Speaker Dependent Model Order Selection of Spectral Envelopes

Matthias Wölfel

Interactive Systems Laboratories
Institut für Logik, Komplexität und Deduktionssysteme, Universität Karlsruhe (TH)
Am Fasanengarten 5, 76131 Karlsruhe, Germany
wolfel@ira.uka.de

Abstract

This work introduces a maximum-likelihood based *model order* (MO) selection technique for spectral envelopes to apply speaker dependent adaptation in the feature-space similar to vocal tract length normalization.

Speech recognition systems based on spectral envelopes are using a fixed MO for the underlying linear parametric model. Using a fixed MO over different speakers or channels might not be optimal. To address this problem we investigated the use of warped and scaled minimum variance distortionless response spectral estimation techniques with speaker dependent MOs based on a maximum-likelihood criteria. Comparing experimental results on the Translanguage English Database we can show an improvement by 1,9% relative compared to the word error rate by the fixed MO and 3,5% relative to the traditional Mel-frequency cepstral coefficients.

1. Introduction

Common speech recognition systems which are based on perceptual linear prediction or linear prediction cepstral coefficients are using a fixed *model order* (MO) for the underlying linear parametric model. Using a fixed MO over different speakers or channels might not be optimal, therefore we propose the use of speaker dependent MOs. Commonly the MO with the smallest error (with arbitrary cost function; e.g., squared error) in comparison to the Fourier spectrum is considered the best. In speech recognition the case is different in such as the represented envelope of our test speaker should fit best to the initial acoustic models of the recognizer. This can be established by optimizing the likelihood of the recognizer for every single speaker in dependence on the MO.

2. Theoretical Considerations

In this chapter we briefly repeat the warped *minimum variance distortionless response* (MVDR) spectral estimation technique and its relation to warped linear prediction, as well as the scaling of the MVDR envelope. Furthermore we give a brief summary about speaker normalization based on the vocal tract length which then can be readily extended to estimate the MO for every single speaker in a maximum likelihood manner.

2.1. The Warped & Scaled Minimum Variance Distortionless Response Spectral Envelope

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 extended this approach by *warping* the frequency axis with the bilinear transformation prior to MVDR spectral estimation, therefore dubbed *warped-MVDR*, 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 and by *scaling* of the spectral envelope as a means for extracting robust features [4, 5].

Similar to Burg [6] who has shown the relationship between the MVDR- and the *linear prediction* (LP) envelope, we can write the relationship between the warped-MVDR and the warped-LP envelope as follows:

$$\frac{1}{S^{(p)}_{\mathrm{warped-MVDR}}(\omega)} = \sum_{k=0}^{p} \frac{1}{S^{(k)}_{\mathrm{warped-LP}}(\omega)} \tag{1}$$

This implies that the warped-MVDR spectrum $S^{(p)}_{\text{warped-MVDR}}(\omega)$ of order p is the harmonic mean of the LP spectra $S^{(p)}_{\text{warped-LP}}(\omega)$ of order 0 to p, and explains why the (warped) MVDR spectrum generally exhibits a smoother frequency response with decreased variance than the corresponding (warped) LP spectrum [2]. This attribute makes the (warped) MVDR envelope also more interesting for our considerations as we can adjust the 'resolution' of our envelope due to the MO in smaller steps.

For a fast computation of the warped-MVDR spectrum we have extended Musicus' [7] algorithm to calculate the MVDR spectrum from the LP coefficients as follows (for more details see [8]):

1. **Computation of the warped-*LP coefficients* (LPC)**
   For our experiments we used an algorithm by Matsumoto et al. [9] to calculate the warped-LP coefficients, but any other algorithm should work similarly well.

2. **Correlation of the warped-LPC $\tilde{a}^{(N)}_{0\cdots N}$ of order N**

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

3. **Computation of the warped-MVDR spectrum**

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

   $\epsilon = 1/P_e$ where P_e stands for the prediction error variance.

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.

Spectral peaks have been shown to be particular robust to additive noise in the logarithmic domain, since $\log(a+b) \approx \log(\max\{a,b\})$ [10]. Therefore we match the warped-MVDR envelope to the highest spectral peak of the Fourier spectrum to get the *warped&scaled-MVDR* envelope resulting in features which are less distorted by noise [5].

After having repeated some basics of the warped&scaled-MVDR envelope we want to investigate how the *fundamental frequency* (FF), generated by the vibration of the vocal folds and rating from 60Hz for a large man up to 300Hz for a small woman or child, influences the estimate of our envelope. Therefore we generate spectral envelopes with different MOs of impuls trains with different FFs which have been filtered by a transfer function $H(z)$. The used transfer function has two formants; the first at 1000Hz and the second at 2000Hz. Comparing this spectral envelopes to the spectral envelope of the transfer function $H(z)$ itself we see that different MOs approximate the transfer function differently well. Furthermore we realize that high MOs in combination with a high FF uncover the excitation frequency together with their harmonics. To find the best MO, we are interested in the prediction error in dependency on the MO and the FF. Therefore we plot the squared error in dependency on the MO for different FFs, see Figure 4. For some cases we see more than one minimum, but following the absolute minimum we can conclude that a higher FF is modeled better by a low MO and vice versa. This can be explained by the fact that the FF is setting the periodic baseline for all higher-frequency harmonics. Sparse harmonics result in a lower resolution than dense harmonics, therefore the MO should be reduced for sparse harmonics to reach the optimal estimate.

In speech recognition we are interested in the spectral envelope estimation of the transfer function which fits best to the initial acoustic models of the recognizer. The problem statement is similar to those of estimating the best vocal tract length which should be regarded next.

2.2. Speaker Dependent Model Order Selection

A commonly used approximation in speaker normalization is that the *spectral representations* (SR)s, e.g, the power spectra or the *spectral envelopes* (SE)s, of the same phoneme by two different speakers are uniformely scaled versions of each other

$$\mathrm{SR}(f)_{\text{Speaker A}} \approx \mathrm{SR}(\alpha_{AB} \cdot f)_{\text{Speaker B}} \tag{3}$$

where α_{AB} is an uniform scaling parameter depending on the difference in vocal tract length between *Speaker A* and *Speaker B*. We now want to extend this approach by using the free parameter of the SE, the MO, which is commonly optimized to maximise the word accuracy of a speech recognition system *a priori* where the free MO is set to the same parameter for all speakers in the training and test set. Similar to (3) we can adjust the MO m of *Speaker B*, holding the MO l of *Speaker A* fixed, to find the 'best fit' between the SEs of the same phoneme by two different speakers.

$$\mathrm{SE(l)}_{\text{Speaker A}} \approx \mathrm{SE(m)}_{\text{Speaker B}} \tag{4}$$

Instead of adapting the MO of our investigated speaker to a single speaker, we have to adapt to 'best fit' the models of our trained speech recognition system. To do so we have to calculate the cepstral feature c_m of the SE of MO m. As a sequence of different MOs we can write the cepstal feature as the vector $C = (c_m, c_{m+1}, .., c_n)^T$. Let λ_l denote a set of given hidden Marcov models trained on a broad variety of speakers with a *fixed* MO l. The optimal MO $\hat{m}$ for the given speaker is then obtained by maximising the likelihood of the adaptation data C given the corresponding word string W:

$$\hat{m} = \arg\max_m \Pr(C|\lambda_l, W) \tag{5}$$

The optimal MO can then be obtained by a grid search over a range of values. As more than one minimum might exist, see Figure 1, gradient descent should not be used as it might not lead to the global minimum.

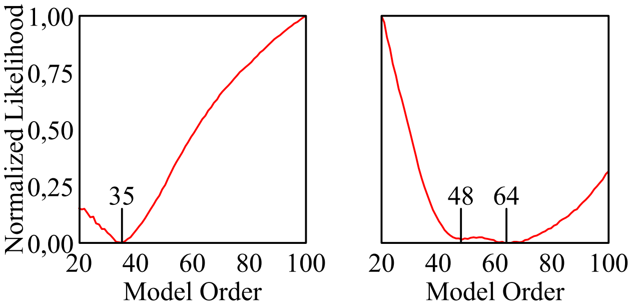

Figure 1: The normalized likelihoods in dependence on the model order for two different speakers are shown.

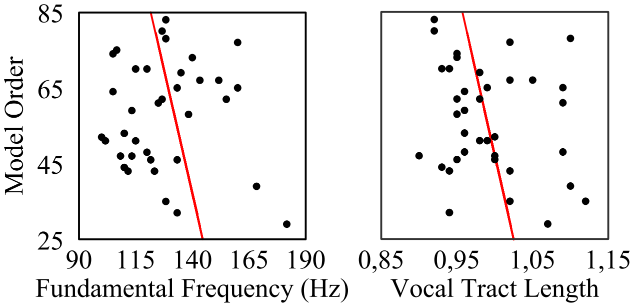

Figure 2: The relationship between the model order and the fundamental frequency (left image) and vocal tract length (right image) are shown for all 39 speakers of the Translanguage English Database corpus. Every single point stands for a speaker while the red line represents the regression line.

To compare the FF (calculated by the average magnitude difference function [11]) as well as the vocal tract length to the MO we want to determine regression lines. Our linear regression problem which must be solved is to fit a straight line to a number of points so that the squared deviations of the observed points from that line are minimized. Comparing for all 39 speakers from the Translanguage English Database corpus, as described in the following chapter, we can see, Figure 2, that the MO might depend on the FF. A correlation also exists for the vocal tract length value which is not surprising as the FF is also correlated to the vocal tract length value. That means (in average) a male speaker (lower FFs, warp factor < 1) should have a higher MO than a female speaker (higher FFs, warp factor > 1).

We couldn't find any statistical relevant correlation between the MO and the signal to noise ratio, which varied between 10dB and 22dB. This seams to be a contradiction to Tierney [12] who has claimed that corrupted speech has to be modeled using a higher MO of the all-pole model, to model both, speech and noise. But as we are only interested in the best prediction of the physical excitation of the vocal tract we have no interest in modeling the noise and therefore we shouldn't expect an increase in MO.

3. 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 Laborato-

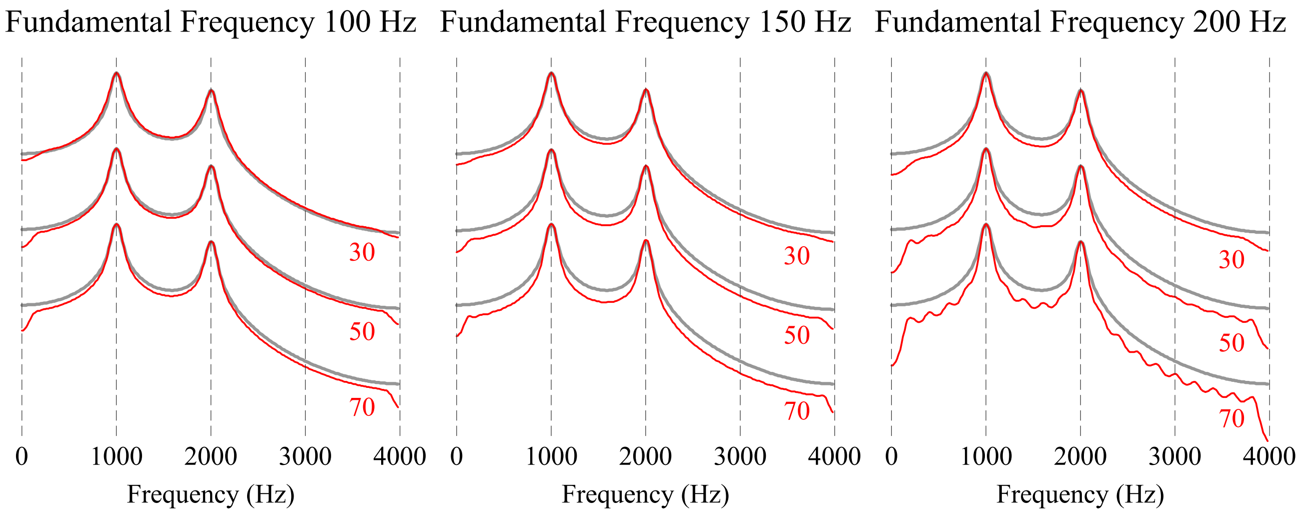

Figure 3: Shown are MVDR spectral envelopes (red lines) for different model orders (30, 50 and 70) and different fundamental frequencies (100, 150 and 200Hz) in comparison to the MVDR spectral envelope with model order 80 of the transfer function $H(z)$ (gray lines).

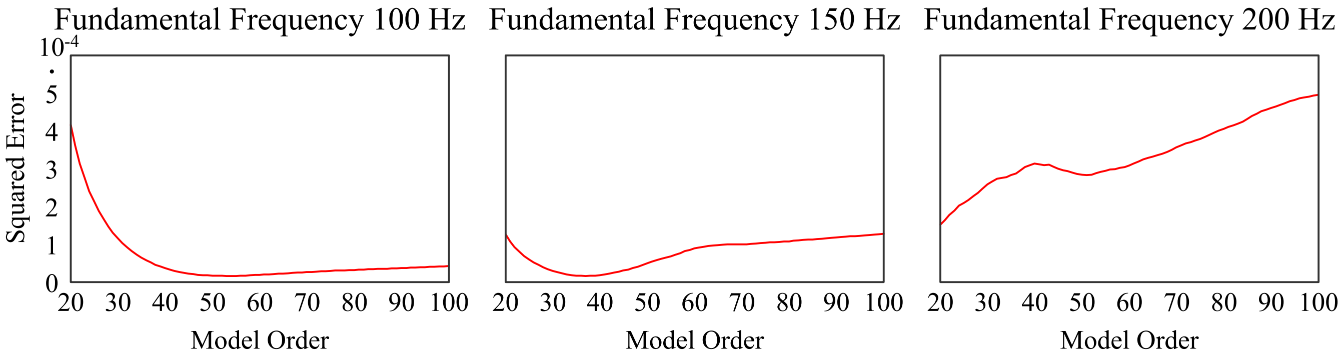

Figure 4: Shown are the squared error of the spectral envelope belonging to different model orders (between 20 and 100) and fundamental frequencies (100, 150 and 200Hz) in comparison to the MVDR spectral envelope with model order 80 of the transfer function $H(z)$. We see that for every fundamental frequency the minimal error is reached by a different model order.

	Fourier Trans.		**Warped&Scaled-MVDR**								
			Fixed Order			Variable Order					
						Test			Adaptation&Test		
Speaker	WER	VTLN	WER	VTLN	order	WER	VTLN	order	WER	VTLN	order
1	37,6%	0,95	38,5%	0,95	60	38,2%	0,94	73	37,4%	0,94	74
2	39,5%	0,96	38,3%	0,97	60	38,4%	0,97	63	38,0%	0,97	62
3	60,7%	0,98	58,2%	1,00	60	57,4%	1,00	65	57,0%	0,99	47
4	38,5%	1,07	36,3%	1,08	60	36,3%	1,08	60	35,9%	1,08	61
5	36,6%	1,04	35,9%	1,07	60	35,3%	1,07	30	35,6%	1,07	29
6	28,5%	0,90	28,8%	0,92	60	28,9%	0,92	79	28,2%	0,92	80
7	22,1%	0,94	22,6%	0,95	60	22,4%	0,95	64	22,2%	0,95	64
8	43,3%	0,99	43,2%	1,02	60	42,9%	1,02	35	41,9%	1,02	35
Average	38,4%	0,98	37,7%	1,00	60	37,5%	0,99	59	37,0%	0,99	57

Table 1: Given are *word error rates* (WER)s for eight test speakers together with their *vocal tract lenght normalization* (VTLN) factor and model order. For the spectral representation we used the Fourier transform and the warped&scaled-MVDR spectral envelope. In the *test* case the model order was only estimated for the eight test speakers while in the *adaptation&test* case the model order was estimated for the 31 adaptation speakers and for the eight test speakers.

ries at the Universität Karlsruhe in Karlsruhe, Germany and at the Carnegie Mellon University in Pittsburgh, Pennsylvania, USA.

Our recognition experiments were conducted on the *Translanguage English Database* (TED) corpus [13] which presents several kind of problems to cope with: Speakers are often nonnative, have a strong accent or are not even fluent, spontaneous speech phenomena occur quite frequently and the recordings were made with a lapel microphone, hence the signal often contains noise. As relatively little supervised data is available for acoustic modeling we have trained our acoustic models on the *Broadcast News* corpus [14] (104 hours of speech collected from speakers of both sexes) and adapted on 31 speakers (8 hours) out of the 39 transcribed speakers from the TED corpus using *Maximum likelihood linear regression* (MLLR) [15]. Our test set contained the final 8 speakers (6 male speakers, Sp.4 and Sp.5 female) of the TED corpus with a wide variety of mother tongues (Sp.1: English, Sp.2: Italian, Sp.3: French, Sp.4: French, Sp.5: Danish, Sp.6: German, Sp.7: Dutch, Sp.8: Japanese).

Our baseline model consisted of 4.139 codebooks with 32 Gaussians each. The features used for speech recognition were obtained by calculating 13 static cepstral coefficients for each frame of speech which have been normalized by cepstral mean subtraction. Thereafter, linear discriminant analysis was used to reduce the current features plus 3 left and right adjacent features to a final feature length of 40. MLLR was used to adapt the means and covariances of the speaker-independent model for every speaker in the test set. *Feature Space Adaptation* was not used as it only improve the likelihood, but not the *word error rate* (WER) of our system which stayed the same. The static cepstral coefficients were obtained through a discrete cosine transform from the warped&scaled-MVDR envelope followed by a filterbank consisting of 30 so adapted filters to compensate for the differences between the bilinear transform and the Mel-frequency. All features were calculated every 10 ms from speech data sampled at 16 kHz, using a 16 ms Hamming window. The Fourier transform case was comprised of a fast Fourier transform followed by a Mel-filerbank instead of the warped&scaled-MVDR envelope followed by the adapted filterbank; everything else stayed the same. The 3-gram language model was generated by proceedings from conferences such as ICSLP, Eurospeech and ICASSP with a dictionary containing 40.000 words (the most frequently words in the proceedings) resulting in an out of vocabulary rate below 0,5%.

Comparing our results in Table 1 we can confirm our former finding on the *Swichboard* corpus [5], that the warped&scaled-MVDR performs better than the Fourier transform as a spectral estimate. The proposed model order adaptation can further improve this performance by 1,9% relative WER reulting in a relative WER improvement by 3,5% in comparison to the traditional Mel-frequency cepstral coefficients.

4. Conclusions

We found the use of unsupervised *maximum-likelihood* (ML) estimation helpful to determine a speaker dependent MO and could show an improvement in word accuracy over a fixed MO which was set *a-priori* to optimise the word accuracy. Even though an error reduction was achieved, two speakers of the *test* case performed worse in comparison to the *fixed order* case. This means that the ML did not always converge properly, leaving space for further improvement which will be considered in our future works.

5. Acknowledgement

The work presented here was partly funded by the *European Union* (EU) under the project FAME (Grant number IST-2000-28323) and CHIL (Grant number IST-506909).

6. References

[1] 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.

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

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

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

[5] M.C. Wölfel, J.W. McDonough, and A. Waibel, “Warping and scaling of the minimum variance distortionless response,” *ASRU*, 2003.

[6] J.P. Burg, “The relationship between maximum entropy and maximum likelihood spectra,” *Geophysics*, vol. 37, pp. 375–376, Apr. 1972.

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

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

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

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

[11] L.R. Rabiner, “On the use of autocorrelation analysis for pitch detection,” *ASSP*, 1977.

[12] J. Tierney, “A study of LPC analysis of speech in additive noise,” *ASSP*, vol. 28, no. 4, 1980.

[13] Linguistic Data Consortium (LDC), “Translanguage english database,” www.ldc.upenn.edu/Catalog/LDC2002S04.html.

[14] Linguistic Data Consortium (LDC), “English broadcast news speech (Hub-4),” www.ldc.upenn.edu/Catalog/LDC97S44.html.

[15] 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.