Higher Order Spectral Phase Features for Speaker Identification

Vinod Chandran and Sridha Sridharan

Speech and Audio Research Lab
School of Electrical and Electronic Systems Engineering
Queensland University of Technology,
GPO Box 2434
Brisbane Qld 4001
{v.chandran,s.sridharan}@qut.edu.au

Abstract

This paper investigates the use of higher order spectra (HOS) phase features in the task of speaker identification. Within the speech processing community, short time spectral phase information is widely regarded as unimportant for speaker recognition. Features are generally defined from the magnitude spectrum only. This paper utilises 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. Initial experiments used unconstrained, microphone speech from a 20 male speaker database to construct Gaussian mixture models (GMM) for each speaker. 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%). Other experiments were conducted on the larger YOHO database of 138 speakers. Average correct identification rates of above 95% were achieved for varying populations sizes up to the full 138 speakers.

1. Introduction

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 (Reynolds and Rose 1995; Reynolds 1995; Liu, He., and Palm 1996; Cordella, Foggia, Sansone, and Vento 2003). 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 (Pobloth and Kleijn 1999).

In this paper, we utilise a set of features derived from higher order spectra (HOS) (Chandran and Elgar 1993). The performance of these features is compared with MFCC features in identical speaker identification systems. The sensitivity of each feature set to additive white Gaussian noise (AWGN) under mismatched training and testing conditions is also compared.

Section 2. introduces HOS and section 3. describes the HOS phase parameters used as features. Section 4. describes the setup of the speaker identification experiments. Section 5. presents the results on each of the experiments performed, accompanied by a brief discussion. 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 correlations[1] of the signal respectively. If $x(t)$ is a stationary random process, then its n^{th} order moments, $m_n(\tau_1, \tau_2, \ldots, \tau_{n-1})$, can be defined as

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

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

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

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

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

and

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

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.

[1] This refers to moment based spectra as opposed to cumulant based spectra. For more information, see (Chandran 1994).

Proceedings of the 10th Australian International Conference on Speech Science & Technology
Macquarie University, Sydney, December 8 to 10, 2004.

Accepted after full paper review

Another important property of the bispectrum is that is 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 (Chandran and Elgar 1993) 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 (Elgar and Chandran 1993) and (Chandran and Elgar 1993).

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

$$B(f_1, f_2) = X(f_1)X(f_2)X^*(f_1 + f_2) \tag{5}$$

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 (Chandran and Elgar 1993), 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

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

where

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

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 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 (Chandran and Elgar 1993). 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.

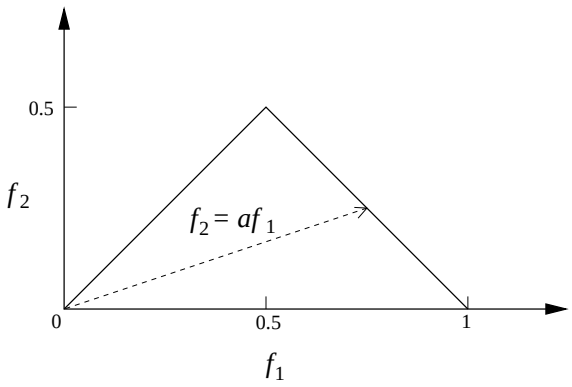

Figure 1: *Region of computation of the bispectrum for real signals. Features are calculated by integrating the bispectrum along the dashed line with slope=a.*

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 the first 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. 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. Each session was recorded using the same microphone and sampled at 16 kHz, but we first filter and downsample each of the speech files to 8 kHz before processing. A total of 20 male speakers was used in this experiment.

The speech for our second set of experiments was obtained from the YOHO voice verification corpus (Campbell Jr. 1995). This corpus consists of 138 speakers (106 males, 32 females). Each speaker has 4 enrollment 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.

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 (Wells 1985; B.Boyanov, Hadjitodorov, and Ivanov 1991; Fackrell and McLaughlin 1994). The voicing decision is determined using the algorithm from the LPC-10E speech coder (Campbell Jr and Tremain 1986). 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 (Reynolds and Rose 1995). 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 (Reynolds and Rose 1995). 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.

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:

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

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

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

where

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

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

$$\overbrace{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_T}^{\text{Segment 1}}, \vec{x}_{T+1}, \vec{x}_{T+2}, \dots$$

$$\vec{x}_1, \overbrace{\vec{x}_2, \dots, \vec{x}_T, \vec{x}_{T+1}}^{\text{Segment 2}}, \vec{x}_{T+2}, \dots$$

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

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

5. Results and Discussion

5.1. Speaker Identification Using NIST Speech Data

The first set of experiments used the NIST speech data to evaluate 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 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%.

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. 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 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, our next objective was to extend these results to larger populations. For these next experiments we use the YOHO speech database. This allows us to test the HOS phase parameters on a population of up to 138 speakers. 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 with the procedure described in section 4.5. however, we have concatenated all the individual test utterances from a particular speaker into one single test utterance. We can then evaluate the performance of our system in the same way that we have 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 figure 2. 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 1. 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 $L = 2s$, we obtained a correct identification rate of only 91.1% (SD=1.4) 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. Ideally we would have liked to use a large population microphone database with unconstrained speech, however, the YOHO database was the closest we had access to. 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 figure 2, we can see 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%.

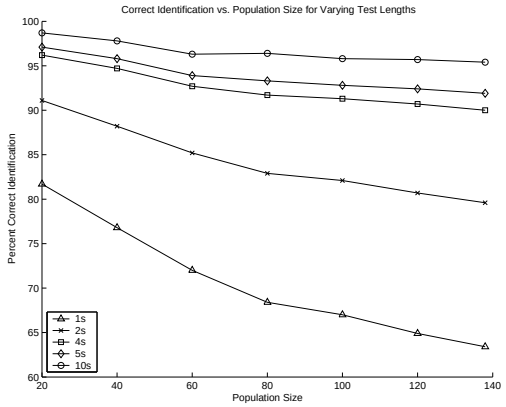

Figure 2: Average correct identification rates for the YOHO database using HOS phase parameters with varying population size and varying lengths of voiced test speech. (All the test utterances from a single speaker are concatenated into a single utterance). For each population size, 20 tests were performed using a different set of randomly selected speakers (except when S=138).

Table 1: Data from figure 2. 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

We should also mention 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 2. Note that we did not perform multiple tests for each population size and average the results like in table 1. 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. We believe that this was a major influence in the low correct identification rates. It is interesting to note, however, that the performance increases by 6–10% when the number of mixtures used in the GMM is increased to 128.

Table 2: 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

5.3. 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, we performed additional tests using 'phase only' and 'magnitude only' speech data. In these tests, preprocessing was performed on each frame of speech, $x(n)$, to discard either magnitude or phase information before the features were extracted. The steps involved in discarding magnitude information while preserving phase information are given in figure 3(a)).

This procedure is identical to that used by Oppenheim *et al*. (Oppenheim, Lim, Kopec, and Pohlig 1979) except that we are processing short time segments of speech as opposed to long time segments. The steps involved in discarding phase information while preserving magnitude information are given in figure 3(b)).

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 3. 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.

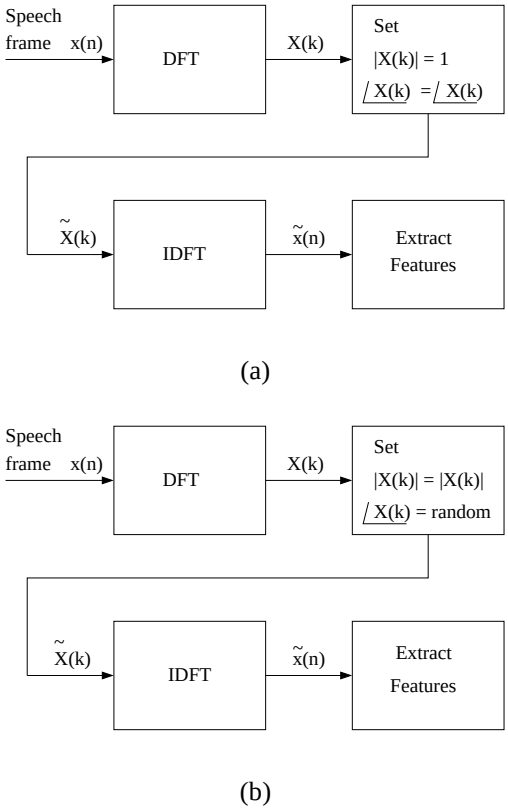

Figure 3: *Preprocessing required to (a) preserve the phase spectral information while discarding the magnitude information, and (b) preserve the magnitude spectral information while discarding the phase information.*

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.

5.4. 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 4. 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 were made to the original setup described in section 4.. Even though techniques exist, such as cepstral mean subtraction (Atal 1974) and RASTA (Hermansky and Morgan 1994) 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 cor-

Table 3: *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

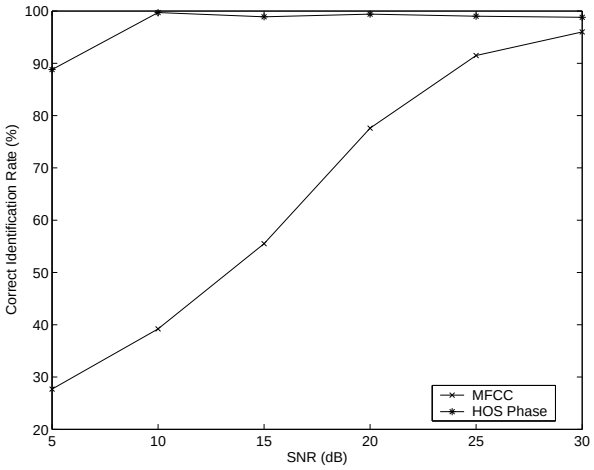

Figure 4: *% Correct identification rate versus SNR (for AWGN) when using the MFCC and HOS phase feature sets. The system operates under mismatched conditions.*

rect 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.

6. Conclusion

We have shown that the HOS phase based parameters derived in (Chandran and Elgar 1993), contain useful information for discerning speakers. On clean microphone speech and under identical conditions, they perform on a similar level as the widely used MFCC parameters. Further, they are more robust to additive Gaussian noise. A fused classification system combining the two type of features may provide improved accuracy.

7. 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 and to Dr. Daryl Ning for performance evaluation and documentation.

References

Atal, B. S. (1974, June). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *Journal of the Acoustical Society of America 55*, 1304–1312.

B.Boyanov, S. Hadjitodorov, and T. Ivanov (1991). Analysis of voiced speech by means of bispectrum. *Electronic Letters 27*(24), 2267–2268.

Campbell Jr., J. (1995). Testing with the YOHO CD-ROM voice verification corpus. *ICASSP 1*, 341–344.

Campbell Jr, J. P. and T. E. Tremain (1986). Voiced/unvoiced classification of speech with application to the u.s. government LPC-10E algortihm. *ICASSP*, 473–476.

Chandran, V. (1994). On the computation and interpretation of auto– and cross–trispectra. *ICASSP 4*, 445–448.

Chandran, V. and S. L. Elgar (1993, January). Pattern recognition using invariants defined from higher order spectra–one dimensional inputs. *IEEE Transactions on Signal Processing 41*(1), 205–212.

Cordella, L., P. Foggia, C. Sansone, and M. Vento (2003). A real–time text independent speaker identification system. *Proc of the 12th Intl Conference on Image Analysis and Processing*, 632–637.

Elgar, S. and V. Chandran (1993, January). 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 3*(1), 19–34.

Fackrell, J. W. A. and S. McLaughlin (1994). The higher–order statistics of speech signals. *IEE Colloquium on Techniques for Speech Processing and their Applications*, 7/1–7/6.

Hermansky, H. and N. Morgan (1994, October). RASTA processing of speech. *IEEE Transactions on Speech and Audio Processing 2*(4), 578–589.

Liu, L., J. He., and G. Palm (1996). Signal modeling for speaker identification. *ICASSP 2*, 665–668.

Oppenheim, A. V., J. S. Lim, G. Kopec, and S. C. Pohlig (1979). Phase in speech and pictures. *ICASSP 4*, 632–637.

Pobloth, H. and W. B. Kleijn (1999). On phase perception in speech. *ICASSP 1*, 29–32.

Reynolds, D. A. (1995, March). Large population speaker identification using clean and telephone speech. *IEEE Signal Processing Letters 2*(3), 46–48.

Reynolds, D. A. and R. C. Rose (1995, January). Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing 3*(1), 72–83.

Wells, B. B. (1985). Voiced/unvoiced decision based on the bispectrum. *ICASSP 10*, 1589–1592.