

The Effectiveness of Higher Order Spectral Phase Features in Speaker Identification

Daryl Ning, Vinod Chandran

Speech, Audio, Image, and Video Technology Programme
School of Electrical and Electronic Systems Engineering
Queensland University of Technology,
GPO Box 2434
Brisbane Qld 4001

{d.ning, v.chandran}@qut.edu.au

Abstract

This paper studies the effectiveness of higher order spectra (HOS) phase features in the task of speaker identification. Within the speech processing community, short time spectral phase information is generally regarded as unimportant for speaker recognition. In fact, the most commonly used features for speaker recognition are the Mel frequency cepstral coefficients (MFCC), which are defined from the magnitude spectrum only. By discarding the phase, however, we lose the ability to discriminate between two different signals with the same amplitude spectrum, for example, a time-reversed version of the input and itself. In our experiments, we utilise features that contain both magnitude and phase spectral information. These HOS phase features are derived by integrating points along a straight line in bifrequency space. Clean microphone speech from a 20 male speaker database is used, and Gaussian mixture models (GMM) are constructed from the set of extracted features. The HOS phase features achieve a correct identification rate of 98.5%, which is similar to the rate achieved by the MFCC feature set (99.4%). The usefulness of short time phase spectral information is also verified by performing experiments after removing the magnitude spectral information from the speech data. The HOS phase features are also shown to be more robust to additive white Gaussian noise in mismatched training and testing conditions than MFCCs.

1. Introduction

A speech signal conveys more information than just the words being spoken. It also contains information about the identity of the person speaking those words. Speaker identification is concerned with extracting the correct identity of the person (from a known set) speaking a given utterance. This is generally performed by extracting features from the given speech signal, and comparing them with a stored set of feature models belonging to known speakers. Applications include security access and forensic science.

Most speech features used in speaker recognition (identification or verification) systems are derived from second order statistics, such as linear prediction and the power spectrum. Mel frequency cepstral coefficients (MFCC), for example, are derived from the power spectrum and have been shown to provide good results in speaker recognition [1–4]. These features, however, ignore phase information in the Fourier spectrum. While most of the perceptual information about speech resides in the

amplitude, phase information has also been shown to be important [5]. At the very least, there is loss of the ability to discriminate between a time-reversed version of the input and itself if all Fourier phase information is discarded.

In this paper, we utilise a set of features derived from higher order spectra (HOS) [6]. The performance of these features is compared with MFCC features in an identical speaker identification system. We also compare the sensitivity of each feature set to additive white Gaussian noise (AWGN) in mismatched training and testing conditions.

Section 2 introduces HOS and section 3 describes the HOS phase parameters used as features in our speaker identification experiments. Section 4 describes the setup of the speaker identification system which is common to all our experiments. Section 5 presents the results on each of the experiments performed, accompanied by a brief discussion. Finally, a conclusion is given in section 6.

2. Higher Order Spectra

While the power spectrum is derived from *second* order statistics, HOS are derived from *higher* order statistics. The bispectrum and trispectrum, for example, are the Fourier transforms of the third and fourth order correlations¹ of the signal respectively. If $x(t)$ is a stationary random process, then its n^{th} order moments, $m_n(\tau_1, \tau_2, \dots, \tau_{n-1})$, can be defined as

$$m_n(\tau_1, \tau_2, \dots, \tau_{n-1}) = E[x(t)x(t + \tau_1) \dots x(t + \tau_{n-1})] \quad (1)$$

where $E[\cdot]$ is the expected-value operator. The power spectrum is defined as the Fourier transform of $m_2(\tau_1)$. The power spectrum at frequency f_1 can be estimated by

$$P_e(f_1) = E[X(f_1)X^*(f_1)] \quad (2)$$

where $X(f)$ is the Fourier transform of a windowed realisation of $x(t)$. Similarly, the bispectrum and trispectrum can be estimated by

$$B_e(f_1, f_2) = E[X(f_1)X(f_2)X^*(f_1 + f_2)] \quad (3)$$

and

$$T_e(f_1, f_2, f_3) = E[X(f_1)X(f_2)X(f_3)X^*(f_1 + f_2 + f_3)] \quad (4)$$

¹This refers to moment based spectra as opposed to cumulant based spectra. For more information, see [7].

respectively.

Equation 2 shows that the power spectrum is completely defined by the magnitude of the Fourier coefficients. Equations 3 and 4, however, show that the bispectrum and trispectrum retain both the phase and amplitude information from the Fourier transform. This is true for HOS in general.

Another important property of the bispectrum is that it has zero expected value for Gaussian signals. Features that are derived from the bispectrum will therefore have high immunity to AWGN when the bispectra are averaged from multiple realisations of the signal. In fact, even with a single realisation, it was shown [6] that noise rejection still results from the averaging that occurs if we integrate many bispectral values along a radial line in bifrequency space. This process of integration is explained further in section 3.

References to seminal and review papers in the field of HOS can be found in the reference lists of [8] and [6].

3. HOS Phase Features

The features used in our experiments are derived from the discrete bispectrum of deterministic signals. The bispectrum of a one-dimensional, deterministic, discrete time signal, $x(n)$, is defined here by

$$B(f_1, f_2) = X(f_1)X(f_2)X^*(f_1 + f_2) \quad (5)$$

where $X(f)$ is the discrete time Fourier transform of $x(n)$ at normalised frequency, f . Note that this is a deterministic framework and there is no expectation operation on the right hand side. If the one-dimensional signal is divided into blocks, the triple products above can be averaged to yield the more conventional estimate of the bispectrum used in higher-order statistics, i.e., the Fourier transform of the third-order correlation of the signal. A set of features based on bispectral phases was derived by Chandran and Elgar [6], and is described briefly below.

Assuming there is no bispectral aliasing, the bispectrum of a real signal is uniquely defined within the triangle $0 \leq f_2 \leq f_1 \leq f_1 + f_2 \leq 1$. Parameters are obtained by integrating along straight lines passing through the origin in bifrequency space. The region of computation and line of integration are depicted in Fig. 1. The bispectral invariant, $P(a)$, is the phase of the integrated bispectrum along the radial line with slope equal to a . This is defined by

$$P(a) = \arctan \left(\frac{I_i(a)}{I_r(a)} \right) \quad (6)$$

where

$$\begin{aligned} I(a) &= I_r(a) + jI_i(a) \\ &= \int_{f_1=0^+}^{\frac{1}{1+a}} B(f_1, af_1) df_1 \end{aligned} \quad (7)$$

for $0 < a \leq 1$, and $j = \sqrt{-1}$.

The variables I_r and I_i refer to the real and imaginary parts of the integrated bispectrum respectively. The HOS phase parameters exploit the relationship between the shape of a deterministic signal (or block of speech) and the phase of its deterministic bispectrum. This shape contains information about speech and speaker, as do Mel-Cepstral features. A statistical model of features, such as a Gaussian mixture, that is trained over many speech blocks from the speaker will tend to become speech independent and can be used for speaker identification. For good discriminability between speakers, the feature

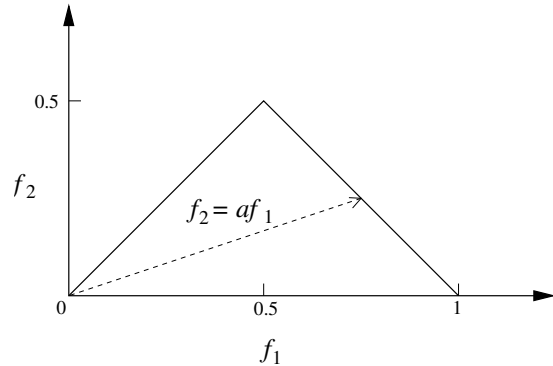


Figure 1: *Region of computation of the bispectrum for real signals. Features are calculated by integrating the bispectrum along the dashed line with slope= a .*

set must be sensitive to small changes in the shape of the signal between speakers for the same speech. At the same time, the features must be invariant or robust to changes in amplitude (decibel-level) and time-shifts caused by changes in sampling or segmentation. The phase of the integral in equation 7 is shown to be invariant to translation, scaling, amplification, and DC-shifting [6]. If features are robust to such transformations, there is less intra-class variance and the probability density will be more dependent on changes that discriminate between speakers. In the remainder of this paper, any mention of ‘HOS phase parameters (or features)’ refers to the set of parameters defined by equation 6.

4. Experimental Setup

Speaker identification experiments were performed using HOS phase parameters, but experiments were also performed using MFCC parameters for comparative purposes. Apart from the features used, the experimental setup for both were identical. The following subsections describe the setup of these experiments.

4.1. Speech data

The speech for this experiment was obtained from the multi-modal task evaluation data used in the 2002 National Institute of Standards and Technology (NIST) speaker recognition evaluation. We used the training data from the spontaneous speech recorded via a high quality tabletop microphone. The data consists of 4 sessions of speech, each being 29 seconds in length. 3 sessions were used for training, and the final session was used for testing. Not all of the speech data is actually used for feature extraction, however, and this is explained in section 4.2. Each session was recorded using the same microphone and sampled at 16 kHz, but we first filter and down-sample each of the speech files to 8 kHz before processing. A total of 20 male speakers was used in this experiment.

4.2. Feature Extraction

Before features are calculated, the input speech frame, $x(n)$, is first classified as voiced, unvoiced, or silence. Only the voiced speech frames are utilised in these experiments, since voiced segments contain the appropriate harmonic structure that give rise to significant bispectral values [9–11]. The voicing decision

is determined using the algorithm from the LPC-10E speech coder [12]. Since the speech data from each of the speakers have varying amounts of voiced speech, the amount of data used for training and testing is not the same for each of the speakers.

Each frame of speech, $x(n)$, consists of 256 samples with a frame advance (hop) of 80 samples. This equates to 32 ms and 10 ms respectively, hence 100 frames are processed every second. For the HOS phase features, the bispectrum is calculated from each $x(n)$ and the parameters, $P(a_i)$, are determined, where $a_i = i/D$ and $i = 1 \dots D$. In this work we choose $D = 16$, therefore, we obtain a *feature vector* of 16 integrated phase parameters for each $x(n)$. These phase parameters are not unwrapped such that $-\pi < P(a_i) \leq \pi$. A total of 12 parameters are calculated for each MFCC feature vector.

4.3. Speaker Modelling

Each speaker's collection of feature vectors needs to be modelled in a manner that will allow us to effectively distinguish one speaker from another. We choose a probabilistic model, specifically a Gaussian mixture model (GMM), to represent the distribution of these feature vectors. A GMM is simply a weighted sum of M Gaussian densities, and in this work, the densities are multivariate. GMM's are popular in speaker recognition systems for two reasons. Firstly, it is assumed that the individual components are capable of modelling some underlying set of broad phonetic events, e.g. vowels, fricatives [1]. Secondly, a GMM is capable of smoothly approximating many arbitrarily shaped densities. An explanation of GMM's and procedures for estimation of their mixture weights and densities are given in [1]. After estimating the GMM from a particular speaker's training speech, he/she is represented by the model, $\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\}$, where $i = 1 \dots M$, and p_i , $\vec{\mu}_i$ and Σ_i are the mixture weight, mean vector and covariance matrix of the i^{th} mixture respectively. In this work, diagonal covariance matrices are used, and a value of 32 was chosen for M .

4.4. Speaker Identification

To perform speaker identification, the goal is to find which, out of a group of S speaker models, is most likely to produce the observation sequence, $X = \{\vec{x}_1, \dots, \vec{x}_T\}$. X is simply a sequence of T feature vectors extracted from the given speech. Assuming equally likely speakers and noting that $p(X)$ is the same for all speaker models, we classify the speaker based on the following:

$$\hat{S} = \arg \max_{1 \leq k \leq S} p(X|\lambda_k) \quad (8)$$

where λ_k is the GMM for the k^{th} speaker. Assuming independence between observations, this becomes:

$$\hat{S} = \arg \max_{1 \leq k \leq S} \sum_{t=1}^T \log p(\vec{x}_t|\lambda_k) \quad (9)$$

where

$$p(\vec{x}_t|\lambda) = \sum_{i=1}^M p_i b_i(\vec{x}_t) \quad (10)$$

and $b_i(\vec{x}_t)$ is the i^{th} component of the multivariate Gaussian mixture density evaluated for the feature vector, \vec{x}_t , and p_i 's are the priors (mixture weights).

4.5. Performance Evaluation

To evaluate the speaker identification system, the test speech is divided into overlapping segments of observation sequences. Each sequence consists of $T = 200$ feature vectors, corresponding to 2 second utterances. An observation advance of 10 ms is used, hence each sequence differs from the previous sequence by only one feature vector. For example, the first two segments would be

$$\begin{array}{c} \text{Segment 1} \\ \overbrace{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_T, \vec{x}_{T+1}, \vec{x}_{T+2}, \dots} \\ \text{Segment 2} \\ \overbrace{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_T, \vec{x}_{T+1}, \vec{x}_{T+2}, \dots} \end{array}$$

For each segment, equation 9 is used to determine the speaker which gives the maximum probability for the observation sequence. This is repeated for all the possible segments in the length of the test speech and across all the speakers in the database. The correct identification rate is then calculated as

$$\% \text{ Correct Identification} = \frac{\text{Total \# of correctly identified segments}}{\text{Total \# of segments}} \times 100. \quad (11)$$

5. Results and Discussion

5.1. Speaker Identification Using Original Speech Data

The first set of experiments performed evaluated the basic speaker recognition system described in section 4. Speaker models were trained using the 16 integrated phase parameters extracted from each voiced frame of the training speech. The test speech from each speaker was then stochastically compared with each of these models. The correct identification rate was computed using the procedure outlined in sections 4.4 and 4.5. Using the HOS phase parameters as features, the 20 male speaker system achieved a correct identification rate of 98.5%. It must be noted, however, that this result is biased since the amount of training and testing speech differs between the speakers.

From the above result, it is evident that phase based parameters can be useful for speaker identification. Even if the short time Fourier phase spectrum is not directly useful for the task, parameters can be defined from higher-order spectra that capture useful information including that from the phase spectrum. The simple speaker identification experiment above shows that the integrated bispectral phase parameters hold important information capable of discerning known speakers within a database.

For comparison, a second speaker identification experiment was performed. This experiment was almost identical to the first, with the only difference being the choice of feature vectors. A set of 12 MFCC parameters were used instead of the 16 integrated phase parameters. The MFCC based system achieved a correct identification rate of 99.4%.

From this result we can see that the HOS phase parameters can produce comparable results to the more widely used MFCC parameters when utilised as features in a simple speaker identification task. In fact, when the test segment used for identification is increased from 2 seconds to 4 seconds, the correct identification rate improves to 100% for both feature sets.

5.2. Speaker Identification Using ‘Phase Only’ and ‘Magnitude Only’ Speech Data

Since the HOS integrated phase parameters contain both phase and magnitude spectral information, one may suggest that the high correct identification rate for the HOS parameters can be attributed to the magnitude information. To disprove this suggestion, we performed additional tests using ‘phase only’ and ‘magnitude only’ speech data. In order to perform these tests, additional preprocessing was performed on each frame of speech, $x(n)$, before the features were extracted. To preserve the phase spectral information only, while discarding the magnitude information, we perform the following procedure (illustrated in figure 2(a)) :

1. Take the DFT of $x(n)$ to obtain $X(k)$.
2. Set all the amplitudes of $X(k)$ to unity while preserving all the phases. $X(k)$ becomes $\tilde{X}(k)$.
3. Take the inverse DFT of $\tilde{X}(k)$ to obtain $\tilde{x}(n)$.
4. Extract the features from $\tilde{x}(n)$ as normal.

This procedure is identical to that used by Oppenheim *et al.* [13] except that we are processing short time segments of speech as opposed to long time segments. To preserve the magnitude spectral information only, while discarding the phase information, we perform the following procedure (illustrated in figure 2(b)):

1. Take the DFT of $x(n)$ to obtain $X(k)$.
2. Replace all the phases of $X(k)$ with values chosen from a set of normally distributed random phases, while preserving all the amplitudes. $X(k)$ becomes $\tilde{X}(k)$.
3. Take the inverse DFT of $\tilde{X}(k)$ to obtain $\tilde{x}(n)$.
4. Extract the features from $\tilde{x}(n)$ as normal.

The overall correct identification rates using the ‘magnitude only’ and ‘phase only’ speech data for both MFCC and HOS phase parameter sets are given in table 1. The results using the original speech are also included for comparison.

The results for the MFCC parameters are to be expected. Since MFCC parameters are derived from the magnitude spectrum, the loss of phase spectral information should have no effect on its performance, and this was the case. The loss of magnitude spectral information, however, should cause the system to fail, and this was also the case. A correct identification rate of only 5.65% was obtained when using the ‘phase only’ data. This is roughly equivalent to guessing 1 person from a group of 20 speakers.

The correct identification rate for the HOS phase parameters using ‘magnitude only’ and ‘phase only’ speech was 16.4 % and 77.3 % respectively. Since the magnitude information is not neglected completely in the calculation of the HOS phase parameters (they provide a weighting for the integrated phases), the ‘magnitude only’ data performs better than the equivalent guessing rate of 5 %. The integrated magnitudes tend to provide a *radial* spectrum across the different values of a in $P(a)$. These magnitudes on their own, however, do not provide us with sufficient information to discern between speakers. The integrated phase information, on the other hand, can alone provide a feature set that performs reasonably well for speaker identification.

The success of the integrated phase information, independent of the magnitude information, is an important result when considering the effects of channel degradation on the performance of the system. It is well known that Fourier phase is

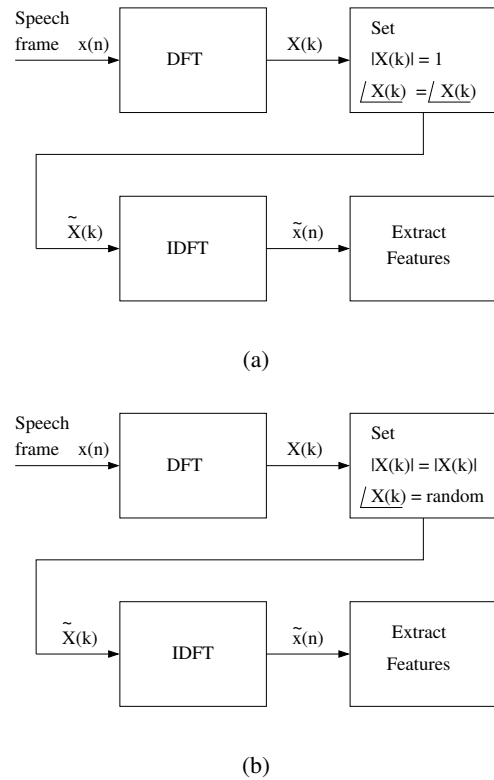


Figure 2: Preprocessing required to (a) preserve the phase spectral information while discarding the magnitude information, and (b) preserve the magnitude spectral information while discarding the phase information.

more robust to channel fading and additive noise than Fourier magnitude. In such cases, the HOS phase feature set could theoretically provide better performance than the Fourier magnitude based MFCCs.

It should be noted that the results for the HOS feature set using magnitude or phase only data would change significantly with the type of input signal. Using knowledge of the preprocessing procedure, the given results suggest that the input data is from a broadband source. If the source had been narrowband, such as a single sinusoid in noise, the accuracy for the ‘phase only’ data would decrease. Using a narrowband source, the phases best representative of the speaker (with most weight) would lie in the bandpass region. Setting the spectrum to unity magnitude, however, places equal weighting across all the phases to be integrated in bispectrum space. The phases from outside the original bandpass region would now act as greater noise in the feature space, leading to poorer results.

5.3. Effects of AWGN

Experiments were also performed to compare the effects of AWGN at varying signal-to-noise ratios (SNR), on the correct identification rate for each feature set. The results of these tests are illustrated in figure 3. In each of these tests, the speaker models remained trained on clean speech, however, WGN was added to each of the test speech utterances. We therefore have mismatched training and testing conditions, which is more typical of real operating conditions. No other changes

Table 1: Correct identification rates using ‘magnitude only’ and ‘phase only’ speech data.

Input	MFCC Feature Set	HOS Phase Feature Set
Original speech	99.4	98.5
Magnitude only	99.2	16.4
Phase only	5.65	77.3

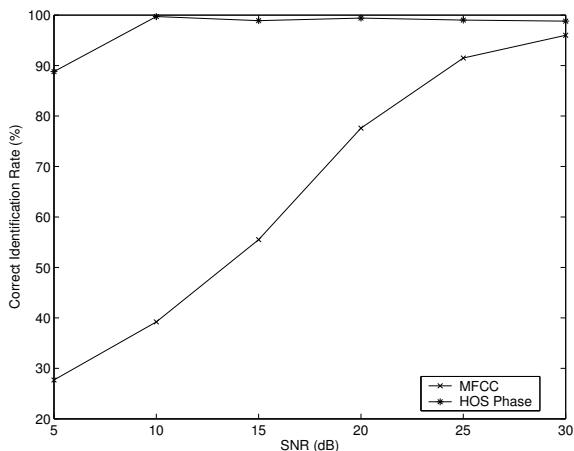


Figure 3: % Correct identification rate versus SNR (for AWGN) when using the MFCC and HOS phase feature sets. The system operates under mismatched conditions.

were made to the original setup described in section 4. Even though techniques exist, such as cepstral mean subtraction [14] and RASTA [15] processing, to make MFCC’s more robust to channel mismatch, we wished to keep both systems identical apart from the feature set. It should also be noted that less testing speech is available in the presence of AWGN, especially at low SNR’s. This is because the system only utilises the voiced speech frames of each speaker.

As the SNR is decreased from 30dB to 5dB, the correct identification rates using MFCC parameters decreases almost linearly from 96% down to 27.7%. The correct identification rates using HOS phase parameters remains around 99% for SNR’s above 10dB. At 5dB the accuracy drops to a, still reasonable, 88.8%. The HOS phase features are, therefore, more robust to AWGN than the MFCC feature set. On the other hand, the MFCC’s may provide a much more robust feature set in the presence of a varying channel with a non-linear phase response. Since both feature sets can potentially complement each other, it would be advantageous to investigate the use of both sets of parameters in a fused classification system for improved accuracy.

6. Conclusion

In this paper, we have shown that the HOS phase based parameters derived in [6], contain information that is useful in discern-

ing speakers within a small speaker set (20 male speakers). On clean microphone speech and under identical conditions, they perform on a similar level as the widely used MFCC parameters. Further experiments showed that the phase spectral information plays a crucial role in the performance of the speaker identification system. Since phase-based parameters are more robust to additive noise than magnitude based parameters, there is the possibility to combine the HOS phase and MFCC feature sets in a fused classification system for improved accuracy. This will be the focus of future work.

Acknowledgments This research was supported by the Australian Research Council through the Large Grants Scheme, Grant A00106132, 2001-2003. We are grateful to NIST for making speaker recognition evaluation data available.

7. References

- [1] D. A. Reynolds and R. C. Rose, “Robust text-independent speaker identification using gaussian mixture speaker models,” *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, January 1995.
- [2] D. A. Reynolds, “Large population speaker identification using clean and telephone speech,” *IEEE Signal Processing Letters*, vol. 2, no. 3, pp. 46–48, March 1995.
- [3] L. Liu, J. He., and G. Palm, “Signal modeling for speaker identification,” *ICASSP*, vol. 2, pp. 665–668, 1996.
- [4] L.P. Cordella, P. Foggia, C. Sansone, and M. Vento, “A real-time text independent speaker identification system,” *Proceedings of the 12th International Conference on Image Analysis and Processing*, pp. 632–637, 2003.
- [5] H. Pobloth and W. B. Kleijn, “On phase perception in speech,” *ICASSP*, vol. 1, pp. 29–32, 1999.
- [6] V. Chandran and S. L. Elgar, “Pattern recognition using invariants defined from higher order spectra—one dimensional inputs,” *IEEE Transactions on Signal Processing*, vol. 41, no. 1, pp. 205–212, January 1993.
- [7] Vinod Chandran, “On the computation and interpretation of auto- and cross-trispectra,” *ICASSP*, vol. 4, pp. 445–448, 1994.
- [8] S. Elgar and V. Chandran, “Higher order spectral analysis to detect nonlinear interactions in measured time series and an application to chua’s circuit,” *International Journal of Bifurcation and Chaos*, vol. 3, no. 1, pp. 19–34, January 1993.
- [9] B. B. Wells, “Voiced/unvoiced decision based on the bispectrum,” *ICASSP*, vol. 10, pp. 1589–1592, 1985.
- [10] B. Boyanov, S. Hadjitodorov, and T. Ivanov, “Analysis of voiced speech by means of bispectrum,” *Electronic Letters*, vol. 27, no. 24, pp. 2267–2268, 1991.
- [11] J. W. A. Fackrell and S. McLaughlin, “The higher-order statistics of speech signals,” *IEE Colloquium on Techniques for Speech Processing and their Applications*, pp. 7/1–7/6, 1994.
- [12] J. P. Campbell Jr and T. E. Treiman, “Voiced/unvoiced classification of speech with application to the u.s. government LPC-10E algorithm,” *ICASSP*, pp. 473–476, 1986.

- [13] A. V. Oppenheim, J. S. Lim, G. Kopec, and S. C. Pohlig, "Phase in speech and pictures," *ICASSP*, vol. 4, pp. 632–637, 1979.
- [14] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of the Acoustical Society of America*, vol. 55, pp. 1304–1312, June 1974.
- [15] Hynek Hermansky and Nelson Morgan, "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578–589, October 1994.