Using group delay functions from all-pole models for speaker recognition

Padmanabhan Rajan [1], *Tomi Kinnunen* [1], *Cemal Hanilci*[2], *Jouni Pohjalainen*[3] *and Paavo Alku*[3]

[1]School of Computing, University of Eastern Finland, Finland
[2] Dept. of Electronic Engineering, Uludag University, Turkey
[3] Dept. of Signal Processing and Acoustics, Aalto University, Finland

{paddy,tkinnu}@cs.uef.fi, chanilci@uludag.edu.tr, jpohjala@acoustics.hut.fi, paavo.alku@aalto.fi

Abstract

Popular features for speech processing, such as mel-frequency cepstral coefficients (MFCCs), are derived from the short-term magnitude spectrum, whereas the phase spectrum remains unused. While the common argument to use only the magnitude spectrum is that the human ear is phase-deaf, phase-based features have remained less explored due to additional signal processing difficulties they introduce. A useful representation of the phase is the group delay function, but its robust computation remains difficult. This paper advocates the use of group delay functions derived from parametric all-pole models instead of their direct computation from the discrete Fourier transform. Using a subset of the vocal effort data in the NIST 2010 speaker recognition evaluation (SRE) corpus, we show that group delay features derived via parametric all-pole models improve recognition accuracy, especially under high vocal effort. Additionally, the group delay features provide comparable or improved accuracy over conventional magnitude-based MFCC features. Thus, the use of group delay functions derived from all-pole models provide an effective way to utilize information from the phase spectrum of speech signals.

Index Terms: speaker verification, group delay functions, high vocal effort

1. Introduction

Feature extraction, or front-end, is a critical component in any speech processing system. Typically, a spectral representation (usually via the discrete Fourier transform, DFT) is computed from short-term frames and converted into feature vectors, such as cepstral coefficients. In general, the Fourier transform of a signal is complex, with magnitude and phase spectrum components. Most parametric representations of speech have been derived from the magnitude spectrum. But several studies (for example, [1]) have shown that the phase spectrum contains important information, contributing to speech intelligibility. Moreover, information complementary to the magnitude spectrum may be obtained from the phase spectrum, which may be useful in the classification of speakers or of sound units.

Extracting features from the phase spectrum is challenging due to various reasons, including signal processing difficulties. Although there has been considerable interest in utilizing features derived from the phase spectrum, seamless processing of the phase spectrum has proved elusive [1]. Various methods of processing the phase spectrum have been proposed, including those based on the *group delay function* [2, 3, 4], *instantaneous frequency* [5] and *inter-frame phase difference* [6]. The group delay function is defined as the negative derivative of the phase spectrum. This paper revisits the use of group delay functions for robust feature extraction in speaker recognition.

As explained later in this paper, the presence of spurious high-amplitude spikes in the group delay function makes its processing difficult. In this paper, we tackle this problem by utilizing the group delay function from *all-pole models* of speech, formed by linear predictive analysis. Temporally weighted variants of linear prediction provide robust parameterizations of the speech signal, and the group delay functions from these also can be used to derive features for speaker recognition.

Although group delay functions from all-pole models have been utilized earlier for formant extraction in [7], they have not been previously investigated for speaker recognition. Most studies on group delay based features for speaker recognition have used a non-parametric approach, computing the group delay function directly from the signal. Hence, the goal of the present study is to utilize features from group delay functions of parametric all-pole models of speech signals.

Another motivation is to investigate the effect of group delay features when the train and test speech are mismatched in terms of vocal effort. In noisy environments, the Lombard effect affects several speech production parameters, including vocal effort, pitch, spectral shape and formant locations [8]. Vocal effort mismatch leads to considerable degradation of speaker recognition performance [9].

2. Group delay representations for speech

If $x(n)$ is a frame of speech, its Fourier transform $X(\omega)$ can be written in polar form as

$$X(\omega) = |X(\omega)|e^{j\theta(\omega)} \tag{1}$$

where $|X(\omega)|$ is the magnitude spectrum and $\theta(\omega)$ is the phase spectrum. The group delay function $\tau(\omega)$ is the negative derivative of the continuous phase function,

$$\tau(\omega) = -\frac{d}{dw}\theta(\omega) \tag{2}$$

Early studies on the group delay function have indicated that it appears as a squared-magnitude response curve [7]. In other words, as in the magnitude spectrum, resonances of the vocal tract appear as peaks in the group delay spectrum. Moreover, a multiplication in the magnitude spectrum domain becomes an addition in the group delay domain. Due to this additive property, each resonance peak in the group delay function has little influence on the other peaks. Thus closely spaced formants are better resolved in the group delay domain than in the magnitude domain. This property has been utilized for formant extraction in [7].

This work was supported by Academy of Finland (proj. 253120).

The group delay function can be directly obtained from the speech signal $x(n)$ as derived in [10, 11]:

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2}, \tag{3}$$

where $x(n) \leftrightarrow X(\omega)$ and $y(n) \leftrightarrow Y(\omega)$ are Fourier transform pairs, and $y(n) = nx(n)$. The advantage of the above expression is that no explicit phase unwrapping is required.

2.1. Issues in processing the group delay function

Processing of group delay functions is not straightforward [12]. The presence of zeros of the vocal tract system function (in the Z-transform representation) can lead to an ill-behaved group delay function. It can be seen that a zero (or dip in the spectrum) will result in a small value of the denominator of Equation 3, leading to an indefinitely large value of the group delay function. These zeros can occur as a result of the excitation source, and can also be an artifact of short-term processing [4]. Computation of the group delay function at frequency bins near these zeros thus results in high amplitude spurious peaks, masking out the formant structure [4, 13].

Several methods have been proposed to mitigate the effect of the zeros. The *product spectrum* was proposed in [3], where the denominator term is canceled by multiplying Equation 3 by the power spectrum. The product spectrum thus utilizes information from both the magnitude and phase spectra. The *chirp group delay function* [4] avoids the zeros near the unit circle by evaluating the spectrum on a circle other than the unit circle.

The *modified group delay function* (MODGDF) was proposed in [2]. This method modifies Equation 3 as,

$$\tau_m(\omega) = \text{sign}\, |\tau'(\omega)|^{\alpha}, \tag{4}$$

where,

$$\tau'(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|S(\omega)|^{2\gamma}.} \tag{5}$$

Here, $|S(\omega)|$ is a cepstrally smoothed version of $|X(\omega)|$. Following the notation in [13], the parameters α and γ are introduced to control the dynamic range of the MODGDF, and the window length for the cepstral smoothing is denoted as lifter_{ω}. Features derived from the MODGDF have been used in speech and speaker recognition in [13]. To restrict the dynamic range of the MODGDF without the parameters α and γ, log compressed group delay features were proposed in [11].

2.2. Group delay function of all-pole models

Linear prediction analysis of speech approximates the speech spectrum using an all-pole model [14]. Considering the vocal tract as an all-pole filter allows an equivalent representation as a cascade of several second-order and first order all-pole filters [7]. In this representation, the overall magnitude spectrum is the product of the magnitude spectra of the individual filters. The overall phase spectrum (and hence the group delay spectrum), on the other hand, is a summation of the individual phase spectra.

Denoting

$$H(\omega) = \frac{G}{1 - \sum_{k=1}^{p} a(k) e^{-j\omega k}} \tag{6}$$

the linear prediction formulation can be specified as follows [14]. Given the speech power spectrum $|X(\omega)|^2$, determine the set of coefficients $a(k)$ such that the power spectrum of $H(\omega)$ matches the speech power spectrum in a least-squared sense. Here, G is the signal dependent gain and p is the model order. The filter formed by $H(\omega)$ has both a magnitude response and a phase response. The all-pole group delay function is defined as the group delay function of this filter. Figure 1 shows a frame of speech, its magnitude spectrum and its all-pole group delay function, formed by linear prediction analysis with a model order $p = 20$. Due to the high resolution property of the group delay function, closely spaced higher formants can be captured, unlike in the magnitude spectrum. For speaker recognition, since we are interested only in the resonances of the filter $H(z)$, the gain G is usually set to unity for normalization purposes.

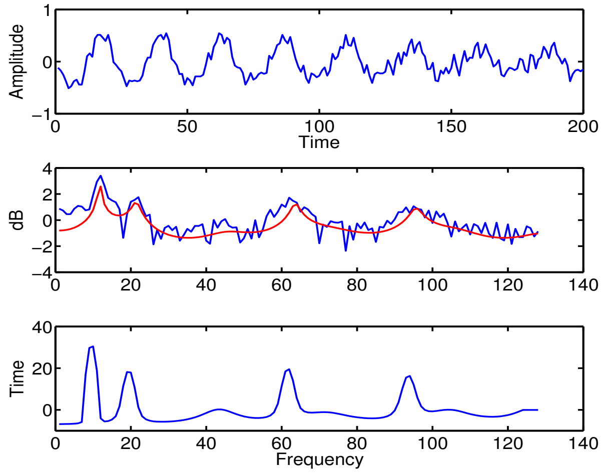

Figure 1: A frame of speech (top panel), corresponding log magnitude spectra (both DFT and all-pole) (middle panel) and all-pole group delay function (bottom panel); model order $p = 20$.

Any speech system function $X(z)$ can be decomposed into *minimum-phase* and *all-pass components* [15]

$$X(z) = X_{\text{min}}(z)\, X_{\text{ap}}(z) \tag{7}$$

The corresponding magnitude spectra and phase spectra are

$$|X(\omega)| = |X_{\text{min}}(\omega)|\, |X_{\text{ap}}(\omega)| \tag{8}$$

$$\theta(\omega) = \theta_{\text{min}}(\omega) + \theta_{\text{ap}}(\omega) \tag{9}$$

When the linear prediction equations are solved by the covariance method, the all-pole filter $H(\omega)$ in Equation 6 is a minimum phase filter [14]. Thus, with respect to the phase response in Equation 9, the filter keeps only the minimum-phase part. This is the loss incurred while working with the phase of the all-pole filter. Our experiments demonstrate that, in spite of this lossy representation, the all-pole group delay function has valuable information which makes it suitable for speaker recognition.

The high resolution property of the group delay spectrum (shown for different vocal efforts in Figure 2) may be helpful in capturing formant information in a more robust manner, compared to the magnitude spectrum. The higher order formants are more pronounced in the group delay spectrum, particularly

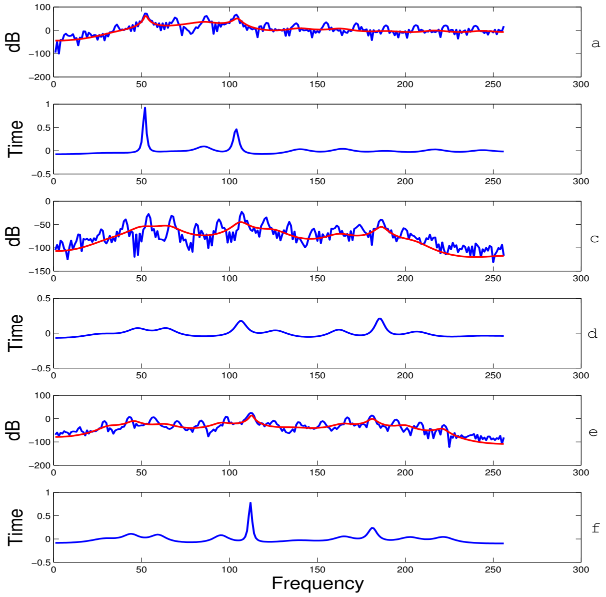

Figure 2: Spectra under different vocal effort for the vowel /a/ spoken by the same female speaker. (a) Log magnitude and all-pole spectrum under normal vocal effort. (b) Group delay spectrum under normal vocal effort. (c), (d) low vocal effort. (e), (f) high vocal effort.

in the low and high vocal effort conditions.

3. Temporally weighted linear prediction

In noisy conditions, the performance of the standard linear prediction model deteriorates and hence, noise-robust versions have been proposed. *Weighted linear prediction* (WLP) applies temporal weighting to the squared residual error during computation of the all-pole filter coefficients [16]. Weighting is applied to emphasize the contribution of samples with higher signal-to-noise ratio (SNR) and de-emphasize samples with low SNR. In [16], the WLP coefficients $a(k)$ are obtained by minimizing the weighted residual energy,

$$E = \sum_{n} \left(x(n) - \sum_{k=1}^{p} a(k)x(n-k) \right)^2 W(n), \qquad (10)$$

where $W(n)$ is the weighting function. The coefficients $a(k)$ are obtained by setting the partial derivatives of E to zero, and solving for each $a(k)$. The WLP normal equations are

$$\begin{aligned} \sum_{k=1}^{p} a(k) \sum_{n} W(n)x(n-k)x(n-i) = \\ \sum_{n} W(n)x(n)x(n-i) \quad 1 \leq i \leq p \end{aligned} \qquad (11)$$

Unlike the autocorrelation method of standard linear prediction analysis, the all-pole model produced by WLP is not guaranteed to be stable. To mitigate this, *stabilized weighted linear prediction* was developed in [17]. Writing the WLP normal equations in terms of partial weights $Z_{n,j}$, we have

$$\begin{aligned} \sum_{k=1}^{p} a(k) \sum_{n} Z_{n,k}x(n-k)Z_{n,i}x(n-i) = \\ \sum_{n} Z_{n,0}x(n)Z_{n,i}x(n-i) \quad 1 \leq i \leq p \end{aligned} \qquad (12)$$

Here, $Z_{n,j} = \sqrt{W(n)}$ for $0 \leq j \leq p$. Typically the temporal weighting function $W(n)$ used in WLP and SWLP is the *short term energy* (STE), computed using a sliding window of M samples around the region of interest as $W(n) = \sum_{i=1}^{M} x^2(n-i)$. A comparison of spectrum estimators from various methods, including WLP and SWLP was made in [18]. Studies have shown improved performance in speech recognition [19] and speaker recognition [20] using magnitude-spectrum based features derived from WLP and SWLP.

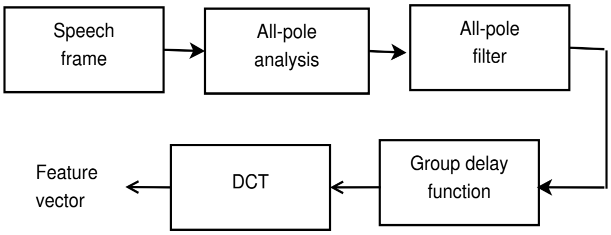

Figure 3: Feature extraction from group delay functions derived from different types of all-pole models.

It is reasonable to assume that the coefficients $a(k)$ obtained from the temporally weighted linear prediction methods provide a more robust representation of the vocal tract filter $H(\omega)$. As in the case of standard linear prediction, the group delay function from these weighted variants can also be utilized to derive features. Collectively, we term these group delay functions as *all-pole group delay functions*.

4. Speaker verification experiments

We apply features derived from all-pole group delay functions to the vocal effort subconditions of the NIST 2010 speaker recognition evaluation (SRE). On account of producing stable filters, the all-pole models considered are the standard linear prediction (denoted LP) and stabilized weighted linear prediction (denoted SWLP). The performance of all-pole group delay features are compared to features from non-parametric group delay representations, namely the standard group delay function (Equation 3), and the modified group delay function (Equation 4). Since several features are being evaluated in the study, due to its simplicity and fast turnaround time, the classical Gaussian mixture model with universal background model (GMM-UBM) [21] speaker recognition system is used for our experiments. Additionally, a state-of-the-art verification system using the i-vector representation [22] is also used to evaluate group delay based features.

4.1. Converting the group delay function into features

Similar to [2], the all-pole group delay function is converted into cepstral coefficients by applying the discrete cosine transform (DCT). The feature extraction process is shown in Figure 3 and is described as follows.

Features are extracted from a frame of speech (frame size =

30 ms, shift = 15 ms), as follows:

1. Perform all-pole modeling (either LP or SWLP) on the frame of speech with prediction order $p = 20$. For SWLP, the short term energy duration M is set to 20 samples as in [20]. Obtain the all-pole filter coefficients $a(k)$.
2. From the $a(k)$, form the frequency response $H(\omega)$ using Equation 6 with $G = 1$.
3. Compute the group delay function by taking the negative derivative of the phase response of $H(\omega)$. In practice, the derivative is computed using the sample-wise difference.
4. Take DCT on the group delay function and keep the first 18 cepstral coefficients, excluding the zeroth coefficient.
5. Velocity and acceleration coefficients are appended to the feature in a conventional manner, to from a 54 dimensional feature vector.

Additionally, mean and variance normalization are applied to the features on an utterancewise basis.

4.2. Speaker verification setup

A 128-mixture GMM-UBM system was utilized, where the UBM utilized telephone data from NIST SRE 2004, 2005 and 2006 corpora. Speaker models are adapted from the UBM using maximum *a posteriori* estimation [21]. Since the data was telephone speech, a simple energy-based voice activity detector was used to remove non-speech frames. The evaluation dataset was the vocal effort conditions of the NIST 2010 (conditions 5, 6 and 8) and is summarized in Table 1 [23].

Table 1: Chosen sub conditions of NIST SRE 2010 [23], consisting of telephone data only. VE = vocal effort.

NIST condition	Train VE	Test VE
CC 5	Normal	Normal
CC 6	Normal	High
CC 8	Normal	Low

Speaker verification accuracy in terms of equal error rate (EER), is evaluated for features derived from parametric all-pole group delay functions (LP or SWLP). Features from non-parametric group delay functions (standard group delay function, MODGDF) are also evaluated. Features from the standard group delay function were obtained by taking a DCT on Equation 3. Finally, the conventional magnitude based mel-frequency cepstral coefficients (MFCC) were also evaluated, to compare how group delay based methods compare to them.

4.3. Results

Results for experiments with the GMM-UBM system are shown in Table 2. It can be seen that features from the standard group delay function do not fare well in speaker recognition. The presence of large spikes in the group delay spectrum masks out the formant information, rendering the unprocessed group delay function useless as a feature. The MODGDF features perform much better, but are sensitive to its parameters α, γ and lifter$_\omega$. The best results we obtained (shown in table 2) were for the values $\alpha = \gamma = 0.1$ and lifter$_\omega = 8$. The parametric models performed better than the non-parametric models, with the standard LP showing higher performance than SWLP. Although both LP and SWLP generate minimum-phase filters, the weighting function used in SWLP possibly introduces some phase distortion. Finally, for the normal and low vocal effort conditions, the magnitude-based MFCC features gave results comparable to the parametric methods. The MFCC features work poorly for the high vocal effort condition, which is more pronounced in the female case.

As a final note, the standard LP group delay features were also evaluated using a state-of-the-art speaker verification system using the i-vector representation for utterances [22] and the Gaussian probabilistic linear discriminant analysis (G-PLDA) classifier [24, 25]. For this setup, 600 dimension length normalised i-vectors were estimated for each utterance. The i-vector extractor (or T-matrix) was estimated using data from NIST SRE 2004, 2005, 2006, Fisher and Switchboard data. The PLDA model was also trained using the same data, utilising 200 dimensions for the speaker subspace and a full-covariance residual term as in [26]. The results of the i-vector PLDA system are presented in Table 3. From these results, we see that group delay features perform well in a state-of-the-art system.

Table 2: Results of speaker verification experiments (GMM-UBM) in terms of EER on the SRE 10 vocal effort conditions.

	Male			Female		
Feature	CC 5	CC 6	CC 8	CC 5	CC 6	CC 8
Group delay based feature						
Std. gdelay	32.9	42.1	32.0	46.9	46.4	44.6
MODGDF	19.7	25.9	13.7	25.0	30.1	20.6
SWLP	15.3	21.3	11.7	16.1	23.6	11.5
LP	11.4	22.1	9.9	15.4	22.8	12.1
Magnitude based feature						
MFCC	13.0	24.0	7.5	16.9	34.9	11.7

Table 3: Verification accuracy in terms of EER, using LP group delay features and MFCC features on an i-vector G-PLDA system.

	Male			Female		
Feature	CC 5	CC 6	CC 8	CC 5	CC 6	CC 8
LP	1.71	3.24	0.28	3.20	5.11	1.03
MFCC	1.78	3.66	0.35	3.17	4.43	1.96

5. Conclusions

In this paper, we investigated the use of phase-based features in speaker recognition. Using both parametric and non-parametric approaches, features from the phase spectrum were derived using the group delay function. Speaker verification experiments on the vocal effort conditions of the NIST 2010 SRE dataset demonstrate that group delay features perform comparable to conventional magnitude spectrum-based MFCC features. In particular, the experiments also reveal that the group delay methods are useful under high vocal effort. Utilizing parametric all-pole models provide an effective mechanism to extract information from group delay functions, which otherwise suffer from signal processing difficulties. Thus, group delay features from all pole models can be used to effectively process phase information for speaker recognition.

Further work in this direction will look at the performance of parametric methods in the presence of noise. Group delay functions derived from the theoretically noise-robust SWLP model are good candidates for investigation as features for processing noisy speech.

6. References

[1] L. D. Alsteris and K. K. Paliwal, “Short-time phase spectrum in speech processing: a review and some experimental results,” *Digital Signal Process.*, vol. 17, pp. 578–616, 2006.

[2] H. A. Murthy and V. R. R. Gadde, “The modified group delay function and its application to phoneme recognition,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, 2003, pp. 68–71.

[3] D. Zhu and K. Paliwal, “Product of power spectrum and group delay function for speech recognition,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2004, pp. 125–128.

[4] B. Bozkurt, L. Couvreur, and T. Dutoit, “Chirp group delay analysis of speech signals,” *Speech Commun.*, vol. 49, pp. 159–176, 2007.

[5] Y. Wang, J. Hansen, G. K. Allu, and R. Kumaresan, “Average instantaneous frequency AIF and average log-envelopes ALE for asr with the aurora 2 database,” in *Proc. Eurospeech*, 2003, pp. 25–28.

[6] I. McCowan, D. Dean, M. McLaren, R. Vogt, and S. Sridharan, “The delta-phase spectrum with application to voice activity detection and speaker recognition,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 7, pp. 2026–2038, 2011.

[7] B. Yegnanarayana, “Formant extraction from linear-prediction phase spectra,” *Jnl. Acoust. Soc. Amer.*, vol. 63, no. 5, pp. 1638–1640, 1978.

[8] H. Boril and J. H. L. Hansen, “Unsupervised equalization of lombard effect for speech recognition in noisy adverse environments,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 18, no. 6, pp. 1379–1393, 2010.

[9] C. S. Greenberg, A. F. Martin, B. N. Barr, and G. R. Doddington, “Report on performance results in the NIST 2010 speaker recognition evaluation,” in *Proc. Interspeech*, 2011.

[10] H. Banno, J. Lu, S. Nakamura, K. Shikano, and H. Kawahara, “Efficient representation of short-time phase based on group delay,” in *Proc. Int. Conf. Acoust. Speech Signal Process.*, vol. 2, 1998, pp. 861–864.

[11] T. Thiruvaran, E. Ambikairajah, and M. Epps, “Group delay features for speaker recognition,” in *Int. Conf Information, Communications Signal Process.*, 2007, pp. 1–5.

[12] H. A. Murthy and B. Yegnanarayana, “Speech processing using group delay functions,” *Signal Processing*, vol. 22, no. 3, pp. 259–267, 1991.

[13] R. M. Hegde, H. A. Murthy, and V. R. Gadde, “Significance of the modified group delay feature in speech recognition,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, pp. 190–202, Jan. 2007.

[14] J. Makhoul, “Linear prediction: a tutorial review,” *Proc. IEEE*, vol. 64, no. 4, pp. 561–580, 1975.

[15] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time Signal Processing*. Prentice-Hall, 2000.

[16] C. Ma, Y. Kamp, and L. F. Willems, “Robust signal selection for linear prediction analysis of voiced speech,” *Speech Commun.*, vol. 12, no. 1, pp. 69–81, 1993.

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

[18] C. Hanilci, T. Kinnunen, R. Saeidi, J. Pohjalainen, P. Alku, F. Ertas, J. Sandberg, and M. Hansson-Sandsten, “Comparing spectrum estimators in speaker verification under additive noise degradation,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2012, pp. 4769–4772.

[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*, 2009, pp. 1315–1318.

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

[21] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, “Speaker verification using adapted Gaussian mixture models,” *Digital Signal Process.*, vol. 10, pp. 19–41, 2000.

[22] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Trans. Audio, Speech Lang. Process.*, vol. 19, no. 4, pp. 788–798, 2011.

[23] “The NIST year 2010 speaker recognition evaluation plan,” 2010, http://www.itl.nist.gov/iad/mig/tests/sre/2010/.

[24] S. J. D. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *Proc. ICCV*, 2007, pp. 1–8.

[25] P. Kenny, “Bayesian speaker verification with heavy tailed priors,” in *Proc. Odyssey*, 2010.

[26] D. Romero and C. Espy-Wilson, “Analysis of i-vector length normalisation in speaker recognition systems,” in *Proc. Interspeech*, 2011.