Extended Weighted Linear Prediction (XLP) Analysis of Speech and its Application to Speaker Verification in Adverse Conditions

Jouni Pohjalainen[1], *Rahim Saeidi*[2], *Tomi Kinnunen*[2], *Paavo Alku*[1]

[1]Department of Signal Processing and Acoustics, Aalto University
School of Science and Technology, Finland
[2]School of Computing, University of Eastern Finland, Finland

jpohjala@acoustics.hut.fi, rahim.saeidi@uef.fi, tomi.kinnunen@uef.fi, paavo.alku@tkk.fi

Abstract

This paper introduces a generalized formulation of linear prediction (LP), including both conventional and temporally weighted LP analysis methods as special cases. The temporally weighted methods have recently been successfully applied to noise robust spectrum analysis in speech and speaker recognition applications. In comparison to those earlier methods, the new generalized approach allows more versatility in weighting different parts of the data in the LP analysis. Two such weighted methods are evaluated and compared to the conventional spectrum modeling methods FFT and LP, as well as the temporally weighted methods WLP and SWLP, by substituting each of them in turn as the spectrum estimation method of the MFCC feature extraction stage of a GMM-UBM based speaker verification system. The new methods are shown to lead to performance improvement in several cases involving channel distortion and additive noise mismatch between the training and recognition conditions.

Index Terms: linear prediction, speaker verification, mel frequency cepstral coefficients

1. Introduction

Modeling of the short-time magnitude spectrum is a central task in speech and audio signal processing. The two most common methods of spectrum analysis are the discrete Fourier transform implemented as the fast Fourier transform (FFT) and linear prediction (LP). Among other fields, they are used in feature extraction for speech and speaker recognition, in which statistical recognizers are first trained to represent spectral features of the training utterances and subsequently used to recognize other utterances. The conventional spectrum analysis methods used in speech and speaker recognition are known to be sensitive to transmission channel distortions and additive noise. An additional difficulty specific to recognition is *mismatch*, which occurs when the channel and/or environmental noise conditions during recognition differ from those of the training material.

In this paper, we propose a new type of robust all-pole model and apply it to text-independent speaker verification [3]. In speaker verification, robustness with respect to noises and mismatch has traditionally been pursued via feature normalization (e.g. cepstral mean and variance normalization, RASTA filtering, feature warping [9]), speaker model compensation [10] and score normalization [2]. Relatively little effort, however, has been put on making the spectrum estimation itself robust. Typically, these approaches have been based on different variants of LP analysis, e.g. [1] [14].

Temporally weighted LP methods aim to increase the contribution of such samples in the LP analysis that have been less corrupted by distortion and noise. Weighted linear prediction (WLP) [5] and its stabilized version (SWLP) [6], complemented with short-time-energy (STE) weighting, have been successfully used to alleviate the problem of noise in isolated word recognition [6], continuous speech recognition [11] and, most recently, speaker verification [14]. In these studies, the methods have been substituted, in place of the FFT, as the spectrum estimation part of the popular mel frequency cepstral coefficient (MFCC) front-end. They have been shown to improve the robustness compared to baseline FFT-based MFCC features in noisy conditions.

The present study introduces two novel methods related to the previously known temporally weighted LP methods: eXtended weighted Linear Prediction (XLP) and its stabilized version (SXLP). We evaluate these new methods in a speaker verification task, which involves signal corruption due to both channel distortion and additive noise, and compare their performance with that of FFT, LP, WLP and SWLP. The speaker verification system is based on adapted Gaussian mixtures [12] and standard MFCC feature extraction.

2. Linear Predictive Models

2.1. Linear Prediction (LP)

Linear predictive speech spectrum modeling [7] assumes that each speech sample can be predicted as a linear combination of p previous samples, $\hat{s}_n = \sum_{k=1}^{p} a_k s_{n-k}$, where s_n are the samples of the speech signal in a given short-term frame and $\{a_k\}$ are the predictor coefficients. The number of predictor coefficients p is the *order* of linear prediction. The prediction error is denoted as $e_n = s_n - \hat{s}_n = s_n - \sum_{k=1}^{p} a_k s_{n-k}$. Conventional LP analysis minimizes the energy of the prediction error signal $E_{\mathrm{LP}} = \sum_n e_n^2 = \sum_n (s_n - \sum_{k=1}^{p} a_k s_{n-k})^2$ by setting the partial derivatives of E_{LP} with respect to each coefficient a_k to zero. This results in the normal equations [7] $\sum_{k=1}^{p} a_k \sum_n s_{n-k} s_{n-j} = \sum_n s_n s_{n-j}$, $1 \leq j \leq p$. Although not explicitly written, the range of summation of n is chosen in this work to correspond to the *autocorrelation method*, in which the energy is minimized over a theoretically infinite interval, but s_n is considered to be zero outside the actual analysis window [7]. An important benefit of the autocorrelation method is that the LP synthesis model $H(z) = 1/(1 - \sum_{k=1}^{p} a_k z^{-k})$ is guaranteed to be stable, i.e., the roots of the denominator polynomial are guaranteed to lie inside the unit circle [7].

2.2. Weighted Linear Prediction (WLP)

Weighted linear prediction (WLP) [5] is a generalization of LP analysis. In contrast to conventional LP, WLP introduces a temporal weighting of the squared residual in model coefficient optimization. Specifically, in WLP, the predictor coefficients $\{b_k\}$ are solved by minimizing the energy

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

where W_n is the weighting function. The weighting can be used to emphasize the importance of the prediction error in the temporal regions assumed to be less affected by noise, and de-emphasize the importance of the noisy regions. The WLP model is obtained by solving the normal equations

$$\sum_{k=1}^{p} b_k \sum_n W_n s_{n-k} s_{n-i} = \sum_n W_n s_n s_{n-i}, \quad 1 \le i \le p. \quad (2)$$

It is easy to show that conventional LP can be obtained as a special case of WLP: by setting $W_n = d$ for all n, where $d \neq 0$, d becomes a multiplier of both sides of (2) and cancels out, leaving the LP normal equations. Typically, the weighting function W_n in WLP is chosen as the short-time energy (STE) of the immediate signal history [5] [6] [11] [14]: $W_n = \sum_{i=1}^{M} s_{n-i}^2$, where M has previously been chosen close to or equal to the value of p [11] [14]. When compared to conventional spectral modeling methods such as FFT and LP, WLP using STE weighting has been recently shown to improve robustness with respect to additive noise in the feature extraction stages of both large vocabulary continuous speech recognition [11] and speaker verification [14].

2.3. The Proposed XLP Method

The present paper introduces a further generalization of the WLP analysis. In this formulation, the prediction error energy is expressed as follows:

$$E_{\mathrm{XLP}} = \sum_n (s_n Z_{n,0} - \sum_{k=1}^{p} c_k s_{n-k} Z_{n,k})^2. \quad (3)$$

WLP is obtained as a special case when $Z_{n,j} = \sqrt{W_n}$ and LP is obtained when $Z_{n,j} = d$, with $d \neq 0$, for all n and j. However, if $Z_{n,i} = Z_{n,j}$ does not hold for all n, i and j, the result is a novel LP analysis method, in which each lagged sample at each time instant is weighted separately. In other words, the new formulation allows temporal weighting on a finer time scale than WLP. This method is referred to as *eXtended weighted Linear Prediction* (XLP).

The minimization of the error energy in Eq. 3 gives rise to the XLP normal equations

$$\sum_{k=1}^{p} c_k \sum_n Z_{n,k} s_{n-k} Z_{n,j} s_{n-j} = \sum_n Z_{n,0} s_n Z_{n,j} s_{n-j}, \quad 1 \le j \le p. \quad (4)$$

The optimal c_k values from this equation yield the inverse filter of the XLP analysis as follows:

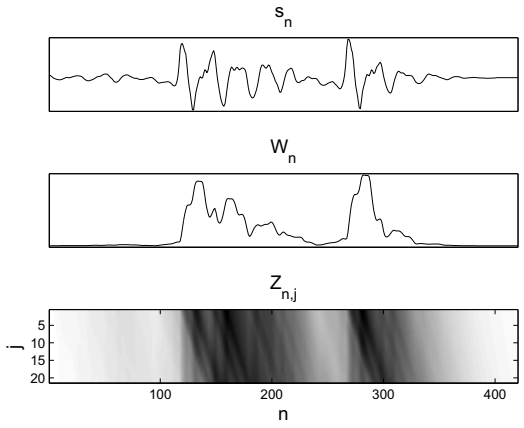

Figure 1: *Upper panel: One frame of a 16 kHz male vowel /a/. Middle panel: The corresponding STE weight function used with WLP. Lower panel: The corresponding two-dimensional AVS weight function used with XLP.*

$$A(z) = 1 - \sum_{k=1}^{p} c_k z^{-k}. \quad (5)$$

In the present study, the following recursion was used to compute the weights:

$$Z_{n,j} = \frac{m-1}{m} Z_{n-1,j} + \frac{1}{m}(|s_n| + |s_{n-j}|), \quad (6)$$

with $Z_{n,j} = 0$ for all j before the beginning of the frame. For the parameter m, which controls the effective length of the moving average memory, $m = p$ has been used. This weighting, referred to as absolute value sum (AVS), emphasizes the predictions of prominent signal samples, and within each prediction, it emphasizes those lags for which the lagged signal sample has a large amplitude. The underlying rationale for this is the same as that for STE weighting in WLP: higher-amplitude samples are arguably likely to contain smaller relative amounts of corruption (such as additive noise) than lower-amplitude samples.

Figure 1 shows one vowel frame sampled at 16 kHz, its corresponding STE weight with $M = 20$, and the two-dimensional AVS weighting matrix computed according to Eq. 6 with $m = 20$. Figure 2 shows the spectra for the same frame computed using four methods: FFT, LP, WLP (using the STE weights) and XLP (using the AVS weights). Order $p = 20$ has been used for the linear predictive methods.

2.4. Stabilized Methods

WLP is not guaranteed to produce a stable all-pole synthesis model $1/(1 - \sum_{k=1}^{p} b_k z^{-k})$ (even when using the autocorrelation method, which in conventional LP always gives a stable model). As a remedy, a stabilized version of WLP, called SWLP, was developed in [6]. Although SWLP is stabilized mainly for synthesis purposes, it has been found, like WLP, to be a robust method in the feature extraction stages of speech recognition [6] [11] and speaker verification [14] — even surpassing WLP in performance in the latter application. As stated in Section 2.3, the WLP normal equations (Eq. 2) can be rewritten as

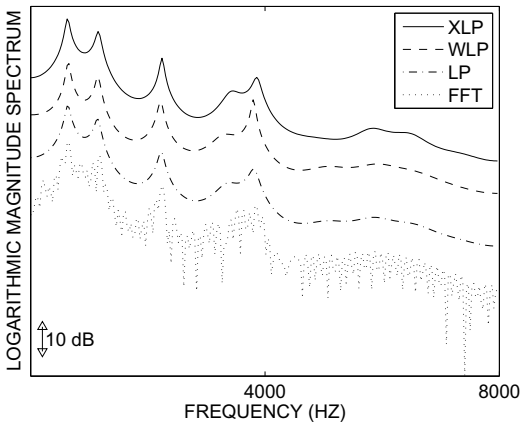

Figure 2: *Spectra of the vowel /a/. LP, WLP and XLP use prediction order* $p = 20$.

$$\sum_{k=1}^{p} b_k \sum_n Z_{n,k} s_{n-k} Z_{n,j} s_{n-j} = \sum_n Z_{n,0} s_n Z_{n,j} s_{n-j}, \quad (7)$$
$$1 \le j \le p,$$

where $Z_{n,j} = \sqrt{W_n}$ for $0 \le j \le p$. As shown in [6] (using a matrix-based formulation), model stability is guaranteed if the weights $Z_{n,j}$ are, instead, defined recursively as $Z_{n,0} = \sqrt{W_n}$ and $Z_{n,j} = \max(1, \frac{\sqrt{W_n}}{\sqrt{W_{n-1}}}) Z_{n-1,j-1}$, $1 \le j \le p$. Substitution of these values in Eq. 7 gives the SWLP normal equations. A similar approach can be utilized for XLP: once the weights $Z_{n,j}$ have been determined, they are replaced with $Z'_{n,j} = \max(Z_{n,j}, Z_{n-1,j-1})$ with $Z_{n,j} = 0$ for $j < 0$. The resulting analysis method will be denoted as stabilized XLP (SXLP).

3. Application to Speaker Verification

3.1. Test setup

Six different methods (FFT, LP, WLP, SWLP, XLP and SXLP) were evaluated by substituting each one as the spectrum estimation technique in the MFCC feature extraction module for speaker verification. The effectiveness of the features was evaluated on the NIST 2002 speaker recognition evaluation (SRE) corpus, which consists of speech samples transmitted over different cellular networks with varying types of handsets. There are 2982 genuine and 36,277 impostor test trials in the NIST 2002 corpus. For each of the 330 target speakers, two minutes of untranscribed, conversational speech is available to train the target speaker model. The duration of the test utterances varies between 15 and 45 seconds.

The experiments were conducted using a standard Gaussian mixture model classifier with a universal background model (GMM-UBM) [12]. Test normalization (T-norm) [2] was applied on the logarithmic likelihood ratio scores.

The (gender-dependent) background models and cohort models for T-norm, having 1024 Gaussians, were trained using the NIST 2001 corpus. This baseline system [13] has comparable or better accuracy than other systems evaluated on this corpus (e.g. [4]).

Features were extracted every 15 ms from 30 ms frames multiplied by a Hamming window. Depending on the feature extraction method, the magnitude spectrum was computed differently. For the baseline method, the FFT of the windowed frame was directly computed. For LP, WLP, XLP, SWLP and SXLP, the model coefficients and the corresponding all-pole spectra were first derived as explained in Section 2. All the five parametric methods used a predictor order of $p = 20$. For WLP and SWLP, STE weighting was used with the energy window duration set to $M = 20$ samples. For XLP and SXLP, AVS weighting was used with averaging parameter $m = 20$. A 27-channel mel-frequency filterbank was used to extract 12 MFCCs. After RASTA filtering, Δ and Δ^2 coefficients were appended. Speech frames were then selected using an energy-based voice activity detector (VAD). Finally, cepstral mean and variance normalization (CMVN) was performed.

Two standard metrics were used to assess recognition accuracy: the equal error rate (EER) and the minimum detection cost function value (MinDCF). EER corresponds to the threshold at which the miss rate (P_{miss}) and false alarm rate (P_{fa}) are equal; MinDCF is the minimum value of a weighted cost function given by $0.1 \times P_{\mathrm{miss}} + 0.99 \times P_{\mathrm{fa}}$. All the reported MinDCF values were multiplied by 10, for ease of comparison.

In terms of robustness issues, the original NIST samples contain training/recognition condition mismatch mostly due to transmission channel and handset variation, and possibly small amounts of additive noise captured by the handsets. To further study robustness with respect to additive noise, some noise was digitally added from the NOISEX-92 database [8] to the speech samples. This study used the *factory2* noise. The background models and target speaker models were trained with clean data, but the noise was added to the test files with a given average segmental (frame-average) signal-to-noise ratio (SNR). Five values were considered: $\mathrm{SNR} \in \{\mathrm{original}, 20, 10, 0, -10\}$ dB, where "original" refers to the original NIST samples (which contain different types of channel distortion as well as possibly already some background noise). In summary, the evaluation data used in the present study contained linear and nonlinear distortion present in the sounds of the NIST 2002 database as well as additive noise taken from the NOISEX-92 database.

3.2. Results

Table 1 shows the speaker verification results. According to the EER measure, both XLP and SXLP are consistently better than the FFT baseline in every mismatch case. Considering both measures of performance, SXLP is the overall best performing method in the "original" case, where the primary source of mismatch is channel variation, as well as in the moderately noisy cases (added noise SNR levels 20 and 10 dB). It is closely followed by XLP in overall performance in these cases, these two being the two best performing methods in two cases according to the MinDCF criterion and one case according to the EER criterion. When noise corruption is further increased (added noise SNR levels 0 and -10 dB), SWLP is still the most robust method, as in an earlier study [14]. In these cases, however, the system performance has severely deteriorated with each method, EERs at 0 dB SNR being roughly almost twice those of the "original" case.

The DET plot in Fig. 3, for the SNR level 0 dB, indicates that the weighted LP models outperform the baseline FFT and LP models by a wide margin. The two overall best methods at this noise level are the stabilized versions, SWLP and SXLP.

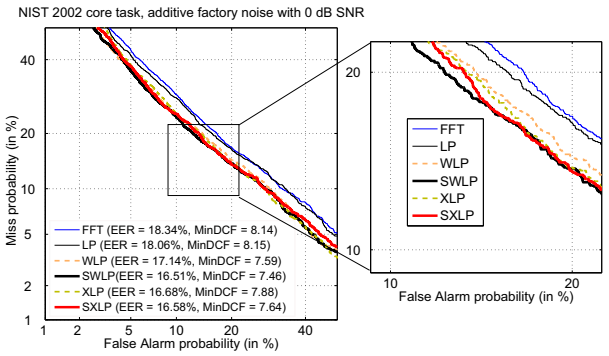

Figure 3: *A detection error tradeoff (DET) curve plot for the case with added factory noise SNR level 0 dB.*

Table 1: *System performance in different scenarios according to two measures of performance.*

Signal-to-added-noise ratio (dB)	Equal error rate (EER %)					
	FFT	LP	WLP	SWLP	XLP	SXLP
original	9.22	8.89	9.15	9.15	9.14	8.80
20	9.71	9.26	9.55	9.20	9.37	9.42
10	10.05	10.29	10.17	9.99	9.89	9.84
0	18.34	18.06	17.14	16.51	16.68	16.58
-10	27.25	26.90	25.54	24.70	25.20	25.94
Signal-to-added-noise-ratio (dB)	MinDCF					
	FFT	LP	WLP	SWLP	XLP	SXLP
original	3.56	3.47	3.50	3.54	3.35	3.36
20	3.69	3.64	3.65	3.63	3.62	3.61
10	4.09	4.25	4.12	4.15	4.25	4.12
0	8.14	8.15	7.59	7.46	7.88	7.64
-10	9.99	9.99	10.00	10.00	9.98	9.99

4. Conclusions

Novel weighted linear predictive methods XLP and SXLP, which generalize earlier temporally weighted methods WLP and SWLP, respectively, were evaluated in MFCC feature extraction for speaker verification. The new methods are based on imposing time-domain weighting separately on each lagged signal sample in the prediction model and then optimizing the prediction coefficients according to the least squares criterion.

The new variants were compared with the conventional methods FFT and LP, as well as with WLP and SWLP. XLP and SXLP used the novel AVS weighting scheme, while WLP and SWLP used their usual STE weighting scheme. According to the evaluation, SXLP was the best performing method for the cases with channel distortion mismatch and low to moderate amount of additive noise. XLP came close to its performance in these cases. Both the XLP and SXLP methods improved upon the FFT baseline in each evaluated mismatch scenario in terms of the EER measure.

In summary, while the previously proposed SWLP method showed the best overall performance for speech utterances heavily corrupted by additive noise, XLP and SXLP were, in general, the best methods in the cases where additive noise corruption was slight and channel variation was the primary source of mismatch. Interesting research topics for future include the search for new two-dimensional weighting schemes for XLP and SXLP, as well as the application of these methods to new tasks.

5. Acknowledgements

The work of J. Pohjalainen was supported by Academy of Finland project no 127345. The work of R. Saeidi was supported by a scholarship from the Finnish Foundation for Technology Promotion (TES). The work of T. Kinnunen was supported by Academy of Finland project no 132129, "Characterizing individual information in speech".

6. References

[1] Assaleh, K.T. and Mammone, R.J., "New LP-Derived Features for Speaker Identification", IEEE Trans. Speech and Audio Processing, 2(4):630–638, 1994.

[2] Auckenthaler, R., Carey, M. and Lloyd-Thomas, H., "Score Normalization for Text-Independent Speaker Verification Systems", Digital Signal Processing, 10(1-3):42-54, 2000.

[3] Kinnunen, T. and Li, H., "An Overview of Text-Independent Speaker Recognition: from Features to Supervectors", Speech Communication, 52(1):12–40, 2010.

[4] Longworth, C. and Gales, M. J. F., "Combining Derivative and Parametric Kernels for Speaker Verification", IEEE Trans. Audio, Speech and Language Processing, 17(4):748–757, 2009.

[5] Ma, C., Kamp, Y. and Willems, L. F., "Robust signal selection for linear prediction analysis of voiced speech", Speech Communication, 12(2):69–81, 1993.

[6] Magi, C., Pohjalainen, J., Bäckström, T. and Alku, P.,"Stabilised weighted linear prediction",Speech Communication, 51(5):401–411, 2009.

[7] Makhoul, J., "Linear prediction: a tutorial review", Proceedings of the IEEE, 63(4):561–580, 1975.

[8] NOISEX-92 database samples available online: http://spib.rice.edu/spib/select_noise.html, accessed on 29 Apr 2010.

[9] Pelecanos, J. and Sridharan, S., "Feature Warping for Robust Speaker Verification", in Proc. Speaker Odyssey, Crete, Greece, 2001.

[10] Pillay, S.G., Ariyaeeinia, A., Pawlewski, M. and Sivakumaran, P., "Speaker Verification under Mismatched Data Conditions", Signal Processing, 3(4):236–246, 2009.

[11] 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.

[12] Reynolds, D.A., Quatieri, T.F. and Dunn, R.B., "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, 10(1):19–41, 2000.

[13] Saeidi, R., Mohammadi, H. R. S., Ganchev, T. and Rodman, R. D., "Particle swarm optimization for sorted adapted Gaussian mixture models", IEEE Trans. Audio, Speech and Language Processing, 17(2):344–353, 2009.

[14] 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), 2010.