Combining Amplitude and Phase-based Features for Speaker Verification with Short Duration Utterances

Md Jahangir Alam, Patrick Kenny, Themos Stafylakis

CRIM, Montreal (QC) Canada

{jahangir.alam, patrick.kenny, themos.stafylakis}@crim.ca

10.21437/Interspeech.2015-94

Abstract

Due to the increasing use of fusion in speaker recognition systems, one trend of current research activity focuses on new features that capture complementary information to the MFCC (Mel-frequency cepstral coefficients) for improving speaker recognition performance. The goal of this work is to combine (or fuse) amplitude and phase-based features to improve speaker verification performance. Based on the amplitude and phase spectra we investigate some possible variations to the extraction of cepstral coefficients that produce diversity with respect to fused subsystems. Among the amplitude-based features we consider widely used MFCC, Linear frequency cepstral coefficients, and multitaper spectrum estimation-based MFCC (denoted here as MMFCC). To compute phase-based features we choose modified group delay- and all-pole group delay-, linear prediction residual phase-based features. We also consider product spectrum-based cepstral coefficients features that are influenced by both the amplitude and phase spectra. For performance evaluation, text-dependent speaker verification experiments are conducted on the a proprietary dataset known as Voice Trust-Pakistan (VT-Pakistan) corpus. Experimental results show that the fused system provide reduced error rate compared to both the amplitude and phase-based features. On the average fused system provided a relative improvement of 37% over the baseline MFCC systems in terms of EER, DCF (detection cost function) of SRE 2008 and DCF of SRE 2010.

Index Terms: speaker verification, modified group delay, product spectrum, LP residual, GMM-UBM

1. Introduction

Diversity can be used as a means for improving the performance of speaker recognition systems comprising subsystems incorporating different configurations (different front-ends and/or different back-ends). This enables the integration of different information when fused at the score or feature level. In the speaker recognition context, combination of such information is typically known as fusion. In fusion, scores from different models trained for a speaker are combined when fused at the score level, or features from different front-ends are combined when fused at the feature level [1, 3, 33]. Examples of this kind of variation in front-ends includes the use of different feature extractors and/or voice activity detectors (VADs) across systems; and in back-ends includes the use of different classifiers and/or compensation techniques [2, 3] across systems [3]. In this work, our aim is to combine amplitude spectrum-, phase spectrum-, and joint amplitude-phase-based front-ends at score level to incorporate complementary information for improving speaker verification accuracy.

In the majority of speech processing applications such as speaker/speech recognition systems and speech enhancement, cepstral features are always computed from short-time amplitude spectra. Phase related information is generally discarded. This is because human ears are considered largely insensitive to phase and so, speech communication and recording equipment do not preserve original waveform's phase structure [16]. Another reason is phase wrapping, a key problem with the phase spectrum. This results in an intractable, noise-like, and chaotic shape lacking any informative trend. A solution to this problem is to unwrap the phase using unwrapping methods [4-5]. Another solution is to work with phase-derived representations such as the group delay function (GDF). Short-time phase spectra (analysis frame length 20 to 40 ms) do not contain much intelligibility information. Although increasing analysis frame length results in increasing information content [6-12] the stationary assumption no longer remain valid due to the relative non-stationary nature of speech.

Despite these drawbacks, various methods for processing the phase spectrum have been proposed. They include, group delay function[13-15], all-pole group delay function [16-17], instantaneous frequency [18] and inter-frame phase difference-based methods [19]. It has been shown in [20-21] that phase-based features provide good performance when fused with MFCC-based systems.

In this paper, we combine amplitude spectrum-based features with phase spectrum-based features and amplitude plus phase spectrum-based features to obtain improved speaker verification performance. The MFCC, LFCC (Linear frequency cepstral coefficients) and multitaper MFCC (MMFCC) belong to the amplitude-based feature category. The phase-based features are: MGDCC (modified group delay cepstral coefficients), LP-GDCC (linear prediction-group delay cepstral coefficients), SWLP-GDCC (stabilized weighted LP-GDCC), and LPRPC (LP residual phase cepstra). Product spectrum-based MFCC (PS-MFCC) is in the joint amplitude-phase-based feature category.

2. Amplitude-Based Features

For amplitude-based features we choose conventional MFCC, LFCC (both computed from the DFT-based direct power spectrum), and multitaper MFCC, i.e., MFCCs computed from a multiple windowed direct power spectrum. The generalized form for the estimation of a windowed direct power spectrum can expressed as [26]:

$$X_p(\omega) = \sum_{t=1}^{T} \lambda(t) \left| \sum_{i=0}^{N-1} w_t(i) x(i) e^{-j\frac{i\omega}{N}} \right|^2,$$

where N is the frame length and $w_t(i)$ is the t-th data taper ($t = 1, 2, ..., T$) used for the spectral estimate $X_p(\omega)$, also known as the t-th *eigenspectrum*, T denotes the number of

tapers and $\lambda(t)$ is the weight of the *t*-th taper. For Hamming windowed direct spectrum estimates $t = T = 1$, $\lambda(t) = 1$, $w_t(i)$ is the Hamming window. For $t > 1$, i.e., for the multitaper method, the tapers $w_t(i)$ are typically chosen to be orthonormal so that, for all t_1 and t_2,

$$\sum_i w_{t_1}(i) w_{t_2}(i) = \delta_{pq} = \begin{cases} 1, & t_1 = t_2 \\ 0, & \text{otherwise.} \end{cases}$$

In this work we choose the Hamming windowed direct spectrum estimator and Thomson multitaper spectrum estimator for the computation of amplitude-based MFCC features. For the Thomson method, the orthonormal tapers with $T = 6$ can be generated using the *MATLAB* built in function *dpss* as follows:

$$[\mathbf{w}\ \boldsymbol{\lambda}] = \text{dpss}(N, 3.5, T);$$

Figure 1 presents a block diagram to extract MFCC, LFCC and MMFCC features from the single taper and multitaper spectra.

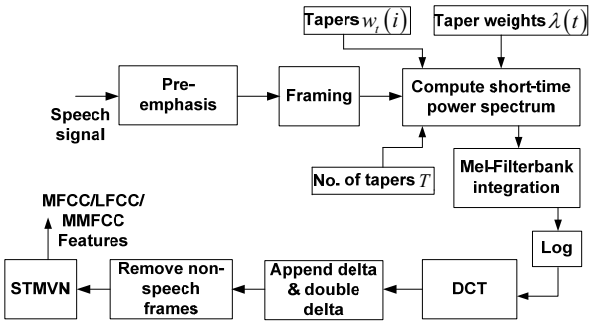

Figure 1: General block diagram showing various steps to extract Amplitude spectrum-based Mel-frequency cepstral coefficients (MFCC), linear frequency cepstral coefficients (LFCC), and multitaper MFCC (MMFCC) features. For MFCC and LFCC, $t = T = 1$, $\lambda(t) = 1$, and $w_t(i)$ is the Hamming window. For LFCC a linear filterbank is used instead of a Mel-scale filterbank.

3. Phase-Based Features

3.1. Feature based on Modified Group Delay

The Fourier transform $X(\omega)$ of a speech signal $x(n)$ in polar form can be expressed a

$$X(\omega) = Ae^{j\theta(\omega)}, \tag{1}$$

where $A = |X(\omega)|$ is the amplitude and $\theta(\omega)$ is the phase of $X(\omega)$. The group delay function can be defined as the negative derivative of the phase spectrum and can be expressed mathematically, as:

$$\tau_g(\omega) = -\frac{d}{d\omega}\left(\theta(\omega)\right) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|X(\omega)|^2}, \tag{2}$$

where $Y(\omega)$ is the Fourier transform of $y(n) = nx(n)$, and the subscripts *R* and *I* denote the real and imaginary parts, respectively.

The group delay function is well-behaved only if the zeros of the system transfer function are not close to the unit circle [13, 22, 23]. The standard modification [13] of the group delay function is performed by suppressing the zeros of the transfer function. This is done by replacing the magnitude spectrum $X(\omega)$ by its cepstrally smoothed version $S(\omega)$ and two parameters α ($0 < \alpha \le 1$) and γ ($0 < \gamma \le 1$) are introduced to control the dynamic range. The modified function is defined as

$$\tau_m(\omega) = \frac{\tau(\omega)}{|\tau(\omega)|}\left(|\tau(\omega)|^{\alpha}\right), \tag{3}$$

where

$$\tau(\omega) = \frac{X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)}{|S(\omega)|^{2\gamma}}. \tag{4}$$

$P(\omega) = X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega)$ is known as the product spectrum. Cepstrally smoothed spectra $S(\omega)$ are obtained using the following the steps [24]:

- Compute the log amplitude spectra from $X(\omega)$ and apply median filter for smoothing the log spectra.
- Apply a DCT to the log spectra and take the first 30 cepstral coefficients.
- Apply the inverse DCT to the cepstral coefficients to obtain cepstrally smooth spectra $S(\omega)$.

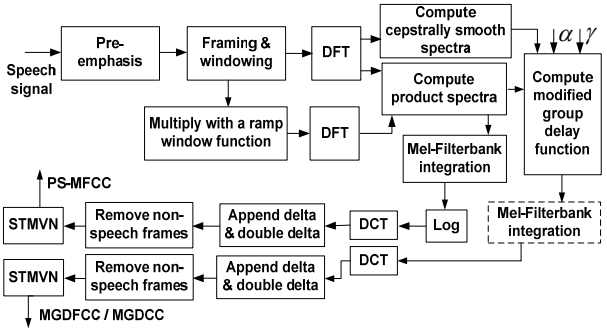

Figure 2: Computation of Product spectrum-based MFCC and modified group delay-based cepstral coefficients (MGDCC or MGDFCC (when a filterbank is used) features from speech signal.

Fig. 2 shows the complete block diagram to compute the modified group delay function (MGDF) and modified group delay cepstral coefficients (MGDCC) or modified group delay Mel-filterbank cepstral coefficients (MGDFCC) from it. After computing $P(\omega)$ and $S(\omega)$, the modified group delay function (MGDF) is obtained using eqns. (3)-(4). Apply a DCT to the MGDF and take the first *q* coefficients (excluding c0) to obtain MGDCC features. Append delta and double delta features. After removing the non-speech frames using the VAD label files, features are normalized using a short-time mean and variance normalization (STMVN) technique with a window of 1.5s. In this work we experimentally choose $\alpha = \gamma = 0.1$.

3.2. Product Spectrum-Based MFCC

The product spectrum was first introduced in [14] to mitigate the effect of zeros in the group delay function. The product spectrum $P(\omega)$, the product of power spectrum $|X(\omega)|^2$ and the group delay function $\tau_g(\omega)$, can be expressed as:

$$P(\omega) = |X(\omega)|^2 \tau_g(\omega) = X_R(\omega)Y_R(\omega) + X_I(\omega)Y_I(\omega). \quad (5)$$

Eqn. (5) indicates that the product spectrum incorporates information from both the amplitude spectrum and phase spectrum. Product spectrum features can be a good candidate for spoofing detection and speaker recognition. Figure 2 provides an overview of the MFCC feature extraction procedure from the product spectrum.

3.3. Cepstral Features from All-Pole Group Delay

In the case of a speech signal, some zeros can occur in the vicinity of the unit circle due to the excitation source and also due to an artifact of short-time processing [16, 17]. Computation of the group delay function using eqn, (2) at frequency bins near these zeros thus results in high amplitude spurious peaks, masking out the formant structure [15, 23]. Modified group delay [13], product spectrum [14], and chirp group delay [15], all-pole group delay [16-17] functions have been proposed to alleviate the problem associated with group delay i.e., eqn. (2). The idea behind all-pole group delay is to keep only the vocal tract (filter) component of the speech signal and discard the contribution due to excitation source. This can be approximated by extracting the spectral envelope of the speech signal via all-pole modeling. Fig. 4 presents different steps for the extraction of cepstral features from the all-pole group delay function.

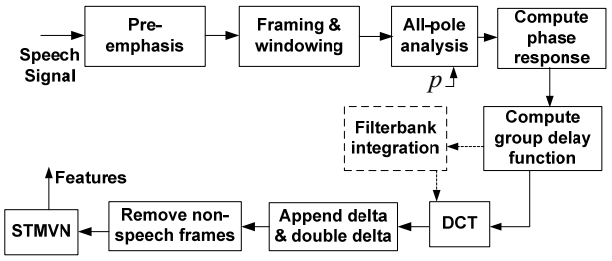

Figure 3: Computation of cepstral features from all-pole group delay function.

Once pre-processing (i.e., pre-emphasizing, framing and windowing) is done we perform all-pole modeling (e.g., linear prediction (LP) with a model order p and obtain the autoregressive (AR) coefficients $\mathbf{a} = \{a(k,m)\}$, $k = 1,2,...,p$; $m = 1,2,...,M$; k is the index for AR coefficients and m is the frame index. Compute the phase response from the AR coefficients **a**. Here, we use both the standard LP and stabilized weighted LP (SWLP) for all-pole analysis with prediction order $p = 12$. The group delay function is then calculated by taking the negative derivative of the phase response. Cepstral coefficients are obtained by applying a DCT to the group delay function. We keep the first q coefficients (here, q = 12) excluding the 0*th* cepstrum coefficient and append delta and double delta features to form $3q$ dimensional features. The STMVN technique is applied to normalize the features after removing the non-speech frames using the VAD label files. Note that no compression (logarithmic or power-law nonlinearity) is applied for computing cepstral features from the phase spectra or group delay function because multiplying the Fourier transform of two signals (e.g., source and filter) corresponds to addition of their phase spectra [16].

3.4. LP Residual Phase-Based Features

The change in the phase polarities is the main reason behind the rapid fluctuations in the amplitudes of the linear prediction (LP) residual signal. It has been demonstrated by speaker recognition experiments that the phase of the LP residual signal contains speaker specific information [31, 32]. Now the phase of the LP residual signal can be obtained as the cosine of the analytic signal phase function and can be given as:

$$\cos(\theta(n)) = \frac{r(n)}{\sqrt{r^2(n) + r_h^2(n)}}, \quad (6)$$

where LP residual $r(n)$ is the prediction error obtained as the difference between the predicted speech $\hat{x}(n)$ and the current speech sample as $r(n) = x(n) - \hat{x}(n) = x(n) - \sum_{k=1}^{p} a_k x(n-k)$, $r_h(n)$ is the Hilbert transform of $r(n)$, a_k is the k-th prediction coefficient. Cepstral coefficients are then obtained by applying a DCT to the LP residual phase and keeping the first q coefficients (here, q = 12) excluding the 0*th* cepstrum. We append delta and double delta features to form $3q$ dimensional features. The STMVN is applied to normalize the feature after removing nonspeech frames.

4. Experiments and Discussion

4.1. Corpora used

In order to train a gender-independent universal background model (UBM) following background data were used: RSR2015 [28], CSLU [27], and a proprietary dataset collected at the CONCORDIA University [29-30]. For evaluation we used a proprietary dataset, known as the Voice Trust-Pakistan (VT-Pakistan) corpus, collected over land line and cellular channels consisting of recordings of two common passphrases, one in ENGLISH and the other in URDU. In this work only ENGLISH data were used [29-30].

4.2. Experimental setup and Results

A Gaussian Mixture Model - Universal Background Model (GMM-UBM) framework [25] was used as the backend for the speaker verification task on the text-dependent corpus. A 512-component UBM with diagonal covariances was trained on all the background data (approximately 94 hours of speech). Nine variants of MFCC features computed from the amplitude and phase spectra were used for performance evaluation and for doing fusion in the score domain. The Feature dimension in each front-end is 36 (12 static + 12 delta + 12 double delta coefficients) without log energy or 0th cepstrum. If log energy or 0th cepstrum is included then the feature dimension is 39. The evaluation metrics used for performance evaluation are: Equal error rate (EER), DCF (detection cost function) of NIST SRE 2008 (DCF08) and DCF of NIST SRE 2010 (DCF10).
Speaker models were trained with an adapted UBM using three enrolment utterances per speaker from the VT-Pakistan corpus. For UBM adaptation we used 1 iteration with a relevance factor of 2. Each trained speaker model is scored against all the test utterances. For score normalization s-norm was used to reduce the variation in likelihood scores in making speaker verification decisions. Based on the nine variants of MFCC features nine systems and two fused systems denoted as **FUSED** and **FUSED5**, as mentioned in Table 1, were used

for text-dependent speaker verification task. The number of trials reported in Table 2. Results for experiments using different front-ends with the GMM-UBM framework and the results of the fused systems are shown in tables 3 & 4 for female and male, respectively. The MGDCC features are sensitive to the parameters α , γ and the cepstrally smooth spectra $S(\omega)$. It is observed from Table 2 that, if implemented correctly and if the parameters of the modified group delay function (MGDF) are tuned properly, phase-based cepstral features (e.g., MGDCC) can provide comparable results to those of the amplitude-based features, e.g., MFCC. Comparing the results of **MFCC** (1st row of table 2) and **MFCC_E** (2nd row of table 2) it can be seen that the inclusion of prosodic feature (e.g., log energy) with the cepstral coefficients helped to boost the speaker verification performance.

Table 1. Description of the features used in this work.

MFCC	36-dimensional MFCC features computed from the amplitude spectra (without c0 and log energy)
MFCC_E	39-dimensional MFCC features (with log energy)
MGDCC	36-dimensional modified group delay cepstral coefficients (MGDCC) features computed using MGDF (without c0 and log energy).
PS-MFCC	36-dimensional product spectrum - based MFCC features (without c0 and log energy).
LP-GDCC	36-dimensional Linear Prediction (p = 12) based group delay cepstral coefficients feature (without c0 and log energy).
SWLP-GDCC	36-dimensional Stabilized Weighted Linear Prediction (p = 12) based group delay cepstral coefficients feature (without c0 and log energy).
MMFCC	36-dimensional MFCC features computed from the multitaper amplitude spectra (without c0 and log energy)
LFCC	39-dimensional LFCC features computed from the amplitude spectra (with log energy)
LPRPC	39-dimensional LPRPC features computed from the LP residual phase spectra (with log energy)
FUSED	Fusion of above systems excluding the LPRPC
FUSED5	Fusion of **MFCC, MFCC_E, MGDCC, PS-MFCC,** and **LP-GDCC** systems

Table 2. Trial statistics for VT-Pakistan eval set per gender

gender	# target	# nontarget	gender	# target	# nontarget
male	579	122332	female	173	13861

Fusing amplitude and phase-based systems resulted in improved speaker verification performances both on male and female trials of the VT-Pakistan corpus. **FUSED5** system represents fusion of selected five best systems. Although group delay features from all pole models (e.g., LP-GDCC, SWLP-GDCC) can be used to effectively process phase information for speaker recognition, their performances is not as good as that of the MGDCC or MFCC features. After observing the results of fused system **FUSED5** it is evident that the all-pole model-based group delay features, for example LP-GDCC, modified group delay-based features and PS-MFCC features possessed complementary information to that of the amplitude-based features. Therefore, combining them at the score level resulted in a huge improvement of the speaker verification performance. The discrepancy between the male and female results in Tables 3 & 4 can be accounted for by the imbalance in the test set indicated in Table 2.

5. Conclusions

In this work we evaluated the performance of various amplitude-, phase-, and combined amplitude and phase -based cepstral features for text dependent speaker verification task. In order to capture complementary information we combined amplitude and phase -based features in the score level. Speaker verification experiments on the Voice Trust - Pakistan (VT-Pakistan) text-dependent corpus demonstrated that phase features perform comparable to conventional magnitude spectrum-based MFCC features. It also observed that the fusion of amplitude and phase-based features yielded improved speaker verification performance with a relative improvement of 37%.

Table 3. Speaker verification results on female trials of the ENGLISH part of the VT-Pakistan corpus in terms of Equal error rates (EER), normalized minimum detection cost functions of NIST-SRE 2008 (DCF08) NIST-SRE 2010 (DCF10).

		FEMALE		
		EER	**DCF08**	**DCF10**
1	**MFCC**	2.91	0.144	0.512
2	**MFCC_E**	2.14	0.096	0.436
3	**MGDCC**	2.83	0.151	0.696
4	**PS-MFCC**	2.83	0.146	0.556
5	**LP-GDCC**	5.75	0.226	0.601
6	**SWLP-GDCC**	5.17	0.211	0.494
7	**MMFCC**	1.98	0.099	0.344
8	**LFCC**	2.79	0.139	0.616
9	**LPRPC**	5.84	0.327	0.612
10	**FUSED**	**0.670**	**0.0296**	**0.0635**
11	**FUSED5**	**0.573**	**0.0469**	**0.109**

Table 4. Speaker verification results on male trials of the ENGLISH part of the VT-Pakistan corpus in terms of Equal error rates (EER), normalized minimum detection cost functions of NIST-SRE 2008 (DCF08) NIST-SRE 2010 (DCF10).

		MALE		
		EER	**DCF08**	**DCF10**
1	**MFCC**	2.78	0.091	0.317
2	**MFCC_E**	2.57	0.086	0.319
3	**MGDCC**	4.88	0.182	0.544
4	**PS-MFCC**	3.07	0.094	0.357
5	**LP-GDCC**	5.35	0.181	0.488
6	**SWLP-GDCC**	4.90	0.164	0.453
7	**MMFCC**	3.22	0.107	0.354
8	**LFCC**	3.01	0.096	0.372
9	**LPRPC**	8.57	0.320	0.736
10	**FUSED**	**2.36**	**0.0812**	**0.286**
11	**FUSED5**	**2.32**	**0.0781**	**0.290**

6. Reference

[1] Ravi P. Ramachandran, Kevin R. Farrell, Roopashri Ramachandran, and Richard J. Mammone, “Speaker recognition–general classifier approaches and data fusion methods,” Pattern Recognition , vol. 35, no. 12, pp. 2801–2821, 2002.

[2] H. Li, B. Ma, K. A. Lee, H. Sun, D. Zhu, C. S. Khe, C. You, R. Tong, I. Karkkainen, C.L. Huang, V. Pervouchine, W. Guo, Y. Li, L. Dai, M. Nosratighods, T. Tharmarajah, J. Epps, E. Ambikairajah, E.S. Chng, T. Schultz, and Q. Jin, “The i4u system in nist 2008 speaker recognition evaluation,” in Proceedings of ICASSP, pp. 4201 – 4204, 2009.

[3] J. M. K. Kua, J. Epps, E. Ambikairajah, M. Nosratighods, "Front-end diversity in Fused Speaker Recognition Systems," in Proc. of APSIPA, pp. 59-63, Singapore, 2010.

[4] J. M. Tribolet, “A new phase unwrapping algorithm,” IEEE Trans. Acoust. Speech, Signal Process., vol.25, pp. 170–177, 1977.

[5] G. Nico and J. Fortuny, “Using the matrix pencil method to solve phase unwrapping,” IEEE Trans. Signal Process., vol. 51, pp. 886–888, 2003.

[6] A. V. Oppenheim, J. S. Lim, G. E. Kopec, and S. C. Pohlig, “Phase in speech and pictures,” in Proc. IEEE Int. Conf. Acoust., Speech and Signal Processing, pp. 632-637, Apr. 1979.

[7] A. V. Oppenheim and J. S. Lim, “The Importance of Phase in Signals,” Proceedings of the IEEE, Vol. 69, No. 5, pp. 529-541, May 1981.

[8] D. L. Wang and J. S. Lim, “The unimportance of phase in speech enhancement,” IEEE Trans. Acoust., Speech, Signal Processing, vol. 30, pp. 679–681, Aug. 1982.

[9] L. Liu, J. He, and G. Palm, “Effects of phase on the perception of intervocalic stop consonants,” Speech Communication, Vol. 22, pp. 403-417, 1997.

[10] K. K. Paliwal and L. D. Alsteris, “Usefulness of phase spectrum in human speech perception,” in proc. of Eurospeech-2003, Geneva, Switzerland, pp. 2117–2120, 2003.

[11] L. D. Alsteris and K. K. Paliwal, “Importance of window shape for phase-only reconstruction of speech,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Processing, Montreal, Canada, pp. 573–576, 2004.

[12] L. D. Alsteris and K. K. Paliwal, “Short-time phase spectrum in speech processing: A review and some experimental results,” Digital Signal Processing, Vol. 17, pp. 578-616, May 2007.

[13] H. Murthy and V. Gadde. The modified group delay function and its application to phoneme recognition. In Proc. of ICASSP, vol. 1, p. 68–71, 2003.

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

[15] B. Bozkurt, L. Couvreur, and T. Dutoit, "Chirp group delay analysis of speech signals," Speech Commun., vol. 49, p. 159–176, 2007.

[16] E.Loweimi, S.M.Ahadi, T.Drugman, A New Phase-based Feature Representation for Robust Speech Recognition, IEEE International Conference on Audio Speech and Signal Processing (ICASSP), Vancouver, Canada, 2013

[17] P. Rajan, T. Kinnunen, C. Hanilci, J. Pohjalainen, P. Alku, ”Using group delay functions from all-pole models for speaker recognition”, Proc. Interspeech 2013, pp. 2489--2493, Lyon, France, August 2013.

[18] 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, pp. 25–28, 2003.

[19] 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.

[20] T. Thiruvaran, E. Ambikairajah, and J. Epps, “Extraction of fm components from speech signals using all-pole model,” Electronics Letters , vol. 44, no. 6, pp. 449–50, 2008.

[21] J.M.K Kua, J. Epps, E. Ambikairajah, and E. Choi, “Ls regularization of group delay features for speaker recognition,” Proceedings of the 10th Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 2887 – 2890, 2009

[22] 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., volume 2, p. 861–864, 1998.

[23] R. Hegde, H. Murthy, and V. Gadde, "Significance of the modified group delay feature in speech recognition," IEEE Transactions on Audio, Speech, and Language Processing, , vol. 15, no. 1, p.190–202, 2007.

[24] Zhizheng Wu, Xiong Xiao, Eng Siong Chng, and Haizhou Li, “Synthetic speech detection using temporal modulation feature,” in IEEE Acoustics, Speech and Signal Processing (ICASSP), pp. 7234–7238, 2013.

[25] 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.

[26] Md. J. Alam, T. Kinnunen, P. Kenny, P. Ouellet, D. O'Shaughnessy, “Multitaper MFCC and PLP Features for Speaker Verification Using i-Vectors”, Speech Communication, 55(2): 237--251, February 2013.

[27] Patrick Kenny, Themos Stafylakis, Md. Jahangir Alam, Pierre Ouellet and Marcel Kockmann, "In-Domain versus Out-of-Domain Training for Text-Dependent JFA," Proc. INTERSPEECH, Singapore, September 2014.

[28] A. Larcher, A.-K. Lee, B. Ma, H. Li, “Text-dependent speaker verification: Classifiers, databases and RSR2015”, Speech Communication, , March 2013.

[29] Patrick Kenny, Themos Stafylakis, Md. Jahangir Alam, Pierre Ouellet and Marcel Kockmann, "JFA features with joint density model back-end for HMM Text-dependent speaker recognition," submitted to ICASSP, 2015.

[30] Stafylakis, Patrick Kenny, Md. Jahangir Alam, Multi-tier JFA - Report 12, CRIM, 2014

[31] S. R. Krothapalli, S. G. Koolagudi, Emotion Recognition using Speech Features, Springer, New York, 2013.

[32] K. S. R. Murthy, B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," IEEE Sig. Proc. Letters, vol. 13, pp. 52-55, January 2006.