Autoregressive Modeling based Feature Extraction for Aurora3 DSR Task

Petr Motlíček[1,2], Jan Černocký[1]

[1]Faculty of Information Technology, Brno University of Technology
Božetěchova 2, Brno, 612 66, Czech Republic
[2]OGI School of Science and Engineering at OHSU
20000 NW Walker Road, Beaverton, OR 97006, USA

{motlicek,cernocky}@fit.vutbr.cz

Abstract

Techniques for analysis of speech, that use autoregressive (all-pole) modeling approaches, are presented here and compared to generally known Mel-frequency cepstrum based feature extraction. In the paper, first, we focus on several possible applications of modeling speech power spectra that increase the performance of ASR system mainly in case of large mismatch between training and testing data. Then, the attention is payed to the different types of features that can be extracted from all-pole model to reduce the overall word error rate. The results show that generally used cepstrum based features, which can be easily extracted from all-pole model, are not the most suitable parameters for ASR, where the input speech is corrupted by different types of real noises. Very good recognition performances were achieved e.g., with discrete or selective all-pole modeling based approaches, or with decorrelated line spectral frequencies. The feature extraction techniques were tested on SpeechDat-Car databases used for front-end evaluation of advanced distributed speech recognition (DSR) systems.

1. Introduction

The purpose of feature extraction (often referred to as signal modeling algorithms) is to transform audio data into a space where observations from the same class will be grouped together and observations from different classes will be pushed apart. For their derivations, psychological studies of the human auditory and articulatory systems were used.

The short-time Fourier spectrum is usually examined as the first preprocessing block of the feature extraction. Usually, the length of analyzed frames is 25ms with 10ms time shift and weighted by Hamming window.

Most feature extraction methods employ cepstral analysis to extract the vocal tract component from the speech signal. Many algorithms have been proposed to compute the cepstrum, e.g., as well known MFC (Mel Frequency cepstrum). The second widely used technique, Perceptual Linear Prediction (PLP) analysis, combines several engineering approximations of psychology of human hearing processes. PLP is built on all-pole modeling of critical band warped power spectrum of speech. The critical band analysis simulated by an auditory-based warping of the frequency axis is derived from the frequency sensitivity of human hearing. In original approach, Bark scale warping function is employed.

In an all-pole model based analysis, we examined several possibilities to obtain appropriate features that are less sensitive to the noise that corrupts the input speech signal. At first,

the goal is to find the optimal parameters of system with classical usage of all-pole modeling. Then, we focus on frequency selective and discrete all-pole modeling approaches, where the preceding operations were kept unchanged.

Linear system, represented by an all-pole model approximating the envelope of given signal spectrum (in our case perceptually warped spectrum), is fully described by set of linear prediction coefficients (LPCs) with additional information about the total energy represented by the gain factor. Such system can obviously be defined by different types of coefficients that might be more or less suitable for ASR. Once we have estimated LPCs, we are unrestricted to use such one type of representation of linear system. This is great advantage of PLP versus MFC analysis. In our experiments, we were interested (besides cepstral coefficients) in line spectral frequencies (LSFs), reflection coefficients (RCs) and log-area ratio coefficients (LARs).

It has been shown [10] that the noise cancellation [3] and Voice Activity Detection (VAD) algorithms play important roles to achieve high recognition performance in SpeechDat-Car task. However, our goal is to find better representation of speech that outperforms standard feature extraction algorithms and that eventually can be preceded or followed by such systems.

2. All-pole model approaches

2.1. Classical method

Graphical description of PLP analysis [5], which contains several preprocessing blocks, is given in Fig. 1. At the beginning, the speech power spectrum is integrated within overlapping critical band filter responses. In order to compensate the unequal sensitivity of human hearing at different frequencies, the next processing stage simulates equal loudness (EQL) curve, such as:

$$E(\omega) = \frac{(\omega^2 + 56, 8.10^6)\omega^4}{(\omega^2 + 6, 3.10^6)^2(\omega^2 + 0, 38.10^9)}. \quad (1)$$

It can be replaced by preemphasis that is applied in the time-domain using first-order high pass filter:

$$H(z) = 1 - \alpha z^{-1}, \quad \alpha \in \langle 0.9, 0.95 \rangle. \quad (2)$$

The next stage, called intensity-loudness power law, models a non-linear relation between the intensity of sound and its perceived loudness. In PLP analysis, a cubic root compensation of critical band energies is applied and resulting power spectrum $P(\omega)$ is obtained. Such $P(\omega)$ is then approximated by the frequency response of the all-pole model obtained by Levinson-Durbin algorithm [6]. The input autocorrelation coefficients
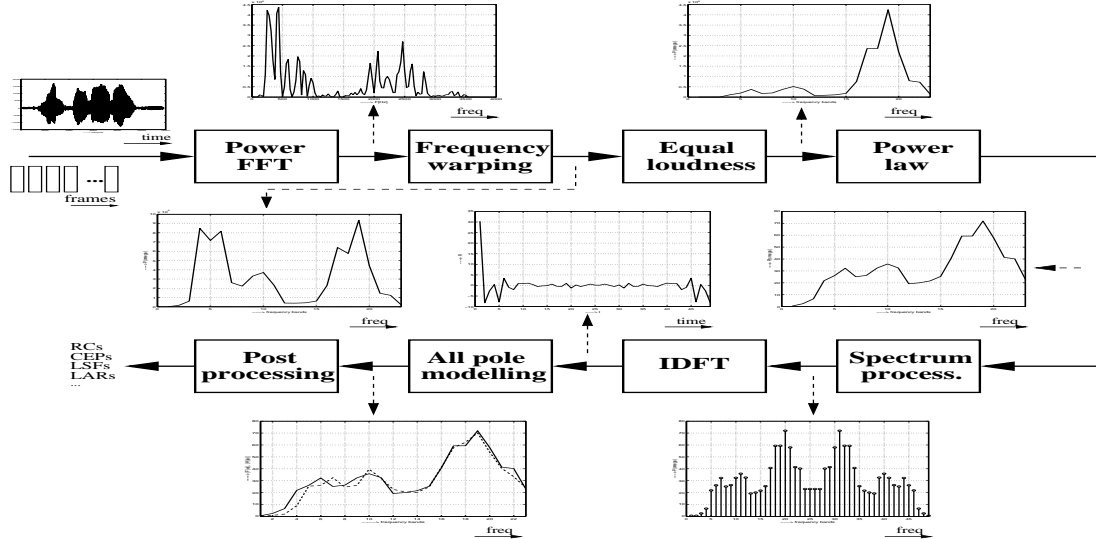
Figure 1: PLP analysis with time and frequency trajectories of voiced speech for $F_{sampl} = 8\text{kHz}$.

$\{R_k\}$ for Levinson-Durbin algorithm come from application of inverse discrete Fourier transform (IDFT) to the warped power spectrum $P(\omega)$. The model power spectrum is given as:

$$\hat{P}(\omega) = \frac{G^2}{\left|1 + \sum_{k=1}^{p} a_k e^{-jk\omega}\right|^2},$$ (3)

where p is the order of the model, $\{a_k\}$ are the model coefficients (LPCs) given by:

$$\sum_{k=1}^{p} a_k R_{i-k} = -R_i, \quad 1 \le i \le p,$$ (4)

and G^2:

$$G^2 = R_0 + \sum_{k=1}^{p} a_k R_k = \sum_{k=0}^{p} a_k R_k,$$ (5)

is power of the gain factor. Because of such model property that $\frac{\partial \hat{P}(\omega)}{\partial \omega} = 0$ for $\omega = 0, \pi$; it is reasonable to repeat the first and last frequency bands of warped spectrum before its symmetrization, as can be seen in Fig. 1 (spectrum processing block).

Normalized minimum error V_p defined as:

$$V_p = \frac{G^2}{R_0},$$ (6)

can be used to determine the optimal value of p [6].

2.2. Frequency selective all-pole modeling

In SpeechDat-Car task, where the speech is corrupted by different types of real noises, it can be very advantageous to process regions of power spectrum independently [10]. Applying to our feature extraction algorithm, we can approximate different part of warped power spectrum by separate all-pole model. There can be several other reasons to use frequency selective all-pole modeling in speech recognition that can be justified by speech perception studies [6].

2.3. Discrete all-pole modeling (DAP)

In extended all-pole modeling of discrete signal spectra [7], Itakura-Saito error measure is used to find the coefficients of an all-pole model. Standard approach based on linear prediction error criterion is the wrong approach to envelope estimation for discrete spectra, since it does not account for the aliasing caused by spectral sampling. DAP is superior to the standard LP analysis. Coefficients of an all-pole model $\{d_k\}$ are (instead of Eq. 4) found as:

$$\sum_{k=0}^{p} d_k R_{i-k} = \hat{h}_{-i}, \quad 0 \le i \le p,$$ (7)

where $\hat{h}_{-i}$ is the (time-reversed) impulse response of the discrete frequency sampled all-pole model, and $\{d_k\}$ are coefficients of the all-pole model. However, the set of equations is not linear and has to be solved iteratively.

3. Extraction of parameters

As mentioned above, linear system approximating the envelope of given speech warped spectrum can be described by set of different coefficients than LPCs.

The derivation of cepstral coefficients $\{c_k\}$ from given set of $\{a_k\}$ is simple. There exists direct transformation from $\{a_k\}$ to $\{c_k\}$ [12]. The cepstral coefficients are highly used in ASR, because they are well decorrelated.

Coefficients $\{a_k\}$, as well as $\{c_k\}$ are known to be inappropriate for quantization. They have large dynamic range. Therefore, by applying the quantization, the all-pole model can get unstable. This might be drawback in feature extraction, especially for DSR system, where the feature stream from the terminal side is quantized, encoded and transmitted to the server side [4]. Hence, different set of parameters representing the same spectral information, and also having good quantization properties, were proposed. Between the most popular belong line spectral frequencies (LSFs) [8]. There are several other good properties of LSFs (besides suitability for quantization), such as: LSFs allow interpretation in terms of formants; their shifting has a localized spectral effect so that quantization errors will primarily affect the region of the spectrum around that

SDC-Accuracy [%]	Italian			Finnish			Spanish			overall
test	hm	mm	wm	hm	mm	wm	hm	mm	wm	
MFC_c	37.17	85.18	93.83	37.53	66.69	92.01	37.47	75.37	84.52	**71.90**
PLP_c	38.14	85.18	94.26	41.2	65.73	91.86	38.92	73.01	87.15	**72.42**
PLP_{slc}	45.56	83.78	93.55	48.62	63.54	90.18	42.23	75.79	87.55	**73.55**
PLP_{dap}	39.27	85.06	93.20	37.56	72.3	90.26	41.2	71.63	88.39	**72.62**

Table 1: Word recognition accuracies from different all-pole modeling experiments.

frequency; minimum phase property of an all-pole model is preserved after quantization of LSFs.

Reflection coefficients $\{k_k\}$ that are a by-product of Levinson-Durbin algorithm can be recursively computed from the filter coefficients $\{a_k\}$ [12]. On contrary to LSFs, they have poor linear quantization properties.

This drawback is attenuated in case of Log-area ratios (LARs). They are computed from $\{k_k\}$ using non-linear transformation [9] so that their spectral sensitivity curves are approximately flat (the shape is very similar to spectral sensitivity shapes of cepstral coefficients).

In our experiments, we have used basic feature normalization (mean and variance normalization applied online), that is usually the last preprocessing block in feature extraction stream, and which allows us to keep similar statistical properties over all experiments. The online mean and variance normalization (OMVN) [10] is based on the estimation of local mean and variance of the features.

It is known that cepstrum based features are well decorrelated. This property is not satisfied in case of LSFs, RCs, and LARs. Hence, such feature streams are linearly transformed using Karhunen-Loève transformation (KLT) before application of OMVN.

4. Experimental setup

The feature extraction algorithms proposed for speech recognition system were tested on three SpeechDat - Car (SDC) databases used for Advanced DSR Front-End Evaluation: Italian [1], Spanish [2] and Finnish SDC. The recordings were taken from the close-talk microphone and from one of the hands-free microphones. Data were recorded at 16kHz, but downsampled to 8kHz. The databases contain various utterances of digits. During experiments, the robustness has been tested under three different conditions. For each of these three conditions 70% of the files were used for training, 30% for testing.

- *Well-matched (wm):* All the files (close-talk and hands-free microphones) were used for training and testing.

- *Medium mis-matched (mm):* Recordings made with the hands-free microphone were used for training while for testing close-talk recordings were taken.

- *Highly mis-matched (hm):* For the training only close-talk microphone recordings were used, whereas for testing the hands-free files were taken.

In all experiments, the output features for speech recognizer were 15 dimensional vectors completed by 15 Δ and 15 $\Delta\Delta$ coefficients. An ASR back-end based on HMM-HTK recognizer defined for evaluation purposes [11] has been used in our work.

The overall results of the experiments are obtained so that the **wm** conditions are weighted by 0.4, **mm** by 0.35, and **hm** by

0.25 over average of all three databases, as defined for Aurora3 task.

5. Experiments and results

In first experiments, the goal was to estimate the optimal parameters of all preprocessing blocks of PLP analysis (Fig.1). The experimental baseline were MFCCs (MFC_c). The partial results (given in [9]) with PLP based feature extraction point out that: **a)** Warping of input power spectra using Mel filter bank performs slightly better than using Bark filter bank. **b)** Equal loudness - does not play an important role in our experiments. **c)** Power law - allows better fitting of an all-pole model the signal spectrum. Value of power root varying in range of $\frac{1}{3} - \frac{1}{2}$ does not affect the recognition performance. **d)** Spectrum processing - symmetrization of the warped power spectrum with repetition of side frequency bands results into a better fitting of an all-pole model on the sides of frequency axis of the signal spectrum. Such operation brings small improvements over all training conditions. **e)** The optimal value of p has been estimated using V_p (Eq. 6) and is equal to 14. This value was proved experimentally, as well.

The final results of PLP analysis with optimal parameters (cepstrum based features - PLP_c experiment) are in Tab. 1.

In experiments with frequency selective all-pole modeling, a warped power spectrum was split into lower and upper parts (initial preprocessing operations were kept untouched, as in PLP_c). These two frequency parts were approximated by all-pole models separately, and concatenated either on the level of approximating frequency responses of these models, or on the level of their cepstral coefficients. The best performance (PLP_{slc}) was achieved with $f_{cutoff} = 1.8$kHz, where lower and upper parts of warped spectra were approximated by linear system with order $p_1 = 12$, and $p_2 = 5$, respectively. 10 and 4 cepstral coefficients were computed independently from each all-pole model and linearly concatenated (one stream with global energy information added) into resulting feature stream.

DAP has been applied on the same warped spectrum as in PLP_c experiment (PLP_{dap}). In the first iteration, standard LP analysis was performed. Then, several iterations of DAP (Eq. 7) re-estimate $\{d_k\}$ coefficients of an all-pole model. The iterative approach is computationally expensive. However, the values of $\hat{h}_{-i}$ quickly converge to zero so that the number of iterations was usually less than 10.

The second branch of experiments (Tab. 2) was focused on extraction of parameters from given all-pole model. In order to keep the same recognition setup, all extracted features were normalized by online mean and variance normalization (OMVN). The baseline are MFCCs, followed by OMVN ($MFCC\_O$). PLP analysis (optimally chosen set of parameters) with cepstrum based output features were also normalized by OMVN (PLP_O). OMVN was applied on top of the LSFs, LARs and reflection based (RCs) output features (LSF_O,

SDC-Accuracy [%]	Italian			Finnish			Spanish			overall
test	hm	mm	wm	hm	mm	wm	hm	mm	wm	
$MFCC\_O$	73.83	74.35	92.00	59.12	59.37	88.91	73.68	86.57	92.01	**79.31**
PLP_O	56.9	84.26	94.21	62.61	86.18	93.73	74.02	81.14	89.76	**82.51**
LSF_O	64.72	86.9	94.09	61.31	86.39	94.8	70.68	82.21	90.83	**83.48**
$REFL\_O$	56.88	69.12	89.16	56.25	84.13	93.32	62.83	77.06	88.52	**77.67**
LAR_O	61.18	86.78	93.64	61.52	85.84	94.48	65.95	80.45	88.88	**82.18**

Table 2: Word recognition accuracies related to different types of features extracted from the all-pole model. In overall, the highest accuracy on SDC task was achieved with LSFs normalized by OMVN.

LAR_O, $REFL\_O$ experiments, respectively) that were computed from given all-pole model (according to PLP_c experiment). More detailed results are given in [9].

6. Conclusions

In the paper, we have discussed the advantages of all-pole modeling employed in ASR. The relevance of elementary operations of PLP analysis was mentioned. The results, given in Tab. 1, show that PLP based features provide generally better performance than MFCCs on SDC task. Frequency selective and discrete all-pole modeling approaches, in overall, also increase the final recognition performance.

However, the best results on SDC tasks were achieved with LSFs, decorrelated by KLT and normalized by OMVN. This type of features outperforms MFCCs as well as PLP-cepstrum based features (also normalized by OMVN), which is evident from Tab. 2. Generally, achieved results indicate good properties of all-pole model based speech features. Although, cepstrum based features are very popular in speech recognition due to their good decorrelation properties, they can be very well substituted by LSFs or LARs decorrelated by standard KTL.

For easy evaluation of individual feature extraction systems the overall results were primarily taken into account. The results mentioned in Tab. 1 give also consistently better recognition performances over all three different training conditions (**hm, mm, wm**). However, similar conclusions can not be explicitly made for results shown in Tab. 2. If we look for example at the worst ($MFCC\_O$) and the best feature extraction systems (LSF_O), we can easily observe different performances over different SDC databases. This is definitely a weakness of the weighting system defined in the Aurora3 task.

In the future work we intend to tie LSF based features with frequency selective or discrete all-pole modeling approaches. In spite of the fact that OMVN improves global recognition accuracies (already verified in several other experiments [10]), we have observed that achieved results are inconsistent across different SDC databases. The influence of OMVN will be further investigated.

7. Acknowledgments

8. References

[1] U. Knoblich. "Description and Baseline Results for the Subset of the SpeechDat-Car Italian Database used for ETSI STQ Aurora WI008 Advanced DSR Front-End Evaluation", Alcatel, April 2000.

[2] D. Macho. "Spanish SDC-Aurora Database for ETSI STQ Aurora WI008 Advanced DSR Front-End Evaluation, Description and Baseline Results", UPC, November 2000.

[3] P. Motlicek, L. Burget. "Noise Estimation for Efficient Speech Enhancement and Robust Speech Recognition", *in proceedings of the ICSLP*, pp. 1033-1036, Denver, USA, September 2002.

[4] H. G. Hirsch & D. Pearce. "The AURORA Experimental Framework for the Performance Evaluations of Speech Recognition Systems under Noisy Conditions", ISCA ITRW ASR2000 "Automatic Speech Recognition: Challenges for the Next Millennium", Paris, France, September 2000.

[5] H. Hermansky. "Perceptual linear predictive (PLP) analysis of speech", *Journal of the Acoust. Soc. of Am.*, vol. 87, pp. 1738-1752, 1990.

[6] J. Makhoul. "Spectral linear prediction: properties and applications", *IEEE Trans. on ASSP*, vol. 23, pp. 283-296, June 1975.

[7] A. El Jaourdi , J. Makhoul. "Discrete all-pole modeling", *in proceedings of the ICASSP*, pp. 320-323, Dallas, USA, 1987.

[8] F. Itakura. "Line spectrum representation of linear predictive coefficients of speech signals", *Journal of the Acoust. Soc. of Am.*, vol. 57, p. S35, April 1975.

[9] P. Motlíček. "Feature extraction in speech coding and recognition", *Technical Report of PhD research internship in ASP Group, OGI-OHSU*, <http://www.fit.vutbr.cz/~motlicek/publi/2002/rp.pdf>, 2002.

[10] A. Adami, L. Burget, S. Dupont, H. Garudari, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, S. Sivadas. "QUALCOMM-ICSI-OGI FEATURES FOR ASR", *in proceedings of the ICSLP*, pp. 21-24, Denver, USA, September 2002.

[11] "Advanced DSR Front-end: Definition of required performance characteristics", Technical report, version 3, source: Motorola, October 2001.

[12] L. R. Rabiner and R. W. Schafer. "Digital Processing of Speech Signals", Prentice-Hall, 1978.