Temporally Weighted Linear Prediction Features for Tackling Additive Noise in Speaker Verification

Rahim Saeidi, *Student Member, IEEE,* Jouni Pohjalainen, Tomi Kinnunen and Paavo Alku

***Abstract*—Text-independent speaker verification under additive noise corruption is considered. In the popular mel-frequency cepstral coefficient (MFCC) front-end, the conventional Fourier-based spectrum estimation is substituted with weighted linear predictive methods, which have earlier shown success in noise-robust speech recognition. Two temporally weighted variants of linear predictive modeling are introduced to speaker verification and they are compared to FFT, which is normally used in computing MFCCs, and to conventional linear prediction. The effect of speech enhancement (spectral subtraction) on the system performance with each of the four feature representations is also investigated. Experiments by the authors on the NIST 2002 SRE corpus indicate that the accuracy of the conventional and proposed features are close to each other on clean data. For factory noise at 0 dB SNR level, baseline FFT and the better of the proposed features give EERs of 17.4 % and 15.6 %, respectively. These accuracies improve to 11.6 % and 11.2 %, respectively, when spectral subtraction is included as a pre-processing method. The new features hold a promise for noise-robust speaker verification.**

***Index Terms*—Speaker verification, additive noise, stabilized weighted linear prediction (SWLP)**

I. Introduction

Speaker verification is the task of verifying one's identity based on the speech signal [1]. A typical speaker verification system consists of a short-term spectral feature extractor (front-end) and a pattern matching module (back-end). For pattern matching, Gaussian mixture models [2] and support vector machines [3] are commonly used. The standard spectrum analysis method for speaker verification is the discrete Fourier transform, implemented as the fast Fourier transform (FFT). Linear prediction (LP) is another approach to estimate the short-time spectrum [4].

Research in speaker recognition over the past two decades has largely concentrated on tackling the channel variability problem, that is, how to normalize the adverse effects due to differing training and test handsets or channels (e.g. GSM versus landline speech) [5]. Another challenging problem in speaker recognition, and speech technology in general, is that of additive noise, that is, degradation that originates from other sound sources and adds to the speech signal.

Rahim Saeidi and Tomi Kinnunen are with the School of Computing, University of Eastern Finland, FI-80101 Joensuu, Finland (e-mail: rahim.saeidi@uef.fi, tomi.kinnunen@uef.fi).

Jouni Pohjalainen and Paavo Alku are with the Department of Signal Processing and Acoustics, Aalto University School of Science and Technology, FI-00076 Aalto, Finland (e-mail: jpohjala@acoustics.hut.fi, paavo.alku@hut.fi).

The work of Rahim Saeidi was supported by a scholarship from the Finnish Foundation for Technology Promotion (TES). The work of Jouni Pohjalainen and Tomi Kinnunen was supported by Academy of Finland projects (127345, 135003, 132129 Lastu-programme).

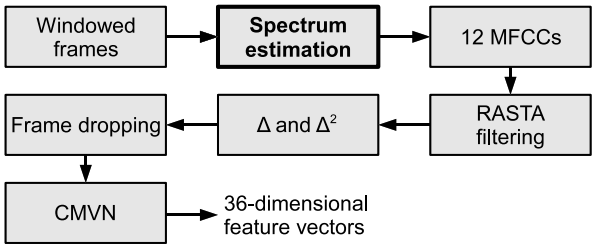

Fig. 1. Front-end of the speaker recognition system. While standard mel-frequency cepstral features derived through a mel-frequency spaced filterbank placed on the magnitude spectrum are used in this work, the way how the magnitude spectrum is computed varies (FFT = Fast Fourier transform, baseline method; LP = Linear prediction; WLP = Weighted linear prediction; SWLP = Stabilized weighted linear prediction).

Neither FFT nor LP can robustly handle conditions of additive noise. Therefore, this topic has been studied extensively in the past few decades and many speech enhancement methods have been proposed to tackle problems caused by additive noise [6], [7]. These methods include, for example, spectral subtraction, Wiener filtering and Kalman filtering. They are all based on forming a statistical estimate for noise and removing it from the corrupted speech. Speech enhancement methods can be used in speaker recognition as a pre-processing stage to remove additive noise. However, they have two potential drawbacks. First, noise estimates are never perfect, which may result in removing not only the noise but also speaker-dependent components of the original speech. Second, additional pre-processing increases processing time which can become a problem in real-time authentication.

Another approach to increase robustness is to carry out feature normalization such as cepstral mean and variance normalization (CMVN), RASTA filtering [8] or feature warping [9]. These methods are often stacked with each other and combined with score normalization such as T-norm [10]. Finally, examples of model-domain methods, specifically designed to tackle additive noise, include model-domain spectral subtraction [11], missing feature theory [12] and parallel model combination [13], to mention a few. Model-domain methods are always limited to a certain model family, such as Gaussian mixtures.

This paper focuses on short-term spectral feature extraction (Fig. 1). Several previous studies have addressed robust feature extraction in speaker identification based on LP-derived methods, e.g. [14]–[16]. All these investigations, however, use vector quantization (VQ) classifiers and some of the feature extraction methods utilized are computationally intensive, because they involve solving for the roots of LP polynomials. Differently from these previous studies, this work (a) compares two straightforward noise-robust modifications of LP and (b)

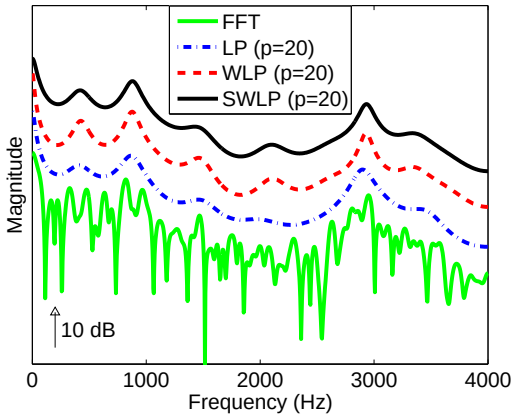

Fig. 2. Examples of FFT, LP, WLP and SWLP spectra for a voiced speech sound taken from the NIST 2002 speaker recognition corpus and corrupted by factory noise (SNR -10 dB). The spectra have been shifted by approximately 10 dB with respect to each other.

utilizes them in a more modern speaker verification system based on adapted Gaussian mixtures [2] and MFCC feature extraction. The robust linear predictive methods used for spectrum estimation (Fig. 1) are weighted linear prediction (WLP) [17] and stabilized WLP (SWLP) [18], which is a variant of WLP that guarantees the stability of the resulting all-pole filter. Rather than removing noise as speech enhancement methods do, the weighted LP methods aim to increase the contribution of such samples in the filter optimization that have been less corrupted by noise. As illustrated in Fig. 2, the corresponding all-pole spectra may preserve the formant structure of noise-corrupted voiced speech better than the conventional methods. The WLP and SWLP features were recently applied to automatic speech recognition in [19] with promising results; the authors were curious to see whether these improvements would translate to speaker verification as well.

II. Spectrum Estimation Methods

In linear predictive modeling, with prediction order p, 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 is the digital speech signal and $\{a_k\}$ are the prediction coefficients. The difference between the actual sample s_n and its predicted value $\hat{s}_n$ is the residual $e_n = s_n - \sum_{k=1}^{p} a_k s_{n-k}$. WLP is a generalization of LP. In contrast to conventional LP, WLP introduces a temporal weighting of the squared residual in model coefficient optimization, allowing emphasis of the temporal regions assumed to be little affected by the noise, and de-emphasis of the noisy regions. The coefficients $\{b_k\}$ are solved by minimizing the energy of the weighted squared residual [17] $E = \sum_n e_n^2 W_n = \sum_n (s_n - \sum_{k=1}^{p} b_k s_{n-k})^2 W_n$, where W_n is the weighting function. The range of summation of n (not explicitly written) 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 [4]. By setting the partial derivatives of E with respect to each b_k to zero, the WLP normal equations arrived at are

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

which can be solved for the coefficients b_k to obtain the WLP all-pole model $H(z) = 1/(1 - \sum_{k=1}^{p} b_k z^{-k})$. It is easy to show that conventional LP can be obtained as a special case of WLP: by setting $W_n = c$ for all n, where c is a finite nonzero constant, c becomes a multiplier of both sides of (1) and cancels out, leaving the LP normal equations [4].

The conventional autocorrelation LP method is guaranteed to produce always a stable all-pole model, that is, a filter where all poles are within the unit circle [4]. However, such a guarantee does not exist for autocorrelation WLP when the weighting function W_n is arbitrary [17], [18]. Because of the importance of model stability in coding and synthesis applications, SWLP was developed [18]. The WLP normal equations (1) can be alternatively written in terms of partial weights $Z_{n,j}$ as

$$\sum_{k=1}^{p} b_k \sum_n Z_{n,k} s_{n-k} Z_{n,i} s_{n-i} = \sum_n Z_{n,0} s_n Z_{n,i} s_{n-i}, \quad 1 \le i \le p, \quad (2)$$

where $Z_{n,j} = \sqrt{W_n}$ for $0 \le j \le p$. As shown in [18] (using a matrix-based formulation), model stability is guaranteed if the partial 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 (2) gives the SWLP normal equations.

The motivation for temporal weighting is to emphasize the contribution of the less noisy signal regions in solving the LP filter coefficients. Typically, the weighting function W_n in WLP and SWLP is chosen as the short-time energy (STE) of the immediate signal history [17]–[19], computed using a sliding window of M samples as $W_n = \sum_{i=1}^{M} s_{n-i}^2$. STE weighting emphasizes those sections of the speech waveform which consist of samples having a large amplitude. It can be argued that these segments of speech are likely to be less corrupted by stationary additive noise than the low-energy segments. Indeed, when compared to traditional spectral modeling methods such as FFT and LP, WLP and SWLP using STE-weighting have been shown to improve noise robustness in automatic speech recognition [18], [19].

III. Speaker Verification Setup

The effectiveness of the features is evaluated on the NIST 2002 speaker recognition evaluation (SRE) corpus, which consists of realistic speech samples transmitted over different cellular networks with varying types of handsets.

The experiments are conducted using a standard Gaussian mixture model classifier with a universal background model (GMM-UBM) [2]. The GMM-UBM system was chosen since it is simple and may outperform support vector machines under additive noise conditions [13]. Test normalization (T-norm) [10] is applied on the logarithmic likelihood ratio scores. 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

TABLE I
SYSTEM PERFORMANCE UNDER WHITE NOISE.

Signal-to-noise ratio (dB)	Equal error rate (EER %)								MinDCF							
	Without spectral subtraction				With spectral subtraction				Without spectral subtraction				With spectral subtraction			
	FFT	LP	WLP	SWLP	FFT	LP	WLP	SWLP	FFT	LP	WLP	SWLP	FFT	LP	WLP	SWLP
clean	9.22	8.89	9.15	9.15	9.29	8.92	9.26	9.19	3.56	3.47	3.50	3.54	3.59	3.51	3.53	3.60
20	9.76	9.43	9.46	9.39	9.52	9.35	9.39	9.19	3.83	3.77	3.69	3.82	3.77	3.60	3.69	3.69
10	12.37	12.04	12.01	12.11	10.73	10.19	10.32	10.09	5.12	5.10	5.09	5.20	4.17	4.10	4.18	4.14
0	26.27	26.19	25.15	25.39	13.22	12.71	12.91	12.71	9.34	9.51	9.50	9.44	5.28	5.14	5.15	5.10
-10	37.66	37.73	37.06	37.16	23.51	22.77	23.44	22.50	10.00	10.00	10.00	10.00	8.57	8.29	8.56	8.27

TABLE II
SYSTEM PERFORMANCE UNDER FACTORY NOISE.

Signal-to-noise ratio (dB)	Equal error rate (EER %)								MinDCF							
	Without spectral subtraction				With spectral subtraction				Without spectral subtraction				With spectral subtraction			
	FFT	LP	WLP	SWLP	FFT	LP	WLP	SWLP	FFT	LP	WLP	SWLP	FFT	LP	WLP	SWLP
clean	9.22	8.89	9.15	9.15	9.29	8.92	9.26	9.19	3.56	3.47	3.50	3.54	3.59	3.51	3.53	3.60
20	9.57	9.22	9.22	9.29	9.69	9.26	9.46	9.35	3.71	3.70	3.70	3.71	3.72	3.65	3.64	3.68
10	10.13	10.26	10.13	10.03	10.47	10.20	10.26	10.03	4.05	4.20	4.16	4.16	4.09	4.00	4.15	4.09
0	17.40	17.04	16.03	15.59	11.64	11.57	11.57	11.17	7.62	7.82	7.24	7.04	4.54	4.64	4.76	4.60
-10	26.19	25.63	24.41	23.68	16.84	16.60	16.87	15.55	9.80	9.84	9.75	9.69	6.99	6.70	6.72	6.34

the target speaker model. The duration of the test utterances varies between 15 and 45 seconds. The (gender-dependent) background models and cohort models for Tnorm, having 1024 Gaussians, are trained using the NIST 2001 corpus. This baseline system [20] has comparable or better accuracy than other systems evaluated on this corpus (e.g. [21]).

Features are extracted every 15 ms from 30 ms frames multiplied by a Hamming window. Depending on the feature extraction method, the magnitude spectrum is computed differently. For the baseline method, the FFT of the windowed frame is directly computed. For LP, WLP and SWLP, the model coefficients and the corresponding all-pole spectra are first derived as explained in Section II. All the three parametric methods use a predictor order of $p = 20$. For WLP and SWLP, the short-term energy window duration is set to $M = 20$ samples. A 27-channel mel-frequency filterbank is used to extract 12 MFCCs. After RASTA filtering, Δ and Δ^2 coefficients, a standard component in modern speaker verification [1], are appended. Voiced frames are then selected using an energy-based voice activity detector (VAD). Finally, cepstral mean and variance normalization (CMVN) is performed. The procedure is illustrated in Fig. 1.

Two standard metrics are 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 $0.1 \times P_{\mathrm{miss}} + 0.99 \times P_{\mathrm{fa}}$. In addition, a few selected detection error tradeoff (DET) curves are plotted showing the full trade-off curve between false alarms and misses on a normal deviate scale. All the reported minDCF values are multiplied by 10, for ease of comparison.

To study robustness against additive noise, some noise is digitally added from the NOISEX-92 database[1] to the speech samples. This study uses *white* and *factory2* noises (the latter is referred to as "factory noise" throughout the paper). The background models and target speaker models are trained with clean data, but the noise is added to the test files with a given average segmental (frame-average) signal-to-noise ratio (SNR). Five values are considered: $\mathrm{SNR} \in \{\mathrm{clean}, 20, 10, 0, -10\}$ dB, where "clean" refers to the original, uncontaminated NIST samples. In summary, the evaluation data used in the present study contains linear and nonlinear distortion present in the sounds of the NIST 2002 database as well as additive noise taken from the NOISEX-92 database.

Also included in the experiments is the well-known and simple speech enhancement method, spectral subtraction (as described in [6]). The effect of speech enhancement is studied alone as well as in combination with the new features. The noise model is initialized from the first five frames and updated during the non-speech periods with VAD labels given by the energy method.

IV. SPEAKER VERIFICATION RESULTS

The results for white and factory noise are shown in Tables I and II, respectively. In addition, Fig. 3 shows a DET plot that compares the four feature sets under factory noise degradation at SNR of 0 dB without any speech enhancement. Examining the EER and MinDCF scores without speech enhancement, the following observations are made:

- The accuracy of all four feature sets degrades significantly under additive noise; performance in white noise is inferior to that in factory noise [2].
- WLP and SWLP outperform FFT and LP in most cases, with large differences at low SNRs and for factory noise; the best performing methods for white noise and factory noise are WLP and SWLP, respectively.
- WLP and SWLP show minor improvement over FFT also in the clean condition, showing consistency of the new features.
- It is interesting to note that, although SWLP is stabilized mainly for synthesis purposes and WLP has performed

[1]Samples available at http://spib.rice.edu/spib/select_noise.html

[2]Factory noise has an overall "lowpass" spectral slope close to that of speech, whereas the spectrum of white noise is flat. White noise is thus likely to corrupt the higher formants of speech more severely.

better in speech recognition [19], SWLP seems to slightly outperform WLP in speaker recognition.

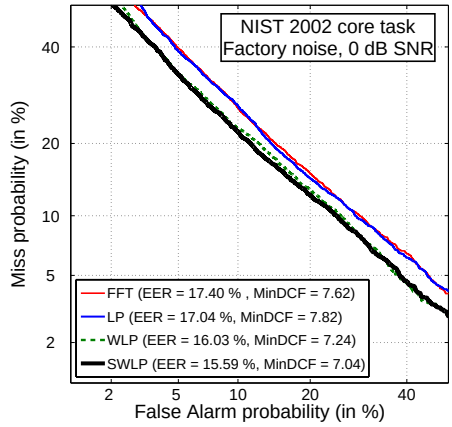

Fig. 3. Comparing the features without any speech enhancement.

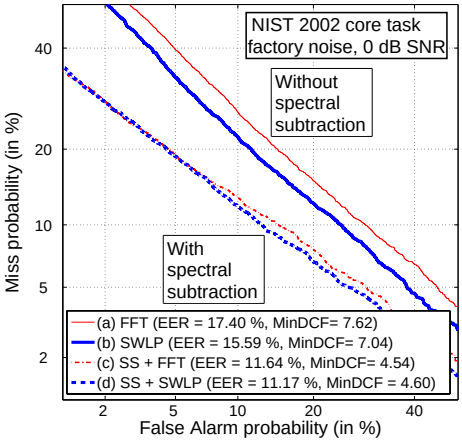

Fig. 4. Comparing FFT and SWLP with and without speech enhancement (SS = Spectral Subtraction).

Considering the effect of speech enhancement, as summarized by the representative DET plot in Fig. 4, speech enhancement as a pre-processing step is seen to significantly improve all the four methods. In addition, according to Tables I and II, the difference becomes progressively larger with decreasing SNR. This is expected since for a less noisy signal, spectral subtraction is also likely to remove other information in addition to noise. After including speech enhancement, even though the enhancement has a larger effect than the choice of the feature set, SWLP remains the most robust method and together with WLP outperforms baseline FFT. Note that here the benefit from spectral subtraction may be quite pronounced due to almost stationary noise types.

V. Conclusions

Temporally weighted linear predictive features in speaker verification were studied. Without speech enhancement, the new WLP and SWLP features outperformed standard FFT and LP features in recognition experiments under additive-noise conditions. The effectiveness of spectral subtraction in highly noisy environments was also demonstrated. However, in the enhanced case, both proposed methods still improved upon the FFT baseline, and SWLP remained the most robust method. In summary, the weighted linear predictive features are a promising approach for speaker recognition in the presence of additive noise.

References

[1] T. Kinnunen and H. Li. An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication*, 52(1):12–40, January 2010.

[2] D.A. Reynolds, T.F. Quatieri, and R.B. Dunn. Speaker verification using adapted gaussian mixture models. *Digital Signal Processing*, 10(1):19–41, January 2000.

[3] W.M. Campbell, D.E. Sturim, and D.A. Reynolds. Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, 13(5):308–311, May 2006.

[4] J. Makhoul. Linear prediction: a tutorial review. *Proceedings of the IEEE*, 64(4):561–580, April 1975.

[5] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel. Joint factor analysis versus eigenchannels in speaker recognition. *IEEE Trans. Audio, Speech and Language Processing*, 15(4):1435–1447, May 2007.

[6] P. C. Loizou. *Speech Enhancement: Theory and Practice*. CRC Press, 2007.

[7] T. Ganchev, I. Potamitis, N. Fakotakis, and G. Kokkinakis. Text-independent speaker verification for real fast-varying noisy environments. *International Journal of Speech Technology*, 7(4):281–292, October 2004.

[8] H. Hermansky and N. Morgan. RASTA processing of speech. *IEEE Trans. on Speech and Audio Processing*, 2(4):578–589, October 1994.

[9] J. Pelecanos and S. Sridharan. Feature warping for robust speaker verification. In *Proc. Speaker Odyssey: the Speaker Recognition Workshop (Odyssey 2001)*, pages 213–218, Crete, Greece, June 2001.

[10] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas. Score normalization for text-independent speaker verification systems. *Digital Signal Processing*, 10(1-3):42–54, January 2000.

[11] J. A. Nolazco-Flores and L. P. Garcia-Perera. Enhancing acoustic models for robust speaker verification. In *Proc. Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP 2008)*, pages 4837–4840, Las Vegas, U.S.A., April 2008.

[12] Ji Ming, T. J. Hazen, J. R. Glass, and D. A. Reynolds. Robust speaker recognition in noisy conditions. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(5):1711–1723, July 2007.

[13] S. G. Pillay, A. Ariyaeeinia, M. Pawlewski, and P. Sivakumaran. Speaker verification under mismatched data conditions. *IET Signal Processing*, 3(4):236–246, July 2009.

[14] K. T. Assaleh and R. J. Mammone. New LP-derived features for speaker identification. *IEEE Trans. on Speech and Audio Processing*, 2(4):630–638, October 1994.

[15] R. P. Ramachandran, M. S. Zilovic, and R. J. Mammone. A comparative study of robust linear predictive analysis methods with applications to speaker identification. *IEEE Trans. on Speech and Audio Processing*, 3(2):117–125, March 1995.

[16] M.S. Zilovic, R.P. Ramachandran, and R.J. Mammone. Speaker identification based on the use of robust cepstral features obtained from pole-zero transfer functions. *IEEE Trans. on Speech and Audio Processing*, 6(3):260–267, 1998.

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

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

[19] J. Pohjalainen, H. Kallasjoki, K.J. Palomäki, M. Kurimo, and P. Alku. Weighted linear prediction for speech analysis in noisy conditions. In *Proc. Interspeech 2009*, pages 1315–1318, Brighton, UK, 2009.

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

[21] C. Longworth and M.J.F. Gales. Combining derivative and parametric kernels for speaker verification. *IEEE Trans. Audio, Speech and Language Processing*, 17(4):748–757, May 2009.