

Bispectrum: Applications to Biometric Identification

Vinod Chandran, *Senior Member, IEEE*

Abstract—The bispectrum retains Fourier phase information unlike the power spectrum. It is well known that there is shape information in the Fourier phase of signals that can be useful for classification. Features can be extracted using the bispectrum that are also invariant to translation and scaling, and robust to noise. For a non-stationary signal, bispectral invariant features can be extracted from segments that are sufficiently short to be quasi-stationary. Probability distributions of the features can be learned using Gaussian Mixture models and the Expectation Maximization algorithm. These models contain information about the source of the signal and can be used for biometric identification provided text (or content) and channel (or sensor) dependence is removed or separated. This paper will present such a methodology and successfully apply it to speaker recognition and online signature verification.

Index Terms—Bispectrum, higher order spectra, biometric identification, voice recognition, handwriting recognition, signature verification.

I. INTRODUCTION

THE bispectrum retains Fourier phase information unlike the power spectrum. It is well known that there is shape information in the Fourier phase of signals that can be useful for classification. Features can be extracted using the bispectrum that are also invariant to translation and scaling, and robust to noise. For a non-stationary signal, bispectral invariant features can be extracted from segments that are sufficiently short to be quasi-stationary. Probability distributions of the features can be learned using Gaussian Mixture Models and the Expectation Maximization algorithm. These models contain information about the source of the signal and can be used for biometric identification provided text (or content) and channel (or sensor) dependence is removed or separated. This methodology has been applied successfully to speaker recognition and online signature verification.

For text-independent speaker recognition, bispectral features are shown to perform better than the traditional Mel-frequency Cepstral Coefficient (MFCC) features with

microphone speech when the Signal to Noise ratio is poor. They perform comparably for clean speech and provide complementary information in tests conducted using the YOHO database and the Switchboard telephone speech Corpus. Models generated from each feature set were also compared using the Bhattacharya distance between mode centre distributions indicating that the bispectral feature set may be more discriminative.

Online signature data can be viewed as variations in time of position, pressure and pen angles. These signals and their derivatives can be processed to extract bispectral invariant features and Gaussian Mixture Models for signatures of each person. The system does not rely on normalization and is robust to changes in size and orientation of the signature. A real-time system for signature verification has been implemented.

II. HIGHER ORDER (BI-)SPECTRAL FEATURES

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:

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

where $X(f)$ is the Fourier transform of a windowed realisation of $x(t)$ and $*$ represents complex conjugation. Unlike the power spectrum, higher order spectra (HOS) retain both the phase and amplitude information from the Fourier transform. The bispectrum, $B(f_1, f_2)$, of a one-dimensional, deterministic, discrete-time signal, $x(n)$, is defined by:

$$B(f_1, f_2) = X(f_1)X(f_2)X^*(f_1 + f_2) \quad (3)$$

where $X(f)$ is the discrete time Fourier transform of $x(n)$ at normalised frequency, f . Equation (2) shows that the power spectrum is completely defined by the magnitude of the

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

Manuscript received April 2, 2004. The work on speaker recognition and face recognition has been supported by the Australian Research Council in recent years Grants A00106132 (2001-4) and DP0558415 (2005-7).

V. Chandran is with the School of Engineering Systems, Queensland University of Technology, Brisbane 4001 Australia (phone: +61 7 3864 2124; fax: +61 7 3864 1516; e-mail: v.chandran@qut.edu.au).

Fourier coefficients because the phase of $X(f)$ gets cancelled in the product with $X^*(f)$, its conjugate. In Eqn. (3), however, the bispectrum retains both the phase and the amplitude information from the Fourier transform. Only the linear (with frequency) phase component is cancelled in the triple product, and this component is related to the shift of a finite duration signal rather than its shape.

Note, that the bispectrum in Eqn. (3) is in 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, which is the Fourier transform of a third-order correlation of the signal. 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 additive white Gaussian noise (AWGN) when the bispectra are averaged from multiple realisations of the signal. In fact, even with a single realisation, it was shown [2] that noise rejection still results from the averaging that occurs if many bispectral values are integrated along a radial line in bifrequency space. References to seminal and review papers in the field of higher-order spectra can be found in [2] and [4].

The normalised magnitude of $E[B(f_1, f_2)]$ gives an indication of the relative degree of phase coupling between triads of Fourier components [3]. However, phase coupling is not the information extracted here because no averaging is performed. A more direct relationship between the shape of a deterministic signal and the phase of its deterministic bispectrum is exploited. The bispectrum retains phase information from the signal unlike the power spectrum and it is sensitive to asymmetry (or irreversibility with respect to the time axis) of the signal. Time reversal of $x(n)$ results in a sign reversal of the phase of the bispectrum in Eqn. 3, as can be shown using simple properties of the Fourier transform [2].

A set of features based on bispectral phases that are translation, time-scaling, amplitude-scaling, mean-shift invariant was derived by Chandran and Elgar [2] and is described briefly here. 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 (or slices) passing through the origin in bifrequency space. The region of computation and line of integration are depicted in Fig. 1. The phase and bispectral invariant, $P(a)$, is the phase of the integrated bispectrum along the radial line with slope equal to a . This is defined by:

$$P(a) = \angle I(a) = \arctan\left(\frac{I_i(a)}{I_r(a)}\right) \quad (4)$$

where

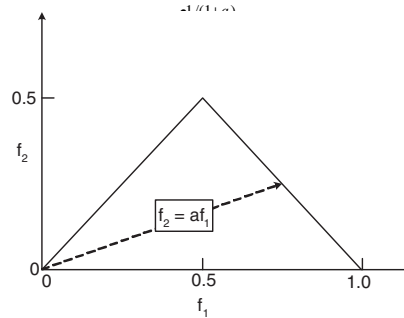


Fig. 1. Region of computation of the bispectrum for real signals. Features are calculated by integrating the bispectrum along the dashed line with slope = a . Frequencies are normalized by the Nyquist frequency. in the discrete bifrequency plane as done in [2, 4] or by computing the DFT exactly at points on polar rasters as explained in methods in [5]. The latter approach is preferred when interpolation error can be a factor. The computational complexity is increased when many slopes are considered but for time segments that are short, verification in real-time is possible despite of it. In the work described here, the former approach is used in section III and [28],[29] while the latter is used for signature verification in section IV where segments consist of fewer samples.

III. SPEAKER IDENTIFICATION USING HOS

A. Background

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 establishing the correct identity of the person (from a known set) using speech as a biometric. 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 of speaker identification include secure access systems and forensic investigation.

Most speech features used in speaker recognition identification or verification) systems are derived from second order statistics, using linear prediction or the power spectrum. Mel frequency cepstral coefficients (MFCC), for example, are derived from the power spectrum. MFCC have been shown to provide good results in speaker recognition [6, 7, 8, 9]. These features, however, ignore spectral phase information which has been considered within the speech processing community as unimportant for speaker recognition. Since MFCC features are derived from Fourier magnitude they are also quite sensitive to additive noise.

While most of the perceptual information about speech resides in the amplitude, phase information has also been suggested to be important [10, 11]. At the very least, from a pattern recognition point of view at a speech frame or block level, 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. Whether this loss results in a loss of intelligibility or a loss of discrimination between speakers is investigated here. The effectiveness of higher-

order spectral (HOS) phase features in speaker recognition is investigated by comparison with Mel Cepstral features on the same speech data. Telephone speech and microphone speech were used in the experiments. However, channel and handset variations were intentionally kept to a minimum in the former because they were not the object of the study. Three different speech corpuses were used.

Additional experiments are conducted to compare the performance of the two feature sets using magnitude-only and phase-only information, and in the presence of varying amounts of additive white Gaussian noise.

B. Experiments

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.

1) Speech data

The first set of experiments used speech data obtained from the Switchboard-2 Phase I telephone speech corpus. Since the intention is to investigate variability of the features between speakers and not owing to channel, handset or recording condition variations, speech files were selected keeping the channel quality ratings as similar as possible. A total of 19 speakers (10 male and 9 female) were selected from this database, and from each speaker, one conversation was selected. This conversation was then split into two non-overlapping parts. One part was used for training, and the other was used for testing.

The speech for the second set of experiments was obtained from the multi-modal task evaluation data used in the 2002 National Institute of Standards and Technology (NIST) speaker recognition evaluation. The primary purpose of this test was to confirm that the discrimination achieved in the first experiment was not owing to channel or handset variations, and only owing to voice, because these variations are not present with recordings from the same microphone. Training data from the spontaneous speech recorded via a high quality tabletop microphone was used. 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. Each session was recorded using the same microphone and sampled at 16 kHz, but each of the speech files was filtered and down-sampled to 8 kHz before processing. Data from 20 male speakers was used in this experiment.

The speech for the third set of experiments was obtained from the YOHO voice verification corpus [12] and the intention was to test how the performance of the two feature sets scale with increasing number of speakers. This corpus consists of 138 speakers (106 males, 32 females). Each speaker has 4 enrolment sessions of 24 utterances each (total of 96), and 10 verification sessions of 4 utterances each (total of 40). These are all prompted combination lock phrases, so only digits are spoken. Speech is recorded via a high quality

telephone handset (but not passed through a telephone channel) and sampled at 8 kHz.

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 [13, 14, 15]. Whether unvoiced segments behave differently for the two feature sets is not the subject of this study. The voicing decision is determined using the algorithm from the LPC-10E speech coder [16]. 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 $D = 16$, and therefore, a *feature vector* of 16 integrated phase parameters is obtained 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.

3) Speaker Modelling

The relationship between the shape of the speech signal and the phase of its deterministic bispectrum is exploited. This shape contains information about speech and speaker, as do MFCCs. 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 (SI). For good discriminability between speakers, the feature 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. 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.

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. A probabilistic model, specifically a Gaussian mixture model (GMM), is chosen 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 [6]. 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 [6]. After estimating the GMM from a particular speaker's training

speech, he/she is represented by the model, $\lambda = \{p_i, \bar{\mu}_i, \Sigma_i\}$, where $i = 1 \dots M$, and p_i , $\bar{\mu}_i$ 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 $M = 32$.

4) Speaker identification

The task in SI is to correctly identify a speaker (from a group of known speakers) given some of his/her speech. This is achieved by finding which, out of a group of S speaker models, is most likely to produce the observation sequence, $X = \{\mathbf{x}_1, \dots, \mathbf{x}_T\}$. 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, the speaker is identified based on the following:

$$\hat{X} = \arg \max_{1 \leq k \leq S} p(X | \lambda_k) \quad (6)$$

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

$$\hat{X} = \arg \max_{1 \leq k \leq S} \sum_{i=1}^T \log p(\bar{x}_i | \lambda_k) \quad (7)$$

where

$$p(\bar{x}_i | \lambda_k) = \sum_{i=1}^M p_i b_i(\bar{x}_i) \quad (8)$$

Table 1. Percent correct ID for tests using MFCC and HOS phase parameters.

| Speaker | MFCC Feature Set | HOS Phase Feature Set | Speaker | MFCC Feature Set | HOS Phase Feature Set |
|---------|------------------|-----------------------|---------|------------------|-----------------------|
| 1 | 100 | 100 | 11 | 100 | 100 |
| 2 | 100 | 100 | 12 | 100 | 78.1 |
| 3 | 90.0 | 100 | 13 | 100 | 100 |
| 4 | 100 | 100 | 14 | 100 | 100 |
| 5 | 100 | 97.6 | 15 | 99.6 | 99.8 |
| 6 | 100 | 100 | 16 | 100 | 100 |
| 7 | 100 | 91.8 | 17 | 100 | 100 |
| 8 | 80.7 | 92.1 | 18 | 100 | 100 |
| 9 | 100 | 100 | 19 | 100 | 100 |
| 10 | 100 | 90.5 | | | |

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

5) Performance evaluation

To evaluate the SI 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:

$$\begin{array}{c} \text{Segment 1} \\ \bar{x}_1, \bar{x}_2, \dots, \bar{x}_T, \bar{x}_{T+1}, \bar{x}_{T+2}, \dots \\ \text{Segment 2} \\ \bar{x}_1, \bar{x}_2, \dots, \bar{x}_T, \bar{x}_{T+1}, \bar{x}_{T+2}, \dots \end{array} \quad (9)$$

For each segment, Eqn. (7) 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:

$$\% \text{ Correct Identification} = \frac{\text{Total \# of correctly identified segments}}{\text{Total \# of segments}} \times 100 \quad (10)$$

C. Results and Discussion

The following results show that HOS phase parameters contain information that can be used to recognize speakers. They are shown to perform at the same level as MFCCs on the same data under identical conditions. They contain information that can complement MFCCs in a fused classifier. The widely accepted notion that Fourier phase information is unimportant in speech processing is questioned by experimental evidence.

1) Speaker Identification Using Telephone Speech

The first set of experiments used speech data obtained from the Switchboard-2 Phase I telephone speech corpus. Using the 16 integrated phase parameters as features, a correct identification rate of 97.46% was achieved. This simple SI experiment shows that the integrated bispectral phase parameters hold information capable of discerning known speakers within a database. A set of 12 MFCC parameters were used in the comparison experiment, (instead of the 16 integrated phase parameters), and all other aspects of the system remained the same. The MFCC based system achieved a correct identification rate of 97.95%. From this result it can be seen that the HOS phase parameters can produce comparable results to the more widely used MFCC parameters.

Since the MFCC parameters can already deliver similar identification rates, there is no real need for using HOS phase parameters as an alternative feature set. It may be useful, however, to use both sets of parameters in a fused classification system for improved performance. The main diagonal of the confusion matrix for each of the two tests is given in Table 1. This indicates the percentage of correct identifications for each individual speaker in the database. This table suggests that the two feature sets may be used to complement each other. For speakers 3 and 8, the MFCC features obtain correct identification rates of 90% and 80.7% respectively, whereas the HOS phase features obtain higher rates of 100% and 92.1% respectively. Likewise, the MFCC features outperform the HOS phase features for speakers 7, 10 and 12.

In order to understand the differences between the feature sets that give rise to the small differences in recognition

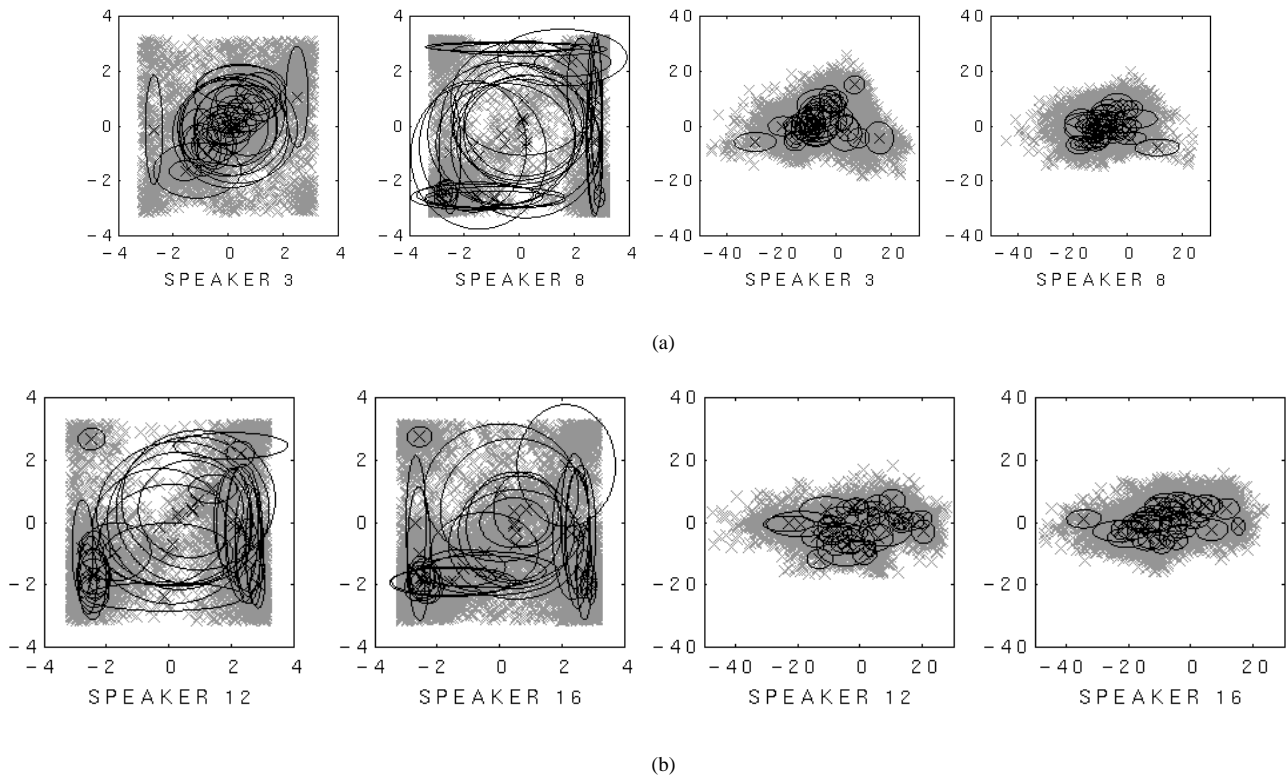


Fig. 2. Scatter plot of 2 dimensions for (a) HOS phase parameters and their corresponding GMMs for speakers 3 and 8 (left) and MFCC parameters and their corresponding GMMs for speakers 3 and 8 (right); and (b) HOS phase parameters and their corresponding GMMs for speakers 12 and 16 (left) and MFCC parameters and their corresponding GMMs for speakers 12 and 16 (right).

accuracy, scatter plots are used to model parameter distributions in two arbitrary dimensions. This is a reduced representation but it allows visualization of the feature spaces and compare models. Speakers that were significantly misclassified as each other as observed from the confusion matrix were selected. Speakers 3 and 8 caused problems for the MFCC features, with misclassifications of 10% and 19.7% from 3 to 8 and vice-versa. Speakers 12 and 16 caused problems for the HOS phase feature set with misclassification of 6.8% from speaker 12 to speaker 16. Feature distributions for both feature sets were plotted for these sets of speakers.

Fig. 2(a) shows scatter plots of the two feature sets for speakers 3 and 8. The HOS phase features are on the left and MFCC features on the right. The centres of the Gaussian mixtures and elliptical contours corresponding to unit standard deviations along the major and minor axes are overlaid on the cluster plots to give some indication of the models in each case. The principal axes of these ellipses are along the x and y dimensions because the covariance matrices of the models are diagonal by choice. The weights (or priors) of each mode cannot be inferred from these plots. Differences in feature distributions between speakers are evident in both sets of plots, and there is also a marked difference between the clusters for the two feature sets.

However, it is still difficult to understand why one feature set may perform better than the other or why there would be

more misclassifications between the two particular speakers from one of the feature sets. The feature distributions are therefore compared indirectly using the distribution of the centres (or mean values) of the 32 modes in the GMM in an attempt to quantify the differences even if the procedure may be a gross approximation. The distribution of these mode centres is approximated by a unimodal Gaussian and a Bhattacharya distance [17] given by:

$$\frac{1}{8}(M_2 - M_1)^T \left[\frac{\Sigma_1 + \Sigma_2}{2} \right]^{-1} (M_2 - M_1) + \frac{1}{2} \ln \frac{|\Sigma_1 + \Sigma_2|}{2 \sqrt{|\Sigma_1| |\Sigma_2|}} \quad (11)$$

where M_i is the mean of the set of mean vectors for speaker i , and Σ_i is the covariance of the set of mean vectors for speaker i , is computed. This distance gives a quantitative measure of the difference between two probability distributions.

Fig. 3(a) shows the centres and the unit standard deviation ellipses of the mode centre distributions (in the reduced 2 dimensional case) for speakers 3 and 8 – for HOS phase features on the left and MFCC features on the right. The Bhattacharya distances for the HOS phase set and the MFCC set for speakers 3 and 8 are given in Table 2. The greater

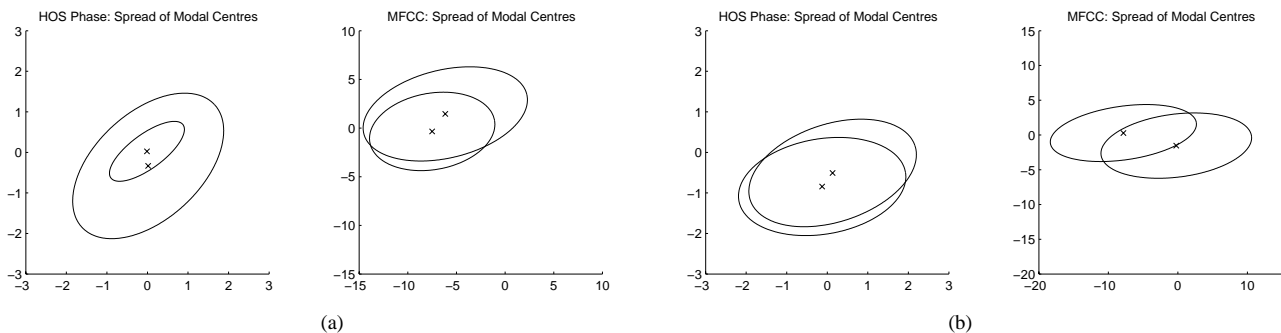


Fig. 3. A representation of the distribution of centres of the modes in the GMM models for (a) Speakers 3 and 8, and (b) Speakers 12 and 16. The x marks are the mean values and the ellipses correspond to unit standard deviations along the principal axes. The plot on the left is for HOS phase features, on the right for MFCC features.

Table 2. Bhattacharya distances between mode centre distributions of the GMMs.

| Dimensions | Speakers 3 and 8 | | Speakers 12 and 16 | |
|------------|------------------|-----------------------|--------------------|-----------------------|
| | MFCC Feature Set | HOS Phase Feature Set | MFCC Feature Set | HOS Phase Feature Set |
| All | 1.46 | 7.62 | 1.73 | 3.12 |
| 2 | 0.04 | 0.4 | 0.12 | 0.01 |

distances, indicating better separation of speaker models, corresponds to the better performance of HOS features for these two speakers. Note that distances are given when calculating (11) using the reduced two-dimensional case (corresponding to Fig. 3(a)), as well as using the full 16 dimensional feature space.

Scatter plots for features from speakers 12 and 16 are given in Fig. 2(b), and the GMM modal centre distributions represented by unimodal Gaussians are depicted in Fig. 3(b). Fig. 3(b) shows that the overlap between the centre distributions is less for the MFCC than for HOS parameters in the reduced two-dimensional case. The Bhattacharya distances between the mode centre distributions for the models of these two speakers are given in Table 2.

The distance measures for the two-dimensional case, corresponds to Fig. 3(b) and with the fact that MFCC features performed better. However, for the full feature set in each case, HOS features show a greater distance, and this was surprising. Although these distances are only a rough indication, they seem to suggest that the HOS parameter set may contain more information than the MFCC set for discrimination between the speakers. This is not surprising, if it is recalled that they do contain phase as well as magnitude spectral information. This does not necessarily translate to identification accuracy, however, because the HOS phase distributions may be more difficult to capture in a statistical model such as the GMM.

2) Speaker Identification Using NIST Microphone speech

It is possible that the system described in the subsection above using telephone speech was classifying speakers based on individual channel or handset differences, rather than speaker differences. In order to remove these influences on SI

performance, another experiment was performed using speech obtained from a tabletop microphone (as opposed to through a telephone channel). 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 evaluations.

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.

Using the HOS phase parameters as features, the 20 male speaker system achieved a correct identification rate of 98.5%. Since the same microphone is used for each of the speakers in the database, it can be concluded that the HOS phase features are successful in identifying the speakers based on their speech as opposed to channel variations. It must be noted, however, that this result is biased since the amount of training and testing speech differs between the speakers.

The above result confirms that HOS features are indeed capable of discriminating between speakers capturing voice information. For comparison, a second speaker identification experiment was performed using the same NIST speech data. 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%.

This result shows 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 using microphone speech. 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.

3) Speaker Identification Using YOHO Speech Data

Having shown that the HOS phase parameters are an effective set of features for microphone quality speech from a small speaker database, the next objective was to extend these results to larger populations. This was attempted using the YOHO speech database. This allowed comparison and performance evaluation on a population of up to 138 speakers.

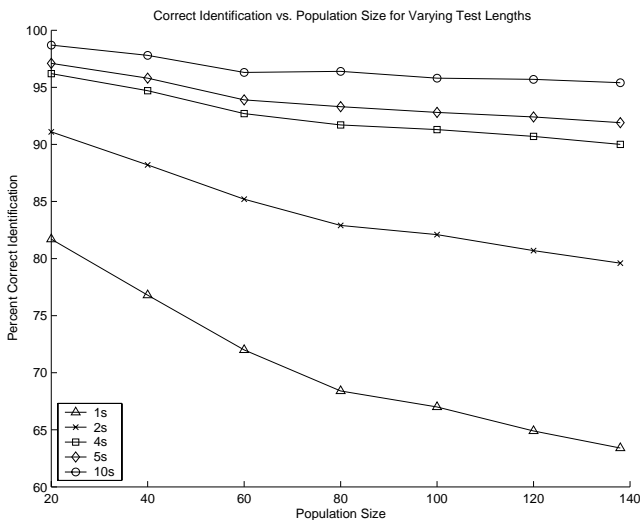


Fig. 4. Average correct identification rates for the YOHO database using HOS phase parameters with varying population size and varying lengths of voiced test speech. (All the test utterances from a single speaker are concatenated into a single utterance). For each population size, 20 tests were performed using a different set of randomly selected speakers (except when $S=138$).

There is, however, a major difference between this YOHO data and the aforementioned NIST data. The NIST data is comprised of continuous unconstrained speech, whereas the YOHO data consists solely of dual digit numbers within combination lock phrases. An example of such a phrase is “twenty-six, eighty-one, fifty-seven”. In addition, the YOHO database was designed to be evaluated by classifying each individual test utterance independently of each other. In order to remain consistent, all the individual test utterances from a particular speaker were concatenated into one single test utterance. The performance of the system can then be evaluated in the same way as done with the previous experiment.

The correct identification rates for HOS phase parameters for varying lengths of voiced test speech, L , and varying population sizes, S , are given in Fig. 4. Each point on the graph is the average of 20 different tests (except when $S = 138$), and in each test, the speakers are randomly chosen from the complete database of 138 speakers. The mean values and their standard deviations are given in Table 3. No effort has been made to bias the number of male or female speakers within the different population sizes.

For the case when $S = 20$ and $L = 2s$, a correct identification rate of only 91.1% (standard deviation = 1.4) was achieved for this YOHO data as opposed to 98.5% for the NIST data. This may be due to the differing nature of the speech within each database although the technique employed is text-independent. It must also be noted that for short test segments, the text-independence may not strictly hold true. A large population microphone database with unconstrained speech would be ideal, however, the YOHO database was a close compromise. It was still useful to study the effects of varying test speech length and increasing population sizes on the correct identification rates. From the graph in Fig. 4, it can be seen

Table 3. Data from Fig. 4. The standard deviation for each set of tests is given in brackets below the mean.

| Test Length | Population Size | | | | | | |
|-------------|-----------------|---------------|---------------|---------------|---------------|---------------|------|
| | 20 | 40 | 60 | 80 | 100 | 120 | 138 |
| 1s | 81.7 (2.0) | 76.8 (3.4) | 72.0 (2.2) | 68.4 (1.3) | 67.0 (0.9) | 64.9 (0.6) | 63.4 |
| 2s | 91.1 (1.9) | 88.2 (2.3) | 85.2 (1.9) | 82.9 (1.2) | 82.1 (0.9) | 80.7 (0.6) | 79.6 |
| 4s | 96.2 (1.4) | 94.7 (1.4) | 92.7 (1.7) | 91.7 (1.2) | 91.3 (0.9) | 90.7 (0.7) | 90.0 |
| 5s | 97.1 (1.6) | 95.8 (1.4) | 93.9 (1.7) | 93.3 (1.2) | 92.8 (0.9) | 92.4 (0.7) | 91.9 |
| 10s | 98.7 (1.6) | 97.8 (1.7) | 96.3 (1.7) | 96.4 (1.1) | 95.8 (0.8) | 95.7 (0.5) | 95.4 |

that for small T , the percent correct identification rate decreases more rapidly as S increases than for when T is large. With $L=10s$, the HOS phase parameters maintains an average percent correct identification rate above 95%.

It should be mentioned that preliminary tests were performed using the YOHO database as it was intended to be used, i.e., classifying each individual test utterance separately. The results from a single test with varying population size are given in Table 4. Note that multiple tests were not performed for each population size and results were not averaged here like in Table 3. Since each test utterance consists only of 3 double digit numbers, the actual amount of test data after extracting voiced frames was sometimes as little as 0.5 seconds. This may have been a major influence in the low correct identification rates. It is interesting to note, however,

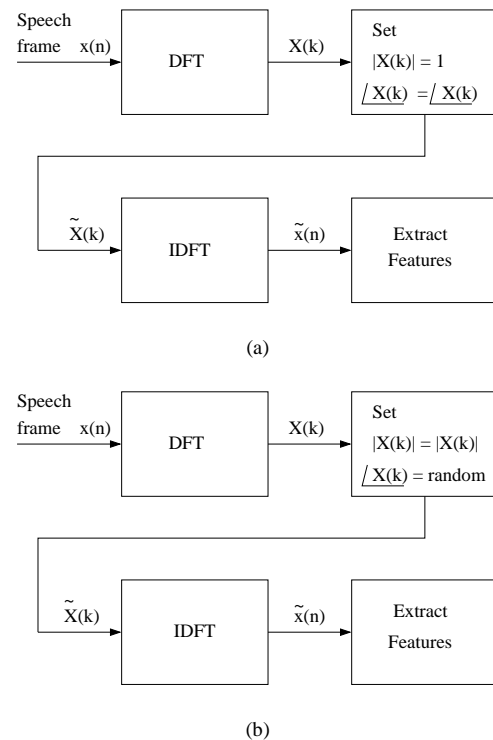


Fig. 5. Preprocessing required to (a) preserve the phase spectral information while discarding the magnitude information, and (b) preserve the magnitude spectral information while discarding the phase information.

that the performance increases by 6-10% when the number of mixtures used in the GMM is increased to 128.

4) Speaker Identification Using 'Phase Only' and 'Magnitude Only' Speech Data

HOS integrated phase parameters contain both phase and magnitude spectral information. In order to establish that there is indeed a contribution to speaker discrimination from the phase, additional tests using 'phase only' and 'magnitude only' speech data were performed. In these tests, each frame of speech, $x(n)$, was pre-processed to discard either magnitude or phase information before the features were extracted. The steps involved in discarding magnitude information while preserving phase information are (see Fig. 5(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.

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 5. The results using the original speech are also included for comparison.

Since MFCC parameters are derived from the magnitude spectrum, the loss of phase spectral information does not have an effect on its performance. The loss of magnitude spectral information, however, causes the MFCC system to fail. The correct identification rate for MFCC was 5.65% when using 'phase only' data. This is approximately the same as 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 their calculation (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 on its own provide a feature set that performs reasonably well for speaker identification, even when the power spectrum has been whitened.

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 more robust to additive noise than Fourier magnitude. The results in this section suggest that the HOS phase feature set would be more robust to such noise.

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

Table 4. Percent correct identification rates for the YOHO database using HOS phase parameters with varying population sizes. Each utterance is classified independently of each other, as the database was intended to be used. Results are given when using both 32 and 128 mixtures for the GMM's..

| # of Speakers | # of Mixtures | |
|---------------|---------------|------|
| | 32 | 128 |
| 20 | 77.0 | 84.0 |
| 40 | 72.8 | 82.7 |
| 60 | 69.4 | 78.4 |
| 80 | 68.3 | 76.9 |
| 100 | 66.0 | 75.1 |
| 120 | 65.0 | 74.6 |
| 138 | 64.3 | 73.9 |

accuracy for the 'phase only' data would decrease. Using a narrowband source, the phases best representative of the

Table 5. 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 |

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) 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 Fig. 6. In each of these tests, the speaker models remained trained on clean speech, however, AWGN was added to each of the test speech utterances. Therefore training and testing conditions are mismatched, which is more typical of real operating conditions. No other changes were made to the original experimental setup. Even though techniques exist, such as cepstral mean subtraction [18] and RASTA [19] processing, to make MFCC's more robust to channel mismatch, it was decided 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.

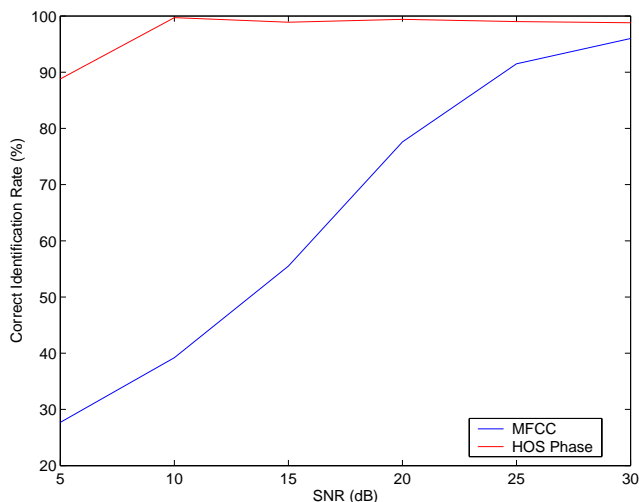


Fig. 6. % Correct identification rate versus SNR (for AWGN) when using the MFCC and HOS phase feature sets. The system operates under mismatched conditions.

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 MFCCs may provide a much more robust feature set in the presence of a varying channel with a non-linear phase response.

6) Computational Cost

The computational price paid for incorporating phase information is not exorbitant. A 29 second segment of speech can be processed for MFCC feature extraction in about 1 second and for HOS feature extraction in roughly 12 seconds, using the procedures described in this work. These results are for Matlab programs without any optimization and averaged over 100 runs on the same platform. The computational complexity of HOS feature extraction depends also on whether the bispectrum is computed by bilinear interpolation or over polar rasters. Roughly, the bispectral phase features method is a factor of 12 more computationally expensive when processing 32 millisecond blocks. This does not limit practical implementation in real-time systems even up to audio sampling rates.

IV. SIGNATURE VERIFICATION USING HOS

This section presents a system for recognition and verification of on-line signatures using higher order spectral features in a manner similar to that described in the previous section for speech processing. The sampling rate, however, for this system is only 100 Hz and the blocks are much smaller in the number of samples. Further, the on-line signature data is not one time series but a multidimensional one comprising positions, velocities and pressures from the pen tip. Because the grid of frequencies is coarser, a polar raster computation [5] of the bispectrum is employed such that interpolation

errors do not pose problems. Bilinear interpolation in the bifrequency domain [2],[4] can also be employed.

The motivations behind the use of this approach is that (a) robustness to variations in the size and orientation of the signature can be achieved, (b) the method will degrade gracefully with loss of partial signature information or addition of superfluous detail such as underlining, and (c) small local variations in the signature are confined to the features extracted from these segments only and thus have a small effect on the overall feature distribution. The bispectral phase features used have been proved to be translation, scale, amplification invariant in [2]. These invariance properties help in achieving property (a) above. Property (b) is achieved similar to property (c) as explained there.

These features are used to train Gaussian Mixture Models for each person using the methodology described in the previous section. Maximum likelihood is used for recognition and the log-likelihood separation between the best and second best matches used to provide a measure of confidence for verification. A threshold is applied to this separation in order to generate DET curves of performance. This strategy is applied because of the small size of the database collected which was unsuitable for the generation of background models for verification. The system described here is a text-dependent handwriting verification system where the text happens to be a signature. For a true signature verification system, an HMM will need to be used or the segments from the signature treated as an ordered set of parts in some other manner such that global shape is also utilized and the system becomes a text recognizer as well.

The performance of the system has only been tested on a small locally collected database and has not been optimized. A real-time implementation is available and preliminary results reported here are promising at around 6% EER. It is not among the best reported results in literature. However, it is the author's belief that signatures contain significant intra-class variations as a biometric and performance better than single digit EER will largely be owing to text recognition or significant global shape differences between signatures in the data used.

An international benchmarking exercise for signature verification was held at the International Conference on Biometric Authentication (ICBA), Hong Kong, in 2005 [29]. A system similar to the one described here achieved about 5% EER on random forgeries and 20% EER for skilled forgeries for the seen data. This system was also only a handwriting recognition system and not fully optimized. The database comprised of 40 persons and 10 genuine signatures were used to train the model for each person.

For the sake of putting the system in perspective, a brief description of signature verification is given here. There has been a great deal of research in this area over more than forty years and the interested reader is referred to [20-27] and references therein. The purpose here is only to illustrate how the bispectrum can be effectively applied to this problem.

A. Background

Signatures are required on credit card transaction records and cheques, and manual verification is often performed over the check-out counter. Some commercial automated signature verification products are available but a large number of transactions are still only manually verified using visual similarity in shape. Technology capable of acquiring on-line signatures has progressed tremendously over the last few decades and there are many relatively inexpensive pen-tablet systems on the market. These systems can capture not only static or shape information but also dynamic information about the speed at which different parts of a signature are produced. This information is available from the coordinates of the pen-tip and the pressure exerted as a function of time. These signals are sampled by the pen-tablet interface at around 1/100-th of a second. Such information makes it possible to verify not just the shape but also the writing style from the pen-tip velocity and pressure with an appropriately designed algorithm.

Many signature verification algorithms have been proposed in the past two decades and a number of them have used pen dynamics rather than shape. Signature data obtained and processed this way is called online as opposed to signature data obtained from images referred to as static or offline. Online signature analysis can use pressure and pen angles as well, and does not need any image processing operations to obtain position information. Segmentation into strokes is also easier with online data because pen up and pen down events result in significant changes in pen tip pressure. It is consequently faster and more accurate than static signature analysis.

Global features in the context of signature verification are considered to be those extracted from entire pen-down segments or the whole signature. Local features on the other hand are extracted from equally spaced sub-segments or around every sample point.

B. System Description

The signature data was collected using a Wacom Pen tablet system. The device had a resolution of 393 points/cm and sampling rate of 100 Hz. An example of a signature like scribble in a two-dimensional 'static' image is given in Figure 7. Pen tip coordinates, $x(n)$ and $y(n)$, and pressure, $p(n)$, were used, and their time derivatives, where n refers to the (time) sample index, were used to extract features. These six signals are processed to extract bispectral invariant features from 50% overlapping blocks of 130 milliseconds. Strokes less than 40 millisecond were deleted. 8 bispectral phase parameters were extracted from each block using a polar raster of 64 points along each line of integration. An 8 mixture GMM model was used. It must be emphasized that these are by no means optimal settings. The system is designed to be independent of the script but deletion of short strokes as done above may be a problem with Chinese, for example. Adaptations to the base system described here can eliminate such drawbacks and improve performance.

The system described here is motivated by the following objectives:

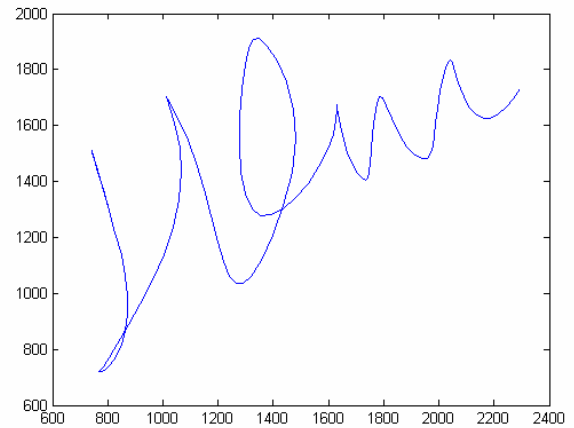


Fig. 7. A signature-like scribble comprising one stroke.

1. Avoid relying on segmentation other than through strokes defined by pen-up and pen-down.
2. Discard unreliable and very short strokes as would result from dots and dashes
3. Use pen-dynamics because it is more difficult to forge than signature shape
4. Use features that provide invariance or significant robustness to scale and orientation and thereby avoid the need for normalization
5. Extract local features and rely on the statistics of these features over reliable strokes in the signature, thereby providing robustness to small intra-personal variations in parts of a signature. Pen dynamics may change over small portions between two signatures by the same person but they would change over considerably larger portions in a forgery.
6. Use a classifier that provides a likelihood or confidence measure in the decision.

A maximum likelihood approach based on log-likelihood scores obtained from the GMMs of every person in the gallery was used for identification. Verification was performed by accepting the claim only if the person was identified (as the most likely candidate from the gallery) and the next best score was not closer than a set threshold. This threshold was varied to obtain false acceptance and false rejection error rates. This strategy was adopted in the absence of a background model and the small size of the database.

C. Results and Discussion

There were 14 persons in the database varying in age at the time of data collection from 15 to 81. There were 8 males and 6 females. Some of the persons in the database shared the same surname and four of these persons had signatures that differed only in the initial. 6 signatures from each person were used for training. Both training and test signatures varied significantly in height, length and orientation. Only 10 iterations of the K-means and 40 iterations of the EM algorithm were used in the GMM training. This may again not be sufficient. In trials with a similar system it has been found

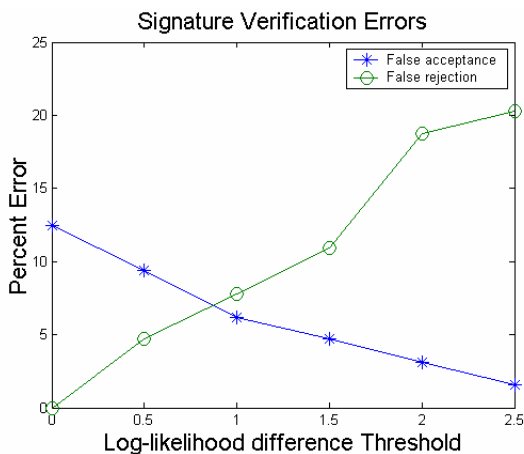


Fig. 8. Signature verification using HOS performance.

that convergence rates can be considerably different for different signature data.

The system achieved an equal error rate of about 6% as shown in the DET curve in Figure 9. It must be noted that no skilled forgeries were used in this test. The system is implemented in real-time on a Windows platform and takes only seconds to train with 6 signatures from a person or do a full identification comparing against 14 models. Storage requirements for the models are in the order of 8 Kilobytes for the parameters chosen without any compression.

V. CONCLUSION

This paper shows the the bispectrum can be effectively used for feature extraction in biometric identification and verification systems to achieve discrimination and robustness without compromising the capacity for real-time implementation. Further, it is possible to adopt identical feature extraction and classification strategies to different biometrics using the procedures described here, which can make classifier combination easier. In future work, the strategies will be used for such classifier combination in prompted multi-modal systems to achieve better performance.

ACKNOWLEDGMENT

Dr. Chandran is grateful to QUT for travel and conference support. Support from and many postgraduate students in the Speech, Audio, Image and Video Research Laboratory at QUT are also acknowledged. Daryl Ning conducted speaker recognition experiments. Mark Monsour, Darren Moore and Michael Mason assisted with programming the pen tablet system signature verification system.

REFERENCES

- [1] V. Chandran, "On the computation and interpretation of auto- and cross-trispectra," *International Conference on Acoustics, Speech and Signal Processing*, pp. 445-448, 1994.

- [2] V. Chandran and S.L. Elgar, "Pattern Recognition Using Invariants Defined from Higher Order Spectra—One-Dimensional Inputs", *IEEE Trans. on Signal Processing*, vol. 41, no. 1, pp. 205—212, January 1993.
- [3] S. Elgar, 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, 1993.
- [4] V. Chandran, et. al., "Pattern Recognition Using Invariants Defined from Higher Order Spectra: 2-D Image Inputs", *IEEE Trans. on Signal Processing*, vol. 6, no. 5, pp. 703—712, May 1997.
- [5] A.G. Bessios and C.L. Nikias, "FFT-based Bispectrum computation on polar rasters," *IEEE Trans. On Signal Processing*, vol. 39, no. 11, pp. 2535-2539, Nov. 1991.
- [6] D. Reynolds and R. C. Rose, "Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, pp. 72-83, 1995.
- [7] D. A. Reynolds, "Large Population Speaker Identification using Clean and Telephone Speech," *IEEE Signal Processing Letters* vol. 2, no. 3, pp. 46-48, 1995.
- [8] L. Liu, J. He, and G. Palm, "Signal Modeling for Speaker Identification," *International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 665-668, 1996.
- [9] L. P. Cordella, P. Foggia, C. Sansone, 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.
- [10] R. D. Patterson, "A Pulse Ribbon Model of Monaural Phase Perception," *Journal of the Acoustical Society of America*, vol. 82, no. 5, pp. 1560-1586, 1987.
- [11] H. Poblath and W. B. Kleijn, "On Phase Perception in Speech," *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 29-32, 1999.
- [12] J. Campbell Jr., "Testing with the YOHO CD-ROM voice verification corpus," *International Conference on Acoustics, Speech and Signal Processing*, pp. 341-344, 1996.
- [13] B. B. Wells, "Voiced/Unvoiced Decision based on the Bispectrum," *International Conference on Acoustics, Speech and Signal Processing*, vol. 10, pp. 1589-1592, 1985.
- [14] 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.
- [15] 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.
- [16] J. Campbell Jr., and T. E. Tremain, "Voiced/unvoiced classification of speech with application to the U.S. government LPC-10E algorithm," *International Conference on Acoustics, Speech and Signal Processing*, pp. 473-476, 1986.
- [17] K. Fukunaga, "Introduction to Statistical Pattern Recognition," Academic Press, Boston (1990)
- [18] 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.
- [19] H. Hermansky and N. Morgan., "RASTA processing of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 578-589, October 1994.
- [20] R. Plamondon and G. Lorette, "Automatic signature verification and writer identification – The state of the art," *Pattern Recognition*, vol. 22, pp. 107-131, 1989.
- [21] R. Plamondon, "The design of an on-line signature verification system: from theory to practice," *Intl. Journal of Pattern Recognition and Artificial Intelligence*, vol. 8, no. 3, pp. 796-811, 1994.
- [22] D. Z. Lejtman and S. E. George, "On-line handwritten signature verification using wavelets and back-propagation neural networks," *Proc. Of the sixth Intl. Conf. on Document Analysis and Recognition*, 10-13 Sept. 2001, pp. 992-996, 2001.
- [23] R. Kashi and W. Nelson, "Signature verification: benefits of multiple tries," in *Proc. Of the eighth Intl. workshop on Frontiers of handwriting recognition*, pp. 424-427, 2002.

- [24] L. Lee, T. Berger and E. Aviczer, "Reliable online human signature verification systems", IEEE Trans. On PAMI, vol. 18, no. 6, pp. 643-647, June 1996.
- [25] B. Kashi, J. Hu, W. L. Nelson, W. Turin, "Hidden Markov Model approach to online handwritten signature verification," Proc. of the Fourth International Conf. On Document Analysis and Recognition, vol. 1, pp. 253-257, 1997.
- [26] V. S. Nalwa, "Automatic On-line Signature Verification," Proc. IEEE, vol. 85, no. 2, pp. 215-239, Feb. 1997.
- [27] N. Mohankrishnan, W. Lee and M. Paulik, "A performance evaluation of a new signature verification algorithm," Proc. ICIP-99, pp. 25-29, Oct. 1999.
- [28] D. Ning and V. Chandran, "The effectiveness of Higher Order Spectral Phase Features in Speaker Identification," Proc. Of Odyssey 2004 – Speaker and Language Identification Workshop, Toledo, Spain, May 31-June 4, 2004.
- [29] V. Chandran, D. Ning and S. Sridharan, "Speaker identification Using Higher Order Spectral Phase Features and their effectiveness vis-à-vis Mel Cepstral Features," Proc. Of the First International Conference on Biometric Authentication (ICBA-2004), Hong Kong, July 15-17, 2004.

Vinod Chandran (S'85–M'90–SM'01) received the B.Tech. degree in electrical engineering (electronics) from the Indian Institute of Technology, Madras, India, in 1982, the M.S. degree in electrical engineering from Texas Tech University, Lubbock, in 1985, and the Ph.D. degree in electrical and computer engineering and the M.S. degree in computer science from Washington State University, Pullman, WA, in 1990 and 1991, respectively.

He is currently an Associate Professor at the Queensland University of Technology, Brisbane, Australia, in the School of Engineering Systems. His research interests include pattern recognition, higher order spectral analysis, speech processing, and image processing.

Associate Professor Chandran is a Senior Member of the Institute of Electrical and Electronic Engineers (IEEE) and the Chairman of the Queensland chapter of the IEEE Computer Society. He is also a member of the Association for Computing Machinery (ACM).