Minimum Variance Distortionless Response on a Warped Frequency Scale

Matthias Wölfel, John McDonough, Alex Waibel

Institut für Logik, Komplexität und Deduktionssysteme, Universität Karlsruhe (TH), Germany
Interactive Systems Laboratories, Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

In this work we propose a time domain technique to estimate an all-pole model based on the *minimum variance distortionless response* (MVDR) using a *warped* short time frequency axis such as the Mel scale. The use of the MVDR eliminates the overemphasis of harmonic peaks typically seen in medium and high pitched voiced speech when spectral estimation is based on *linear prediction* (LP). Moreover, warping the frequency axis prior to MVDR spectral estimation ensures more parameters in the spectral model are allocated to the low, as opposed to high, frequency regions of the spectrum, thereby mimicking the human auditory system. In a series of speech recognition experiments on the *Switchboard Corpus* (spontaneous English telephone speech), the proposed approach achieved a *word error rate* (WER) of 32.1% for female speakers, which is clearly superior to the 33.2% WER obtained by the usual combination of Mel warping and linear prediction.

1. Introduction

It is well known that an all-pole model applies equal resolution to all frequency bands. Human hearing, on the other hand, has greater resolution for low frequencies than for high. To mimic human hearing, the spectrum obtained from all-pole analysis is typically warped with a Mel-filterbank. This "postprocessing" does not, however, lead to an improvement of the resolution of the envelope in lower frequencies. To achieve this higher resolution, Strube [9] proposed a method based on the *bilinear transform* wherein the short-time frequency axis is warped *prior* to all-pole analysis. Such warping can be applied to *linear prediction* (LP) using the Mel-frequency as a warping factor. For male speakers, this *Mel-warped-LP* (MWLP) provides a significant reduction in *word error rate* (WER) over standard LP, and a slight reduction with respect to *Mel-frequency cepstral coefficients* (MFCC) [5]. For female speakers, MWLP performs either approximately as well or slightly worse than the other methods. The poorer performance of MWLP for female speakers can be explained as follows: The spectral envelope obtained from LP tends to overestimate and overemphasize sparsely spaced harmonic peaks. Female speakers have more sparsely spaced harmonic peaks than male speakers, because their voices in general have higher fundamental frequencies. To overcome this problem, Murthi and Rao [7] proposed the use of high order *minimum variance distortionless response* (MVDR) all-pole models and showed the superiority of such models to LP all-pole models for medium and high pitched (i.e., female) voiced speech.

Based on the foregoing, we are led to consider how best to combine the desireable emphasis of the lower frequency bands seen in the human auditory system with superior spectral estimate provided by the MVDR. To achieve this combination, we propose in this work a refinement of MVDR all-pole models hereafter known as *Mel-warped-MVDR* (MWMVDR) all-pole models. Furthermore, adapting the approach by Musicus [8] for a fast MVDR derivation from the LP coefficients, we show that the MWMVDR can be readily obtained form the MWLP coefficients.

2. Theoretical Background

MVDR spectral estimation, from the point of view of filterbank analysis, is a problem of filter design subject to the *distortionless constraint* which is stated as [2]:

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

$$H(e^{j\omega_{\text{foi}}}) = \sum_{k=0}^{M} h^*(k) e^{-jk\omega_{\text{foi}}} = 1$$

where $h^*(k)$ are components in impulse response of $H(e^{j\omega})$. This can also be written in vectorform:

$$\mathbf{s}^H(\omega_{\text{foi}}) \cdot \mathbf{h}^*{}_{\text{foi}} = 1$$

where $\mathbf{s}(\omega)$ is the *fixed frequency vector*

$$\mathbf{s}(\omega) = [1, e^{-j\omega}, \ldots, e^{-jM\omega}]^T$$

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

This scheme may be generalized by replacing the unit delay elements $e^{-jm\omega}$ of the fixed frequency vector $\mathbf{s}(\omega)$ with all-pass selections; i.e., the *first order all-pass filter*

$$e^{-j\tilde{\omega}} = D_1(e^{-j\omega}) = \frac{e^{-j\omega} - \alpha}{1 - \alpha \cdot e^{-j\omega}}$$

where α is a *warping parameter* and $D_1(e^{-j\omega})$ is a *warped delay element*. The phase function of $D_1(e^{-j\omega})$ is [5]

$$\arg\left(D_1(e^{-j\omega})\right) = \tilde{\omega} = \omega + 2\arctan\frac{\lambda\sin\omega}{1-\lambda\cos\omega}$$

which is also known as the *frequency mapping function*. Therefore, the linear frequency axis ω is transformed to the warped frequency axis $\tilde{\omega}$, resulting in the frequency-warped spectrum $\tilde{S}(e^{j\tilde{\omega}})$. Using a particular warp factor enables the approximation of the *Mel-frequency* as shown in Figure 1.

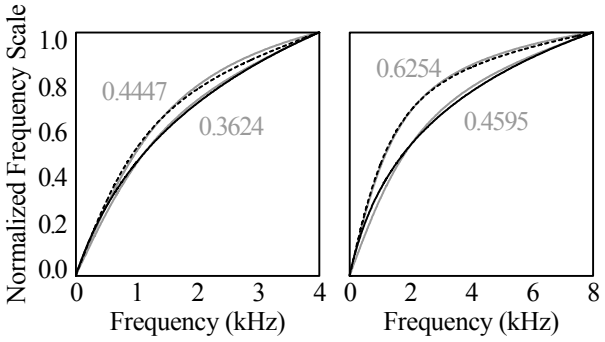

Figure 1: The approximations of Mel-frequency (black lines) and Bark-frequency (dotted black lines) by the bilinear transformation (gray lines including the warping factor in gray digits) are demonstrated for 8 and 16 kHz sampling rates.

This generalization results in the *warped frequency vector*:

$$\tilde{\mathbf{s}}(\omega) = \left[1, \frac{e^{-j\omega}-\alpha}{1-\alpha\cdot e^{-j\omega}}, \ldots, \frac{e^{-jM\omega}-\alpha}{1-\alpha\cdot e^{-jM\omega}}\right]^T \quad (1)$$

The distortionless filter $\mathbf{h}_{\text{foi}}$ can now be obtained by the *warped constrained minimization problem* which minimizes the output power of the overall warped frequency domain:

$$\min_{\mathbf{h}_{\text{foi}}} \mathbf{h}_{\text{foi}}^H \phi_{M+1} \mathbf{h}_{\text{foi}} \;\text{ subject to }\; \tilde{\mathbf{s}}^H(\omega_{\text{foi}})\mathbf{h}_{\text{foi}} = 1$$

where ϕ_{M+1} is the $(M+1)\cdot(M+1)$ Toeplitz autocorrelation matrix of the filter output:

$$y(i) = \sum_{l=0}^{M} h^*(l)u(i-l)$$

The solution of the warped constrained minimization problem is very similar to its unwarped counterpart, as given in [2]:

$$\tilde{S}_{\text{MV}}(\omega) = \frac{1}{\tilde{\mathbf{s}}^H(\omega)\phi^{-1}\tilde{\mathbf{s}}(\omega)}$$

Under the assumption that the $(M+1)\cdot(M+1)$ Hermitian Toeplitz correlation matrix ϕ is positive definite and thus invertible, Musicus [8] has derived a fast algorithm to calculate the MVDR spectrum from the LP coefficients. As the warped-MVDR spectrum can be obtained from the warped-LP coefficients, Musicus algorithm can be readily extended to compute the warped-MVDR spectrum as follows:

1. **Calculation of the warped-LP coefficients**
   For our experiments we used an algorithm by Matsumoto et al. [5] to calculate the warped-LP coefficients.

2. **Correlation of the warped prediction coefficients**

$$\tilde{\mu}_k = \begin{cases} \sum_{i=0}^{N-k}(N+1-k-2i)\tilde{a}_i^{(N)}\tilde{a}_{i+k}^{*(N)} \\ \qquad\qquad : k = 0, \cdots, N \\ \tilde{\mu}_{-k}^* \qquad\qquad : k = -N, \cdots, -1 \end{cases}$$

3. **Fast warped MVDR spectrum computation**

$$S_{\text{warped MV}}(\omega) = \frac{\epsilon}{\sum_{k=-M}^{M}\tilde{\mu}_k e^{-j\omega k}} \quad (2)$$

Note that the spectrum calculated through (2) is on the warped frequency scale and therefore we have to replace the Mel-filterbank with a filterbank of uniformly half overlapping triangular filters in the acoustic preprocessing of an automatic speech recognizer. If we are directly interested in a spectral envelope on the linear frequency scale we can use

$$\tilde{S}_{\text{MV}}(\omega) = \frac{\epsilon}{\sum_{k=-M}^{M}\tilde{\mu}_k \frac{e^{-jk\omega}-\alpha}{1-\alpha\cdot e^{-jk\omega}}}$$

instead of (2). This envelope is different from the conventional MVDR envelope inasmuch as it uses more parameters to describe the lower frequencies and fewer parameters to describe the higher; the conventional MVDR uses an equal number of parameters for both.

Figure 2 illustrates the difference between the MVDR and MWMVDR spectral envelopes. The warp factor for the MWMVDR was set to 0.4595 so as to simulate the Mel-scale for a 16 kHz sampling frequency. While the MVDR exhibits frequency-independent spectral resolution, the warped-MVDR provides high resolution for frequencies below 2 kHz and decreasing resolution for higher frequencies. The warping of the MVDR provides an interesting property which cannot be achieved when the MVDR is *followed* by frequency-warping: The residuals show spectral flattening and level compensation similar to the adaptation of the firing rate in the auditory nerve. This results in information of the MWMVDR residuals which resembles the overall information in the

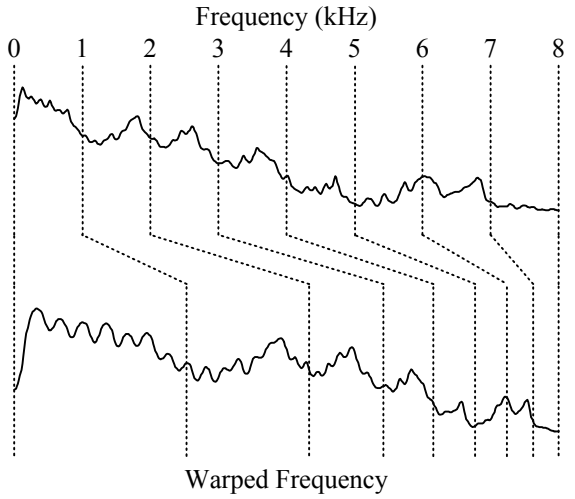

Figure 2: Comparison of MVDR (top) and Mel-warped-MVDR (bottom) spectral envelopes, both of same model order 120.

auditory nerve firing similar to MWLP [4], but without the negative effect of overestimating and overemphasizing the harmonic peaks in medium and high pitched voiced speech.

3. Speech Recognition Experiments

The speech recognition experiments described below were conducted with the *Janus Recognition Toolkit* (JRTk), which is developed and maintained jointly at the Universität Karlsruhe (TH), in Karlsruhe, Germany and at the Carnegie Mellon University in Pittsburgh, Pennsylvania, USA.

Our recognition experiments were conducted on the *Switchboard Corpus* using 548 speakers of both sexes for training, and 5 male and 5 female speakers for testing. We used a baseline model with 32 Gaussians for each of 4,166 codebooks for a total of 133,312 Gaussians. All features were calculated every 10 ms from speech data sampled at 8 kHz, using a 20 ms Hamming window. To compare our proposed method 13 cepstral components, along with their first and second differences were derived by a discrete cosine transform using *cepstral coefficients* (CC) from different spectral representations:

- The *fast Fourier transform* (FFT), the LP and the MVDR, all followed by a Mel-filterbank consisting of 30 half-overlapping Mel-spaced triangularly shaped filters.
- The MWLP and the MWMVDR, both followed by a filterbank consisting of 30 half-overlapping uniformly-spaced triangularly shaped filters.

In order to provide a good comparison of all investigated spectral estimation techniques, all spectral envelopes were reconstructed and scaled to the maximum peak of the Fourier spectrum [10]. The difference between the MVDR and warped-MVDR acoustic preprocessing is also shown in Figure 3. To compensate for channel variations cepstral mean normalization was used. Linear discriminant analysis was used to reduce the final feature length to 32.

It should be noted that *vocal tract length normalization* (VTLN) has to be implemented differently between the methods with and without warping, one in the linear frequency domain and the other in the warped frequency domain and therefore, for the sake of a better comparison between the different methods, VTLN was not used.

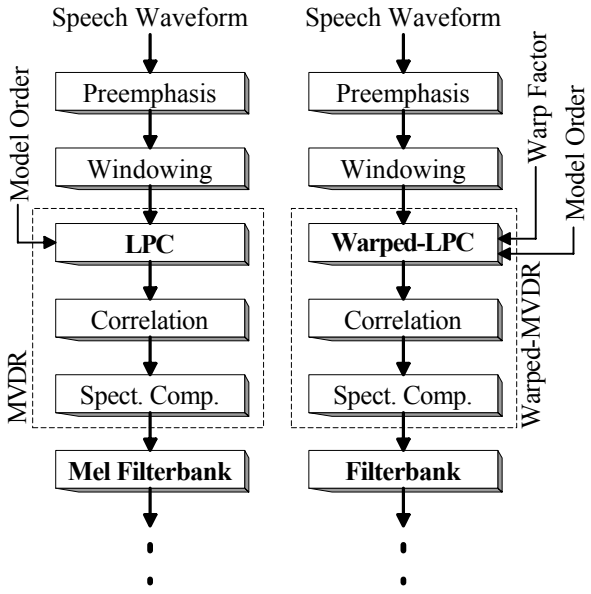

Figure 3: Extract of the MVDR (left) and the warped-MVDR (right) acoustic preprocessing used in our experiments.

4. Discussion

With the results of our experiments, Table 1, we can confirm that MWLP provides an improvement in recognition accuracy over LP in general and MFCC for male speakers. We can also confirm that the suggested use of MVDR is superior or equal to the aforementioned approaches for male speakers while for female speakers its performance is superior. Furthermore, our proposed method is able to further improve the results, at least for female speakers, as reached through the use of MVDR alone which confirms our theoretical considerations.

The gain of spectral envelope techniques, except for LP because of its limited ability to approximate the spectrum, over the Fourier approach can be explained by the way in which they differ in the representation of spectral peaks and valleys: While Fourier spectra describes spectral peaks and valleys equally well spectral envelopes

		FFT	*LP(13)*	*MWLP(13)*	*MVDR(80)*	*MWMVDR(40)*
Five Male Speakers	*1*	26.8	27.2	26.5	26.8	25.9
	2	31.4	30.8	29.9	33.7	30.1
	3	40.0	41.1	40.8	39.2	41.1
	4	42.8	43.4	40.7	41.1	42.4
	5	46.9	44.7	47.6	44.9	45.8
Average Male		37.6	37.4	37.1	37.1	37.1
Five Female Speakers	*1*	23.8	24.2	24.2	24.4	23.0
	2	39.8	41.5	40.8	37.7	39.2
	3	40.7	39.6	39.1	40.7	38.9
	4	29.0	29.6	29.5	28.3	27.9
	5	32.8	32.8	32.3	32.6	31.7
Average Female		33.2	33.5	33.2	32.7	32.1
Average Overall		**35.4**	**35.5**	**35.1**	**34.9**	**34.6**

Table 1: Comparison of word error rates. The numbers in brackets show the optimal and used model order.

provide an accurate description only for spectral peaks. For the representation of spectral valleys no information of the fine structure of the spectrum is considered, limiting the description more or less to the energy levels. As noise, in the logarithmic magnitude domain, is most evident in spectral valleys, spectral envelopes are more robust to noise than there Fourier counterpart.

5. Conclusions

This paper has presented a method for warped all-pole modelling of the MVDR leading to a higher spectral resolution of the envelope in low frequency bands while the use of MVDR instead of LP provides a better modelling of medium and high pitched voice speech.

It has been shown that the performs of the proposed method, Mel-warped-MVDR, is superior to all other spectral envelope techniques presented in this paper as well as the FFT-based MFCC approach. Particular in the case of female speakers the MWMVDR clearly demonstrates its ability to provide a good spectral envelope.

Further work will focus on the implementation of preemphasis, intensity-loudness conversion and a filterbank refinement to further improve the system performance; e.g., to compensate for the approximated Mel-spectrum or/and by replacing the triangular filterbank by a critical-band filter which is flat topped and non-symmetric as typically used in the perceptual linear predictive approach [3]. Furthermore, as the bilinear transform was shown to be successful in speaker-dependent vocal tract length normalization [6], we want to use a varying warping factor, instead of the one fixed to the Mel-frequency.

6. Acknowledgement

The work presented here was partly funded by the *European Union* (EU) under the Grant number IST-2000-28323. Any opinions, findings and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the EU.

7. References

[1] Capon, J. High-resolution frequency-wavenumber spectrum analysis. *Proc. IEEE*, vol. 57:pp. 1408–1418, August 1969.

[2] Haykin, S. *Adaptive filter theory—3th ed.* Prentice Hall, 1991.

[3] Hermansky, H. Perceptual linear predictive. *Journal Acoustic Society of America*, vol. 87, no. 4:pp. 1738–1752, 1990.

[4] Karjalainen, M. Auditory interpretation and application of warped linear prediction. *Proceedings of Consistent & Reliable Acoustic Cues for Sound Analysis*, 2001.

[5] Matsumoto, H. and Moroto, M. Evaluation of Mel-LPC cepstrum in a large vocabulary continuous speech recognition. *IEEE International Conference on Acoustic, Speech, and Signal Processing*, vol. 1:pp. 117–120, 2001.

[6] McDonough, J.W. *Speaker compensation with all-pass transforms.* Ph.D. thesis, Johns Hopkins University, Baltimore, USA, 2000.

[7] Murthi, M.N. and Rao, B.D. All-pole modeling of speech based on the minimum variance distortionless response spectrum. *IEEE International Conference on Acoustic, Speech, and Signal Processing*, vol. 8(no. 3):pp. 221–239, 2000.

[8] Musicus, B.R. Fast MLM power spectrum estimation from uniformly spaced correlations. *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 33:pp. 1333–1335, 1985.

[9] Strube, H.W. Linear prediction on a warped frequency scale. *Journal Acoustic Society of America*, vol. 68(no. 8):pp. 1071–1076, 1980.

[10] Wölfel, M.C. *Minimum variance distortionless response spectral estimation and subtraction for robust speech recognition.* Diploma thesis, Universität Karlsruhe (TH), Karlsruhe, Germany, January 2003.