Odyssey 2014:
The Speaker and Language Recognition Workshop
16-19 June 2014, Joensuu, Finland

Allpass modelling of Fourier phase for speaker verification

Karthika Vijayan, Vinay Kumar and K. Sri Rama Murty

Department of Electrical Engineering
Indian Institute of Technology Hyderabad, India
{ee11p011, ee10b039, ksrm}@iith.ac.in

Abstract

This paper proposes features based on parametric representation of Fourier phase of speech for speaker verification. Direct computation of Fourier phase suffers from phase wrapping and hence we attempt parametric modelling of phase spectrum using an allpass (AP) filter. The coefficients of the AP filter are estimated by minimizing an entropy based objective function motivated from speech production process. The AP cepstral coefficients (APCC) derived from the group delay response of estimated AP filter are used as features for speaker verification. An i-vector based speaker verification system is employed to evaluate the performance of the proposed APCC features on NIST 2003 speaker recognition evaluation database. The equal error rates (EER) obtained from speaker verification systems built using APCC features and baseline mel-frequency cepstral coefficients (MFCC) are reported. A relative improvement of 12% was obtained over MFCC features by combining evidences from both MFCC and APCC based systems.

1. Introduction

Speaker verification refers to verifying a person's claimed identity by employing a machine [1]. The accuracy of speaker verification systems critically depend on the features extracted from speech signals. Most of the speaker verification systems use features derived from the magnitude spectrum of speech signal. Mel-frequency cepstral coefficients (MFCC) [2] and linear prediction cepstral coefficients (LPCC) [3], which represent the envelope of magnitude spectrum, are the commonly used features for speaker recognition. Another commonly used set of features of speech is frequency domain linear prediction (FDLP) coefficients, which model the magnitude envelope of speech signals in analytic domain [4]. The phase information is neglected in all these feature sets, which is equally important as magnitude in speaker characterization. It is evident from informal listening that identification of speaker from phase distorted speech[1] is difficult, which points to the relevance of phase in conveying speaker specific characteristics. Apart from speaker recognition, the phase information in speech signals has also been applied for signal estimation [5], speech enhancement [6], speech separation [7] etc. Also the significance of phase spectral characteristics of signals in speech as well as image processing is elaborately described in [8].

Several attempts have been made for feature extraction from phase characteristics of speech. Features derived from amplitude weighted instantaneous frequency of a set of bandpass components of speech signal were proposed with application to speaker identification in [9]. Another set of features extracted from subband instantaneous frequencies obtained using a gammatone filter-bank was proposed in [10] for speech recognition. Features based on amplitude modulation-frequency modulation decomposition of speech signals were proposed in [11, 12]. All these features deal with phase of analytic domain representation of speech.

Importance of short-time phase spectrum in speech processing was extensively studied in [13], mentioning its applications to pitch determination, formant extraction, epochs estimation etc. The significance of phase in human speech recognition was further explored in [14] based on subjective studies. Computation of phase based features was attempted directly from short time Fourier transform representation of speech signals in [15]. In this work, ambiguity in calculation of phase values based on clipping positions in speech signal was nullified by calculating phases of each frequency component with respect to a fixed basis frequency. But the phase wrapping problem associated with computation of Fourier phase was not addressed.

Feature extraction from speech by capturing phase spectral information using a group delay function for speaker identification was presented in [16]. The computation of group delay function in [16] was based on a minimum phase assumption [17] on speech signals. Since speech signal is the output of a nonminimum phase vocal tract system, these group delay features represent only partial phase spectral information. A set of post processed group delay features was proposed in [18], which dealt with the spiky nature of group delay of speech signals. The spikiness of group delay function was addressed using zeros of short term z-transform representation of speech signals in [19]. Phase based features for speaker recognition were derived from linear prediction (LP) residual, since the unmodelled phase spectrum of speech after LP analysis will be contained in it [20]. This set of features also represent incomplete phase spectral information from residual phase alone.

Unlike all the methods mentioned above, we propose parametric modelling of phase spectrum of speech signals for feature extraction. The modelling of phase spectrum of speech signals is attempted as the response of an allpass (AP) system. AP modelling of LP residual was attempted to extract features for speaker recognition in [21]. In this paper, we propose AP modelling of a phase signal which is derived from speech upon removal of magnitude spectrum. The modelling is performed by imposing constraints on the signal input to the AP system based on speech production process. The estimated AP coefficients (APC) are transformed to cepstral domain to obtain allpass cepstral coefficients (APCC) which serve as features for speaker verification. The verification system is built on the state-of-the-art i-vector based architecture [22]. A GMM-UBM based modelling is employed to obtain speaker specific statistical models [23]. The speaker verification studies are conducted for male speakers in NIST 2003 speaker recognition evaluation database.

[1] Speech signals synthesized with distorted phase spectra are given in http://www.iith.ac.in/~ee11p011.

The equal error rate (EER) for APCC based speaker verification is reported. Also the combination of APCC and the baseline MFCC features are observed to deliver an EER lesser than their individual EERs, indicating the existence of complementary speaker specific information in both magnitude and phase of speech signals.

The rest of the paper is organized as follows: In section 2, our motivation to model phase spectrum and the prerequisites for modelling are explained, section 3 discusses extraction of APCCs and building speaker verification system and section 4 demonstrates the speaker verification studies portraying the effectiveness of APCCs. In section 5, we summarize the contributions of this work towards speaker verification.

2. Modelling of phase spectrum

The Fourier transform representation of a discrete time signal, $s[n]$ is [17]:

$$S(j\omega) = \sum_{n=-\infty}^{\infty} s[n]e^{-j\omega n} \quad (1)$$

which can be represented in polar form as:

$$S(j\omega) = |S(j\omega)|e^{j\angle S(j\omega)} \quad (2)$$

where $|S(j\omega)|$ is the magnitude spectrum and $\angle S(j\omega)$ is the phase spectrum of $s[n]$. This expression will completely represent $s[n]$ only when both $|S(j\omega)|$ and $\angle S(j\omega)$ are specified. Features like MFCC and LPCC derive information from $|S(j\omega)|$ alone. The features representing only a portion of speech characteristics cannot completely capture information about the speaker.

Direct computation of $\angle S(j\omega)$ is affected with phase wrapping problem. In this paper, we attempt to parametrize $\angle S(j\omega)$ and derive features from it. The parametric modelling of phase spectrum can be attempted by suppressing magnitude spectrum in order to highlight phase spectral characteristics in speech signal. A magnitude suppressed signal, termed as phase signal $y[n]$ is generated from speech signal, $s[n]$ as follows:

$$y[n] = \mathscr{F}^{-1}\left\{\frac{S(j\omega)}{|S(j\omega)|}\right\} \quad (3)$$

where $\mathscr{F}^{-1}$ denotes the inverse discrete-time Fourier transform.

The phase signal will clearly have a flat magnitude spectrum and hence negligible autocorrelations, since power spectral density and autocorrelation function are Fourier transform pairs. Fig. 1(b) portrays the phase signal derived from the speech segment in Fig. 1(a). Even though $y[n]$ is uncorrelated, its samples are not statistically independent. It posses higher order statistical relationships between samples due to the phase spectrum contained in it. The cumulants are the most commonly used measures for representing higher order statistics [24]. The rich 3^{rd} and 4^{th} order relations existing in $y[n]$ can be expressed in terms of 3^{rd} and 4^{th} order cumulants- $C_{3,y}(k_1, k_2)$ and $C_{4,y}(k_1, k_2, k_3)$ respectively. The contour plots of estimates of cumulants [25] computed for a $25ms$ segment of $y[n]$ in Fig. 1(b) are shown in Fig. 2. The 4^{th} order cumulant, $C_{4,y}(k_1, k_2, k_3)$ is plotted for single slice of k_3. Modelling a suitable system with $y[n]$ as its output, is expected to capture these higher order relations in $y[n]$ and thus model the phase spectrum $\angle S(j\omega)$.

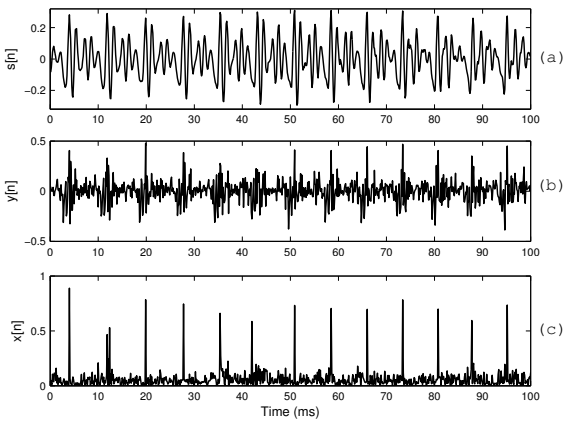

Figure 1: Signals at different stages of AP modelling: (a) Speech signal, $s[n]$ (b) Phase signal, $y[n]$ and (c) AP residual, $x[n]$.

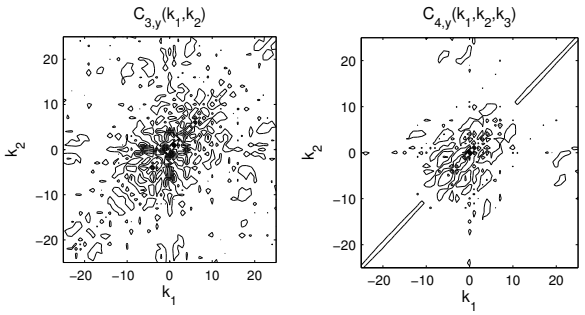

Figure 2: Illustration of higher order relations existing in phase signal: Contour plots of estimated cumulants of $y[n]$.

2.1. Allpass systems

The allpass (AP) system is an autoregressive moving average (ARMA) model capable of modelling phase characteristics of systems, including nonminimum phase systems. AP systems are capable of generating uncorrelated and dependent output samples when excited with non-Gaussian, independent and identically distributed (i.i.d) input sequence [26]. Since these characteristics match with those of $y[n]$, AP system is expected to be a suitable choice for phase modelling. The system transfer function of an M^{th} order AP system is:

$$H(z) = \frac{a_M + a_{M-1}z^{-1} + ... + a_1 z^{-M+1} + z^{-M}}{1 + a_1 z^{-1} + ... + a_{M-1}z^{-M+1} + a_M z^{-M}} \quad (4)$$

Such a system has flat magnitude response ($|H(j\omega)| = 1$) with its poles lying at conjugate reciprocal locations of its zeros. Both numerator and denominator polynomials of $H(z)$ are represented by the same set of coefficients and thus $H(z)$ can be uniquely described in terms of $\mathbf{a} = [a_1 a_2 ... a_M]^T$, termed as AP coefficients (APC).

The input-output relationship of $H(z)$ is given by:

$$y[n] = -\sum_{l=1}^{M} a_l y[n-l] + x[n-M] + \sum_{l=1}^{M} a_l x[n-M+l] \quad (5)$$

where $y[n]$ is constituted by uncorrelated and dependent samples when $x[n]$ is a non-Gaussian i.i.d sequence.

For a stable and causal $H(z)$, all the poles will lie inside unit circle in z plane. Consequently all zeros will lie outside the unit circle, which act as poles of the inverse filter $H^{-1}(z)$. Thus $H^{-1}(z)$ will be unstable unless it is a noncausal filter [27]. Hence the criterion of noncausality is forced on $H^{-1}(z)$ and the input signal $x[n]$, given $y[n]$ and $\mathbf{a}$, should be computed in a noncausal manner as:

$$x[n] = -\sum_{l=1}^{M} a_l x[n+l] + y[n+M] + \sum_{l=1}^{M} a_l y[n+M-l] \quad (6)$$

The APCs are estimated by imposing constraints over input signal $x[n]$ based on speech production process.

3. Estimation of allpass cepstral coefficients

AP modelling of phase signal aims to model the phase spectrum of speech by capturing higher order statistical relationships in $y[n]$ and generate an AP residual, $x[n]$ with maximally independent samples. This is an ill posed problem in which both the system parameters $\mathbf{a}$ and input signal $x[n]$ are unknown. Hence it requires some assumptions or prior knowledge of either the system or input to make the modelling problem tractable. In [26], Chi et al. proposed an AP modelling method which maximizes q^{th} order cumulant of a non-Gaussian i.i.d. input sequence. This approach requires knowledge of the optimum value of q. Modelling based on least absolute deviations for AP modelling is proposed in [28] by imposing a Laplacian distribution on $x[n]$. A maximum likelihood approach for AP modelling is presented in [29] when $x[n]$ follows any arbitrary probability distribution with known parameters. These methods require prior knowledge of probability distribution followed by $x[n]$. In this paper we attempt to model AP system by imposing constraints on $x[n]$ from a speech production perspective.

The excitation signal to the vocal tract system while producing voiced speech, will have significant values only at glottal closure instants (GCI) [30]. Hence the input signal can be viewed as train of impulses, in which energy distribution is constrained to GCIs alone. The excitation at GCIs are not spiky in nature. Still such a consideration of impulse excitation at GCIs renders an advantage of mathematical simplification for the modelling process. The APCs are needed to be estimated such that input signal $x[n]$ should have its energy localized to GCIs. Since $x[n]$ is generated by inverse filtering $y[n]$ through AP system and the magnitude response of AP system is flat, the total energy in $x[n]$ is same as that of $y[n]$. Hence frames of $y[n]$ can be normalized to have unit energy so as to obtain frames of $x[n]$ containing unit energy. The sample-wise energy content in $x[n]$ can be represented as:

$$e[n] = x^2[n] \quad (7)$$

Since $e[n]$ has positive sample values and sum of samples is unity, it can be viewed as a valid probability mass function. The energy $e[n]$ can be concentrated to a few samples by minimizing its entropy function. The entropy associated with $e[n]$ can be defined as a function of APCs as [31]:

$$J(\mathbf{a}) = -\sum_{n=1}^{N} e[n]\ log\ e[n] \quad (8)$$

We choose $J(\mathbf{a})$ as the objective function to be minimized for estimating $\mathbf{a}$. The minimum entropy deconvolution strategy

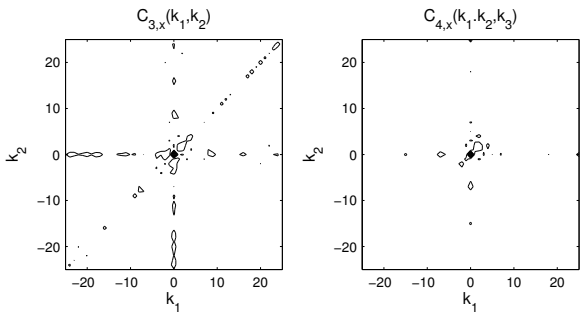

Figure 3: Illustration of absence of higher order relations in AP residual: Contour plots of estimated cumulants of $x[n]$.

[32] has been successfully used for applications like period estimation of periodic/quasi periodic signals [33]. In this paper, a gradient based iterative procedure is followed for minimization of $J(\mathbf{a})$. APCs, $\mathbf{a}$ should be initialized with small random values such that the locations of poles of $H(z)$ are within unit circle. The values of $\mathbf{a}$ will be updated iteratively with the gradient of $J(\mathbf{a})$. At iteration k, the APCs $\mathbf{a}_k$ will be computed as:

$$\mathbf{a}_k = \mathbf{a}_{k-1} - \mu \frac{\partial J(\mathbf{a})}{\partial \mathbf{a}}|_{\mathbf{a}=\mathbf{a}_{k-1}} \quad (9)$$

where $\mathbf{a}_{k-1}$ is the APCs estimated at $(k-1)^{th}$ iteration and μ is the learning rate parameter (μ is empirically chosen as 0.005 in this study). The gradient of $J(\mathbf{a})$ with respect to $\mathbf{a}$ can be computed using chain rule and is given as:

$$\begin{aligned} \frac{\partial J(\mathbf{a})}{\partial \mathbf{a}} &= \frac{\partial J(\mathbf{a})}{\partial e[n]} \frac{\partial e[n]}{\partial x[n]} \frac{\partial x[n]}{\partial \mathbf{a}} \\ &= -\sum_{n=1}^{N} [1 + log(e[n])](2x[n]) \left(\frac{\partial x[n]}{\partial \mathbf{a}} \right) \end{aligned} \quad (10)$$

The input signal $x[n]$ can be computed as in (6). The derivative of $x[n]$ with respect to $\mathbf{a}$ can be computed as:

$$\frac{\partial x[n]}{\partial a_p} = -\sum_{l=1}^{M} a_l \frac{\partial x[n+l]}{\partial a_p} - x[n+p] + y[n+M-p] \quad (11)$$

$\forall p \in \{1, 2...M\}$. It is clear from (11) that the derivative of $x[n]$ with respect to a_p can be computed by filtering $y[n+M-p] - x[n+p]$ through an all-pole filter with coefficients $\mathbf{a}$. The updating of coefficients is terminated when value of J saturates to a minimum possible value, i.e.$J(\mathbf{a}_{k-1}) - J(\mathbf{a}_k) < \epsilon$, with ϵ chosen to be 10^{-6}.

The proposed algorithm delivers a set of APCs $\mathbf{a}$ by modelling frames of $y[n]$ and generates AP residual, $x[n]$ with its energy localized to GCIs. The AP residual $x[n]$ generated by modelling a segment of $y[n]$ in Fig. 1(b) is shown in Fig. 1(c), which clearly demonstrates the train of impulse-like nature of $x[n]$. The AP system is expected to capture the higher order statistical relations from $y[n]$ and generate an $x[n]$ with maximally independent samples. The estimates of 3^{rd} and 4^{th} order cumulants for the segment of AP residual in Fig. 1(c) is shown in Fig. 3. From this figure, it is clearly evident that the 3^{rd} and 4^{th} order relationships existed in $y[n]$ (See Fig. 2) are absent in $x[n]$.

The estimated APCs $\mathbf{a}$, are expected to have captured speaker specific information from speech signals. To illustrate

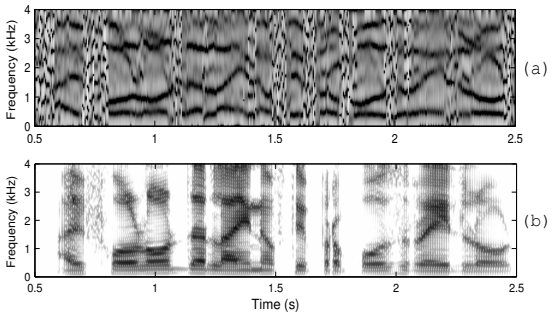

Figure 4: Illustration of effectiveness of AP coefficients in capturing speaker characteristics: (a) Groupdelaygram from estimated APCs and (b) Spectrogram of speech signal.

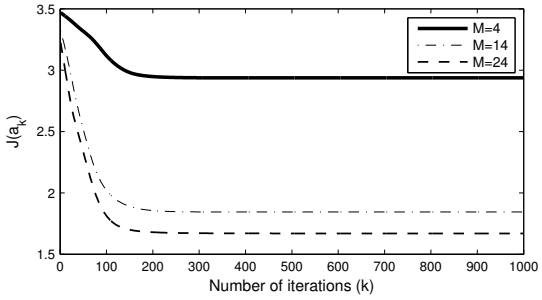

Figure 5: Minimization of entropy function with respect to order of AP system, M.

the information captured by APCs, groupdelaygram is plotted by computing the group delay of estimated AP system (group delay computation is explained in section.3.1) for segments of phase signal. The groupdelaygram in Fig. 4(a) demonstrates all formant tracks as clearly as in the spectrogram of speech given in Fig. 4(b). This clearly illustrates the efficiency of estimated APCs in capturing speaker characteristics.

The minimization of entropy function attained for different orders of AP system (M) is shown in Fig. 5. For small values of M, the minimization of entropy function is not effective and hence resultant APCs are unreliable. While for large values of M, the modelling procedure tries to overfit the phase signal $y[n]$ and destroys the energy localization in $x[n]$. The model orders between 10 and 15 are observed to be providing faithful AP models. The estimated APCs are observed to be capturing speech characteristics effectively and hence are used to derive features for speaker verification.

3.1. Allpass cepstral coefficients

The estimated APCs are coefficients of a polynomial and are not stabilized. Small changes in values of APCs will result in huge changes in the poles and zeros of $H(z)$. Hence APCs should be transformed to a more stabilized set of coefficients. The group delay of the estimated AP system is computed as [17]:

$$\tau(\omega) = \mathscr{I}\left\{\frac{H'(j\omega)}{H(j\omega)}\right\} \tag{12}$$

where $\mathscr{I}\{.\}$ denotes the imaginary part of a complex quantity and $H'(j\omega)$ is the derivative of $H(j\omega)$. The $H'(j\omega)$ can be computed using Fourier transform relationship [17]:

$$H'(j\omega) = -j\mathscr{F}\{n\ h[n]\}. \tag{13}$$

where $h[n]$ is the impulse response of AP system and $\mathscr{F}$ denotes Fourier transform. Discrete cosine transformation (DCT) is performed over the group delay to account for redundancies existing in it. The first few coefficients from DCT output are retained, which we call as allpass cepstral coefficients (APCC). The logarithmic transformation involved in cepstrum computation is neglected, owing to the additive nature of group delay response. The APCC together with its first and second order differences constitute the APCC feature vectors. Speaker verification system is built using these APCC feature vectors.

3.2. Speaker verification system

An i-vector based speaker verification system is used to evaluate performance of the proposed APCC features. Speaker independent distribution of APCC features is captured using a universal background model (UBM) [34], which is essentially a very large Gaussian mixture model (GMM) trained by pooling features from several speakers. The parameters of the UBM are adapted to reference/test utterance to capture the speaker-dependent distribution [23]. The mean vectors of the adapted GMM are concatenated to form a GMM supervector, which essentially represents speaker and session specific information [35]. The dimensionality of the GMM supervectors is reduced using total variability matrix (T-matrix), to obtain a low dimensional representation referred to as the i-vectors [22]. Probabilistic linear discriminant analysis (PLDA) is performed on the i-vectors to compensate for session/channel effects and to yield efficient classification [36]. The i-vectors from reference and test utterances are used to compute a likelihood based confidence score Λ, which is used to take decision on claimed identity.

4. Speaker Verification Studies

4.1. Database for building UBM

Speech data of male speakers from switchboard cellular part-2 database was used as training data for building the UBM, T-matrix and PLDA model. There are 1446 male speaker files in the database. The total available database is divided into 2 subsets- one for training the UBM and other for training T-matrix and PLDA model.

4.2. Database for performance evaluation

The performance of speaker verification system based on proposed APCC features is carried out on male speakers from NIST-2003 speaker recognition evaluation database [37]. We have chosen NIST-2003 over more recent versions of NIST databases due to its relatively small data size, which is suitable for the very limited computational resources available to us. There are 1492 male test utterances in the database. The average duration of reference utterance is 2 minutes and that of test utterance is 30 seconds [38]. There are a 14773 test pair utterances out of which 1214 are genuine and 13559 are impostor tests.

4.3. Speaker verification evaluation

The speech signals are segmented into frames of 25ms duration with 10ms shift. Phase signals are derived from these speech

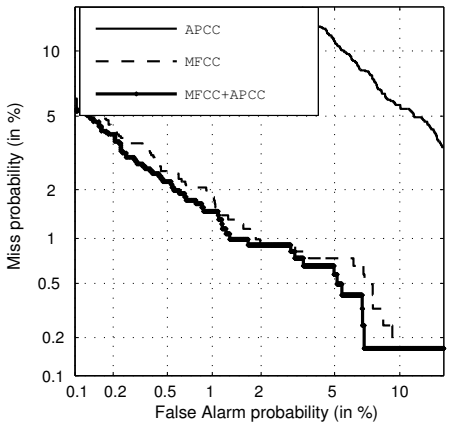

Figure 6: Detection error trade-off curves for speaker verification systems.

Table 1: Equal error rates obtained using different features in speaker verification.

Feature used	EER (%)
MFCC	1.32
APCC	7.58
MFCC + APCC	1.15

frames and a 10^{th} order AP modelling is performed to obtain APCs. 39 dimensional APCC feature vectors (13 cepstra + 13 Δ + 13 $\Delta\Delta$) are extracted from each speech segment and speaker verification system is trained. The confidence scores between the reference and test utterances are obtained and the performance of the proposed speaker verification is demonstrated as detection error trade-off [39] curve in Fig.6. The equal error rate (EER) of 7.58% was obtained, illustrating the effectiveness of proposed APCC features in speaker characterization. The speaker recognition system built as an autoassociative neural network based on features derived from LP residual phase resulted in an EER of 22% for NIST 2003 [20]. Also the EER obtained with modified group delay features [16] evaluated upon NIST 2003 was reported as 11.92% [40].

The performance of APCC features is compared with the baseline MFCC features (13 cepstra + 13 Δ + 13 $\Delta\Delta$) extracted from magnitude spectrum of speech signals. Table. 1 shows the performance of both systems based on MFCC and APCC. MFCC features are observed as superior to APCC in speaker characterization. Since MFCC and APCC features represent magnitude and phase spectral information in speech respectively, the confidence scores obtained from both are combined to investigate their complementary nature. The scores from both systems are fused as a convex combination:

$$\Lambda_{fused} = \alpha\Lambda_{MFCC} + (1-\alpha)\Lambda_{APCC} \qquad (14)$$

The weight of combination α is chosen as 0.7, giving more emphasis to MFCC features. The resulting EER is observed to be lesser than that of both MFCC and APCC features, delivering a relative improvement of 12% compared to MFCCs. This points to the existence of complementary speaker specific information in both MFCC and APCC features.

5. Conclusions

A set of features derived from phase spectrum of speech signals for speaker verification was proposed in this paper. The phase spectrum of speech was parametrically modelled to explore speaker specific information in it. Allpass systems were chosen to model phase spectrum. Cepstral features were derived from allpass coefficients (APCC) to represent phase information in speech. APCC features were used to build speaker verification system whose performance was evaluated on male speakers from NIST 2003 speaker recognition evaluation database. An equal error rate (EER) of 7.58% was obtained from APCC features as opposed to 1.32% from mel cepstral features (MFCC). The combination of systems based on MFCC and APCC features delivered an EER of 1.15% which is clearly lesser than the individual EERs of both systems. The speaker verification studies in this paper illustrated the effectiveness of APCCs in capturing phase spectral characteristics and marked the presence of complementary speaker information in magnitude and phase spectrum of speech signals.

6. References

[1] Joseph P. Campbell Jr, "Speaker recognition: A tutorial," in *Proceedings of the IEEE*, September 1999, vol. 85, pp. 1437–1462.

[2] Xuedong Huang, Alex Acero, and Hsiao-Wuen Hon, *Spoken language processing: A guide to theory, algorithm and system development*, Prentice Hall PTR, Upper Saddle River, NJ, USA, 1st edition, 2001.

[3] John Makhoul, "Linear prediction: A tutorial review," in *Proceedings of the IEEE*, April 1975, vol. 63, pp. 561–580.

[4] Marios Athineos and Daniel P.W. Ellis, "Frequency-domain linear prediction for temporal features," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2003, pp. 261–266.

[5] D. Griffin and Jae S. Lim, "Signal estimation from modified short time fourier transform," in *IEEE Transactions on Acoustics, Speech and Signal Processing*, April 1984, vol. 32, pp. 236–243.

[6] P. Mowlaee and R. Saeidi, "Iterative closed-loop phase-aware single-channel speech enhancement," in *IEEE Signal Processing Letters*, December 2013, vol. 20, pp. 1235–1239.

[7] P. Mowlaee, R. Saeidi, and R. Martin, "Phase estimation for signal reconstruction in single-channel speech separation," in *Interspeech 2012*, September 2012, pp. 1548–1551.

[8] Alan V. Oppenheim and Jae S. Lim, "The importance of phase in signals," in *Proceedings of the IEEE*, May 1981, vol. 69, pp. 529–550.

[9] Marco Grimaldi and Fred Cummins, "Speaker identification using instantaneous frequencies," in *IEEE Transactions on Audio, Speech and Language processing*, August 2008, vol. 16, pp. 1097–1111.

[10] Hui Yin, Volker Hohmann, and Climent Nadeu, "Acoustic features for speech recognition based on gammatone filterbank and instantaneous frequency," in *Speech Communication*, 2011, vol. 53, pp. 707–715.

[11] Dimitrios Dimitriadis, Petros Maragos, and Alexandros Potamianos, "Robust AM-FM features for speech recognition," in *IEEE Signal Processing Letters*, September 2005, vol. 12, pp. 621–624.

[12] Tharmarajah Thiruvaran, Julien Epps, Eliathamby Ambikairajah, and Edward Jones, "An investigation of subband FM feature extraction in speaker recognition," in *IET Irish Signals and Systems Conference, ISSC*, June 2008, pp. 32–36.

[13] Leigh D. Alsteris and Kuldip K. Paliwal, "Short-time phase spectrum in speech processing: A review and some experimental results," in *Digital Signal Processing*, May 2007, vol. 17, pp. 578–616.

[14] Guangji Shi, Maryam Modir Shanechi, and Parham Aarabi, "On the importance of phase in human speech recognition," in *IEEE Transactions on Audio, Speech and Language processing*, September 2006, vol. 14, pp. 1867–1874.

[15] Longbiao Wang, Shinji Ohtsuka, and Seiichi Nakagawa, "High improvement of speaker identification and verification by combining MFCC and phase information," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'09)*, April 2009, pp. 4529–4532.

[16] Rajesh M. Hegde, Hema A. Murthy, and Gadde V. Ramana Rao, "Application of the modified group delay function to speaker identification and discrimination," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'04)*, 2004, pp. 517–520.

[17] Alan V. Oppenheim, Ronald W. Schafer, and John R. Buck, *Discrete-time signal processing*, Signal processing series. Prentice Hall Inc., Upper Saddle River, NJ, USA, 2nd edition, January 1999.

[18] Tharmarajah Thiruvaran, Eliathamby Ambikairajah, and Julien Epps, "Group delay features for speaker recognition," in *International Conference on Information, Communications & Signal Processing, ICICS*, December 2007, pp. 1–5.

[19] Baris Bozkurt, Laurent Couvreur, and Thierry Dutoit, "Chirp group delay analysis of speech signals," in *Speech communication*, March 2007, vol. 49, pp. 159–176.

[20] K. Sri Rama Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," in *IEEE Signal Processing Letters*, January 2006, vol. 13, pp. 52–56.

[21] K. Sri Rama Murty, Vivek Boominathan, and Karthika Vijayan, "Allpass modelling of LP residual for speaker recognition," in *International Conference on Signal processing and communications (SPCOM)*, July 2012, pp. 1–5.

[22] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," in *IEEE Transactions on Audio, Speech and Language processing*, May 2011, vol. 19, pp. 788–798.

[23] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted gaussian mixture models," in *Digital Signal Processing*, January 2000, vol. 10, pp. 19–41.

[24] Jerry M. Mendel, "Tutorial on higher-order statistics (spectra) in signal processing and system theory: theoretical results and some applications," in *Proceedings of the IEEE*, March 1991, vol. 79, pp. 278–305.

[25] Ananthram Swami, *HOSA- Higher order spectral analysis toolbox*, [Online], www.mathworks.in/matlabcentral/fileexchange/3013-hosa-higher-order-spectral-analysis-toolbox, February 2003.

[26] C.-Y. Chi and J.-Y. Kung, "A new identification algorithm for allpass systems by higher-order statistics," in *Signal Processing*, January 1995, vol. 41, pp. 239–256.

[27] Alan V. Oppenheim, Alan S. Willsky, and S. Hamid Nawab, *Signals and systems*, Pearson Education Inc., Upper Saddle River, NJ, USA, 2nd edition, 1997.

[28] F. J. Breidt, R. A. Davis, and A. A. Trindade, "Least absolute deviation estimation for all-pass time series models," in *Annals of Statistics*, 2001, vol. 29, pp. 919–946.

[29] B. Andrews, R. A. Davis, and F. J. Breidt, "Maximum likelihood estimation for all-pass time series models," in *Journal of Multivariate Analysis*, August 2006, vol. 97, pp. 1638–1659.

[30] Thomas F. Quatieri, *Discrete-time speech signal processing: principles and practice*, Prentice Hall Press, Upper Saddle River, NJ, USA, 2001.

[31] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, Telecommunications and signal processing. Wiley-Interscience, 2006.

[32] Ralph A. Wiggins, "Minimum entropy deconvolution," in *Geoexploration*, April 1978, vol. 16, pp. 21–35.

[33] G. González, R.E. Badra, R. Medina, and J. Regidor, "Period estimation using minimum entropy deconvolution," in *Signal Processing*, January 1995, vol. 41, pp. 91–100.

[34] Douglas A. Reynolds and Richard C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," in *IEEE Transactions on Speech and Audio Processing*, January 1995, vol. 3, pp. 72–83.

[35] Patrick Kenny, Pierre Ouellet, Najim Dehak, Vishwa Gupta, and Pierre Dumouchel, "A study of inter-speaker variability in speaker verification," in *IEEE Transactions on Audio, Speech and Language processing*.

[36] Patrick Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Proc of Odyssey 2010: The speaker and language recognition workshop*, Brno, Czech Republic, 2012.

[37] Linguistic Data Consortium, Philadelphia, *NIST 2003 speaker recognition evaluation*, 2003.

[38] Linguistic Data Consortium, Philadelphia, *The NIST year 2003 speaker recognition evaluation plan*, February 2003.

[39] A. Martin, G. Doddington, T. Kamm, and M. Ordowski, "The DET curve in assessment of detection task performance," in *Eur. Conf. Speech Processing Technology*, Rhodes, Greece., 1997, pp. 1895–1898.

[40] R. Padmanabhan and Hema A. Murthy, "Acoustic feature diversity and speaker verification," in *Interspeech 2010*, September 2010, pp. 2110–2113.