ROBUST SPEECH ANALYSIS BY LAG-WEIGHTED LINEAR PREDICTION

Jouni Pohjalainen and Paavo Alku

Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland

ABSTRACT

This study introduces an approach for linear predictive spectrum analysis based on emphasizing selected time-domain properties in the analyzed signal in combination with a stabilization operation. A stable weighted linear predictive method based on a novel autocorrelation-based weighting scheme is described and its spectral properties are demonstrated. The robustness of the proposed method is compared with conventional techniques in terms of an Euclidean MFCC distortion measure in different additive noise conditions. In the experimental evaluation, the novel speech analysis technique outperforms the other evaluated methods.

Index Terms— linear prediction, spectrum analysis

1. INTRODUCTION

Short-time spectrum modeling is a central task in speech and audio processing applications. In automatic recognition applications, both in speech and speaker recognition, the short-time magnitude spectrum is used as a basis for the most prevalent feature representations, such as mel frequency cepstral coefficients (MFCCs) [1]. Many studies related to recognition tasks have addressed the usefulness of linear predictive models, e.g., [3] [2] [4]. These models depict the magnitude spectrum envelope of speech and particularly its local peaks, the formants, which arguably comprise the primary discriminative information in most speech-related recognition problems.

Recently, temporally weighted linear prediction [5] with its many variants has been applied (by the present authors) to text independent speaker verification [4] [6] and large vocabulary continuous speech recognition [2] [7] [8] in mismatched recognition conditions. In weighted linear prediction, a temporal weighting function is utilized in the filter optimization in order to emphasize the contribution of those speech samples that contain the most relevant information from the point of view of the underlying modeling problem. The exact nature of temporal weighting depends on the chosen weighting scheme. For example, weighting can be applied to focus on the glottal closed phase, thus obtaining linear predictive models with more prominent formant structures. In the mentioned studies, the linear predictive methods in general have outperformed basic Fourier spectrum modeling, while the temporally weighted methods have succeeded to further improve the robustness with respect to mismatch due to additive noise corruption and channel variation.

The present study proposes a new weighted linear predictive spectrum analysis method and, using an objective distortion measure, compares it against previous methods in terms of robustness against additive noise. The new method is based on our earlier work [6] but it introduces a novel efficient weighting scheme and is guaranteed to produce stable all-pole models. The stability property makes it suitable also for coding and synthesis applications. From a more general viewpoint, the present study highlights an approach for creating new, stable lag-weighted linear predictive analysis methods with arbitrary focus on the underlying information.

This work is supported by Academy of Finland project 127345.

2. LINEAR PREDICTIVE SPECTRUM ESTIMATION

2.1. Linear prediction (LP)

Linear prediction (LP) is a well-known method for modeling the short-time spectrum envelope of speech and audio signals by an all-pole model $H(z) = 1/(1 - \sum_{k=1}^{p} a_k z^{-k})$ [9]. It is assumed 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. To obtain the coefficients, conventional LP minimizes the energy of the prediction error signal $e_n = s_n - \sum_{k=1}^{p} a_k s_{n-k}$ by setting the partial derivatives of $E_{\mathrm{LP}} = \sum_n e_n^2$ with respect to each coefficient a_k to zero. This gives the normal equations $\sum_{k=1}^{p} a_k \sum_n s_{n-k} s_{n-j} = \sum_n s_n s_{n-j}$, $1 \leq j \leq p$, for solving the coefficients $\{a_k\}$. The range of summation of n is typically chosen 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. The LP synthesis model $H(z)$ given by the autocorrelation method is guaranteed to be stable, meaning that the roots of the denominator polynomial $1 - \sum_{k=1}^{p} a_k z^{-k}$ lie inside the unit circle [9].

2.2. Weighted linear prediction (WLP)

Weighted linear prediction (WLP) is a generalization of LP, originally introduced by Ma et al. [5]. In WLP, the prediction coefficients $\{b_k\}$ are obtained by minimization of a quantity which can be termed the weighted prediction error energy $E_{\mathrm{WLP}} = \sum_n e_n^2 W_n = \sum_n (s_n - \sum_{k=1}^{p} b_k s_{n-k})^2 W_n$, in which W_n denotes the temporal weighting function. The weighting function for WLP is usually chosen as the short-time energy of the signal in the delay line, $W_n = \sum_{i=1}^{p} s_{n-i}^2$. In the case of stationary background noise, temporal weighting enables emphasizing the contribution of high-energy samples in the computation of the spectral model. These high-energy samples are likely to have a better local signal-to-noise ratio (SNR) than low-energy segments, thus resulting in more robust models. The WLP normal equations are $\sum_{k=1}^{p} b_k \sum_n W_n s_{n-k} s_{n-j} = \sum_n W_n s_n s_{n-j}, \quad 1 \le j \le p$. From these equations, it is easy to see that by simple substitution of W_n constant across all n, the conventional LP equations are obtained as a special case. The weighting is only meaningful for W_n that varies with n. 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 (LVCSR) [2] and speaker verification [4].

2.3. Stabilized extended weighted linear prediction (SXLP)

2.3.1. General formulation

In a further generalization of LP analysis, the quantity to be minimized for solving the prediction coefficients $\{c_k\}$ can be expressed as

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

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 different LP analysis method, in which each lagged sample (with lag j) at each time instant n (i.e., each lag at each prediction) is weighted separately using weight $Z_{n,j}$. The formulation, referred to as *eXtended weighted Linear Prediction* (XLP), allows temporal weighting on a finer time scale than WLP. It was first evaluated in the context of speaker verification [6] and subsequently in LVCSR [8].

The minimization of the error energy in Eq. 1 gives rise to the XLP normal equations

$$\sum_{k=1}^{p} c_k \sum_n Z_{n,k} s_{n-k} Z_{n,j} s_{n-j} = \sum_n Z_{n,0} s_n Z_{n,j} s_{n-j}, \quad 1 \le j \le p. \qquad (2)$$

2.3.2. Weighting scheme

There are theoretically innumerable ways to determine the weights for XLP. A scheme based on absolute values of the samples was used with the first applications of XLP [6] [8]. In the present study, a new scheme, based on instantaneous autocorrelation structure, is described. When used together with the stabilization operation described in a subsequent section, it is outperforming the previous formulation.

The first step in the weight computation is to determine

$$Y_{n,j} = \frac{s_n s_{n-j}}{\frac{1}{\min(p+1,n+1)} \sum_{k=0}^{p} s_{n-k}^2}, \quad 0 \le j \le p, \qquad (3)$$

with $s_n = 0$ for $n < 0$, such that $(Y_{n,0}, Y_{n,1}, \ldots, Y_{n,p})^T$ will be a vector depicting the normalized "instantaneous autocorrelation" at time n. The values $Y_{n,j}$ are filtered along the n dimension by a highpass FIR filter $1 - 0.99z^{-1}$. Next, the filtered values are replaced by their absolute values. Finally, the absolute values are filtered along the n dimension by a lowpass IIR filter $(1/p)/(1 - ((p-1)/p)z^{-1})$ to yield the weights $Z_{n,j}$. The main motivation is to emphasize, at each time instant, the lags that are associated with formant-related autocorrelation structure that is persisting for even a short time (according to the lowpass filtering), regardless of signal energy (because the autocorrelation is normalized).

2.3.3. Stabilization

Unlike autocorrelation LP, but similarly to WLP, XLP is not guaranteed to produce a stable filter. However, filter stability is required at least in coding and synthesis applications. The stabilization technique described here was originally developed and proven correct by Magi [3] for the special case in which $Z_{n,j} = \sqrt{W_n}$, i.e., for the stabilization of WLP, giving rise to an analysis method known as stabilized weighted linear prediction (SWLP). SWLP has been successfully used to tackle noise robustness issues in automatic speech recognition [3] [7] and speaker recognition [4]. By applying the same stabilization operation to the general case of XLP, in which there are no mutual constraints on the weights $Z_{n,j}$, the SXLP method is obtained [6].

Once the weights $Z_{n,j}$ have been determined, they are replaced with $Z'_{n,j} = \max(Z_{n,j}, Z_{n-1,j-1})$, where $Z_{n,j} = 0$ for $j < 0$. This yields a stabilized model.

Figure 1 illustrates the unstabilized and stabilized weight matrices ($Z_{n,j}$ and $Z'_{n,j}$, respectively) for a voiced frame.

2.4. Spectral properties

Figures 2 and 3 show short-time spectra over two utterances as obtained by four methods: LP, WLP, XLP and SXLP. It can be noted that WLP, which with its usual weighting scheme focuses on the glottal closed phase, produces prominent formants. The formants produced by XLP (using the weighting scheme as detailed in Section 2.3.2) are even more prominent

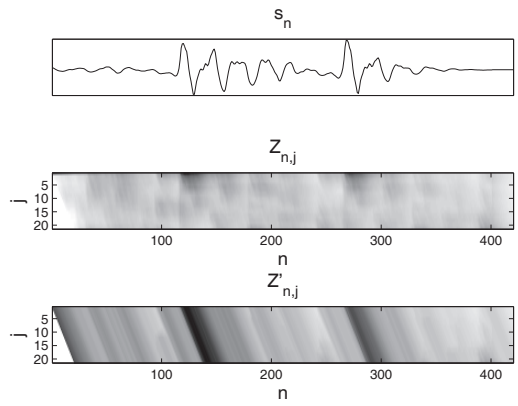

Fig. 1. Upper panel: One frame of a 16 kHz male vowel /a/. Middle panel: XLP weighting matrix $(Z_{n,j})$ (unstabilized). Lower panel: SXLP weighting matrix $(Z'_{n,j})$ (stabilized).

at the higher frequencies. In comparison to the LP reference, XLP with this weighting scheme exaggerates high-frequency formants and even produces spurious formants. However, as seen from the rightmost panels, the stabilization operation described in Section 2.3.3 clearly smoothes the formant structure, making it even somewhat smoother than that of LP. The predictably smooth formant structure and the controlled temporal behavior of SXLP may also reflect robustness against various adverse conditions and sources of variation.

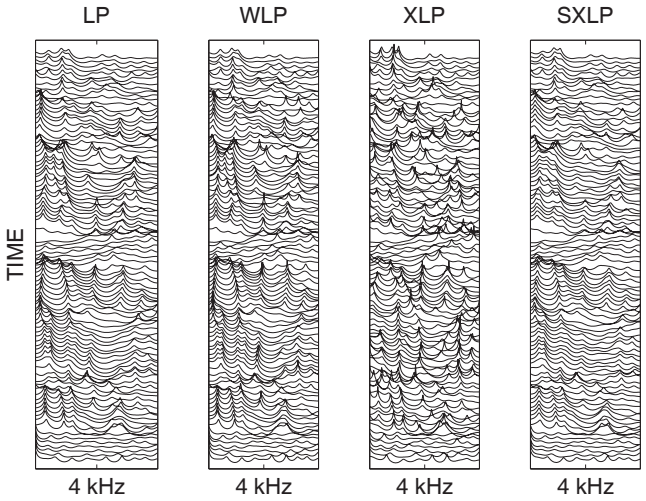

Fig. 2. Spectra from an utterance (length 1.93 s, male speaker), plotted every 20 ms with four spectrum analysis methods.

3. EXPERIMENTAL EVALUATION

The purpose of the evaluation was to study the noise robustness of the spectrum models by analyzing the average distortion in the mel frequency cepstral coefficient (MFCC) domain. The experiment was carried out using speech material from the TIMIT American English database, artificially corrupted by *factory1* and *babble* noise from the NOISEX-92 database at a given frame-averaged signal-to-noise ratio (SNR) level. The *factory1* noise is machinery noise from a factory, including frequent transient impulsive sounds. The *babble* noise contains many people talking simultaneously.

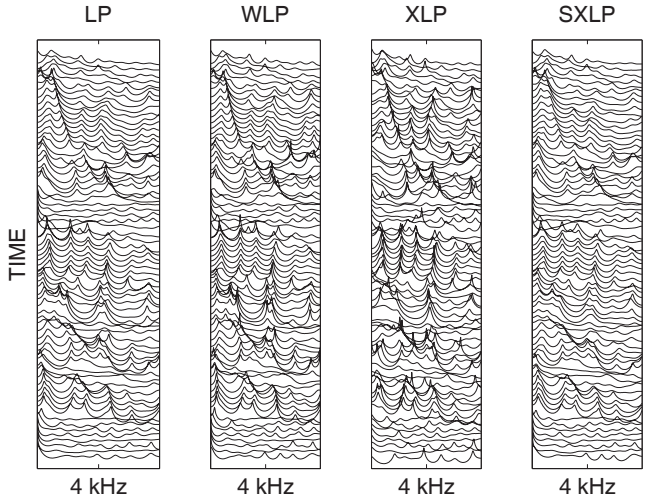

Fig. 3. Spectra from an utterance (length 1.50 s, female speaker), plotted every 20 ms with four spectrum analysis methods.

Each analyzed signal was pre-emphasized by a FIR filter $1 - 0.97z^{-1}$. A total of 183462 speech frames (25 ms frame length, 10 ms frame shift interval, Hamming window) was used as material. These frames were the non-silent frames (according to the TIMIT transcription and excluding voiced stop closures), from a total of 800 sentences, spoken by 50 male and 50 female speakers. They comprised 73.88 % of the total duration of the 800 utterances. This material was used to evaluate noise degradation in terms of MFCC squared Euclidean distance. At each frame location, for the uncorrupted clean frame and each of the noise-corrupted versions of the same frame, an MFCC vector was obtained based on each of the four spectrum analysis methods FFT, LP, WLP and SXLP (followed by the standard MFCC computation chain of mel filterbank analysis, logarithm and discrete cosine transform [1]). Utilization of the MFCC representation is justified because it is a widely used, auditorily motivated representation of the short-time magnitude spectrum as well as a popular feature representation for recognition applications.

Figures 4 and 5 show, for each spectrum analysis method and each case of noise corruption, the squared Euclidean distances between the noisy and clean MFCC vectors, averaged over all the frames. This analysis using an objective MFCC distortion measure is in line with the results obtained in recognition studies, which were referred to earlier. Specifically, the

linear predictive methods outperform the FFT spectrum analysis method in terms of robustness, with the WLP method offering some robustness advantage over conventional LP (the importance of this will vary from application to application). Interestingly, the new SXLP method clearly outperforms the other methods in terms of the MFCC distortion measure.

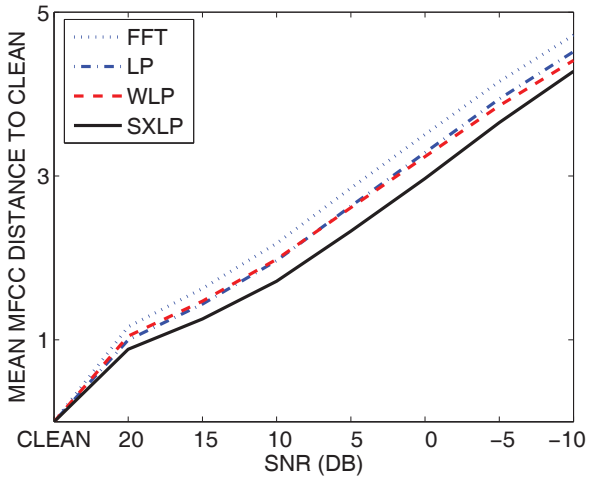

Fig. 4. Noisy-to-clean MFCC distances, averaged over 183462 speech frames, with factory noise corruption.

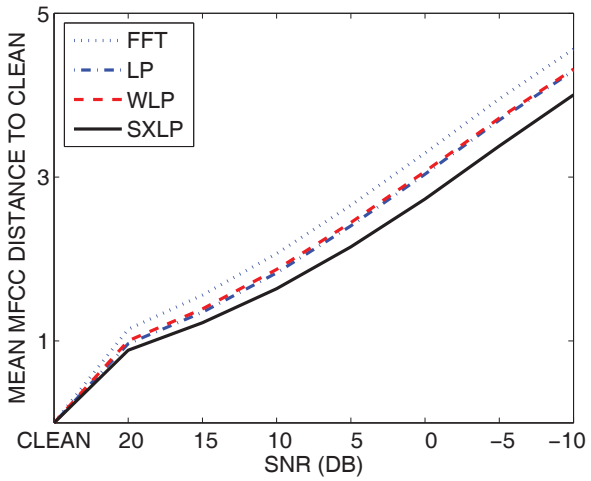

Fig. 5. Noisy-to-clean MFCC distances, averaged over 183462 speech frames, with babble noise corruption.

4. CONCLUSIONS

A new, stable lag-weighted linear predictive analysis method, in the framework of the XLP formulation [6], was described, demonstrated and evaluated. The method is based on a novel XLP weighting scheme which emphasizes instantaneous, but at least to some degree persistent, autocorrelation structures. Although this approach by itself was observed to produce spectra with spurious and exaggerated formants, the stabilization operation of SXLP made the spectra well behaved. This was reflected in the robustness as measured by average distortion between noisy and clean MFCC representations. In this analysis, the proposed method outperformed the conventional FFT, LP and WLP.

The proposed speech analysis method shows promise to be used next in different applications. More generally, the approach outlined in this study can be used for creating weighting schemes that focus on the desired information in the analysis frame and yet, when followed by the stabilization operation, result in well-behaved spectrum models. A potential direction for future work is thus the analysis of weighting schemes according to the specific requirements of different applications, for example in speech and speaker recognition.

5. REFERENCES

[1] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing*, Prentice Hall PTR, 2001.

[2] J. Pohjalainen, H. Kallasjoki, K.J. Palomäki, M. Kurimo, and P. Alku, "Weighted linear prediction for speech analysis in noisy conditions," in *Proc. Interspeech*, Brighton, UK, September 2009.

[3] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilised weighted linear prediction," *Speech Communication*, vol. 51, no. 5, pp. 401–411, 2009.

[4] R. Saeidi, J. Pohjalainen, T. Kinnunen, and P. Alku, "Temporally weighted linear prediction features for tackling additive noise in speaker verification," *IEEE Signal Processing Letters*, vol. 17, no. 6, pp. 599–602, 2010.

[5] C. Ma, Y. Kamp, and L.F. Willems, "Robust signal selection for linear prediction analysis of voiced speech," *Speech Communication*, vol. 12, no. 2, pp. 69–81, 1993.

[6] J. Pohjalainen, R. Saeidi, T. Kinnunen, and P. Alku, "Extended weighted linear prediction (XLP) analysis of speech and its application to speaker verification in adverse conditions," in *Proc. Interspeech*, Makuhari, Japan, September 2010.

[7] H. Kallasjoki, K.J. Palomäki, C. Magi, P. Alku, and M. Kurimo, "Noise robust LVCSR feature extraction based on stabilized weighted linear prediction," in *Proc. SPECOM'2009*, St. Petersburg, Russia, June 2009.

[8] S. Keronen, J. Pohjalainen, P. Alku, and M. Kurimo, "Noise robust feature extraction based on extended weighted linear prediction in LVCSR," in *Proc. Interspeech*, Florence, Italy, August 2011.

[9] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, no. 4, pp. 561–580, 1975.