Noise Robust Feature Extraction Based on Extended Weighted Linear Prediction in LVCSR

Sami Keronen[1], *Jouni Pohjalainen*[2], *Paavo Alku*[2], *Mikko Kurimo*[1]

[1]Department of Information and Computer Science,
Aalto University School of Science and Technology, Finland
[2]Department of Signal Processing and Acoustics,
Aalto University School of Science and Technology, Finland
firstname.lastname@tkk.fi

Abstract

This paper introduces extended weighted linear prediction (XLP) to noise robust short-time spectrum analysis in the feature extraction process of a speech recognition system. XLP is a generalization of standard linear prediction (LP) and temporally weighted linear prediction (WLP) which have already been applied to noise robust speech recognition with good results. With XLP, higher controllability to the temporal weighting of different parts of the noisy speech is gained by taking the lags of the signal into account in prediction. Here, the performance of XLP is put up against WLP and conventional spectrum analysis methods FFT and LP on a large vocabulary continuous speech recognition (LVCSR) scheme using real world noisy data containing additive and convolutive noise. The results show improvements over the reference methods in several cases.

Index Terms: linear prediction, temporal weighting, noise robust, speech recognition

1. Introduction

Extraction of relevant features of speech signal is a key issue in automatic speech recognition (ASR). Short-time spectrum analysis plays an important part in the feature extraction process as frequency is a vital information source. Typically this analysis is done by fast Fourier transform algorithm (FFT) of discrete Fourier transform as a part of mel-frequency cepstral coefficient (MFCC) [1] feature extraction. MFCCs are powerful features if there is no mismatch between the environmental noise in the training material and the recognition conditions. However, MFCCs are not particularly noise robust if a mismatch occurs and therefore they are not well suited, as such, in changing noise condition ASR.

Various feature extraction methods have been designed to address poor recognition performance in mismatch conditions with high noise levels. Such methods include, for example, perceptual linear prediction (PLP) analysis [2], based on explicit modeling of the main phenomena of peripheral auditory processing and usually converted to a cepstral representation. A recently developed feature extraction algorithm called power-normalized cepstral coefficients (PNCC) is also based on auditory processing in a different way from MFCC analysis [3]. However, apart from careful emulation of human auditory perception, there is another approach to short-time feature robustness. Underlying many feature extraction methods, including MFCCs, is the problem of short-time spectrum analysis. So far, relatively little research has been made to make the short-time spectrum approximation robust.

It is possible to replace the spectrum analysis in MFCC computation by linear predictive methods. Temporal weighting in linear predictive analysis aims to emphasize the regions of speech that are relatively less corrupted by noise. Weighted linear prediction (WLP) [4] used in conjunction with short-time-energy (STE) weighting in the MFCC feature extraction has been shown to improve the noise robustness of continuous speech recognition [5] compared to the standard MFCC based on short-time FFT spectral approximation. WLP and its new variants have also led to improved robustness in other recognition applications, including speaker verification [6] [7].

In this work, extended weighted linear prediction (XLP) [7] is introduced to the MFCC feature extraction process to perform the short-time spectral analysis in a large vocabulary continuous speech recognition scheme. XLP is evaluated using speech recorded in demanding authentic noisy conditions and its performance is compared to those of FFT, standard linear prediction and WLP.

2. Methods

In this section, three linear predictive methods are described: standard linear prediction, a well-known temporally weighted method WLP and a new temporally weighted method XLP. Common to all of these is the number of prediction coefficients, denoted by p below, and the use of the autocorrelation criterion in error minimization. In addition, all linear predictive spectral models are finally converted to MFCC feature vectors.

2.1. Linear prediction

Linear prediction (LP) [8] in speech signal processing is based on assumption that speech samples $\hat{x}_n$ can be predicted as a linear combination of p previous samples $\hat{x}_n = \sum_{k=1}^{p} a_k x_{n-k}$, where x_n are the samples of the speech signal in a given frame and a_k are the prediction coefficients. The number of prediction coefficients p is the order of the LP model. In this work, $p = 20$ is used for all three linear prediction methods. The prediction error is defined as $e_n = x_n - \hat{x}_n$ and in LP analysis, the energy of the prediction error signal is minimized by setting the partial derivatives of $E_{LP} = \sum_n e_n^2 = \sum_n (x_n - \sum_{k=1}^{p} a_k x_{n-k})^2$ with respect to each coefficient a_k to zero, resulting in normal equations

$$\sum_{k=1}^{p} a_k \sum_n x_{n-k} x_{n-j} = \sum_n x_n x_{n-j}, \qquad 1 \leq j \leq p. \quad (1)$$

The autocorrelation method [8], which guarantees the stability of the LP synthesis model, is used in this work for solving the normal equations.

Although LP has been shown in [5] and [9] to improve the noise robustness of an ASR system when a mismatch occurs between training and evaluation environments, the noise robustness of LP-based MFCC feature extraction can be further increased by temporal weighting.

2.2. Weighted linear prediction

Weighted linear prediction (WLP) [4] extends the standard LP by adding temporal weighting to the squared residual in model coefficient optimization. In WLP, the prediction coefficients b_k are solved by minimizing the energy of prediction error signal using

$$E_{WLP} = \sum_n e_n^2 W_n = \sum_n (x_n - \sum_{k=1}^{p} b_k x_{n-k})^2 W_n, \quad (2)$$

where W_n is the weighting function. In WLP analysis, the model is computed by solving normal equations

$$\sum_{k=1}^{p} b_k \sum_n W_n x_{n-k} x_{n-i} = \sum_n W_n x_n x_{n-i}, 1 \le i \le p. \quad (3)$$

Standard LP can be seen as a special case of WLP by denoting $W_n = d$ for all n, where $d \neq 0$. In WLP, the weighting function W_n is typically chosen as the short-time energy (STE) of the local signal

$$W_n = \sum_{i=1}^{M} x_{n-i}^2, \quad (4)$$

where M has previously been chosen close or equal to the value of p [9]. Even though WLP does not guarantee a stable synthesis model even if the autocorrelation method is used, WLP has been shown to outperform the stabilized version of WLP in a LVCSR testing scheme [5].

2.3. Extended weighted linear prediction

Extended weighted linear prediction (XLP) [7] further generalizes the LP and WLP analyses by enabling two-dimensional temporal weighting. That is, each lagged sample at each time instant is weighted separately. In XLP, the prediction error energy is expressed as follows

$$E_{XLP} = \sum_n (x_n Z_{n,0} - \sum_{k=1}^{p} c_k x_{n-k} Z_{n,k})^2. \quad (5)$$

LP can be seen as a special case when $Z_{n,j} = d$, with $d \neq 0$, for all n and j, and Eq. 5 is reduced to WLP prediction error when $Z_{n,j} = \sqrt{W_n}$.

Minimizing the error energy in Eq. 5 leads to solving the following XLP normal equations

$$\sum_{k=1}^{p} c_k \sum_n Z_{n,k} x_{n-k} Z_{n,j} x_{n-j} = \sum_n Z_{n,0} x_n Z_{n,j} x_{n-j}, \quad (6)$$
$$1 \le j \le p.$$

The optimal c_k values from Eq. 6 yield the inverse filter of the XLP analysis as follows

$$C(z) = 1 - \sum_{k=1}^{p} c_k z^{-k}. \quad (7)$$

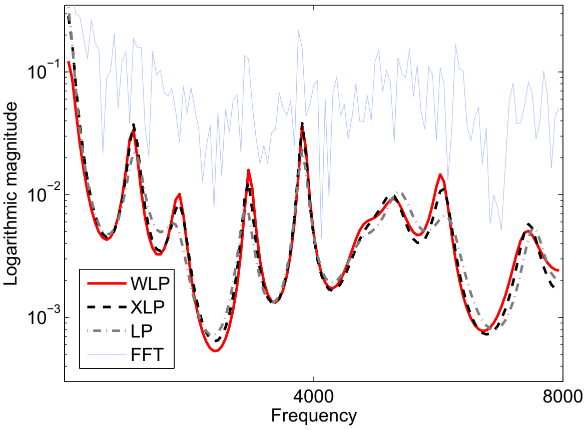

Figure 1: *Spectra of transition from /t/ to /ä/ in noisy conditions. Prediction order* $p = 20$ *is used.*

In this work, the following recursive equation, referred to as absolute value sum (AVS), is used to compute the weights

$$Z_{n,j} = \frac{m-1}{m} Z_{n-1,j} + \frac{1}{m}(|x_n| + |x_{n-j}|), \quad (8)$$

where $Z_{n,j} = 0$ for all j before the beginning of the frame and m is a parameter controlling the effective length of the moving average memory. Here, $m = p$ (the number of prediction coefficients) has been used. This weighting emphasizes the predictions of high amplitude signal samples and within each prediction, the lags for which the lagged signal samples has a large amplitude are also emphasized. The assumption behind this weighting is the same as in the STE weighting of WLP; high amplitude samples are more likely to contain smaller relative amounts of corruption than low amplitude samples. Similarly to WLP, XLP is not guaranteed to produce a stable synthesis filter.

3. Evaluation

3.1. Test setup

The basis of all systems used in this work is a large vocabulary continuous speech recognizer [10] based on hidden Markov models (HMM) with state likelihoods modeled by Gaussian mixture models (GMM). The acoustic models are state-tied triphones constructed with a decision-tree method. Each state is modeled with a maximum of 100 Gaussians and the states are associated with gamma probability functions to model the state durations [11]. The speech signal is represented with a power and 12 MFCC features concatenated with their first and second order differentials. Cepstral mean subtraction is applied before scaling and mapping with maximum likelihood linear transformation (MLLT) [12] optimized in training. Finally, the covariance matrix of each Gaussian is diagonalized.

The system uses a morph-based variable length n-gram language model trained on 150 million word corpus containing text from books, magazines and newspaper data. The decoder, utilizing a beam-pruned Viterbi token-pass system, combines the language and acoustic models using scaling factor on the language model log-probability. The scaling factor is optimized independently for each microphone position and noise type of each method with respect to the letter error rate (LER).

Table 1: *Public place noise evaluation set error rates (LER/WER %).*

Mic	FFT	LP	WLP	XLP
Close	3.3/13.9	4.6/17.7	4.6/16.8	4.5/17.4
Mid	23.9/46.7	19.6/43.8	19.9/44.5	18.8/40.7
Far	40.8/65.8	34.0/64.5	34.3/58.8	32.5/57.8

3.2. Data

Data used in this work was taken from SPEECON [13] Finnish language corpus which contains both spontaneous and read speech. Three speaker exclusive sets were constructed from the database. A training set, comprising 293 speakers and containing approximately 19.5 hours of speech, was used to train all the systems. The training set contained speech recorded in virtually noiseless environments (estimated average SNR was 26 dB) and it was the only set containing spontaneous speech.

System parameters were optimized using development sets containing noisy data from public and car environments. The noisy car development set consisted of approximately 29 minutes of speech from 20 speakers, and the public place development set consisted of 60 minutes of speech from 30 speakers. The final system evaluations were executed using similar noisy data as used in the development sets. The number of speakers in evaluation sets were the same as in the development sets but consisting of approximately 57 and 94 minutes of speech, respectively.

The development and evaluation data sets contained recordings from three microphone distances. The closest microphone (*close*) had an estimated average SNR of 13 dB in the car and 24 dB in public place recordings. The second closest microphone (*mid*) had an estimated average SNR of 5 dB in the car and 14 dB in public place recordings. The farthest microphone (*far*) in the car recordings was located approximately one meter away from the speaker and had an estimated average SNR of 8 dB. The farthest microphone in the public place recordings was placed 0.5–1 meter away from the speaker and had an estimated average SNR of 9 dB. The inconsistency between the SNRs of the mid and far distance microphones in the car recordings is mainly caused by the variant bandwidth characteristics of the respective microphones.

3.3. Parameter optimizations

Although the WLP and XLP systems were trained and evaluated using a fixed STE and AVS window widths $M = m = p = 20$, optimization of the M and m parameters would improve the recognition performance of those systems since the short-time spectral approximation is dependent on those parameters. Different noise conditions are likely to have individual optimum m and M values. This has been investigated in [9] where theoretical lower letter error rate bounds are computed for the similar development sets as used in this work.

Table 2: *Z-scores of pairwise comparisons on public place noise. Bold numbers indicate a statistically significant difference between the systems.*

Pair	Close	Mid	Far
LP-FFT	**-4.33**	**-2.70**	**-3.79**
LP-WLP	-0.55	-1.20	-0.44
LP-XLP	-1.41	**-2.75**	**-2.21**
WLP-FFT	**-4.12**	**-2.54**	**-4.10**
WLP-XLP	0.16	**-3.59**	**-3.49**
XLP-FFT	**-4.68**	**-3.34**	**-4.56**

Table 3: *Car noise evaluation set error rates (LER/WER %).*

Mic	FFT	LP	WLP	XLP
Close	4.0/14.9	5.9/19.3	6.2/20.4	5.9/19.1
Mid	29.8/54.6	36.8/76.5	32.7/62.7	33.4/62.5
Far	66.6/92.9	60.0/95.6	60.1/92.3	62.7/93.7

4. Results

The results of public place noise evaluation set are gathered in Table 1 and the respective Z-scores indicating statistical significances between the letter error rates (LER) of the systems are shown in Table 2. The scores are computed using Wilcoxon signed rank test. An absolute Z-score value exceeding 1.96 is considered significant with a 95 % confidence level. Word error rates (WER) are also shown in Tables 1 and 3, but the actual system ranking is based on letter error rates, which are more realistic peformance measures than word error rates due to the long inflected words in Finnish language. The differences between systems are statistically significant unless stated otherwise.

On public place noise, FFT performs best on close microphone recordings achieving 3.3 % LER, but its performance deteriorates most on mid and far channels with respective LERs of 23.9 % and 40.8 %. On close microphone, the performance of linear predictive systems are close to each other and the differences between systems are statistically insignificant: LP and WLP achieve 4.6 %, and XLP 4.5 % LER. On mid and far recordings, the lowest LERs are obtained by XLP with respective 18.8 % and 32.5 % LERs whereas LP achieves 19.6 % and 34.0 %, and WLP 19.9 % and 34.3 %, respectively. The differences between LP and WLP on mid and far microphones are not statistically significant.

The results of car noise evaluation set are gathered in Table 3 and the respective Z-scores are shown in Table 4. Similarly to the public place noise, the lowest LER on close microphone car noise is obtained by the FFT system (4.0 %) whereas the performance of LP, WLP and XLP are close to each other with respective LERs of 5.9%, 6.2 % and 5.9 %. There is no statistical difference between the linear predictive systems on close microphone data. On mid microphone recordings, LP has the highest error rate of all the systems with LER of 36.8 % whereas FFT has the lowest LER of 29.8 %. WLP and XLP are placed between the two previous systems with respective LERs of 32.7 % and 33.4 %. However, the differences between LP, WLP and XLP on mid microphone are not statistically significant. On far microphone recordings, the lowest LER, 60.0 %, is achieved by the LP system whereas the highest LER, 66.6 %, is obtained by FFT. WLP and XLP are again placed between the two with respective LERs of 60.1 % and 62.7 %. The differences between FFT and XLP, and LP and WLP are not statistically significant on far microphone data.

Table 4: *Z-scores of pairwise comparisons on car noise. Bold numbers indicate a statistically significant difference between the systems.*

Pair	Close	Mid	Far
LP-FFT	**-3.92**	**-2.35**	**-2.69**
LP-WLP	-0.78	**-2.91**	-0.93
LP-XLP	-0.07	**-3.85**	**-3.70**
WLP-FFT	**-3.92**	-1.42	**-3.14**
WLP-XLP	1.64	-1.05	**-2.65**
XLP-FFT	**-3.81**	-1.61	-1.83

5. Conclusions

In this work, FFT and three linear predictive methods (standard LP and temporally weighted methods WLP and XLP) are evaluated as the short-time spectrum analyzer in the MFCC feature extraction process. The linear predictive methods evaluated here are well suited to alleviate the performance degradation due to the mismatch between the training and recognition noise environments, but if no mismatch occurs, the conventional FFT based MFCCs have been shown in [9] to provide the highest performance compared to linear predictive systems. Therefore, only mismatched conditions are evaluated in this work.

The FFT based spectral analysis achieves the lowest letter error rates on close microphone recordings on both the car and public place noises whereas all the three linear prediction systems perform almost identically and slightly inferior to FFT. The lowest error rates on mid and far recordings of public place noise are obtained by the XLP system. On mid recorded car noise, XLP shares the top ranking with FFT and WLP. No direct conclusions can be made from the car noise data since the set is much smaller than the public place set and the speaker dependent error rates are more variant. On an absolute scale, not one of the systems achieve good results on mid and far car noise recordings, which has also been noted in previous studies [9] [14]. The LP and WLP systems have been shown in [5] to achieve as good results as FFT on close microphone recordings. Here, the language model is more compact than in [5] and [9] and thus the respective results are not directly comparable.

WLP and XLP offer two paths to further improve the performance. Firstly, the values of the parameters M and m controlling the STE and AVS window widths during training and evaluation could be optimized in some manner according to the changing noise conditions. Secondly, the STE and AVS weighting schemes in WLP and XLP, respectively, can be replaced with new, potentially more robust schemes, possibly with feedback from the current noise environment. In this respect, the XLP method, by virtue of using two-dimensional weighting, offers considerable freedom and prospects for further improvement. Nevertheless, with a basic weighting scheme and despite the lack of weighting window width optimization, XLP analysis appears to be the most robust method on average.

6. Acknowledgements

This work was supported by the Academy of Finland in projects numbered 135003 (LASTU research programme), 140969 and 127345, and by Langnet Graduate School.

7. References

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

[2] Hermansky, H., “Perceptual Linear Predictive (PLP) Analysis of Speech,”, JASA, 87(4):1738–1752, 1990.

[3] Kim, C. and Stern, R. M., “Feature Extraction for Robust Speech Recognition using a Power-Law Nonlinearity and Power-Bias Subtraction”, in Proc. Interspeech, Brighton, UK, 2009.

[4] Ma, C., Kamp, Y. and Willems, L, “Robust Signal selection for linear prediction analysis of voiced speech”, Speech Communication, 12(2):69–81, 1993.

[5] Pohjalainen, J., Kallasjoki, H., Palomäki, K. J., Kurimo, M. and Alku, P., “Weighted Linear Prediction for Speech Analysis in Noisy Conditions”, in Proc. Interspeech, Brighton, UK, 2009.

[6] Saeidi, R., Pohjalainen, J., Kinnunen, T. and Alku, P., ”Temporally Weighted Linear Prediction Features for Tackling Additive Noise in Speaker Verification”, IEEE Signal Processing Letters, 17(6):599–602, 2010.

[7] Pohjalainen, J., Saeidi, R., Kinnunen, T. and Alku, P., “Extended Weighted Linear Prediction (XLP) Analysis of Speech and its Application to Speaker Verification in Adverse Conditions”, in Proc. Interspeech, Makuhari, Japan, 2010.

[8] Makhoul, J., “Linear Prediction: a Tutorial Review”, in Proc. IEEE, 63(4):561–580, 1975.

[9] Kallasjoki, H., Palomäki, K., Magi, C., Alku, P. and Kurimo, M., “Noise Robust LVCSR Feature Extraction Based on Stabilized Weighted Linear Prediction,” in Proc. SPECOM, St. Petersburg, Russia, 2009.

[10] Hirsimäki, T., Pylkkönen, J. and Kurimo, M., “Importance of High-Order N-Gram Models in Morph-Based Speech Recognition,” in IEEE Trans. ASLP, 17(4):724–732, 2009.

[11] Pylkkönen, J. and Kurimo, M., “Duration Modeling Techniques for Continuous Speech Recognition”, in Proc. Interspeech, Jeju, Korea, 2004.

[12] Gales, M. J., “Semi-tied Covariance Matrices for Hidden Markov Models”, IEEE Trans. SAP, 7:272–281, 1999.

[13] Iskra, D., Grosskopf, B., Marasek, K., van den Heuvel, H. and Kiessling, A., “SPEECON - Speech Databases for Consumer Devices: Database Specification and Validation”, in Proc. LREC, 329–333, 2002.

[14] Keronen, S., Remes, U., Palomäki, K., Virtanen, T. and Kurimo, M., “Comparison of Noise Robust Methods in Large Vocabulary Speech Recognition,” in Proc. EUSIPCO, Aalborg, Denmark, 2010.