.60	0.62	0.73	0.78	0.80	0.81	0.85	0.89	⋯	0.97	0.97	0.97	0.98	0.98	0.98	0.98	0.98	0.98	0.98
spectrum	power spectrum																				
nearest	Z	CH	JH	S	CH	JH	T	T	TH	P	⋯	XL	OW	B	UH	L	N	M	N	N	L
distance	2.41	1.56	0.81	2.27	1.36	1.55	2.36	2.04	1.75	2.33	⋯	3.19	3.55	3.04	2.97	2.94	3.32	2.83	3.59	3.04	4.88
spectrum	warped MVDR																				
nearest	Z	CH	JH	S	CH	JH	T	T	TH	P	⋯	XL	AY	B	UH	L	N	M	N	N	XL
distance	2.32	1.56	0.86	2.21	1.65	1.49	2.26	2.03	1.74	2.36	⋯	3.49	3.8	3.29	3.19	3.18	3.52	3.01	3.65	3.3	5.07
spectrum	warped-twice LP																				
nearest	Z	CH	JH	S	CH	JH	K	T	TH	P	⋯	XL	OW	B	UH	L	N	M	N	N	XL
distance	2.46	1.58	0.87	2.26	1.78	1.5	2.38	2.09	1.72	2.37	⋯	3.22	3.47	3.06	2.93	2.96	3.36	2.77	3.57	3.03	4.97
spectrum	warped-twice MVDR																				
nearest	Z	CH	JH	S	CH	JH	T	T	TH	P	⋯	XL	OW	B	UH	L	N	M	N	N	XL
distance	2.43	1.6	0.85	2.24	1.75	1.58	2.35	2.08	1.74	2.35	⋯	3.26	3.59	3.1	2.99	3	3.36	2.83	3.56	3.12	5.02

Table 2
Average class separability and average word error rates for different front-end types and sanity checks
MO: model order, CC: number of cepstral coefficients, CS: class separability

Spectrum	MO	CC	CS	Word Error Rate %		
Pass				1	2	3
power spectrum	–	13	15.204	48.5	41.6	39.5
power spectrum	–	20	15.995	48.5	41.6	39.2
PLP	13	13	15.075	47.4	41.0	39.2
PLP	20	20	15.625	47.3	41.6	39.6
warped MVDR	60	13	15.199	48.5	41.6	39.6
warped MVDR	60	20	15.821	47.6	40.4	38.5
warped-twice LP	20	13	15.302	48.9	42.1	39.6
warped-twice LP	20	20	15.806	47.6	40.7	38.4
warped-twice MVDR	60	13	15.731	48.1	41.3	38.9
warped-twice MVDR	60	20	16.206	47.4	39.8	37.7

as plotted in Tabel 5. Note that the PLP front-end is excluded from this analysis as it, due to a different scale, can not be directly compared. By comparing the nearest phoneme pairs over different phonemes and spectral estimation methods we observe that different spectral representations result in slightly different phoneme pairs. In addition we observe that, in average, phonemes with a small value of φ are easier confused (smaller distance) with other phonemes than phonemes with a high φ value. This can be explained by the energy of the different phoneme classes where the phoneme classes belonging to small φ values contain less energy and are thus stronger distorted by background noise.

Comparing the power spectrum with the warped MVDR envelope we observe that the power spectrum tends to provide lower confusability for lower φ values and higher confusability for higher φ values. The warped-twice LP and warped-twice MVDR envelopes have a similar distance structure over φ, with in average larger distances for the warped-twice MVDR envelopes. While the warped-twice MVDR envelope, compared to the warped MVDR envelope, provides a lower confusability for small values of φ, the confusability is higher for larger values of φ. While the warped MVDR envelope is not capable to provide a lower confusability over the whole range of φ in comparison to the power spectrum, the warped-twice MVDR envelope provides, in average, a lower confusability over the whole range of φ in comparison to the power spectrum.

7. Conclusion

We have introduced warped-twice MVDR spectral estimation by extending warped MVDR estimation with a second bilinear transformation. With these extensions, it is possible to steer spectral resolution to lower or higher frequencies while keeping the overall resolution of the estimate and the frequency axis fixed. We have demonstrated one possible application in the front-end of a speech-to-text system by steering the resolution of the spectral envelope to classification relevant spectral regions. The proposed framework showed consisted improvements in terms of class separability and WER on a large vocabulary speech recognition task on close talk as well as on distant speech recordings. Further improvements might be expected by a more suitable steering function.

References

[1] J. Olive, Acoustics of American English speech : A dynamic approach, Springer, 1993.

[2] N. Mesgarani, S. David, S. Shamma, Representation of phonemes in primary auditory cortex: Howthe brain analyzes speech, Proc. of ICASSP (2007) 765–768.
[3] V. Driaunys, K. Rudzionis, P. Zvinys, Analysis of vocal phonemes and fricative consonant discrimination based on phonetic acoustics features, Information Technology and Control 34 (3) (2005) 257–262.
[4] M. Wölfel, Warped-twice minimum variance distortionless response spectral estimation, Proc. of EUSIPCO (2006) .
[5] M. Wölfel, J. McDonough, A. Waibel, Minimum variance distortionless response on a warped frequency scale, Proc. of Eurospeech (2003) 1021–1024.
[6] M. Wölfel, J. McDonough, Minimum variance distortionless response spectral estimation, review and refinements, IEEE Signal Processing Magazine 22 (5) (2005) 117–126.
[7] S. Davis, P. Mermelstein, Comparison of parametric representations for monosyllable word recognition in continuously spoken sentences, IEEE Trans. on Acoustics, Speech and Signal Processing 28 (4) (1980) 357–366.
[8] H. Hermansky, Perceptual linear predictive (PLP) analysis of speech, J. Acoustic. Soc. Am. 87 (4) (1990) 1738–1752.
[9] M. Murthi, B. Rao, Minimum variance distortionless response (MVDR) modeling of voiced speech, Proc. of ICASSP (1997) 1687–1690.
[10] M. Murthi, B. Rao, All-pole modeling of speech based on the minimum variance distortionless response spectrum, IEEE Trans. on Speech and Audio Processing 8 (3) (2000) 221–239.
[11] J. Capon, High-resolution frequency-wavenumber spectrum analysis, Proc. of the IEEE 57 (1969) 1408–1418.
[12] B. Musicus, Fast MLM power spectrum estimation from uniformly spaced correlations, IEEE Trans. on Acoustics, Speech, and Signal Processing 33 (1985) 1333–1335.
[13] S. Dharanipragada, B. Rao, MVDR based feature extraction for robust speech recognition, Proc. of ICASSP (2001) 309–312.
[14] S. Dharanipragada, U. Yapanel, B. Rao, Robust feature extraction for continuous speech recognition using the MVDR spectrum estimation method, IEEE Trans. Speech and Audio Processing 15 (1) (2007) 224–234.
[15] S. Stevens, J. Volkman, E. Newman, The mel scale equates the magnitude of perceived differences in pitch at different frequencies, J. Acoust. Soc. Am. 8 (3) (1937) 185–190.
[16] G. U. Yule, On a method of investigating periodicities in disturbed series, with special reference to wolfers sunspot numbers, Phil. Trans. Roy. Soc. 226–A (1927) 267–298.
[17] J. Makhoul, Linear prediction: A tutorial review, Proc. of the IEEE 63 (4) (1975) 561–580.
[18] H. Strube, Linear prediction on a warped frequency scale, J. Acoustic. Soc. Am. 68 (8) (1980) 1071–1076.
[19] H. Matsumoto, M. Moroto, Evaluation of mel-LPC cepstrum in a large vocabulary continuous speech recognition, Proc. of ICASSP (2001) 117–120.
[20] Y. Nakatoh, M. Nishizaki, S. Yoshizawa, M. Yamada, An adaptive mel-LP analysis for speech recognition, Proc. of ICSLP (2004) .
[21] S. Haykin, Adaptive filter theory—3th ed., Prentice Hall, 1991.
[22] A. Oppenheim, D. Johnson, K. Steiglitz, Computation of spectra with unequal resolution using the fast fourier transform, IEEE Proc. Letters 59 (2) (1971) 229–301.
[23] C. Braccini, A. V. Oppenheim, Unequal bandwidth spectral analysis using digital frequency warping, IEEE Trans. Acoust., Speech, Signal Processing 22 (1974) 236–244.
[24] J. O. Smith III, J. S. Abel, Bark and ERB bilinear transforms, IEEE Trans. on speech and audio processing 7 (6) (1999) 697–708.
[25] N. Nocerino, F. Soong, L. Rabiner, D. Klatt, Comparative study of several distortion measures for speech recognition, Proc. of ICASSP (1985) 25–28.
[26] A. Härmä, U. Laine, A comparison of warped and conventional linear predictive coding, IEEE Trans. on Speech and Audio Processing 9 (5) (2001) 579–588.
[27] M. Wölfel, Mel-Frequenzanpassung der Minimum Varianz Distortionless Response Einhüllenden, Proc. of ESSV (2003) 22–29.
[28] M. Matsumoto, Y. Nakatoh, Y. Furuhata, An efficient mel-LPC analysis method for speech recognition, Proc. of ICSLP (1998) 1051–1054.
[29] A. Acero, Acoustical and environmental robustness in automatic speech recognition, Ph.D. thesis, Carnegie Mellon University (September 1990).
[30] A. Oppenheim, R. Schafer, Discrete-time signal processing, Prentice-Hall Inc., 1989.
[31] NIST, Rich transcription 2005 spring meeting recognition evaluation, www.nist.gov/speech/tests/rt/rt2005/spring.
[32] Computers in the human interaction loop, http://chil.server.de.
[33] Linguistic Data Consortium (LDC), Translanguage english database, www.ldc.upenn.edu/Catalog/LDC2002S04.html.
[34] M. Gales, Semi-tied covariance matrices for hidden Markov models, IEEE Trans. Speech and Audio Processing 7 (1999) 272–281.
[35] R. Haeb-Umbach, Investigations on inter-speaker variability in the feature space, Proc. of ICASSP (1999) 397 – 400.
[36] C. Leggetter, P. Woodland, Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models, Computer Speech and Language 9 (2) (1995) 171–185.
[37] M. J. F. Gales, Maximum likelihood linear transformations for HMM-based speech recognition, Computer Speech and Language 12 (1998) 75–98.
[38] L. Welling, H. Ney, S. Kanthak, Speaker adaptive modeling by vocal tract normalization, IEEE Trans. Speech and Audio Processing 10 (6) (2002) 415–426.