Frame Based Model Order Selection of Spectral Envelopes

Matthias Wölfel

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

Abstract

Spectral envelopes, using (warped or perceptual) linear prediction or minimum variance distortionless response for the underlying linear parametric model, are widely used in speech recognition systems where the frequency resolution, namely the *model order* (MO), of the spectrum is kept constant. Modeling different types of phonemes such as vowels or fricatives with the same frequency resolution might not lead to the best possible performance. This could be due to the fact that important parts of various phonemes lie in different frequency regions, that the fundamental frequency varies for different speakers or because of a high variance in the signal to noise ratio. To address this problem we propose to vary the MO frame by frame according to a control factor. In our case, the control factor could be either a relation of autocorrelation coefficients or the spectral entropy. Experimental results on the Translanguage English Database show an improvement by 2.4% relative in word error rate compared to the fixed MO and 4.2% relative to the traditional Mel-frequency cepstral coefficients.

1. Introduction

The selection of the *model order* (MO) is an important, but often difficult, aspect of using all-pole models for a particular application. Intuitively, the optimal MO depends on the length of data over which the MO will be applied. On the one hand, larger MOs can capture the dynamics of a richer class of signals. On the other hand, larger MOs also require proportionally larger data sets for the parameters to be robustly estimated. In the case of speech recognition, the MO is commonly fixed over all speakers, phoneme types and *signal to noise ratio* (SNR). To chose the MO for best recognition performance, different MOs have to be tried out on a development set. Due to high variance in the spectral representation for different speakers, phoneme types or SNR, a fixed order should not lead to the best possible spectral estimation, and therefore recognition performance. To improve the quality of the spectral estimate we have proposed and investigated the use of spectral estimation techniques where the MO, on a frame by frame basis, is steered by a *control factor*. This factor has to be carefully chosen to counteract against the aforementioned unwanted variations; e.g., we could try to reduce the MO of frames with a low SNR to smooth the influence of the noise or where the spectrum is relatively flat like in a fricative.

2. Theoretical Considerations

In this chapter we briefly describe the warped & scaled *minimum variance distortionless response* (MVDR) spectral estimation technique, review previous work on speaker dependent MO selection and frame based techniques on 'resolution' selection before we introduce our proposal on spectral envelope estimation with frame based MO selection.

2.1. Review of 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 have extended this approach by *warping* the frequency axis with the bilinear transformation prior to MVDR spectral estimation [4, 5], therefore dubbed *warped-MVDR*, 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.

For a fast computation of the warped-MVDR spectrum we have extended Musicus' [6] algorithm to calculate the MVDR spectrum of order N from the *linear prediction coefficients* (LPC) $a^{(N)}_{0\cdots N}$ of order N as follows:

1. **Computation of the warped-LPC** $\tilde{a}$
   For our experiments we used an algorithm by Matsumoto et al. [7] to calculate the warped-LP coefficients, but any other algorithm should work similarly well.
2. **Correlation of the warped-LPC**

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

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

$$S_{\mathrm{warped-MVDR}}(\omega) = \frac{\epsilon}{\sum_{k=-N}^{N}\tilde{\mu}_k e^{-j\omega k}} \qquad (1)$$

   ϵ: inverse of the prediction error variance.

Note that the spectrum (1) 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 [8], since

$$\log(a+b) \approx \log(\max\{a,b\}).$$

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 additive noise [9].

2.2. Previous Work on Speaker Dependent Model Order

In previous work [10] we have introduced and investigated a *maximum-likelihood* (ML) based MO selection technique for spectral envelopes to apply speaker dependent adaptation in the feature-space similar to *vocal tract length normalization* (VTLN). In order to choose the optimal MO, a spectral envelope must be estimated that provides features leading to the best possible match to the speaker-independent acoustic models of the recognizer.

For a sequence of different MOs $m \cdots n$ we can write the cepstral features as the matrix $C = (\mathbf{c}_m, \mathbf{c}_{m+1}, .., \mathbf{c}_n)^T$. Let λ_l denote a set of given hidden Markov 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 from a previous recognition run:

$$\hat{m} = \underset{m}{\operatorname{argmax}}\, P(C|\lambda_l, W)$$

2.3. Previous Work on Frame Dependent 'Resolution'

In the work by Nakatoh et al. [11] it was proposed to adjust the warping parameter α of the warped-LPC followed by compensation of pre-emphasis and compensation for the frequency warping. This approach leaves untouched the overall frequency resolution of the envelope as the number of LPCs stay constant, but emphasizes the resolution of lower ($\alpha >$ Mel) or higher ($\alpha <$ Mel) frequencies as more parameters are used to describe the lower frequencies and fewer to describe the higher ones; or vice versa.

To control the frequency resolution frame by frame, indexed by i, a division of the first $R[1]$ by the zero $R[0]$ order autocorrelation coefficient was used:

$$\beta_i = \frac{R_i[1]}{R_i[0]} \leq 1 \tag{2}$$

In the work by Tyagi and Bourland [12] it was proposed to perform 'multi-scale Fourier transform analysis' by differently sized windowing functions. To select the best window size, the window with the lowest spectral entropy criteria

$$H_i = -\sum_{k=0}^{L-1} P_i(k) \cdot ln(P_i(k))$$

was used where P in the equation above stands for the normalized power spectrum:

$$P_i(k) = \frac{|X_i(k)|^2}{\sum_j^{L-1} |X_i(k)|^2} \tag{3}$$

2.4. Frame Dependent Model Order

In this work we will concentrate on modifying the MO of the spectral envelope in contrast to the warp factor as done by Nakatoh et al. or the window size as done by Tyagi and Bourland to control the frequency resolution. Using the MO instead of the warp factor allows to control the overall frequency resolution without modifying the window size as the number of LPCs varies.

For the control factor several functions could be used such as the formant structure, phoneme type or the SNR. Here only two control factors should be investigated to not overstress the experimental part of this paper, concentrating on the feasibility of the suggested approach. A more detailed investigation of possible control factors could follow in future work.

The first approach for the control factor follows Nakatoh et al. by using a relation of the autocorrelation coefficient (2). Furthermore, to compensate for the high variation of the coefficients we use simple smoothing and the absolute of the input coefficients:

$$\beta_{\text{smooth},i} = \frac{1}{4}\beta_{i-1} + \frac{1}{2}\beta_i + \frac{1}{4}\beta_{i+1} \tag{4}$$

With the given smoothed control factor $\beta_{\text{smooth},i}$ we can now easily obtain the MO N as

$$N_i = \max\left\{N_{\max} \cdot \beta_{\text{smooth},i}\ ,\ N_{\min}\right\}$$

where $N_{\max}$ is the maximum value of N and $N_{\min}$ represents the smallest value (we have used a value of 20) to prevent the envelope calculation is using a MO which is too small to give a reasonable approximation. $N_{\max}$ is set to keep the mean of N similar to the value of the fixed MO (in our experiment 60).

The second approach for the control factor uses the spectral entropy calculated on the normalized (3) warped-MVDR envelope with MO 60 instead of the Fourier transformation as used by Tyagi and Bourland, leading to a smoother estimate and therefore resulting in a more reliable entropy value. N is then obtained by smoothing the entropy value, as in (4),and a lower threshold $N_{\min}$ as before:

$$N_i = \max\left\{N_{\text{multiply}} \cdot (O - H_{\text{smooth},i})\ ,\ N_{\min}\right\}$$

The offset O and the multiplication factor N_{multiply} is chosen so that N has a mean value similar to the fixed MO value and a variance of a fourth of the fixed MO value.

Figure 1 presents spectrograms of warped-MVDR envelopes one with a fixed MO, set to 60, and two with a variable MO. Comparing the MOs resulted by the different control factors, namely functions based on the quotient of the first by the zero autocorrelation coefficient or the spectral entropy, we see that the two criteria are very different from each other. While the autocorrelation coefficient based approach has a smooth nature the spectral entropy is rough. The first has a high MO for most of the frames and reduces the MO where more energy is present in high frequency regions than in low frequency regions. The latter reduces the MO of the regions where noise or silence is present as the spectral contour is flat for these frames.

3. Speech Recognition Experiments

The speech recognition experiments described below were conducted with 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.

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 non-native, 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 training, the acoustic models were trained on the *Broadcast News* corpus [14] (104 hours of speech) and adapted on 31 speakers (8 hours) out of the 39 transcribed speakers from the TED corpus using *Maximum likelihood linear regression* (MLLR). Our test set contained the final 8 speakers.

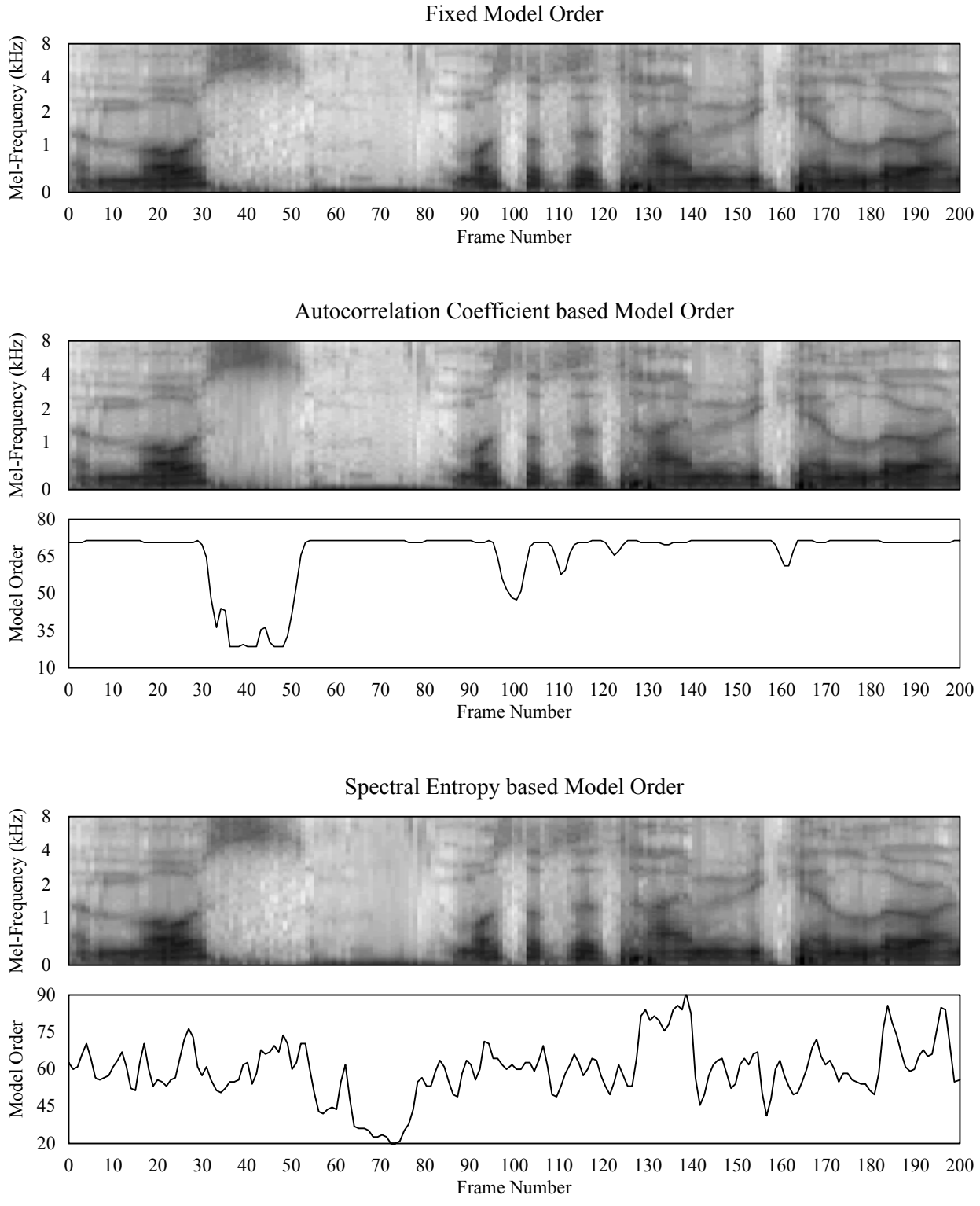

Figure 1: Spectrograms of warped-MVDR envelopes including a fixed and frame by frame model order. The spectrum on the top uses a fixed model order, set to 60. The model order of the center spectrum is based on the quotient of the first by the zero autocorrelation coefficient. The model order of the spectrum on the bottom is based on the spectral entropy.

Our baseline model consisted of 4.139 codebooks with 32 Gaussians each. The 13 static normalized cepstral coefficients were obtained every 10 ms through a discrete cosine transform from different spectral representations using a 16 ms Hamming window to split the speech data sampled at 16 kHz. Thereafter, VTLN was applied before the linear discriminant analysis reduced the current features plus 3 left and right adjacent features to a final feature length of 40. To adapt the means and covariances of the speaker-independent model for every speaker in the test set MLLR was used. As feature space adaptation gave only a small improvement in *word error rate* (WER) we decided to not use it in our system.

The *language model* (LM) consisted of an interpolation of a 3-gram LM based on the talks by the TED adaptation speakers (60k words), a 3-gram LM based on proceedings from conferences such as ICSLP, Eurospeech and ICASSP (17M words) and a class-based 5-gram LM based on broadcast news and the aforementioned conference proceedings (160M words). The overall perplexity was 133 with an out of vocabulary rate of 0.3% for a vocabulary size of 25,000 words, including multiwords and pronounciation variants the dictionary contained 40,000 entries.

Comparing the proposed MO selection, both of mean value 60, with the fixed MO 60, reported in Table 1, we can see 2.4% relative error reduction for the autocorrelation approach. Comparing to the traditional Fourier transformation (Mel-frequency cepstral coefficients) the relative error reduction is 4.2%. Both control factors lead to a gain over the fixed MO, but the autocorrelation coefficient based approach is significantly better than the spectral entropy. Also, the autocorrelation coefficient based approach performs better than the speaker dependent MO, as they are different in nature they might be combined to good effect. For the sake of completeness others feature are also presented in the table, but should not be further discussed here.

	WER	RER
Power Spectrum	38.4%	--
LP	39.7%	-3.4%
Perceptual-LP	38.9%	-1.3%
Warped-LP	38.7%	-0.8%
MVDR	38.6%	-0.5%
Warped-MVDR	38.1%	0.8%
Warped&Scaled-MVDR		
fixed	37.7%	1.8%
ML per speaker	37.0%	3.6%
autocorrelation	36.8%	4.2%
spectral entropy	37.5%	2.3%

Table 1: *Word error rates* (WER)s and *relative error reductions* (RER)s for eight speakers comparing different spectral representations.

4. Conclusions

We have demonstrated the possibility to improve the accuracy of a speech recognition system by modifying the MO frame by frame according to two different control factors on the Translanguage English Database. As both control factors are different in nature and both are leading to an improvement we can conclude that neither of them leads to the best possible performance. Therefore, in future investigations other control factors have to be considered. As the free values of the used control factors have been set a priori a further improvement in the performance of the presented control factors could be expected by tuning the free parameters on a development set. Furthermore, it should be investigated if the frame based MO – and warp factor selection could be combined to good effects.

5. Acknowledgment

The work presented here was partly funded by the *European Union* (EU) under the project 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, isl.ira.uka.de/˜wolfel.

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

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

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

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

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

[10] M.C. Wölfel, "Speaker dependent model order selection of spectral envelopes," *ICSLP*, 2004.

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

[12] V. Tyagi and H. Bourland, "On multi-scale fourier transform analysis of speech signals," *IDIAP Research Report 03-32*, 2004.

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

[14] Linguistic Data Consortium (LDC), "English broadcast news speech (Hub-4)," LDC97S44.