Open access • Journal Article • DOI:10.1109/LSP.2010.2048649

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

Rahim Saeidi, Jouni Pohjalainen, Tomi Kinnunen, Paavo Alku

Institutions: University of Eastern Finland, Aalto University

Published on: 19 Apr 2010 - IEEE Signal Processing Letters (IEEE)

Topics: Speaker recognition, Speech enhancement, Linear prediction, Noise and Mel-frequency cepstrum

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

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

[1]School of Computing, University of Eastern Finland, Finland
[2]Department of Signal Processing and Acoustics, Aalto University, Finland
{rahim.saeidi,tomi.kinnunen}@uef.fi, jpohjala@acoustics.hut.fi, paavo.alku@hut.fi

Abstract

We consider text-independent speaker verification under additive noise corruption. In the popular mel-frequency cepstral coefficient (MFCC) front-end, we substitute the conventional Fourier-based spectrum estimation with weighted linear predictive methods, which have earlier shown success in noise-robust speech recognition. We introduce two temporally weighted variants of linear predictive (LP) modeling to speaker verification and compare them to FFT, which is normally used in computing MFCCs, and to conventional LP. We also investigate the effect of speech enhancement (spectral subtraction) on the system performance with each of the four feature representations. Our experiments on the NIST 2002 SRE corpus indicate that the accuracy of the conventional and proposed features are close to each other on clean data. On 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.

1. 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 by 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 out 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.

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 the 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.

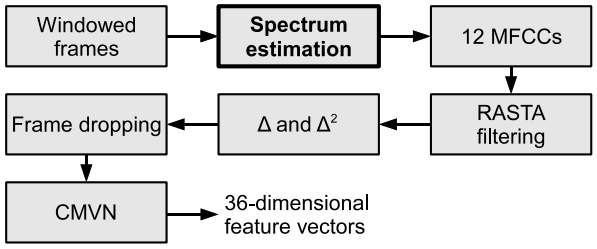
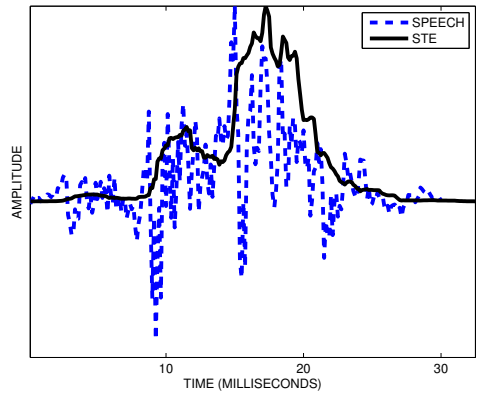


Figure 1: Front-end of the speaker recognition system. While we use standard mel-frequency cepstral features derived through mel-frequency spaced filterbank placed on the magnitude spectrum, the way how the magnitude spectrum is computed varies (FFT = Fast Fourier transform, baseline method; LP = Linear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

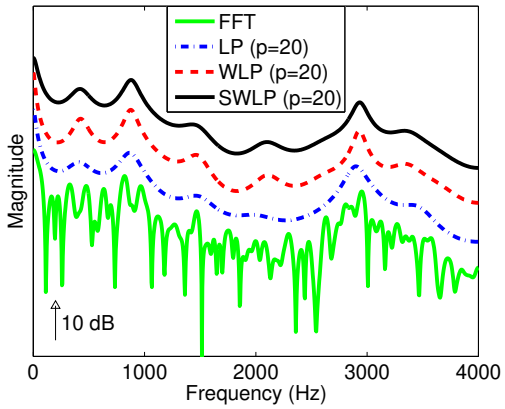
Another approach to increase robustness is to carry out *feature normalization* such as cepstral mean and variance normalization (CMVN), RASTA filtering [8] or feature warping [9]. These methods are often stacked with each other and combined with *score normalization* such as T-norm [10]. Finally, examples 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] [15] [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, we (a) compare two straightforward noise-robust modifications of LP and (b) utilize them in a more modern speaker verification system based on

*Short version of the paper has been accepted to *IEEE Signal Processing Letters*.



(a)



(b)

Figure 2: (a) Short time energy (STE) as it used as the weighting function in WLP and SWLP is shown for a voiced speech sound taken from the NIST 2002 speaker recognition corpus and corrupted by factory noise (SNR -10 dB). (b) Examples of FFT, LP, WLP and SWLP spectra for the speech frame in (a). The spectra have been shifted by approximately 10 dB with respect to each other.

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 modified version 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; we were curious to see whether these improvements would translate to speaker verification as well.

We first introduce the spectrum estimation methods in Section 2. Experimental setup is described in Section 3. We use a robust mel-frequency cepstral coefficient (MFCC) front-end as indicated in Fig. 1 and vary the computation of the magnitude spectrum. The standard FFT and LP form a point of comparison. We expect the temporally weighted LP variants – WLP and SWLP – to perform better under additive noise conditions, which will be demonstrated in Section 4. The paper is concluded in Section 6.

2. Spectrum Estimation Methods

In linear predictive (LP) 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* $e_n = s_n - \sum_{k=1}^{p} a_k s_{n-k}$. Weighted linear prediction (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, we arrive at the WLP normal equations

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

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, for all n, $W_n = c$, 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 always produce 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, *stabilized* WLP (SWLP) was developed [18]. The WLP normal equations (1) can alternatively be written in terms of *partial weights* $Z_{n,j}$ as

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

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] [18] [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 of large amplitude. It can be argued that these segments of speech are likely to have been less corrupted by stationary additive noise than 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 [19] [18].

3. Speaker Verification Setup

We evaluate the effectiveness of the features on the NIST 2002 speaker recognition evaluation (SRE) corpus by using a standard Gaussian mixture model with a universal background model (GMM-UBM) [2]. We chose the GMM-UBM system since it is simple and may outperform support vector machines under *additive* noise conditions [13]. Test normalization (T-norm) [10] is applied on the log 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 for training the target speaker model. 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 NIST 2001 corpus. Our baseline system [20] has comparable or better accuracy to 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, we directly compute the fast Fourier transform (FFT) of the windowed frame. For LP, WLP, and SWLP, the model coefficients and the corresponding all-pole spectra are first derived as explained in Section 2. 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. We use a 27-channel mel-frequency filterbank to extract 12 MFCCs. After RASTA filtering, Δ and Δ^2 coefficients 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.

We use two standard metrics to assess recognition accuracy: equal error rate (EER) and 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 $0.1 \times P_{\text{miss}} + 0.99 \times P_{\text{fa}}$. In addition, we plot a few selected detection error tradeoff (DET) curves which shows the full trade-off curve between false alarms and misses in a normal deviate scale. All the reported minDCF values are multiplied by 10, for ease of comparison.

To study robustness against additive noise, we digitally add some additive noise from the NOISEX-92 database[1] to the speech samples. In this study we use *white*, *pink* and *factory2* noises[2]. The background models and target speaker models are trained on clean data, but the noises are added to the test files with a given average segmental (frame-average) signal-to-noise ratio (SNR). We consider five values: $\text{SNR} \in \{\text{clean}, 20, 10, 0, -10\}$ dB, where "clean" refers to the original, uncontaminated NIST samples[3].

We also include the well-known and simple speech enhancement method, *spectral subtraction* (SS), as described in [6], in the experiments. We study the effect of speech enhancement alone, as well as the combination of speech enhancement 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.

4. Speaker Verification Results

We first study the effects of spectral subtraction and T-norm under white noise corruption in Fig. 3. The results, shown here for the FFT-derived spectrum, are similar for LP, WLP and SWLP. Inclusion of spectral subtraction helps especially in very noisy conditions, and does not degrade the performance even for the clean condition. T-norm helps to reduce the miss rate at small false alarm rates (as reflected by the value of MinDCF), as expected [10]. In the rest of the experiments, we include T-norm unless otherwise stated.

We next study the effect of noise type and noise level to all four feature sets, both with and without spectral subtraction. The equal error rates are presented graphically in Fig. 4, whereas Tables 1, 2 and 3 display more detailed breakdown of the results for white, pink and factory noise, respectively. Finally, Fig. 6 shows a DET plot that compares the four feature sets under factory noise degradation at SNR of 0 dB without any speech enhancement. Comparing the results without speech enhancement, we make the following observations:

- The accuracy of all four feature sets degrades significantly under additive noise; performance in white and pink noises is inferior to that in factory noise.
- WLP and SWLP outperform FFT and LP in most cases, with large differences at low SNRs and for factory noise
- 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 better in speech recognition [19], SWLP seems to slightly outperform WLP in speaker recognition.

In speaker recognition, it is common to fuse FFT- and LP-derived features since that they capture complementary properties of the underlying speech process [22, 23]. Here, we consider fusion of the FFT- and SWLP-based features using two well-known fusion strategies. *Score fusion* is carried out by summing the log-likelihood ratio scores of the two classifiers, $\text{score} = 0.5 \times (\text{LLR}_{\text{FFT}} + \text{LLR}_{\text{SWlP}})$ and *feature fusion* is implemented by training a single GMM-UBM classifier on the concatenated 72-dimensional features. The results for the individual classifiers (FFT, SWLP) and the two types of fusion are given in Fig. 5. Overall, the fusion gains are rather modest and feature fusion is more stable. Since the FFT and SWLP classifiers do not degrade uniformly with decreasing SNR level, for effective score fusion the fusion weight should be adopted for the (estimated) SNR-level; feature fusion seems to be more

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

[2] We will refer this as "factory noise" throughout the paper.

[3] In fact, these samples are far away from "clean" as they have been transmitted over different cellular networks with varying types of handsets and are possibly already contaminated with some additive noise.

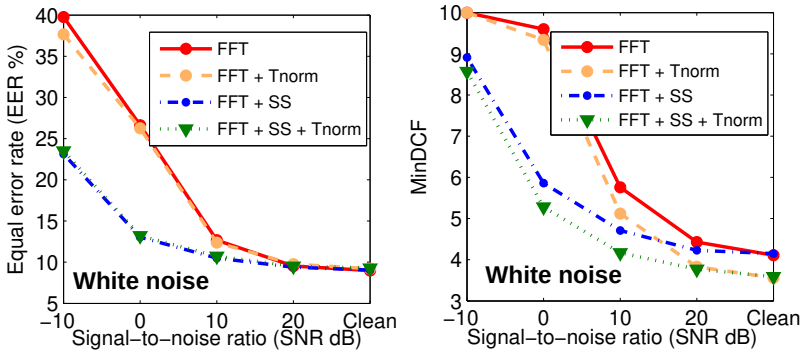


Figure 3: Effects of spectral subtraction (SS) and test normalization (T-norm) to EER (left) and MinDCF (right) on white noise when using features derived from the FFT spectrum. Results for LP, WLP and SWLP spectrum are similar.

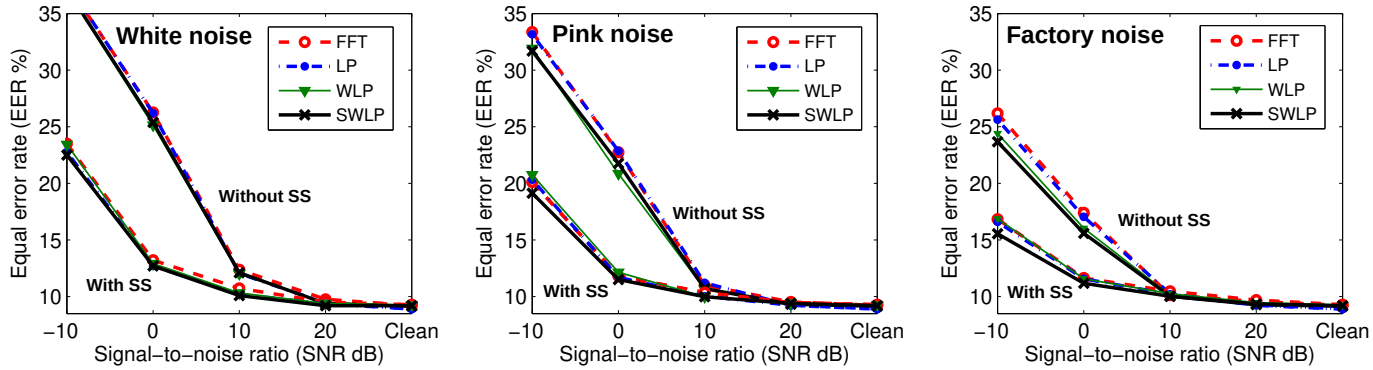


Figure 4: Equal error rates (EER %) of the four spectrum estimation methods on white noise (left), pink noise (middle) and factory noise (right). Test normalization (T-norm) is applied in all cases; SS = spectral subtraction.

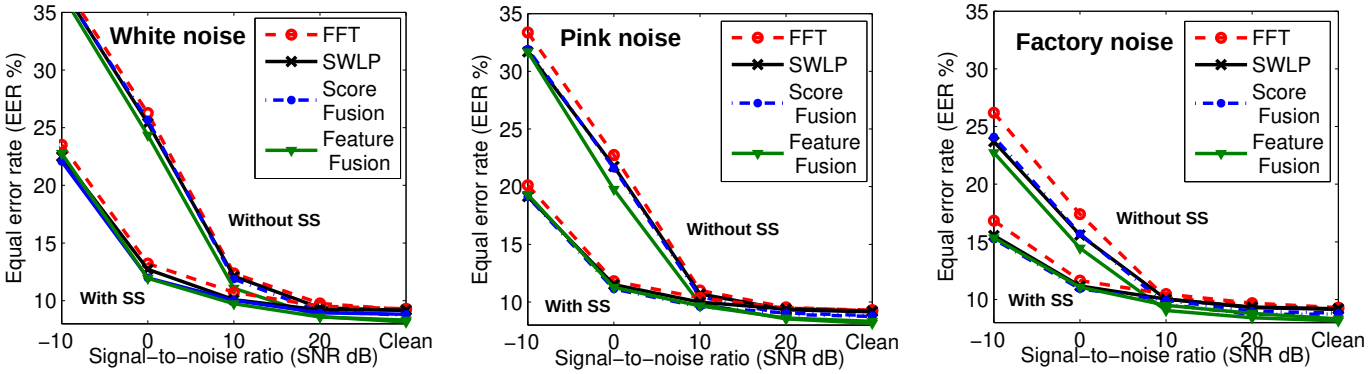


Figure 5: Equal error rates (EER %) of the FFT and SWLP spectrum estimation methods along with score fusion and feature fusion on white noise (left), pink noise (middle) and factory noise (right). Test normalization (T-norm) is applied in all cases; SS = spectral subtraction.

straightforward. The DET plot in Fig. 7 also includes the feature fusion which indicates slight improvements at low false alarm rates.

5. Discussion

Considering the effect of speech enhancement, as summarized by Figs. 4 and 7, we see that speech enhancement as a preprocessing step significantly improves all the four methods. In addition, according to Tables 1 through 3, the difference be-

Table 1: System performance under white noise with T-norm applied.

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
Average	19.08	18.86	**18.57**	18.64	13.25	12.79	13.06	**12.74**	6.37	6.37	**6.36**	6.40	5.08	**4.93**	5.02	4.96

Table 2: System performance under pink noise with T-norm applied.

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	9.53	9.22	9.32	9.32	9.46	9.23	9.42	9.39	3.71	3.72	3.70	3.75	3.77	3.69	3.63	3.70
10	11.00	11.21	10.66	10.70	10.36	10.03	9.99	9.99	4.41	4.62	4.51	4.51	4.12	4.02	4.11	4.05
0	22.74	22.86	20.86	21.76	11.80	11.70	12.14	11.50	8.72	9.07	8.86	8.74	4.76	4.84	4.81	4.77
-10	33.37	33.17	31.92	31.69	20.12	20.32	20.76	19.14	10.00	10.00	10.00	10.00	7.90	7.66	7.94	7.51
Average	17.17	17.07	**16.38**	16.52	12.21	12.04	12.31	**11.84**	**6.08**	6.18	6.11	6.11	4.83	4.74	4.80	**4.73**

Table 3: System performance under factory noise with T-norm applied.

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
Average	14.50	14.23	13.79	**13.55**	11.59	11.31	11.48	**11.06**	5.75	5.81	5.67	**5.63**	4.59	4.50	4.56	**4.46**

Table 4: The effects of spectral subtraction and voice activity detector (VAD) on the noisiest factory noise condition (-10 dB SNR).

Spectral subtraction	VAD labels from	Equal error rate (EER %)				MinDCF			
		FFT	LP	WLP	SWLP	FFT	LP	WLP	SWLP
No	Noisy	26.19	25.63	24.41	23.68	9.80	9.84	9.75	9.69
No	Clean	17.60	17.49	17.54	16.18	7.68	7.57	7.49	7.36
Yes	Noisy	16.84	16.60	16.87	15.55	6.99	6.70	6.72	6.34
Yes	Clean	17.25	16.97	17.42	15.66	7.30	6.68	6.93	6.41

comes progressively larger with decreasing SNR. This is expected, since for a less noisy signal, spectral subtraction is likely to remove also 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.

In analyzing the results further we noticed that the energy-based VAD tends to produce unreliable results at low SNR (0 dB and -10 dB), by declaring most of the frames as speech. To exclude the detrimental effect of the (highly) errorneous VAD and focus on differences of spectrum estimation methods themselves, we performed another experiment on the noisiest (-10 dB) factory noise condition where the VAD labels were derived from the clean signal. The results in Table 4 confirm that the errorneous VAD labels are the main cause of degradation at the low SNRs; spectral subtraction can be seen as a "soft VAD". Interestingly, combination of spectral subtraction and "non-cheating VAD" appears to be the best combination. Further research is required to find good combination of speech enhancement and voice activity detection for non-stationary noises. Comparing the spectrum estimation methods in Table 4, SWLP remains the best method irrespective of the chosen VAD and spectral subtraction.

6. Conclusions

We studied temporally weighted linear predictive features in speaker verification. Without speech enhancement, the new WLP and SWLP features outperformed standard FFT and LP features in recognition experiments under additive noise conditions. The usefulness of robust voice activity detector and spectral subtraction in highly noisy environments was also demonstrated. Overall, the SWLP remained the most robust method. The temporally weighted linear predictive features are a promising approach for speaker recognition in the presence of additive noise.

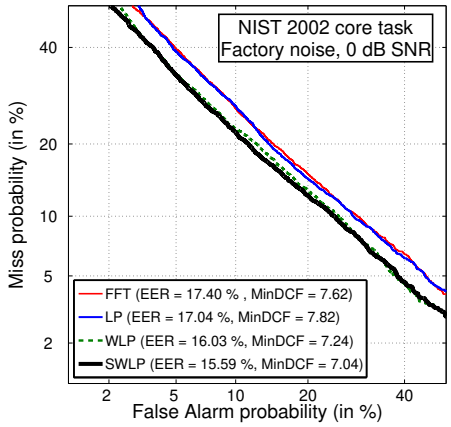


Figure 6: Comparing the features without any speech enhancement.

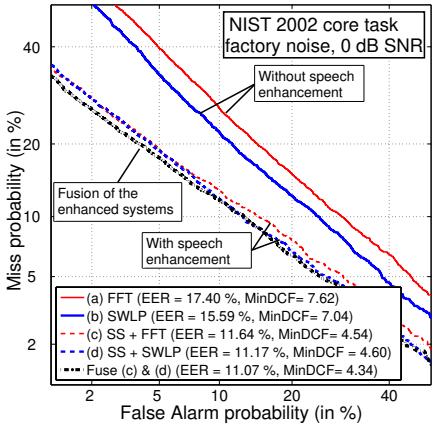


Figure 7: Comparing FFT and SWLP with and without speech enhancement. Feature-level fusion of the enhanced systems is also shown (SS = Spectral Subtraction).

7. Acknowledgment

This work is supported partly by a scholarship from the Finnish Foundation for Technology Promotion (TES) and Academy of Finland, projects no: 132129, 127345, 135003 (Lastu programme). The speaker recognition experiments were performed using computing resources from CSC (http://www.csc.fi/english) under the project no uef4836.

8. References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: from features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 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*, vol. 10, no. 1, pp. 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*, vol. 13, no. 5, pp. 308–311, May 2006.

[4] J. Makhoul, "Linear prediction: a tutorial review," *Proceedings of the IEEE*, vol. 64, no. 4, pp. 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*, vol. 15, no. 4, pp. 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*, vol. 7, no. 4, pp. 281–292, October 2004.

[8] H. Hermansky and N. Morgan, "RASTA processing of speech," *IEEE Trans. on Speech and Audio Processing*, vol. 2, no. 4, pp. 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)*, Crete, Greece, June 2001, pp. 213–218.

[10] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing*, vol. 10, no. 1-3, pp. 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)*, Las Vegas, U.S.A., April 2008, pp. 4837–4840.

[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*, vol. 15, no. 5, pp. 1711–1723, July 2007.

[13] S. G. Pillay, A. Ariyaeeinia, M. Pawlewski, and P. Sivakumaran, "Speaker verification under mismatched data conditions," *IET Signal Processing*, vol. 3, no. 4, pp. 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*, vol. 2, no. 4, pp. 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*, vol. 3, no. 2, pp. 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*, vol. 6, no. 3, pp. 260–267, 1998.

[17] 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.

[18] C. Magi, J. Pohjalainen, T. Bäckström, and P. Alku, "Stabilised weighted linear prediction," *Speech Communication*, vol. 51, no. 5, pp. 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*, Brighton, UK, 2009, pp. 1315–1318.

[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*, vol. 17, no. 2, pp. 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*, vol. 17, no. 4, pp. 748–757, May 2009.

[22] W.M. Campbell, J.P. Campbell, D.A. Reynolds, E. Singer, and P.A. Torres-Carrasquillo, "Support vector machines for speaker and language recognition," *Computer Speech and Language*, vol. 20, no. 2-3, pp. 210–229, April 2006.

[23] T. Kinnunen, V. Hautamäki, and P. Fränti, "Fusion of spectral feature sets for accurate speaker identification," in *Proc. 9th Int. Conf. Speech and Computer (SPECOM 2004)*, St. Petersburg, Russia, September 2004, pp. 361–365.