Author(s): Heikki Kallasjoki, Kalle J. Palomaki, Carlo Magi, Paavo Alku and Mikko Kurimo

Title: Noise robust LVCSR feature extraction based on stabilized weighted linear prediction

Year: 2009

Version: Final published version

Please cite the original version:

Heikki Kallasjoki, Kalle J. Palomaki, Carlo Magi, Paavo Alku and Mikko Kurimo. Noise robust LVCSR feature extraction based on stabilized weighted linear prediction. In Proceedings of the 13th International Conference on Speech and Computer (SPECOM 2009), pages 221-225, St. Petersburg, Russia, June 2009.

This publication is included in the electronic version of the article dissertation:
Kallasjoki, Heikki. Feature Enhancement and Uncertainty Estimation for Recognition of Noisy and Reverberant Speech. Aalto University publication series DOCTORAL DISSERTATIONS, 31/2016.

Noise Robust LVCSR Feature Extraction Based on Stabilized Weighted Linear Prediction

Heikki Kallasjoki[1], *Kalle J. Palomäki*[1], *Carlo Magi*[2], *Paavo Alku*[2], *Mikko Kurimo*[1]

[1]Adaptive Informatics Research Centre, Helsinki University of Technology, Finland
[2]Department of Signal Processing and Acoustics, Helsinki University of Technology, Finland
htkallas@cis.hut.fi

Abstract

In this paper, we evaluate a recently proposed spectral envelope estimation method, stabilized weighted linear prediction (SWLP), in the feature extraction stage of a large vocabulary continuous speech recognizer (LVCSR) system. Using speech recorded in real-world noisy environments, we compare recognition error rates obtained with SWLP to those given by the conventional spectrum estimation methods in feature extraction. We use large vocabulary speech that is simultaneously recorded using a headset, lavalier and a fixed medium-distance microphone. When the recognizer is trained using a multicondition training set the results do not differ significantly. However, when the models are trained with clean speech and evaluated in noisy environments, the SWLP models perform significantly better than the conventional MFCC in all environments and in all recording settings.

1. Introduction

Searching efficient feature representations for speech data is one of the key issues in building speech recognition systems. Evidently, the feature extraction phase should retain only the information necessary for differentiating between potential meanings, and discard as much as possible of the speaker dependent variations such as the harmonic structure.

Mel-frequency cepstral coefficients (MFCCs) have been widely used as a feature set for speech recognition [1]. The method is based on computing the cepstral coefficients of the short-term spectrum estimate, smoothed by a perceptually motivated mel-scaled filterbank. However, the perceptual smoothing is only partially efficient in removing the harmonic structure present in speech. In addition, the resulting features are not particularly robust in the case of noisy speech [2].

Minimum Variance Distortionless Response (MVDR) [3] modeling has been used in feature extraction for speech recognition by replacing the short-term FFT magnitude spectrum in the MFCC computation with the MVDR all-pole spectrum. The MVDR method enables computing smooth spectral envelopes of voiced speech without modeling the harmonic structure. The perceptual frequency representation, such as the mel-frequency scale, can be integrated into the MVDR spectrum estimation by computing MVDR model parameters directly from the mel-filterbank output. This method is advantageous both in terms of its performance and computational cost: compared to the original FFT spectrum, the perceptual mel-scale smoothing improves reliability and reduces dimensionality of the spectrum estimate [4].

Linear prediction (LP) [5] is a traditional method to compute all-pole models for spectral envelopes of speech. It is well-known, however, that the performance of LP deteriorates in modeling of high pitch voices because the spectral models are biased by the relatively sparser harmonic structure of speech. In addition, the performance of LP is vulnerable to noise. The weighted linear prediction (WLP) attempts to improve the noise-robustness of conventional LP by utilizing a temporal weighting function in defining the optimal filter coefficients [6]. With temporal weighting, the contribution of speech samples with a higher signal-to-noise ratio (SNR) can be increased in the computation of LP models. Moreover, in modeling of voiced speech, the use of the short-time energy (STE) weighting function enables emphasizing the role of speech samples located in the closed phase of the glottal cycle thereby concentrating the spectral modeling on a time span during which the speech formants are most prominent. Because the original WLP method does not guarantee the stability of the resulting all-pole model, a new method, Stabilized Weighted Linear Prediction (SWLP), was recently proposed [7].

In this paper, SWLP is used as a method for spectral estimation in feature extraction for a large vocabulary continuous speech recognition (LVCSR) system. This study compares SWLP to other widely used spectral estimation methods: FFT and conventional LP. In the further stages of feature extraction, spectra modeled with FFT, LP or SWLP are used to produce MFCC features for the LVCSR system. Earlier work on SWLP includes spectral distortion measurements and subjective listening tests on speech corrupted with Gaussian white noise, as well as automatic speech recognition tests in an isolated word recognition task using pre-recorded noise data added to clean speech. The current study extends the previous studies [7] by comparing the speech recognition performance of different spectral envelope estimation methods on large vocabulary Finnish continuous speech material recorded under real noisy conditions.

2. Methods

Our speech recognition experiments utilize different feature extraction methods which are described in this section. The emphasis is on the computation of the SWLP based feature representation.

2.1. Stabilized weighted linear prediction

In the linear prediction model, a sample x_n is estimated as

$$\hat{x}_n = -\sum_{i=1}^{p} a_i x_{n-i}, \qquad (1)$$

where p is the model order, and $a_i \in \mathbb{R}$ are the linear prediction coefficients. By denoting $\mathbf{a} = [1\ a_1\ \cdots\ a_p]^T$ and

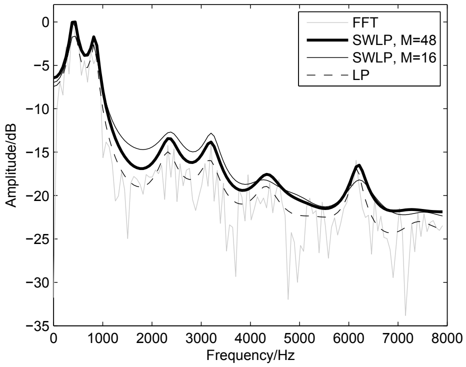

Figure 1: *Spectral envelopes of order* $p = 20$ *computed for a clean vowel with conventional linear prediction (LP), and stabilized weighted linear prediction (SWLP).*

$\mathbf{x}_n = [x_n \cdots x_{n-p}]^T$ we get the prediction error $\varepsilon_n(\mathbf{a})$ in matrix form as

$$\varepsilon_n(\mathbf{a}) = x_n - \hat{x}_n = \mathbf{a}^T\mathbf{x}_n. \tag{2}$$

By utilizing the concept of the weighted linear prediction, the components of the coefficient vector $\mathbf{a}$ are found by minimization of a cost function $\mathscr{E}(\mathbf{a}) = \sum_{n=1}^{N+p}(\varepsilon_n(\mathbf{a}))^2 w_n$. This can be written in matrix form as follows:

$$\mathscr{E}(\mathbf{a}) = \mathbf{a}^T\mathbf{R}\mathbf{a}. \tag{3}$$

In the WLP method the autocorrelation matrix $\mathbf{R} = \sum_{n=1}^{N+p} w_n\mathbf{x}_n\mathbf{x}_n^T$ is *weighted*, while in conventional LP the temporal weights are not used: $w_i = 1, i \in [N+p]$.

To ensure the stability of the resulting model, the stabilized WLP uses a modified autocorrelation matrix $\mathbf{R}$ in equation 3 [7]. The WLP autocorrelation matrix can be written as $\mathbf{R} = \mathbf{Y}^T\mathbf{Y}$, where the columns $\mathbf{y}_k$ of $\mathbf{Y}$ are

$$\begin{aligned} \mathbf{y}_0 &= [\sqrt{w_1}x_1 \cdots \sqrt{w_N}x_N\ 0 \cdots 0]^T \\ \mathbf{y}_{k+1} &= \mathbf{B}\mathbf{y}_k, \quad k = 0, 1, \ldots, p-1, \end{aligned} \tag{4}$$

and matrix $\mathbf{B}$ has the secondary diagonal values $\mathbf{B}_{i+1,i} = \sqrt{w_{i+1}/w_i}$ and is zero elsewhere. For the stabilized version (SWLP), these secondary diagonal entries are modified to be

$$\mathbf{B}_{i+1,i} = \begin{cases} \sqrt{w_{i+1}/w_i}, & \text{if } w_i \le w_{i+1}, \\ 1, & \text{if } w_i > w_{i+1}. \end{cases} \tag{5}$$

The selection of the weight function w_n is important in determining the behavior of the model. Similarly to [6] and [7], this study utilizes the short-time energy (STE) as the weighting function:

$$w_n = \sum_{i=0}^{M-1} x_{n-i-1}^2. \tag{6}$$

The parameter M controls the window length of the STE function. Values of M used in the experiments were chosen based on recognition results for the development data sets. The effect of the M parameter is illustrated in Fig. 1.

2.2. Feature extraction

The three feature extraction methods compared in this paper used different spectrum estimation methods, based on stabilized weighted linear prediction (SWLP), conventional (unweighted) linear prediction (LP) and short term FFT. In all cases the audio data, sampled at 16 kHz, was first pre-emphasized with a filter of the form $1 - 0.97z^{-1}$ and divided into partially overlapping Hamming-windowed frames of 256 samples, at a frame rate of 125 frames per second.

The baseline method employed in this paper, referred to as MFCC, was a straight-forward computation of the mel-frequency cepstral coefficients. The FFT was used to find an estimate of the short-term magnitude spectrum. A filterbank of 23 logarithmically spaced triangular filters was then applied to compute the perceptually smoothed spectrum. The cepstral coefficients were obtained as the discrete cosine transform of the logarithm of the filterbank output.

The proposed feature extraction method, referred to as SWLP-MFCC, was based on the stabilized WLP formulation. In this method, the SWLP model coefficients were first derived from the windowed audio frames, using weights computed with the STE function (equation 6). The impulse response of the constructed model was then used as input for MFCC computation identical to the first method. The chosen SWLP model order was $p = 20$, based on earlier smaller scale experiments.

To evaluate the effectiveness of the weighting in the LP model construction, the SWLP results were also compared to conventional LP models. The algorithm used was identical to the SWLP-MFCC method, except that instead of using the STE function all weights were set to unity. This method is referred to as LP-MFCC. In addition, preliminary development set experiments were made with the MVDR-based methods such as PMCC [3, 4], but these were not included in the final results due to their unexpectedly low performance.

The final feature vectors used were 39-dimensional, containing 12 cepstral coefficients, the logarithmic frame energy and their first and second derivatives. Cepstral mean subtraction (CMS) was applied to the cepstral coefficients and the energy term. The feature vectors were also normalized to have a zero mean and unit variance, and finally a maximum likelihood linear transformation (MLLT) estimated during the training phase was applied.

3. Experimental evaluation

3.1. Speech material

Material from the SPEECON [8] Finnish language corpus was used in the experiments. Two different training sets were constructed. The first training set consisted of approximately 21 hours of clean speech, from 293 separate speakers. The second training set was of similar size, but contained an even split of clean and noisy speech, with noisy recordings both from the public place and car environments. SNR estimates provided by the recording platform give an average SNR of 26 dB for the clean speech and 12 dB for the noisy recordings.

Test data from two different environments were used for the feature extraction method evaluations. In both cases, the recognition tests were performed using three audio channels corresponding to three different microphones which were recorded simultaneously at various distances. Smaller development sets of similar data were used to tune the parameters of the evaluated methods.

The first evaluation set utterances were recorded in a moving car, and the set contained 30 read phonetically rich sentences for each of the 20 speakers, with a total length of 57 minutes including the leading and trailing silences. The corresponding development set length was 29 minutes. Channel 0 microphone was a headset microphone positioned 2–5 centimeters away from the speaker's mouth. Channel 1 audio was recorded with a lavalier microphone positioned between the chin and the shoulder of the speaker. Finally, channel 2 was recorded with a medium-distance microphone mounted at the car ceiling behind the rear-view mirror. Average SNR estimates for the evaluation sets were 14 dB, 5 dB and 8 dB for channels 0, 1 and 2, respectively.

Second evaluation set recordings were done both indoors and outdoors in public places, and contained various types of noise such as speech, footsteps etc. in the background. The evaluation set contained 30 read sentences from 30 separate speakers and had a length of 94 minutes, while the development set had a length of 60 minutes. Channel 0 and 1 microphones were identical to the first evaluation set, but channel 2 was recorded with a different medium-distance microphone placed 0.5–1 meter away from the speaker. The average SNR values for this data set were 24 dB, 14 dB and 9 dB, again for channels 0, 1 and 2, respectively.

3.2. Experiment setup

The speech recognition experiments were performed with our large vocabulary continuous speech recognizer. The language model of the recognizer is an n-gram model trained with a growing method [9] on a Finnish language data set containing book and newspaper data, to a total of approximately 145 million words. The language modeling units used by the n-gram model are statistical morphs learned from the text data with an unsupervised method [10]. The decoder employs a one-pass time-synchronous Viterbi beam search algorithm [11]. The acoustic model is based on cross-word triphones modeled with state-clustered hidden Markov models using Gaussian mixtures. The model states use a mixture of on average 16 Gaussians to model the speech feature space and an additional Gamma probability distribution function for state duration modeling [12]. Separate lmscale values, derived from development set recognition results, were used with the different spectrum estimation methods in clean or noisy environments.

3.3. Parameter optimization

The letter error rate (LER) was used as the primary performance measure, for which optimizations were conducted and statistics were calculated. Word error rates (WER) are shown as a secondary measure for completeness. The WER, despite being a more common measure for other languages, is not well suited to Finnish, because Finnish words are often concatenations of several morphemes and correspond to more than one word in English. As an example, word like 'kahvin+juoja+lle+kin' translates to 'also for a coffee drinker.'

The M parameter values controlling the STE window width in the SWLP-MFCC feature extraction method were derived from development set recognition results. STE window widths ranging from $M = 16$ to $M = 48$ were used in the experiments. Notably, for the clean speech training, it was found advantageous to use a larger window width of $M = 24$ samples when recognizing the noisier speech of channel 2 of both the "car" and "public place" environments, even though the model had been trained using a $M = 16$ sample window.

In this study, fixed M parameter values were used for each data set. The robustness of the spectral envelope extraction clearly depends on the M parameter, which is likely to have different optimal values in different noise conditions. Therefore, we argue that adaptive adjustment of the M parameter towards its optimum would lead to improved recognition results. Considering adaptation where the M value would be selected independently for each recognized sentence, one theoretical lower bound for the average letter error rate can be obtained by using the reference transcripts to select for each sentence the M value leading to the lowest letter error rate. This method, referred to later as oracle-based M-value adaptation, was investigated by using a subset of the development data sets.

3.4. Recognition results

The recognition results for the public place and car evaluation sets for clean speech and multicondition training are collected in Table 1. The three different systems compared here are based on the FFT-MFCC, LP-MFCC and SWLP-MFCC feature extraction methods. Wilcoxon signed rank test was used for pairwise statistical comparisons between the letter error rates of different systems for the public place and car data sets in combination (see Table 2).

Note that the letter error rate results for of the "car" data set are lower for channel 1 (lavalier microphone, SNR 5 dB) than for channel 2 (medium-distance microphone, SNR 8 dB) in spite of the lower SNR estimate given by the recording system. Our spectral analysis revealed as a likely explanation that the low SNR estimate is caused by high noise levels at frequencies below 200 Hz, which are outside the spectral areas important for speech recognition.

When using models trained with clean speech only, both linear prediction based methods SWLP-MFCC and LP-MFCC show improvements in the LER compared to baseline MFCC for channels 1 and 2 which have lower SNR values. Table 3 shows the relative improvements of the linear prediction based approaches over the baseline MFCC method in the letter error rates for the different recording channels of a combined "car" and "public places" data set. The relative improvements can be seen to become more marked as the analyzed speech becomes more affected by noise.

The differences between the linear prediction based methods and the baseline MFCC are statistically significant for both channels 1 and 2. Furthermore, the recognition results of the SWLP-MFCC system are also slightly better when compared to the unweighted LP-MFCC system, reaching statistical significance in case of the more difficult recognition task of channel 2.

With models trained in multiple noise conditions, the differences between systems using the three different feature extraction methods are diminished and none of the statistical pairwise comparisons between the systems reached significance. For multicondition training, the MFCC system is most often the best performing with marginal differences to the linear prediction based systems.

Development set data was also used for a preliminary investigation of adaptively selecting the M parameter values. Table 4 presents average letter error rates for the MFCC and LP baseline systems, the SWLP feature extraction method with a fixed M parameter, and finally the SWLP feature extraction using the per-sentence M values of least errors.

Table 1: *Letter error rate (word error rate) percentages for the compared models.*

"Car" test set, models trained with clean speech.

	channel		
method	0	1	2
MFCC	4.0 *(14.2)*	29.6 *(51.9)*	68.6 *(84.7)*
LP-MFCC	3.9 *(14.4)*	27.2 *(49.7)*	55.0 *(78.9)*
SWLP-MFCC	4.0 *(14.6)*	27.1 *(49.5)*	53.4 *(77.4)*

"Car" test set, models trained with noisy speech.

	channel		
method	0	1	2
MFCC	3.7 *(14.0)*	6.8 *(22.1)*	18.0 *(38.3)*
LP-MFCC	3.9 *(14.8)*	7.2 *(23.4)*	17.6 *(40.1)*
SWLP-MFCC	4.1 *(15.1)*	7.9 *(24.2)*	18.2 *(39.8)*

"Public places" test set, models trained with clean speech.

	channel		
method	0	1	2
MFCC	3.3 *(13.6)*	23.4 *(41.8)*	40.8 *(56.5)*
LP-MFCC	3.4 *(14.2)*	20.8 *(40.4)*	34.9 *(53.2)*
SWLP-MFCC	3.3 *(13.6)*	20.4 *(41.2)*	33.2 *(53.6)*

"Public places" test set, models trained with noisy speech.

	channel		
method	0	1	2
MFCC	3.4 *(14.1)*	6.3 *(21.1)*	11.9 *(28.8)*
LP-MFCC	3.6 *(14.8)*	7.1 *(23.4)*	12.2 *(30.2)*
SWLP-MFCC	3.7 *(15.0)*	6.7 *(22.1)*	12.0 *(30.0)*

Table 2: *Statistical significance results for pairwise comparisons between models using the combined "car" and "public places" data set. The name of the better system is shown with the significance level. No statistically significant differences were found for channel 0 results.*

Channel 1

	LP-MFCC	SWLP-MFCC
MFCC	LP-MFCC ($p<0.001$)	SWLP-MFCC ($p<0.05$)
LP-MFCC		(p=N.S.)

Channel 2

	LP-MFCC	SWLP-MFCC
MFCC	LP-MFCC ($p<0.001$)	SWLP-MFCC ($p<0.001$)
LP-MFCC		SWLP-MFCC ($p<0.05$)

Table 3: *Relative LER improvement (%) of the linear predictive models trained with clean speech with respect to the baseline MFCC system.*

	channel		
method	0	1	2
LP-MFCC	-1.7	9.8	17.3
SWLP-MFCC	-0.7	11.1	20.6

Table 4: *SWLP M parameter adaptation test letter error rate percentages, for models trained with clean speech and tested with development set data. The letter "c" denotes the "car" environment, while "p" denotes the "public places" environment.*

	channel, environment			
method	0, c	0, p	2, c	2, p
MFCC baseline	2.9	3.5	52.0	53.2
LP-MFCC baseline	2.9	3.5	41.2	45.8
SWLP-MFCC, fixed M	2.7	3.4	39.7	42.1
SWLP-MFCC, best M	2.2	2.8	34.5	39.6

4. Discussion

In this study, a stabilized weighted linear prediction (SWLP) based spectral envelope estimation method [7] was employed in the feature extraction stage of a large vocabulary continuous speech recognition (LVCSR) system. Using the LVCSR system, SWLP was compared to two other spectrum estimation methods based on the short-time FFT and conventional linear prediction (LP). The tests were conducted with noisy speech data recorded in different real environments, such as public places and cars. Furthermore, two different training sets for the LVCSR models were tested: a set of clean speech material and a multicondition set including speech recorded in adversely noisy conditions in cars and public places.

Compared to the baseline MFCC system, the linear prediction based feature extraction methods (SWLP-MFCC and LP-MFCC) were found to improve recognition rates of noisy speech significantly, when using LVCSR models trained on clean speech. Furthermore, results for the SWLP method were slightly better than those for conventional LP, reaching statistical significance in the case of the most difficult recording channel. For the multicondition training case, the differences between systems diminished and statistical comparisons did not reach significance.

While our preliminary experiences in using MVDR based methods in the current LVCSR task were disappointing, we believe that this line of study deserves further examination [3, 4] in the near future. Improving the recognition performance of the SWLP feature extraction method by adaptively selecting the model parameters, in particular the STE window width M, has shown some promise in in our small scale test using the oracle-based M adaptation. Our future work will consider potential methods of automatically adapting the STE window width without utilizing knowledge of the correct recognition result required by the oracle approach. Furthermore, the oracle-based M adaptation was limited to using a single M value for each sentence. Clearly, further improvements might be achieved with a more frequent updating of the M parameter, e.g. in a phoneme-wise fashion.

5. Acknowledgements

This work was supported by the Academy of Finland projects “Auditory approaches to automatic speech recognition” and “New adaptive and learning methods in speech recognition”.

6. References

[1] David, S. B. and Mermelstein, P., “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences,” *IEEE Trans. ASSP*, 28(4):357–366, 1980.

[2] Yapanel, U. H. and Hansen, J. H. L., “A New Perspective on Feature Extraction for Robust In-Vehicle Speech Recognition,” *Proc. INTERSPEECH*, pp. 1281–1284, 2003.

[3] Murthi, M. N. and Rao, B. D., “All-Pole Modeling of Speech Based on the Minimum Variance Distortionless Response Spectrum,” *IEEE Trans. Speech, Audio Process.*, 8(3):221–239, 2000.

[4] Dharanipragada, S., Yapanel, U. H. and Rao, B. D., “Robust Feature Extraction for Continuous Speech Recognition Using the MVDR Spectrum Estimation Method,” *IEEE Trans. Audio, Speech, Language Process.*, 15(1):224–234, 2007.

[5] Makhoul, J., “Linear Prediction: A Tutorial Review,” *Proc. IEEE*, 63(4):561–580, 1975.

[6] Ma, C., Kamp, Y. and Willems, L., “Robust Signal Selection for Linear Prediction Analysis of Voiced Speech,” *Speech Communication*, 12(1):69–81, 1982.

[7] Magi, C., Pohjalainen, J., Bäckström, T. and Alku, P., “Stabilised weighted linear prediction,” *Speech Communication*, in press, 2009.

[8] Iskra, D., Grosskopf, B., Marasek, K., van den Heuvel, H., Diehl, F. and Kiessling, A., “SPEECON - speech databases for consumer devices: Database specification and validation,” *Proc. LREC*, pp. 329–333, 2002.

[9] Siivola, V. and Pellom, B., “Growing an n-Gram Language Model,” *Proc. INTERSPEECH*, pp. 1309–1312, 2005.

[10] Hirsimäki, T., Creutz, M., Siivola, V., Kurimo, M., Virpioja, S., Pylkkönen, S., “Unlimited vocabulary speech recognition with morph language models applied to Finnish,” *Computer Speech & Language*, 20(4):515–541, 2006.

[11] Pylkkönen, J., “An Efficient One-pass Decoder for Finnish Large Vocabulary Continuous Speech Recognition,” *Proc. 2nd Baltic Conference on Human Language Technologies (HLT’2005)*, pp. 167–172, 2005.

[12] Pylkkönen, J. and Kurimo, M., “Duration Modeling Techniques for Continuous Speech Recognition,” *Proc. INTERSPEECH*, pp. 385–388, 2004.