Allpass Modeling of LP Residual for Speaker Recognition

K. Sri Rama Murty, Vivek Boominathan and Karthika Vijayan
Department of Electrical Engineering, Indian Institute of Technology Hyderabad, India
email: ksrm@iith.ac.in, vivek@iith.ac.in, ee11p011@iith.ac.in

***Abstract*—The objective of this paper is to demonstrate the usefulness of phase derived from the linear prediction (LP) residual for speaker recognition. Though the sequence of samples in the LP residual are uncorrelated, they are not independent. Since the magnitude spectrum of the LP residual is almost flat, the dependencies among the samples in LP residual are reflected mainly in its phase spectrum. The information in the phase spectrum of the LP residual is captured by modeling LP residual as the output of an allpass filter excited by independent and identically distributed (i.i.d.) nongaussian input. The coefficients of the allpass filter are estimated iteratively using higher order cumulants of the input. The estimated coefficients are used as features to build a speaker recognition system using Gaussian mixture models. The speaker recognition system built from the proposed features resulted in an equal error rate of 6% on a population of 50 speakers.**

***Index Terms*—Allpass filter, higher order cumulants, phase estimation, Gaussian mixture modeling and speaker recognition.**

I. Introduction

Speech is produced as an outcome of a time-varying vocal-tract system driven by a time-varying excitation. Speaker-specific information in the speech signal can be attributed to both the excitation source characteristics as well as the dimensions of the vocal-tract system. Thus the features extracted for speaker recognition should capture the speaker-specific information from both these components. Most of the current day speaker recognition systems use features derived from the magnitude spectrum of the speech signal which grossly represents the vocal-tract system characteristics. These features, which include mel-frequency cepstral coefficients (MFCC) and linear prediction cepstral coefficients (LPCC), do not capture information about the phase spectrum and the excitation source. The information about these components can be enhanced by suppressing the magnitude spectral envelope from the speech signal. The resulting residual will have a near-flat magnitude spectrum, indicating that the information is mainly present in its phase spectrum. Linear prediction (LP) analysis can be employed for suppressing the magnitude spectral envelope [1]. This is achieved by first predicting the magnitude spectral envelope from the signal, and then suppressing it by inverse filter formulation. The resulting signal is termed as the LP residual and contains the information mostly in its phase spectrum. This work deals with extracting speaker specific information from the phase spectrum of the LP residual.

Several attempts have been made to extract the speaker-specific information from the LP residual. Wakita had attempted to use energy of the LP residual for vowel recognition and speaker recognition [2]. It was also shown that a combination of LPCCs and energy of the LP residual gives better speaker recognition performance compared to using only LPCCs [3]. The usefulness of cepstrum computed over the LP residual was also exploited for speaker recognition [4]. It was also observed that a combination of LPCCs and LP residual cepstrum reduce the errors in speaker recognition [5]. Autoassociative neural network models have been employed to capture the speaker specific excitation information directly from the samples of LP residual [6]. In order to minimize the amplitude fluctuations around the glottal closure instants in the LP residual, the analytic phase derived from the LP residual was also exploited for the speaker recognition [7]. In the above mentioned attempts, either the LP residual or its cepstral transformation was used to capture the speaker-specific information. In this work, we attempt to extract speaker-specific features from the phase spectrum of the LP residual, by modeling LP residual as output of an allpass filter.

Since the LP residual is obtained by inverse filtering the speech signal through the magnitude spectral envelope, it does not contain significant second order correlations among its samples. Though the samples of the LP residual are uncorrelated, they are not independent. As a result, the information in the LP residual is mainly due to higher order statistical relations among its samples, which mainly reflects in its phase spectrum. Allpass filters have been used to model the phase spectrum of uncorrelated but dependent time series. Breidt et. al., have modeled serially uncorrelated financial time series as output of an allpass filter driven by a Laplacian distributed input [8]. But this approach is effective only when the input follows Laplacian distribution. Subsequently, a maximum likelihood estimator for allpass modeling has been proposed for arbitrary input densities with known parameters [9]. Chi et. al. have proposed a higher order cumulant based approach for allpass modeling, and thereby, phase estimation [10], [11].

In this work, the LP residual is modeled as output of an allpass filter. The coefficients of the allpass filter are estimated using the method proposed in [10]. Speaker-specific features from the phase spectrum of the LP residual are extracted from the estimated allpass filter coefficients. The distribution of the features is captured using Gaussian mixture models in order to build speaker models. The rest of the paper is organized as follows: In Section II, we describe the allpass modeling of the LP residual using higher order cumulants. Section III demonstrates the effectiveness of the proposed approach for

the speaker recognition. In Section IV, we summarize the contributions of this work.

II. Allpass modeling of LP residual

During the linear prediction analysis of speech signal, each sample $s[n]$ is estimated as linear weighted sum of the past p samples, where p represents the order of prediction [1]. If $s[n]$ is the present sample, then it is predicted from the past p samples as

$$\hat{s}[n] = -\sum_{k=1}^{p} \alpha_k s[n-k], \tag{1}$$

where $\{\alpha_k\}$, $k = 1, 2, \ldots, p$ are the linear prediction coefficients (LPCs). The difference between the actual sample value $s[n]$ and the predicted sample value $\hat{s}[n]$ is termed as prediction error or LP residual, and is given by

$$y[n] = s[n] - \hat{s}[n] = s[n] + \sum_{k=1}^{p} \alpha_k s[n-k]. \tag{2}$$

The linear prediction coefficients $\{\alpha_k\}$ are typically determined by minimizing the mean squared error over an analysis window, which is equivalent to solving a set of p normal equations given by

$$\sum_{k=1}^{p} \alpha_k r[i-k] = -r[i], \quad 1 \le i \le p, \tag{3}$$

where $r[i]$ is the autocorrelation function defined as

$$r[i] = \sum_{n=0}^{N-1-i} s[n]s[n+i] \quad 1 \le i \le p, \tag{4}$$

where N is the length of the analysis window. After estimating the LP coefficients, the LP residual can be obtained by passing the speech signal through the inverse filter $H_{inv}(z)$ given by

$$H_{inv}(z) = 1 + \sum_{k=1}^{p} \alpha_k z^{-k}. \tag{5}$$

A segment of voiced speech signal and its 12th order LP residual are shown in Fig. 1. As the LP analysis extracts the second order statistical relations through the autocorrelation coefficients, the LP residual does not contain any significant second order relations. That is why the autocorrelation function of LP residual has low correlation values for nonzero time lags (except for lags equal to the fundamental period) as shown in Fig. 2. For the same reason, the magnitude spectrum of the LP residual looks nearly flat, and only pitch harmonics can be observed in Fig. 3. This indicates that the information in the LP residual is mainly because of its phase spectrum.

Though the sequence of samples in the LP residual are uncorrelated, they are not independent. In other words, the information in the LP residual is present in the higher order statistical dependencies among the samples of the LP residual. The higher order dependencies in the time domain gets reflected as phase spectrum in the frequency domain. The bipolar fluctuations in the LP residual, especially around the glottal closure instants in Fig. 1(b), may be attributed to the phase

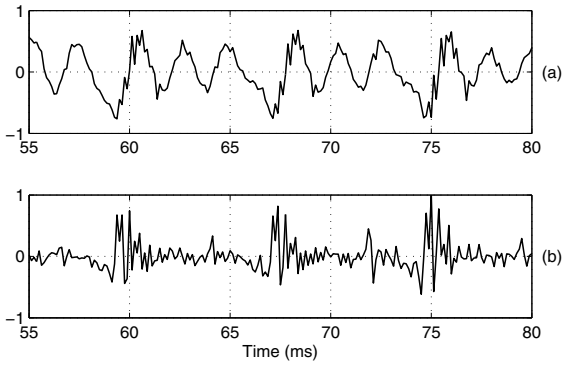

Fig. 1. A 25 ms segment of (a) voiced speech signal sampled at 8 kHz, and its (b) LP residual obtained from a 12th order LP analysis

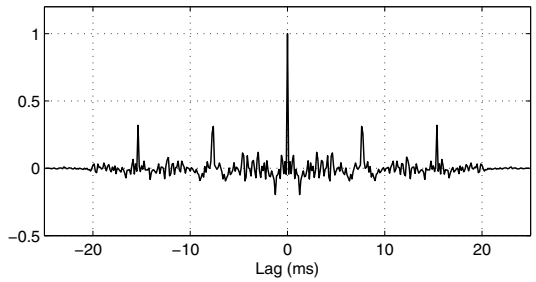

Fig. 2. Normalized autocorrelation function of the segment of LP residual shown in Fig 1(b)

response as only the magnitude response of the vocal-tract system was equalized during the inverse filtering. In order to extract the phase information, LP residual can be assumed as the output of an allpass filter excited by an i.i.d. nongaussian input source, as shown in Fig. 4.

An allpass system is an autoregressive moving average system in which roots of the autoregressive polynomial are the conjugate reciprocals of the roots of the moving average polynomial and vice-versa [12]. The transfer function of an

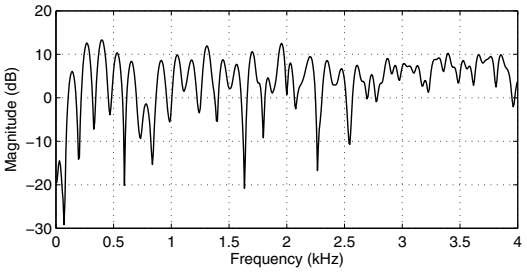

Fig. 3. Magnitude spectrum of the segment of LP residual shown in Fig 1(b)

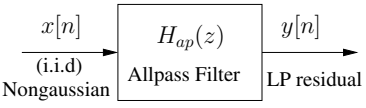

Fig. 4. Block diagram illustrating the production of the LP residual

M^{th} order allpass filter is given by

$$H_{ap}(z) = \frac{a_M + a_{M-1}z^{-1} + \ldots + a_1 z^{-M+1} + z^{-M}}{1 + a_1 z^{-1} + \ldots + a_{M-1}z^{-M+1} + a_M z^{-M}}, \quad (6)$$

where $\mathbf{a} = [a_1 a_2 \ldots a_M]^T$, form the vector of allpass filter coefficients. It can be shown that the magnitude response $|H_{ap}(j\omega)|$ of such a system is a constant [12], and hence the name *allpass* filter. If all the poles of the transfer function $H_{ap}(z)$ (roots of the denominator polynomial of $H_{ap}(z)$) lie inside the unit circle, then it results in a stable and causal allpass filter. All the zeros, being the conjugate reciprocals of poles, of a stable and causal allpass filter lie outside the unit circle. As a consequence, a stable and causal inverse filter does not exist for a stable and causal allpass filter. It can only have a stable but anticausal inverse filter.

An allpass filter, when excited by an i.i.d. nongaussian input sequence, generates uncorrelated but dependent sequence of samples as output. The characteristics of the output sequence of an allpass filter are similar to the characteristics of the LP residual. Hence the LP residual can be modeled as an output of an allpass filter excited by i.i.d. nongaussian input sequence. As the LP residual does not contain significant second order correlations among its samples, higher order statistical dependencies must be exploited for this modeling [13].

A. Estimation of allpass filter coefficients

Let us assume that the LP residual $y[n]$ is the output of an allpass filter with transfer function $H_{ap}(z)$ excited by an input sequence $x[n]$ as shown in Fig. 4. The input-output relation between $x[n]$ and $y[n]$ can be written using allpass filter coefficients as follows:

$$y[n] = -\sum_{k=1}^{M} a_k y[n-k] + x[n-M] + \sum_{k=1}^{M} a_k x[n-M+k] \quad (7)$$

Our goal is to estimate the filter coefficients $[\mathbf{a}]$ such that the higher order dependencies among the samples of $y[n]$ are captured in the phase spectrum of the allpass filter, leading to an allpass residual $x[n]$ whose samples are maximally independent of each other. That is, we need to estimate the filter coefficients $[\mathbf{a}]$ as well as the input sequence $x[n]$ from the output sequence $y[n]$, which is the LP residual in this case. This is an ill-posed inverse problem, and it requires some assumptions to be imposed on either the filter coefficients $[\mathbf{a}]$ or the input signal $x[n]$. Different methods exist in the literature to estimate the filter coefficients by imposing different constraints on the distribution of the input $x[n]$ [8], [9], [10], [11]. In this work, we adopt the method proposed by Chi at. el., to estimate the allpass filter coefficients $[\mathbf{a}]$ by imposing constraints on the higher order moments of the input sequence $x[n]$ [10]. In [10], it was assumed that the input sequence is real, and is derived from a zero-mean, stationary, i.i.d. nongaussian distribution with q^{th} order cumulant γ_q, where $q \geq 3$. For such an input sequence, it was shown that the allpass filter coefficients can be estimated by maximizing the absolute q^{th} order cumulant of the estimated input sequence $x[n]$. When the order of the allpass system $H_{ap}(z)$ is known a priori, the filter coefficients $[\mathbf{a}]$ can be estimated by maximizing the following objective function:

$$J_q(\mathbf{a}) = C^2_{q,x}(k_1 = 0, k_2 = 0, \ldots, k_{q-1} = 0), \quad (8)$$

where $C_{q,x}(k_1, k_2, \ldots, k_{q-1})$ is the q^{th} order cumulant function of the input sequence $x[n]$. For example, the second, third and fourth-order cumulants of zero-mean sequence $x[n]$ are, respectively, given by

$$\begin{aligned} C_{2,x}(k) &= E\{x[n]x[n+k]\} \\ C_{3,x}(k_1, k_2) &= E\{x[n]x[n+k_1]x[n+k_2]\} \\ C_{4,x}(k_1, k_2, k_3) &= E\{x[n]x[n+k_1]x[n+k_2]x[n+k_3]\} \\ &\quad -C_{2,x}(k_1)C_{2,x}(k_2 - k_3) \\ &\quad -C_{2,x}(k_2)C_{2,x}(k_3 - k_1) \\ &\quad -C_{2,x}(k_3)C_{2,x}(k_1 - k_2) \end{aligned} \quad (9)$$

where $E\{.\}$ denotes the expectation operator.

In this work, we estimate the allpass filter coefficients from the LP residual $y[n]$ by maximizing the 4th order cumulant of the input sequence $x[n]$. The objective function for maximizing the 4th order cumulant can be obtained from (8) and (9) as

$$\begin{aligned} J_4(\mathbf{a}) &= C^2_{4,x}(k_1 = 0, k_2 = 0, k_3 = 0) \\ &= \left(E\{x^4[n]\} - 3\left(E\{x^2[n]\}\right)^2\right)^2 \\ &= \left(\frac{1}{N}\sum_{n=0}^{N-1} x^4[n] - 3\left(\frac{1}{N}\sum_{k=0}^{N-1} x^2[n]\right)^2\right)^2 \end{aligned} \quad (10)$$

Notice that the input sequence $x[n]$ used in (10) has to be computed backwards (as the inverse filter is anticausal), and is given by

$$x[n] = -\sum_{k=1}^{M} a_k x[n+k] + y[n+M] + \sum_{k=1}^{M} a_k y[n+M-k] \quad (11)$$

where $n = N-1, N-2, \ldots, 1, 0$. From (11) and (10), we can see that objective function $J_4(\mathbf{a})$ is a highly nonlinear function of the filter coefficients $[\mathbf{a}]$, and it is almost impossible to arrive at a closed-form solution. Instead, Chi et. al., proposed a gradient-type iterative numerical method to search for the optimum set of filter coefficients $[\mathbf{a}]$.

The iterative method starts with an initial estimate for $[\hat{\mathbf{a}}]$, which can be small random values such that all the poles of $H_{ap}(z)$ lies inside the unit circle. With this initial estimate for the filter coefficients, compute the initial estimate of input signal $\hat{x}[n]$ using (11), and then, the evaluate the objective function $J_4(\hat{\mathbf{a}})$ using (10). The gradient of the objective function $J_4(\mathbf{a})$ with respect to $[\mathbf{a}]$ can be used to update the filter coefficients, and the method has to be repeated with the updated filter coefficients. At the ith iteration, the estimate filter coefficients is updated by

$$\hat{\mathbf{a}}(i) = \hat{\mathbf{a}}(i-1) + \mu \mathbf{g}_{i-1} \quad (12)$$

where μ is a positive constant which acts like a learning rate parameter, and $\mathbf{g}_{i-1}$ denotes the gradient of the objective

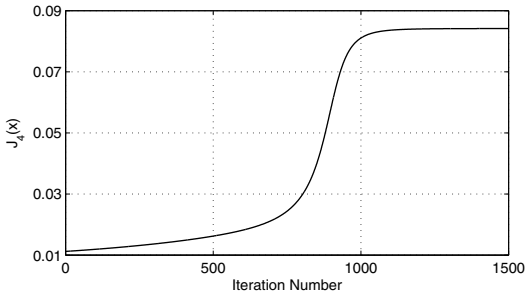

Fig. 5. Incremental change in 4th order cumulant of input $x[n]$ with iterations

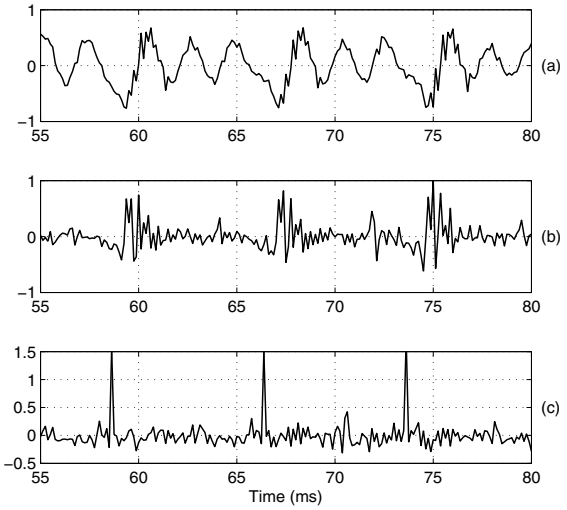

Fig. 6. A 25 ms segment of (a) voiced speech, its (b) LP residual and (c) input to the allpass filter. The speech segment used in this illustration is same as the one used in Fig. 1

function $J_4(\mathbf{a})$ with respect to $[\mathbf{a}]$ at the $(i-1)$th iteration, and is given by

$$\mathbf{g}_{i-1} = \left.\frac{\partial J(\mathbf{a})}{\partial \mathbf{a}}\right|_{\mathbf{a}=\hat{\mathbf{a}}(i-1)}. \tag{13}$$

This procedure has to be iterated until the change in the objective function falls below a predefined positive constant ϵ, i.e., $J(\hat{\mathbf{a}}(i)) - J(\hat{\mathbf{a}}(i-1)) < \epsilon$.

Fig. 5 shows the incremental change in the objective function while modeling the LP residual in Fig. 1(b) using a 20th order allpass filter. In this modeling, the learning rate parameter μ was set to 0.2 and the stopping criterion parameter ϵ was set to 10^{-8}. The estimate of the input signal $\hat{x}[n]$ obtained while modeling the LP residual in Fig. 1(b) is shown in Fig. 6(c). We refer to the estimated input signal $\hat{x}[n]$ as allpass (AP) residual. A quick observation at the LP residual and the AP residual, shown in Fig. 6(b) and Fig. 6(c) respectively, reveal that the samples in the AP residual are more independent than those in the LP residual. Notice that the AP residual exhibits sharper impulse-like excitations around the glottal closure instants than the LP residual. This might be because the phase information present in the LP residual was removed while estimating the AP residual. Hence the speaker-specific information in the phase spectrum of the LP residual might be captured in estimated allpass filter coefficients. The allpass filter coefficients are converted to cepstral domain in order to model them for speaker recognition studies [14]. The cepstral coefficients thus derived are referred to as allpass cepstral coefficients (APCCs), and are used to build speaker-specific models for speaker recognition.

III. Speaker Recognition studies

A. Description of the database

In this study, we have used TIMIT database to illustrate the effectiveness of proposed allpass modeling of LP residual for speaker recognition. TIMIT database consists of a total of 630 speakers with 10 utterances per speaker. In this study, a subset consisting 100 male speakers was considered. Out of 100 speakers, data from 50 speakers was used to build a universal background model (UBM). We have used the remaining 50 speakers for speaker recognition studies. Eight utterances (approximately 20 sec duration) from each speaker are used to adapt the UBM and build a speaker-specific model. The remaining two utterances were used to evaluate performance of the speaker-specific models. A total of 100 (50 x 2) tests, and 4900 (50 x 2 x 49) imposter tests were performed to evaluate the effectiveness of the proposed features. All the speech signals were downsampled to 8 kHz for this study.

B. Speaker modeling

The speech signal is preemphasized, and is divided into frames of 25 ms duration with an overlap of 10 ms. Each frame is multiplied by a hamming window, and 12th order LP analysis is performed on it. LP residual is derived by inverse filtering the speech frame. Allpass cepstral coefficients are derived from each frame of LP residual as described in the Section II-A by maximizing the 4th order cumulant. An energy based speech detector is employed to discard the feature vectors from low energy frames.

The proposed speaker recognition system is based on the assumption that the distribution of the APCCs is unique for each speaker. The underlying probability density function of APCCs of each speaker is approximated as a linear combination of Gaussian density functions, popularly known as Gaussian mixture modeling (GMM) [15]. The parameters, i.e mean vectors, covariance matrices and weights, of the GMM are evaluated iteratively from the training data using maximum-likelihood estimation [16]. Since this method involves estimation several parameters, it typically requires large amount of training data. In order to reduce the data requirements during the training phase, we have used universal background model (UBM) based GMM proposed by Reynolds et. al. [17].

UBM is essentially a very large GMM trained to represent the speaker independent distribution of the APCCs gathered from a large number of speakers. In this work, the UBM is built by training a 128 mixture GMM using the APCCs extracted from 50 speakers (25 minutes of speech data). The parameters of the mixtures are estimated using the maximum likelihood criterion, performing 20 iterations per mixture split [16]. When enrolling a new speaker to the system, the

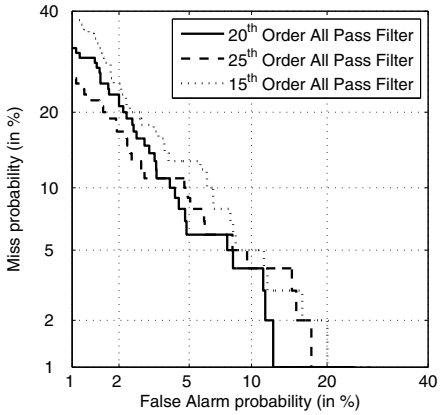

Fig. 7. DET curve showing the performance of speaker recognition system built using the APCCs obtained by maximizing 4th order cumulant

parameters of the UBM are adapted to the feature distribution of the APCCs of the new speaker. Speaker specific adaption of UBM is achieved through classical maximum a-posteriori adaptation (MAP) with one iteration and a relevance factor of 16 [17]. The MAP-adapted speaker model is used as the voice signature of that speaker.

C. Speaker evaluation

During evaluation the MAP-adapted speaker model and the UBM model are coupled, and recognizer is commonly referred to as GMM-UBM. The confidence score is computed by subtracting the average log-likelihood of the test utterance with respect to the UBM from the average log-likelihood of the test utterance with respect to the MAP-adapted speaker model. This subtraction helps in test utterance normalization, and makes the confidence scores across different test utterances comparable. The normalized confidence scores, thus obtained, are used to quantify the performance of the speaker recognition system. The performance of the proposed speaker recognition system is evaluated in terms of equal error rate (EER) which can be obtained from the detection error tradeoff (DET) curve [18]. The performance of the GMM-UBM system built using APCC features extracted from different orders of allpass filter is shown in Fig. 7. In all the three cases, the allpass modeling of LP residual is performed by maximizing the 4th order cumulant. The performance of the system these cases is similar indicating that the order of the allpass filter may not be very critical. The speaker recognition system built using APCCs derived from 20th order allpass model resulted in an EER of 6%. This result validates our hypothesis that speaker-specific information in the phase spectrum of the LP residual can be captured using allpass modeling.

IV. Summary and conclusions

Speech production mechanism can be thought of as a mixed-phase system driven by a sequence of quasi-periodic impulse like excitations. The frequency response of mixed-phase system can be decomposed into a minimum-phase component and an allpass component. LP analysis is popularly used to estimate the minimum-phase component from the speech signal. Hence the LP residual mainly contains the allpass component. In this work, we have modeled the LP residual as an output of an allpass filter excited by i.i.d. nongaussian input. The allpass filter coefficients are estimated by maximizing the higher order cumulants using the method proposed by Chi at. el. [10]. A GMM based speaker recognition system was built using the estimated allpass filter coefficients. The speaker recognition system gave an equal error rate of 6% on a population of 50 speakers, indicating that significant speaker-specific information is present in the allpass component of the speech signal.

References

[1] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, Apr. 1975.

[2] H. Wakita, "Residual energy of linear prediction applied to vowel and speaker recognition," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 24, pp. 270–271, 1976.

[3] M. Faundez and D. Rodriguez, "Speaker recognition using residual signal of linear and nonlinear prediction models," in *International conference on spoken language processing*, 1998.

[4] P. Thevenaz and H. Hugli, "Usefulness of LPC residue in text-independent speaker verification," *Speech Communication*, vol. 17, pp. 145–157, 1995.

[5] J. H. L. Liu and G. Palm, "On the use of features from prediction residual signal in speaker recognition," in *European conference on speech processing technology*, pp. 313–316, 1997.

[6] S. R. M. Prasanna, C. S. Gupta, and B. Yegnanarayana, "Extraction of speaker-specific excitation information from linear prediction residual of speech," *Speech Communication*, vol. 48, pp. 1243–1261, 2006.

[7] K. S. R. Murty and B. Yegnanarayana, "Combining evidence from residual phase and MFCC features for speaker recognition," *IEEE Signal Process. Lett.*, vol. 13, pp. 52–56, Jan. 2006.

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

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

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

[11] H.-M. Chien, H.-L. Yang, and C.-Y. Chi, "Parametric cumulant based phase estimation of 1-D and 2-D nonminimum phase systems by allpass filtering," *IEEE Trans. Signal Processing*, vol. 45, pp. 1742–1762, July 1997.

[12] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*. Prentice Hall, 1999.

[13] J. M. Mendel, "Tutorial on higher order statistics (spectra) in signal processing and system theory," *Proc. IEEE*, vol. 79, pp. 278–305, Mar. 1991.

[14] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, New Jersy: Prentice-Hall, 1993.

[15] D. A. Reynolds, "Speaker identification and verification using Gaussian mixture speaker models," *Speech Communication*, vol. 17, no. 1–2, pp. 91–108, 1995.

[16] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. Wiley-Interscience, 2 ed., 2001.

[17] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 19–41, 2000.

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