WARPING AND SCALING OF THE MINIMUM VARIANCE DISTORTIONLESS RESPONSE

Matthias Wölfel, John McDonough and Alex Waibel

Interactive Systems Laboratories
Institut für Logik, Komplexität und Deduktionssysteme
Universität Karlsruhe (TH)
Am Fasanengarten
76131 Karlsruhe, Germany

ABSTRACT

Spectral estimation based on the *minimum variance distortionless response* (MVDR) is well-known in the signal processing literature and has been shown to be superior to linear prediction for robust speech recognition. In this work we propose two techniques to improve the resolution and the robustness of the MVDR spectral estimate: The first is a time-domain technique to estimate an all-pole model based on the *warped* short time frequency axis such as the Mel-frequency. The second is a method for *scaling* the height of the spectral envelope in order to extract robust features for large vocabulary continuous speech recognition systems which must operate in noisy conditions. Moreover, we show that these two techniques can be combined to good effect. In a series of speech recognition experiments on the Switchboard Corpus, the combination of our proposed approaches achieved a *word error rate* (WER) of 35.9%, which is clearly superior to the 37.0% WER obtained by the common MVDR and the 37.2% WER obtained by the widely used Fourier transform.

1. INTRODUCTION

Large vocabulary continuous speech recognition systems that operate under "real world", e.g. in a meetingroom scenario which we work on in the FAME project, are often confronted with mismatches between training and test conditions. Such mismatches can stem from speaker and speaking style variation, room reverberation, and noise, to name only a few sources, and typically degrade recognition performance. In this work, we seek to reduce the degradation incurred from additive noise with spectral estimation based on the *minimum variance distortionless response* (MVDR). MVDR spectral estimation was previously proposed by Murthi and Rao [1, 2] as a spectral envelope technique, and applied to speech recognition by Dharanipragada and Rao [3]. Moreover, we seek to further enhance the gain provided by the MVDR with respect to the *fast Fourier transform* (FFT) through the use of two refinements:

- *Warping* of the frequency axis prior to MVDR spectral estimation to ensure that more parameters in the spectral model are allocated to the low, as opposed to high, frequency regions of the spectrum, thereby mimicking the frequency resolution of the human auditory system [4, 5].
- *Scaling* of the spectral envelope as a means for extracting robust features for large vocabulary continuous speech recognition systems which must operate under noisy conditions [6].

As we will show, the *word error rate* (WER) reductions provided by both techniques are additive.

The balance of this work is organized as follows. Section 2 summarizes the process of spectral estimation based on the MVDR and describes a method for implementing the frequency warping, which very closely mimics human hearing. In Section 3, we discuss a modification to conventional MVDR spectral estimation that reduces the variance of the amplitudes of the spectral peaks; this modification greatly enhances the utility of the MVDR estimation, especially in the presence of additive noise. The results of initial speech recognition experiments, in which the several types of spectral estimation are combined, are reported in Section 4. Finally, Section 5 discusses our results and Section 6 presents our conclusions and plans for future work.

2. WARPING OF THE MVDR SPECTRUM

The MVDR spectral estimation can be posed as a problem in filter bank design, wherein the final filter bank is subject to the *distortionless constraint* [7]:

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

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

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

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

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 *first order all-pass selections* of the form

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

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 [8]

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

which is also known as the *frequency mapping function*. Thereby, 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.

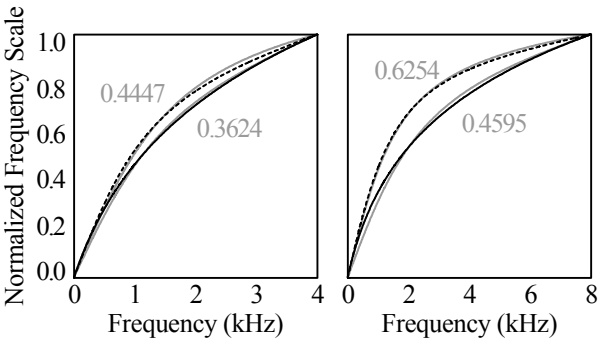

Fig. 1. The approximations of Mel-frequency (black lines) and Bark-frequency (dotted black lines) by the bilinear transformation (gray lines including the warping factor in gray digits) are demonstrated for 8 and 16 kHz sampling rates.

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 [7]. Note that the frequency vector $\mathbf{s}$ is replaced by the warped vector $\tilde{\mathbf{s}}$:

$$\tilde{\mathbf{h}}_{\text{l}} = \frac{\phi^{-1}\tilde{\mathbf{s}}(\omega_l)}{\tilde{\mathbf{s}}^H(\omega_l)\phi^{-1}\tilde{\mathbf{s}}(\omega_{\text{foi}})}$$

That means that the impulse response of the distortionless filter for the frequency ω_{foi} is denoted by $h_{\text{foi}}(n)$. The warped MVDR power spectrum of the signal power spectrum $S(e^{-j\omega})$ at frequency ω_{foi} is than obtained as the output of the optimized constrained filter:

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

Although the MVDR spectral estimation was posed as a problem of designing a distortionless filter for a given frequency ω_{foi}, this was only a conceptual device. The warped MVDR spectrum can in fact be represented in a parametric form for all frequencies and computed very simply as:

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

Under the assumption that the $(M+1)\cdot(M+1)$ Hermitian Toeplitz correlation matrix ϕ is positive definite and thus invertible, Musicus [9] has derived a fast algorithm to calculate the MVDR spectrum from the *linear prediction* (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. [8] 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**

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

Note that the spectrum (2) is in the warped frequency domain. Hence, it is necessary to replace the Mel-filterbank in the front end of an automatic speech recognizer with a filterbank of uniformly half overlapping triangular filters. If we are only interested in a spectral envelope in the linear frequency domain, we can use

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

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.

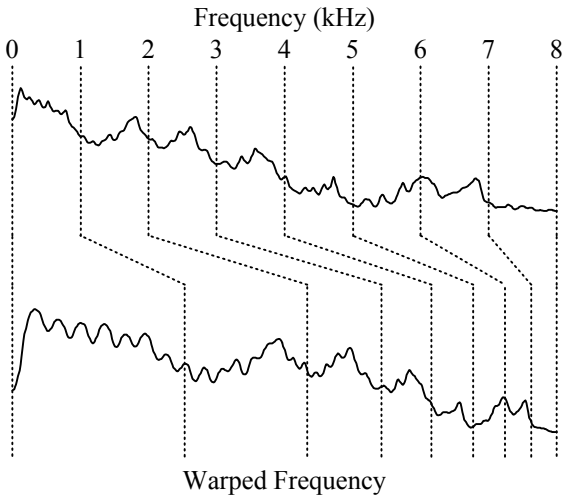

Fig. 2. Comparison of MVDR (top) and Mel-warped-MVDR (bottom) spectral envelopes, both of same model order 120.

Figure 2 illustrates the difference between the MVDR and Mel-warped-MVDR spectral envelopes. The warp factor for the warped-MVDR was set to 0.4595 so as to simulate the Mel-frequency for a signal sampled at 16 kHz. While the MVDR exhibits frequency-independent spectral resolution, the Mel-warped-MVDR provides high resolution for frequencies below 2 kHz and decreasing resolution for higher frequencies. The warping of the MVDR provides interesting properties, similar to Mel-warped-LP [10], which cannot be achieved when the MVDR is *followed* by frequency-warping: The ability to model the spectra similar to the loudnes density spectrum and in given information by the inverse filtered wapred-MVDR residual, resembling the overall information in the auditory nerve firing. But here without the negative effect of overestimating and overemphasizing of the harmonic peaks in medium- and high-pitched voiced speech as seen in Mel-warped-LP.

3. SCALING OF THE MVDR SPECTRUM

Spectral peaks have been shown to be particular robust to additive noise in the logarithmic domain, since $\log(a+b) \approx \log(\max\{a,b\})$ [11]. Therefore we propose to match the MVDR derived spectrum to the highest spectral peak of the Fourier spectrum.

To fully understand why the suggested scaling should be useful, we must first investigate how the power of the features in the logarithmic power spectrum is influenced by additive noise. To do so we have to define two signals s and $\hat{s}$ in the frequency domain and take them into the logarithmic power domain

$$S_{\log} = \log(|s|^2)$$

$$\hat{S}_{\log} = \log(|\hat{s}|^2)$$

For additive noise n we may write

$$S_{\log} + D_{\log} = \log(|s+n|^2)$$

$$\hat{S}_{\log} + \hat{D}_{\log} = \log(|\hat{s}+n|^2)$$

where $D_{\log}$ and $\hat{D}_{\log}$ denote the logarithmic power differences between the clean and noisy signals. Now we can solve the equations for $D_{\log}$ and $\hat{D}_{\log}$

$$D_{\log} = 2\log|s+n| - 2\log|s| = 2\log\left|1+\frac{n}{s}\right|$$

$$\hat{D}_{\log} = 2\log|\hat{s}+n| - 2\log|\hat{s}| = 2\log\left|1+\frac{n}{\hat{s}}\right|$$

Assuming $|n| < |s|$, we can [6] prove that

$$|\hat{s}| > |s| \Rightarrow |\hat{D}_{\log}| < |D_{\log}|$$

This result is also apparent from Figure 3, where the grey plane is getting smaller to the right for $|s| > |n|$.

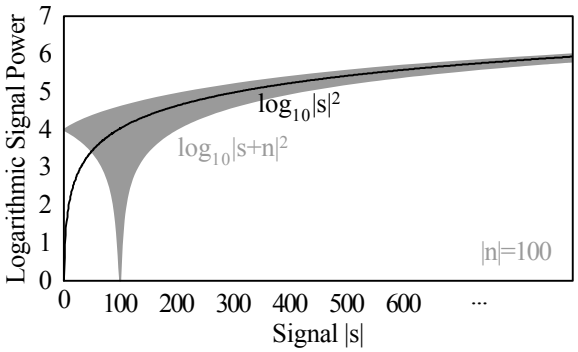

Fig. 3. Comparison of the power to the logarithmic power of the clean signal s, black line, and the influence of additive noise n, gray area.

Furthermore we can show [6] that the *logarithmic spectral distortion* (LSD) is smaller at the highest amplitude than the expected LSD averaged over all frequencies which is commonly considered in the calculation of the envelope.

Deeper insight into this phenomenon can be obtained by plotting the undisturbed energies of the logarithmic power spectrum on the x-axis and the disturbed energies of the logarithmic power spectrum on the y-axis. The gray line in Fig. 4, Theoretical Features, shows the idealistic case of a noise free speech signal; here all points fall on the line $y = x$. In the case of additive noise, black line, the lower values of the power spectrum are lifted to higher energies; i.e., the low-energy components are masked by noise and their information is lost.

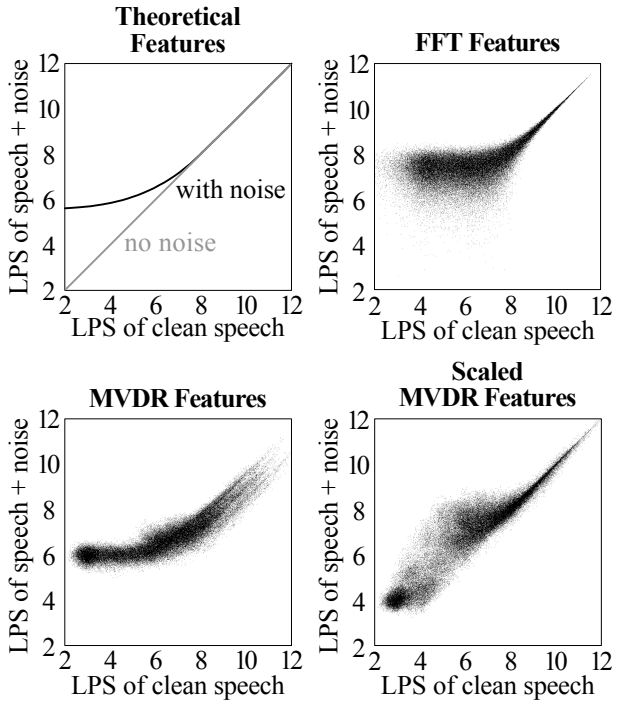

Fig. 4. Shown are the influence of noise (signal to noise ratio = 8dB) in the *logarithmic power spectrum* (LPS) on the features for different spectral estimation methods in dependence of their signal energies.

Comparing the influence of noise on the logarithmic power features derived from the FFT with the MVDR logarithmic power features of Fig. 4, clearly demonstrates the problem which occurs if additive noise is present: Due to the high variance of the maximum amplitude in the MVDR approach, there is a broad band instead of a narrow ribbon even in the high energy regions. The use of the proposed scaling provides more useful features than both conventional MVDR, which is clear upon comparing the MVDR features with the scaled MVDR features of Fig. 4, and the FFT, which can be seen by comparing the FFT features with the scaled MVDR features. The resulting scaled MVDR features are clearly less distorted by noise.

4. SPEECH RECOGNITION EXPERIMENTS

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

Our recognition experiments were conducted on the *Switchboard Corpus*. The training set for these experiments was comprised of 30 hours of speech collected from 548 speakers of both sexes. Speech from 16 speakers of both sexes was used for testing. Two speakers from the test set were recorded using analog cellphone channels resulting in a highly distorted signal. Our baseline model consisted of 4,166 codebooks with 32 Gaussians each. The features used for speech recognition were obtained by calculating 13 static cepstral coefficients for each frame of speech, performinging mean normalization, and then calculating delta and delta-delta cepstra. Thereafter, linear discriminant analysis was used to reduce the final feature length to 32. *Maximum likelihood linear regression* (MLLR) [12] was used to adapt the means and covariances of the speaker-independent model for every speaker and environment condition in the test set.

The static cepstral coefficients were obtained through a discrete cosine transform from different spectral representations:

- The FFT and the MVDR both followed by a Mel-filterbank consisting of 30 half-overlapping Mel-spaced triangularly shaped filters, see Mel-filterbank of Fig. 5.
- The Mel-warped-MVDR and the Mel-warped&scaled-MVDR both followed by a filterbank consisting of 30 so adapted filters to compensates for the differences between the bilinear transform and the Mel-frequency, see adapted filterbank of Fig. 5.
- The Mel-warped-MVDR and the Mel-warped&scaled-MVDR neglecting the use of a filterbank similar to [13] and therefore dubbed perceptual-MVDR and perceptual&scaled-MVDR.

All features were calculated every 10 ms from speech data sampled at 8 kHz, using a 20 ms Hamming window.

To gain a finer appreciation of the differences between MVDR and warped&scaled-MVDR acoustic preprocessing, compare the left and right portions of Fig. 6. In particular, note that the *vocal tract length normalization* (VTLN),

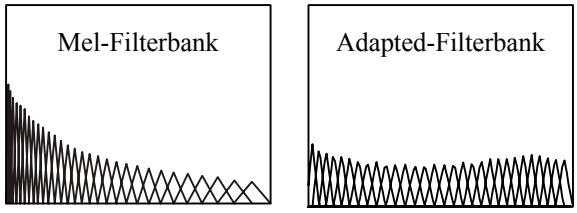

Fig. 5. Mel- and adapted filterbanks.

which was used while training and testing, must be implemented differently between the methods with and without warping, one in the linear frequency domain while the other has to be implemented in the warped frequency domain, see Fig. 7. To compensate for this difference the VTLN of the Mel-warped-MVDR was recalculated to resample the outcome of the linear VTLN approach after Mel-warping was applied.

5. DISCUSSION

The experimental results shown in Table 1 confirm the correctness of the foregoing arguments. The gain of spectral envelope techniques in general 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 describe spectral peaks and valleys equally well, spectral envelopes 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 their Fourier counterparts. The gain of the Mel-warped-MVDR over the MVDR is attained through the better modeling of the human auditory system, while the gain of the Mel-warped&scaled-MVDR over the Mel-warped-MVDR is attained through the scaling of the envelope to the highest point of the Fourier spectrum to reduce the variance of the envelope due to noise.

FFT	37.2%
MVDR(80)	37.0%
Mel-warped-MVDR(50)	36.3%
Mel-warped&scaled-MVDR(50)	35.9%
perceptual-MVDR(25)	36.3%
perceptual&scaled-MVDR(25)	36.1%

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

The parameters of the model order used in the evaluations shown were tuned on a small development set using

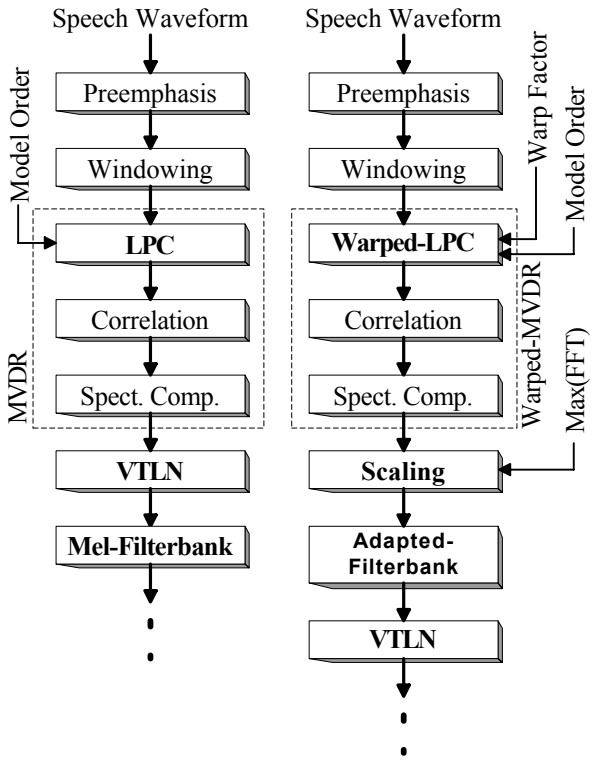

Fig. 6. Extract of the MVDR (left) and the warped&scaled-MVDR (right) acoustic preprocessing as used in our experiments.

no VTLN and MLLR. Small variation of the model order seems to have small effects ($< 0.2\%$ WER) on the word error rate and therefore opimization of the model order seems to be not critical for our results.

The differences in the model order of the different approaches may be explained by the characteristic of the envelope (a reduction in model order increases the smoothness of the envelope) in combination with the following filterbank. The Mel-filterbank provides a stronger smoothing than the adapted-filterbank and therefore a smoother envelope must be provided. In the case where the filterbank is neclected the overall smoothing must be provided only by the envelope and therefore the model order has to be further reduced.

6. CONCLUSIONS

This paper has presented improved feature extraction methods based on the MVDR to address the resolution of the human auditory system and robustness issues.

To provide a good spectral envelope estimate, we have followed Dharanipragada and Rao [3] in using the MVDR instead of LP. Next we have applied the well known techniques of *pre-warping* [8] to the MVDR approach to pro-

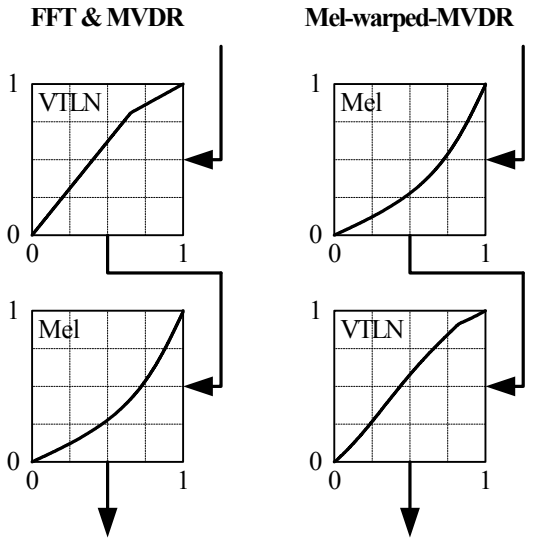

Fig. 7. Shown are the linear and the Mel-warped-VTLN approaches. Note that the VTLN has to be implemented in different manners.

vide a better approximation of the aspects of the human auditory system than the envelope followed by a Mel-filterbank. Furthermore, we have seen that noise added to a signal distorts mainly the spectral valleys while the spectral peaks remain relatively unchanged. Therefore we have introduced a scaling technique which adjusts the highest point of the envelope to the highest point of the Fourier spectrum.

It has been shown that the performance of the proposed methods, Mel-warping and scaling of the MVDR envelope, could improve the accurancy of the used spectral envelope technique. As the MVDR itself already performs at least as well as the widely used FFT-based approach, the Mel-warped and scaled MVDR envelope performs better than the FFT-based approach.

Another interesting feature of the MVDR envelope in comparison to the LP envelope is that the formants do not change positions as a result of changes in the model order [3]. This fact provides for the possibility of setting the model order as a function of the vocal tract length, the signal to noise ratio, or based on a likelihood criterion, in order to further improve the robustness of the analysis or to further increase the accuracy of the speech recognition system.

Further work may focus on larger training and test sets. In particular we want to address speech recorded with distant micropohnes or a microphone array.

7. ACKNOWLEDGEMENT

The work presented here was partly funded by the *European Union* (EU) under the projects FAME, Grant number IST-2000-28323 and PF-STAR, Grant number IST-2001-37599. Any opinions, findings and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the EU.

8. REFERENCES

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

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

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

[4] Strube, H.W., "Linear prediction on a warped frequency scale," *ASA*, vol. 68, no. 8, pp. 1071–1076, 1980.

[5] Wölfel, M.C., "Minimum variance distortionless responce spectral on a warped frequency scale," *Eurospeech*, pp. 1021–1024, 2003.

[6] Wölfel, M.C., "Minimum variance distortionless responce spectral estimation and subtraction for robust speech recognition," *Diploma-Thesis, Universität Karlsruhe (TH), Karlsruhe, Germany*, Jan. 2003
http://www.isl.ira.uka.de/wolfel.

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

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

[9] Musicus, B.R., "Fast MLM power spectrum estimation from uniformly spaced correlations," *ASSP*, vol. 33, pp. 1333–1335, 1985.

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

[11] Barker, J. and Cooke, M.P., "Modelling the recognition of spectrally reduced speech," *Eurospeech*, pp. 2127–2130, 1997.

[12] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech and Language*, pp. 171–185, 1995.

[13] Yapanel, U.H. and Hansen, J.H.L., "A new perspective on feature extraction for robust in-vehicle speech recognition," *Eurospeech*, pp. 1281–1284, 2003.