14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

WARPED-TWICE MINIMUM VARIANCE DISTORTIONLESS RESPONSE SPECTRAL ESTIMATION

Matthias Wölfel

Institut für Theoretische Informatik, Universität Karlsruhe (TH)
Am Fasanengarten 5, 76131 Karlsruhe, Germany
wolfel@ira.uka.de

ABSTRACT

This paper describes a novel extension to warped *minimum variance distortionless response* (MVDR) spectral estimation which allows to steer the resolution of the spectral envelope estimation to lower or higher frequencies while keeping the overall resolution of the estimate and the frequency axis fixed. This effect can be achieved by the introduction of a second bilinear transformation to the warped-MVDR spectral estimation, but now in the frequency domain as opposed to the first bilinear transformation which is applied in the time domain, and a compensation step to adjust for the pre-emphasis of both bilinear transformations. In the feature extraction process of an automatic speech recognition system this novel extension allows to emphasize classification relevant characteristics while dropping classification irrelevant characteristics of speech features according to the characteristics of the signal to analyze, e.g. vowels and fricatives have different characteristics and therefore should be treated differently. We have compared the novel extension on evaluation data of the Rich Transcription 2005 Spring Meeting Recognition Evaluation to warped-MVDR and got an word error rate reduction from 28.2% to 27.5%.

1. INTRODUCTION

To improve phoneme classification it is important to emphasize the relevant characteristics while dropping the irrelevant characteristics for classification. In the traditional feature extraction process of an automatic speech recognition system this is achieved by successive implementations (e.g, a spectral envelope or/and a filterbank followed by cepstral transformation, cepstral normalization and linear discriminant analysis) treating all phoneme types equally. As for different phoneme types the important regions on the frequency axis vary [1], e.g. low frequencies for vowels and high frequencies for fricatives, it is a natural extension to the traditional approach to vary the spectral resolution depending on the phoneme to calculate an acoustic score for. To provide a framework to allow for these adjustments we propose an extension to the warped *minimum variance distortionless response* (MVDR) by a second bilinear-transformation. This novel spectral envelope estimate has two ways of freedom to control spectral resolution. The first is the number of linear prediction coefficients also referred to as model order. E.g. increasing the model order for the underlying linear parametric model is also increasing the overall spectral resolution and vise versa. The second, novel way of freedom, is a compensated warp factor which allows to steer spectral resolution to lower or higher frequency regions without changing the frequency axis.

2. WARPED-TWICE MVDR SPECTRAL ENVELOPE ESTIMATION

The use of MVDR as a spectral envelope technique was previously proposed by Murthi and Rao [2, 3] and applied to speech recognition by Dharanipragada and Rao [4]. Moreover, to ensure that more parameters in the spectral model are allocated to the low, as opposed to the high, frequency regions of the spectrum, thereby mimicking the frequency resolution of the human auditory system, we have extended this approach by *warping* the frequency axis with the bilinear transformation prior to MVDR spectral estimation [5, 6], therefore dubbed *warped MVDR*. Like Nakatoh et al. [7] on the *linear prediction coefficients* (LPC)s, we can further extend the MVDR approach by a second warping, yet in the frequency domain rather than in the time domain as the first warping, dubbed *warped-twice MVDR* (W2MVDR). By *compensating* for the first warp with the second it allows to steer spectral resolution to lower or higher frequencies without changing the frequency axis.

The influence of the model order, the warping and the compensated warping, leaving the warped frequency fixed to the Mel-frequency, is more apparent if we depict an example. While the model order varies the overall resolution of the spectral estimate, see Figure 2a, the warp factor is bending the frequency axis and therefore can be used to apply the Mel-frequency or to implement vocal tract length normalization (not used in our experiments as the traditional approach of piece-wise linear warping is leading to better results), shown in Figure 2b. Even though the bending of the frequency axis can be applied in the time or frequency domain the effect on the spectral resolution differs. Applying the bilinear transformation in the time domain moves more coefficients to lower or higher frequencies before spectral analysis and therefore resulting in an increase or decrease of resolution in lower or higher frequency regions. E.g. let the first warp factor be a value smaller than the Mel-frequency $\alpha < \alpha_{\text{Mel}}$, than spectral resolution improves in high frequency regions and decreases in low frequency regions in comparison to the spectral resolution provided by Mel-frequency. The bilinear transformation applied on the frequency axis, on the other hand, is 'only' bending the already determined spectrum and therefore the spectral resolution is not changed. A new aspect comes into play if the second warp factor is set to compensate for the first warping with the goal to match a particular fixed, warped frequency axis, like the Mel-frequency. In this case the resolution can be moved to low or high frequency regions without changing the frequency axis, see Figure 2c.

2.1 Fast Computation

For a fast computation of the W2MVDR envelope we have extended Musicus' [8] algorithm to calculate the MVDR envelope of model order N from the LPC $a^{(N)}_{0\cdots N}$ of model order N as follows:

1. **Computation of the warped autocorrelation coefficients $\tilde{R}_{0\cdots N+1}$**
   To derive 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*

$$z = \frac{z^{-1} - \alpha}{1 - \alpha \cdot z^{-1}} \tag{1}$$

   Therefore we can derive the warped autocorrelation coefficients by

$$\tilde{R}[m] = \sum_{n=0}^{\infty} \tilde{x}[n]\tilde{x}[n-m] \tag{2}$$

   where the frequency-warped speech signal $\tilde{x}$ is defined by

$$\tilde{X}(\tilde{z}) = \sum_{n=0}^{\infty} \tilde{x}[n]\tilde{z}^{-n} = \sum_{n=0}^{N-1} x[n]z^{-n} \tag{3}$$

   Due to the effect that a bilinear transformed *finite* sequence results in an *infinite* sequence, the direct calculation of warped autocorrelation coefficients is not feasible. To overcome this problem a variety of solutions exists. In our experiments we have used the algorithm proposed by Matsumoto and Moroto [9].
   Note that we have to calculate $N+1$ coefficients as we need the additional coefficient in the compensation step.
2. **Calculation of the compensation warp parameter**
   To fit the final frequency axis to the Mel-frequency α_{Mel} we have to compensate the first warp α by a second warp:

$$\beta = \frac{\alpha - \alpha_{\text{Mel}}}{1 - \alpha \cdot \alpha_{\text{Mel}}} \tag{4}$$

   For a signal sampled at 16 kHz α_{Mel} has to be 0.4595.
3. **Compensation of the pre-emphasis**
   To derive the distortion introduced by the concatenated bilinear transformations with warp values α and β we calculate the phase delay by a frequency derivative of one bilinear transformation (1) with the warp value

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

   and express the result in the *weighting function*:

$$\left|\tilde{W}(\tilde{z})\right|^2 = \frac{1 - \chi^2}{(1 + \chi \cdot \tilde{z}^{-1})^2} \tag{5}$$

   This is a pre-emphasis filter causing the spectrum at the output to be not perfectly flat. To compensate for this unwanted effect, to get a flat spectrum, we have to apply the inverted weighting function

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

   to the warped autocorrelation coefficients $\tilde{r}$. This can be realized as a second order finite impulse response filter:

$$\widehat{R}[i] = \frac{1 + \chi^2 + \chi \cdot \tilde{R}[i-1] + \chi \cdot \tilde{R}[i+1]}{1 - \chi^2} \tag{6}$$

   where $\tilde{R}[-1] = \tilde{R}[1]$.
   The effect of the pre-emphasis of the bilinear-transformations and its complete compensation are depicted in Figure 1.
4. **Computation of the warped-LPCs $\widehat{a}^{(N)}_{0\cdots N}$**
   The warped-LPCs can now be estimated by the Levinson-Durbin recursion [10] replacing the linear autocorrelation coefficients R with their warped and pre-emphasis compensated counterparts $\widehat{R}$.
5. **Correlation of the warped-LPC**

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

6. **Computation of the W2MVDR envelope**

$$S_{\text{W2MVDR}}(\omega) = \frac{\varepsilon}{\sum_{k=-N}^{N} \widehat{\mu}_k \frac{e^{j\omega} - \beta}{1 - \beta \cdot e^{j\omega}}} \tag{7}$$

   ε: inverse of the prediction error variance.

   Note that (7) is already in the Mel-warped frequency domain and therefore we have to replace the Mel-filterbank in the front-end of a speech recognition system by a filterbank of uniformly half overlapping triangular filters.
7. **Scaling of the W2MVDR envelope**
   With the relation $\log(a+b) \approx \log(\max\{a,b\})$ we can conclude that *spectral peaks* are particular robust to additive noise in the logarithmic domain. Therefore, to get features which are less distorted by additive noise we match the W2MVDR envelope to the highest spectral peak of the Fourier spectrum [6].

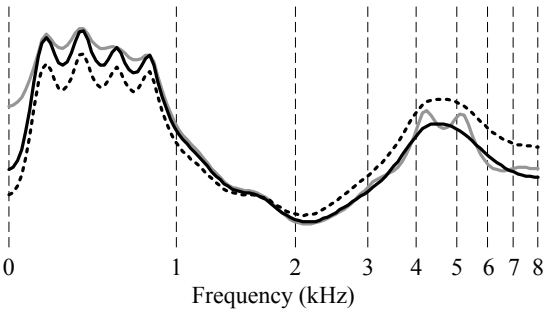

Figure 1: The gray line — the reference — is showing an envelope with model order 30 and warp factor 0.4595. The black lines are spectral envelopes with a warp factor of 0.6595 and same model order as before. The pre-emphasis is not compensated at the dotted line while on the solid line it is. It is clear to see that the solid line is following the amplitude of the gray line while the dotted line is emphasizing high frequencies.

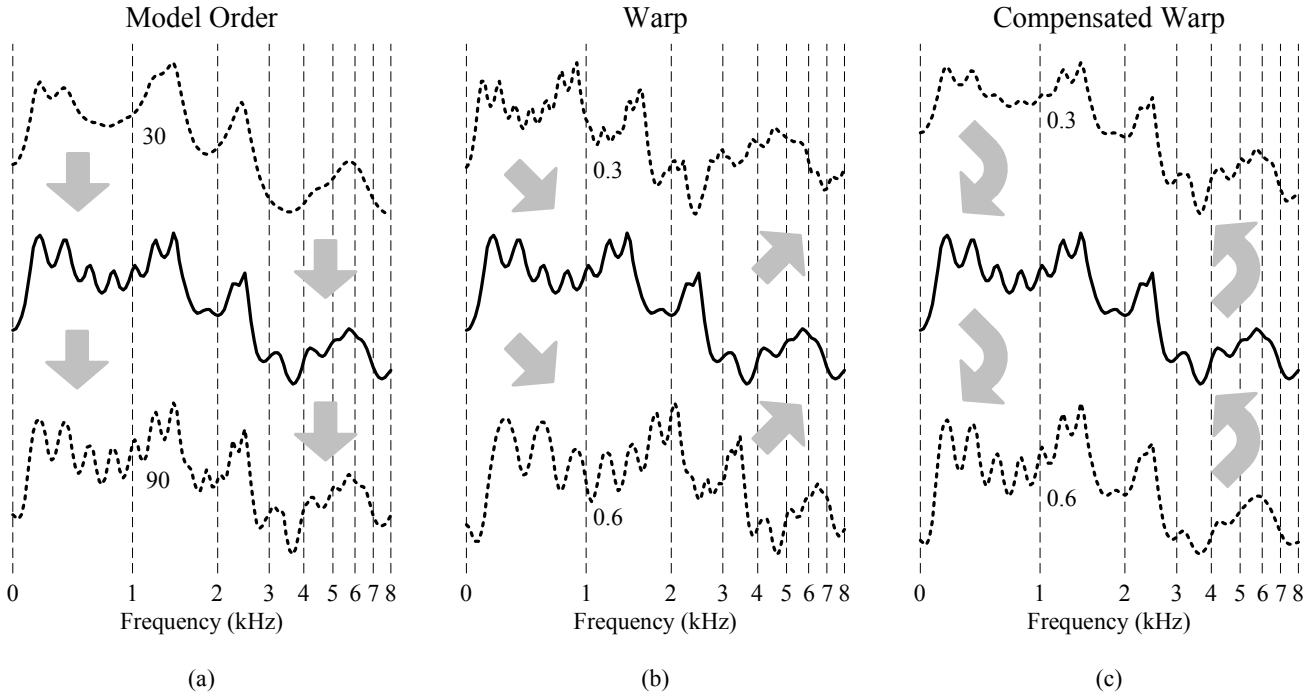

Figure 2: The arrows are showing the influence of the free parameters of the warped-twice minimum variance distortionless response spectral envelope pointing in the direction of higher resolution. The black line is showing an envelope with model order 60 and warp factor 0.4595. Its counterparts with lower and higher (a) model order, (b) warp factor and (c) compensated warp factor are given by pointed lines.

3. STEERING FUNCTION

As we wish to adapt our spectral envelope estimate by the free parameters of the W2MVDR envelope, we have to find a steering function in such a way that classification relevant characteristics are emphasized while less relevant information is suppressed. One promising approach to steer the spectral resolution to lower or higher frequencies was suggested in the work by Nakatoh et al. [7]. There, for every frame indexed i, a division of the first $R[1]$ by the zero $R[0]$ autocorrelation coefficient was used

$$0 \leq \varphi_i = \left| \frac{R_i[1]}{R_i[0]} \right| \leq 1 \tag{8}$$

In combination with γ to adjust the sensibility to the normalized autocorrelation coefficient and the subtraction of the bias to keep the average of α close to α_{Mel} we can write

$$\alpha_i = \gamma \cdot (\varphi_i - 0.5) + \alpha_{\text{Mel}} \tag{9}$$

which is a slight modification to the original proposal by Nakatoh et al. For our experiments we kept γ fix at 0.1. A different value might lead to slightly different results.

4. SPEECH RECOGNITION EXPERIMENTS

In order to evaluate the performance of the proposed W2MVDR spectral estimation in combination with the steering factor we ran experiments on evaluation data of the Rich Transcription 2005 Spring Meeting Recognition Evaluation [11] consisting of five seminars/speakers, collected under the *Computers in the Human Interaction Loop* (CHIL) project[12], providing a total of approximately 130 minutes, sampled at 16 kHz, of continuous, native and non-native, speech material with 16.395 words.

As a speech recognition engine we have used the *Janus Recognition Toolkit* (JRTk), which is developed and maintained by the Interactive Systems Laboratories at two sites: Universität Karlsruhe (TH), Germany and Carnegie Mellon University, USA. Relatively little supervised in domain data is available for acoustic modeling of the recordings. Therefore, we decided to train the acoustic model on the close talking channel of meeting corpora and merge it with the *Translanguage English Database* (TED) corpus [13] summing up to a total of approximately 100 hours of training material. The acoustic model after merge and split training consisted of approximately 3.500 context dependent codebooks with up to 64 Gaussians with diagonal covariances each, summing up to a total of 180.000 Gaussians.

The front-end provided features every 10 ms (first and second pass) or 8 ms (third pass) obtained by the Fourier transformation, warped-LP, warped-twice LP (W2LP) – no particular name is given for the adaptive LP analysis in the work by Nakatoh –,warped-MVDR or W2MVDR spectral estimation followed by a Mel-filterbank (Fourier transformation), no filterbank (warped-LP and W2LP) or a linear-filterbank (warped-MVDR and W2MVDR) and a discrete cosine transformation. Thereafter the 13 cepstral coefficients were mean and variance normalized and after taking 7 adjacent frames the dimension has been reduced to 42 by linear discriminant analysis.

To train a 4-gram language model we have used corpora consisting of broadcast news, proceedings of conferences such as ICSLP, Eurospeech, ICASSP, ACL and ASRU and talks by TED. The vocabulary contains approximately 23,000 words, the perplexity is 120 with an out of vocabulary rate of 0.25%.

The *word error rates* (WER)s of our speech recognition experiments are shown in Table 1 for different spectral estimation techniques and passes. 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, feature space adaptation and vocal track length normalization. Except for the third pass the warped-MVDR spectral envelope leads to better results than the Fourier spectrum (the classical Mel-frequency cepstal coefficients) and looses 0.2% in performance on the third pass. The warped-LP and W2LP perform nearly equally well. On the third pass W2LP shows a 0.2% improvement over the warped-LP. The novel proposed spectral envelope is performing significantly better on all passes compared to all other front-ends under consideration. In a direct comparison to the warped-MVDR an improvement of at least 0.5% in WER can be seen.

Spectral Estimation	WER		
	pass 1	pass 2	pass 3
Fourier	36.1%	30.3%	28.0%
warped-LP	34.9%	30.1%	28.4%
W2LP	35.0%	30.1%	28.2%
warped-MVDR	32.0%	30.0%	28.2%
W2MVDR	31.5%	29.5%	27.5%

Table 1: *Word error rates* (WER)s for five speakers and different spectral estimation methods.

5. CONCLUSION

We have introduced an extension to warped-MVDR spectral estimation by a second bilinear-transformation dubbed warped-twice MVDR which provides two ways of freedom: an overall increase or decrease in resolution and a resolution shift to lower of higher frequencies. We have demonstrated one possible application of the W2MVDR envelope by steering the spectral resolution to emphasize classification relevant features by a steering function based on autocorrelation. The two ways of freedom allow for a variety of adaptation methods which should be investigated in the future. One promising adaptation could be a resolution adaptation in a maximum likelihood fashion, similar to vocal tract length normalization.

6. ACKNOWLEDGMENT

The work presented here was partly funded by the *European Union* (EU) under the integrated project CHIL, Computers In the Human Interaction Loop (Grant number IST-506909).

REFERENCES

[1] N. Malayath, *Data-driven methods for extracting features from speech*, Ph.D. thesis, Oregon Graduate Institute of Science and Technology, January 2000.

[2] M.N. Murthi and B.D. Rao, "All-pole model parameter estimation for voiced speech," *IEEE Workshop Speech Coding Telecommunications Proc., Pacono Manor, PA*, 1997.

[3] M.N. Murthi and B.D. Rao, "All-pole modeling of speech based on the minimum variance distortionless response spectrum," *IEEE Trans. Speech Audio Processing*, vol. 8, no. 3, pp. 221–239, May 2000.

[4] S. Dharanipragada and B.D. Rao, "MVDR based feature extraction for robust speech recognition," *Proc. ICASSP*, vol. 1, pp. 309–312, 2001.

[5] M.C. Wölfel, J.W. McDonough, and A. Waibel, "Minimum variance distortionless response on a warped frequency scale," *Proc. Eurospeech*, pp. 1021–1024, 2003.

[6] M.C. Wölfel and J.W. McDonough, "Minimum variance distortionless response spectral estimation, review and refinements," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 117–126, Sept. 2005.

[7] Y. Nakatoh, M. Nishizaki, S. Yoshizawa, and M. Yamada, "An adaptive Mel-LP analysis for speech recognition," *Proc. ICSLP*, 2004.

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

[9] H. Matsumoto and M. Moroto, "Evaluation of Mel-LPC cepstrum in a large vocabulary continuous speech recognition," *Proc. ICASSP*, vol. 1, pp. 117–120, 2001.

[10] A.V. Oppenheim and R.W. Schafer, *Discrete-time signal processing*, Prentice-Hall Inc., 1989.

[11] National Institute of Standards and Technology (NIST), "Rich transcription 2005 spring meeting recognition evaluation," *http://www.nist.gov/speech/tests/rt/rt2005/spring*, June 2005.

[12] H. Steusslof, A. Waibel, and R. Stiefelhagen, "Computers in the human interaction loop," *http://chil.server.de*.

[13] Linguistic Data Consortium (LDC), "Translanguage english database," LDC2002S04.