BISPECTRUM FEATURES FOR ROBUST SPEAKER IDENTIFICATION

Stanley Wenndt † *and* Sanyogita Shamsunder ‡

† *Rome Laboratory/IRAA, Rome, NY 13441*
tel: (315) 330-4026, fax: (315) 330-2728, e-mail: wenndts@eeyore.ira.rl.af.mil

‡ *Dept. of Electrical Engineering, Colorado State University, Fort Collins, CO 80523*
tel: (970) 491-5767, fax: (970) 491-2249, e-mail: sanyo@engr.colostate.edu

ABSTRACT

Along with the spoken message, speech contains information about the identity of the speaker. Thus, the goal of speaker identification is to develop features which are unique to each speaker. This paper explores a new feature for speech and shows how it can be used for robust speaker identification. The results will be compared to the cepstrum feature due to its widespread use and success in speaker identification applications. The cepstrum, however, has shown a lack of robustness in varying conditions, especially in a cross-condition environment where the classifier has been trained with clean data but then tested on corrupted data. Part of the bispectrum will be used as a new feature and we will demonstrate its usefulness in varying noise settings.

1. INTRODUCTION

Speech is a complex interaction between quasi-periodic puffs of air generated by the larynx opening and closing and the various cavities and anatomical structures of the speaker. Consequently, speaker recognition has been an ongoing effort to solve a very complex problem. Speaker identification is the task of assigning a speech utterance from an unknown speaker to the correct speaker of a known set. Speaker identification techniques have applications in security access, telephone transactions, and forensic science.

One feature that has been used quite successfully in text-independent speaker identification experiments, even with small amounts of training data, is the cepstrum. Measurements from the vocal tract filter are used for distinguishing speakers. The cepstral feature is used in [1] for speaker identification experiments. When the training and testing material is clean speech from the TIMIT database, the speaker identification rate is about 96%. However, when the testing data contains 10 dB additive white Gaussian noise, the success rate drops to about 20%.

Higher-order statistics (HOS) are used in [2] to help combat the drastic fall off rate in cross-conditions. The autoregressive (AR) coefficients are calculated using two different HOS based methods and then the AR coefficients are converted to cepstral coefficients. Sixteen speakers from the King database are used with about 90 seconds of training data and 10 seconds of testing data. When trained with clean data and tested in 20 dB white Gaussian noise, the HOS methods employed yield improvements in speaker identification of up to 40%. However, at 10 dB, the improvement is only about 5% and at 5 dB there is little difference. The drawback to these methods is a lot of extra computations. Interestingly, the performance gains are much less when training and testing are both done with noisy data. The log-bispectrum is used in [3] for speaker identification, but the experiment was limited to two speakers.

Although different databases and conditions were used for these examples, it is clear that current speaker identification techniques severely degrade when the speech samples are noisy or when training and testing are performed under different noise conditions. This research investigates the use of a new feature by using part of the nonredundant region of the bispectrum. The goal is to derive a feature for speaker identification that can be effective when the testing and training data are collected under varying noise levels and channels. Text-independent data will be used in all the tests.

The layout of the paper is as follows: Section 2 reviews HOS with an emphasis on the bispectrum. Some of the advantages of the bispectrum include its immunity to additive Gaussian noise and that phase relations are preserved. Section 3 explains the setup for the speaker identification experiments and gives results under various noise conditions. The cepstrum feature is also included here for comparison. The results are then analyzed in Section 4. Section 5 draws conclusions and gives future work.

2. BISPECTRUM

One of the main motivations in using higher-order statistics is that for Gaussian noise, the kth-order spectrum, $k > 2$, is zero and hence, in theory, the bispectrum of the desired signal will be immune to additive Gaussian noise. Another motivation is that cumulants and polyspectra can preserve phase relations while second order statistics are phase blind. Finally HOS are key tools for analyzing nonlinear processes and could possibly provide additional information regarding the nonlinear speech generation mechanism [4]. The bispectrum is used in this research because it is computationally less expensive than other HOS features.

The third-order cumulant for a zero-mean stationary process $\{x(k)\}$ is defined as:

$$C_{3,x}(\tau_1, \tau_2) = E\{x(k)x(k+\tau_1)x(k+\tau_2)\} \quad (1)$$
$$= E\{x(\tau_1)x(\tau_2)\} - E\{g(\tau_1)g(\tau_2)\} \quad (2)$$

where $\{g(k)\}$ is a Gaussian random process which has

second-order statistics identical to $\{x(k)\}$. Thus, a non-zero third-order cumulant for a process $\{x(k)\}$ indicates deviation from Gaussianity. The bispectrum is defined as the 2-dimensional Fourier transform of the third-order cumulant given by:

$$B(\omega_1, \omega_2) = \sum_{\tau_1=-\infty}^{\infty} \sum_{\tau_2=-\infty}^{\infty} C_{3,x}(\tau_1, \tau_2) e^{-j\omega_1\tau_1} e^{-j\omega_2\tau_2} \quad (3)$$

2.1. Regions of Symmetry

Just like there are regions of symmetry for the autocorrelation (i.e., $r(-\tau) = r^*(\tau)$), there are also regions of symmetry for the third-order cumulant and the bispectrum of a stationary process. By knowing the third-order cumulant or bispectrum in a nonredundant region allows one to know the rest of the coefficients by symmetry. Thus, the number of computations are greatly reduced. The region $\omega_2 > 0, \omega_1 \geq \omega_2, \omega_1 + \omega_2 \leq \pi$ defines the first nonredundant region of $B(\omega_1, \omega_2)$. By using the correct portion of this region, an ARMA(p,q) non-Gaussian signal can be uniquely specified [5]. For more details on the regions of symmetry and techniques for estimating the bispectrum, see [6].

3. EXPERIMENTAL SETUP

Twenty male speakers from the same dialect region were selected from the TIMIT database for the speaker identification experiments. The SX's and SI's sentences were used for training (9 seconds of voiced speech) and the two remaining SA sentences were then used in testing (average of 1.74 seconds of voiced speech per test sentences). The training data and the testing data are kept short to increase the difficulty and to simulate military environments where very limited amounts of data is available. Frame lengths of 16 msec and 32 msec were considered. The effects of smoothing using a 5×5 Rao-Gabor filter was explored along with averaging effects. If averaging was done, a frame length of N was divided into 3 segments of length $N/2$ by using 50% overlap.

The bispectrum in these experiments was calculated using the direct method. Since the bispectrum preserves phase relations, two distances were used: One using the difference between the magnitude of the testing and training data and one that includes the phase. The number of points used from the nonredundant region varied. A symbol like 5⇒15 in the tables indicates that the first five columns in the first nonredundant region of the bispectrum (See Figure 1) are used. Thus, a total of 15 points are used in the distant measure. The frequency spacing for $B(\omega_1, \omega_2)$ is $\omega_1 = (\frac{2\pi F_s}{N})\lambda_1$ and $\omega_2 = (\frac{2\pi F_s}{N})\lambda_2$. For TIMIT data, $F_s = 16000$ Hz and $N = 256$ for a 16 msec window.

3.1. TIMIT Results

A nearest neighbor classifier was used where 20 distances were kept: one for each speaker. For each test frame, the minimum distance to each reference speaker was calculated. The distances were accumulated over the entire test sentences. The speaker with the lowest accumulated distance was declared the winner. Four different noise cases were examined. The first noise case was a cross-condition of training with clean data and testing in data corrupted by 10 dB additive white Gaussian noise. The other noise cases contaminated the training and testing with the same type of noise: 10 dB additive white Gaussian, 10 dB additive colored Gaussian, and multiplicative rayleigh noise to simulate fading effects. Ten Monte Carlo simulations were used with different noise realizations for the training and testing phase.

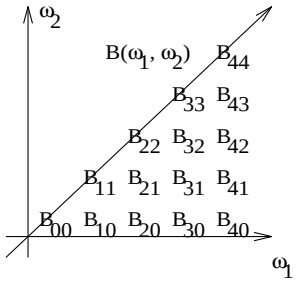

Figure 1. First 15 points in the nonredundant region of the bispectrum.

Table 1 gives the results for a 16 msec frame length when the training and testing data is uncorrupted. The first column gives the number of columns/total points used in the nonredundant region of the bispectrum. The second column gives results when no smoothing is used and just the magnitude of the feature is considered, $||A| - |B||$. The third column gives results when no smoothing is used but the phase is included in the distance measurement, $|A - B|$. The fourth and fifth column is the same as the second and third column except smoothing is used. Table 2 gives the results for a 32 msec frame length. An entry such as 65.0/10.7 in the tables would indicate that 65.0% of the test sentences were scored correctly, but only 10.7% of the individual frames were scored correctly.

	No Averaging			
	No Smoothing		Smoothing	
	$\|\|A\| - \|B\|\|$	$\|A - B\|$	$\|\|A\| - \|B\|\|$	$\|A - B\|$
5⇒15	65.0/10.7	80.0/16.6	75.0/13.2	80.0/17.2
7⇒28	72.5/13.5	85.0/20.7	82.5/16.2	92.5/21.9
9⇒45	75.0/14.6	87.5/20.2	80.0/16.7	85.0/22.3
	Averaging			
	No Smoothing		Smoothing	
	$\|\|A\| - \|B\|\|$	$\|A - B\|$	$\|\|A\| - \|B\|\|$	$\|A - B\|$
5⇒15	70.0/10.0	90.0/18.4	67.5/11.6	75.0/14.4
7⇒28	67.5/16.6	80.0/18.4	67.5/12.3	77.5/15.1
9⇒45	60.0/12.4	80.0/18.3	70.0/13.1	75.0/16.0

Table 1. Bispectrum results for frame length of 16 msec, 20 speakers.

From Tables 1-2, it is seen that some of the best results for the bispectrum occurred for a frame length of 16 msec with 7 columns from the nonredundant region and no averaging. A frame length of 32 msec with 7 columns from the nonredundant region along with averaging also gave decent results. Thus, the focus for testing the bispectrum under various noises was limited to four variations:

1.) 16 msec frame length, no averaging, no smoothing,
2.) 16 msec frame length, no averaging, smoothing,

	No Averaging			
	No Smoothing		Smoothing	
	$\|\|A\|-\|B\|\|$	$\|A-B\|$	$\|\|A\|-\|B\|\|$	$\|A-B\|$
5⇒15	52.5/11.2	55.0/14.1	37.5/12.5	50.0/14.8
7⇒28	35.0/13.1	50.0/15.9	50.0/13.9	60.0/17.4
9⇒45	57.5/14.4	87.5/20.2	45.0/14.6	65.0/19.5
	Averaging			
	No Smoothing		Smoothing	
	$\|\|A\|-\|B\|\|$	$\|A-B\|$	$\|\|A\|-\|B\|\|$	$\|A-B\|$
5⇒15	42.5/10.6	65.0/16.7	75.0/13.3	80.0/17.3
7⇒28	65.0/15.0	85.0/21.5	82.5/16.3	92.5/22.0
9⇒45	77.5/16.7	82.5/21.6	80.0/16.8	85.0/22.4

Table 2. Bispectrum results for frame length of 32 msec, 20 speakers.

3.) 32 msec frame length, averaging, no smoothing,
4.) 32 msec frame length, averaging, smoothing.

When averaging is performed, the frame length is 32 msec, but the DFT length will be the same as the 16 msec frame lengths due to averaging. Thus, the frequency range for ω_1 and ω_2 of the bispectrum is from 0 to 375 Hz.

These four cases should yield insights into how important averaging and smoothing is in regards to estimating the bispectrum, especially in noisy situations. The distance measure will include the phase since it gave better results. Table 3 gives the results. The first line lists the training conditions and the second line lists the testing conditions.

∞	∞	10 dB w	10 dB c	fade
∞	10 dB w	10 dB w	10 dB c	fade
16 msec, No Avg., No Smooth.				
85.0/20.7	86.0/16.9	69.25/14.3	73.0/15.2	81.5/15.0
16 msec, No Avg., Smooth.				
92.5/21.9	88.0/16.7	79.5/14.6	83.25/15.7	87.25/14.3
32 msec, Avg., No Smooth.				
85.0/21.5	84.75/19.3	74.5/17.3	75.75/17.9	79.0/17.4
32 msec, Avg., Smooth.				
92.5/22.0	88.5/16.8	82.5/14.8	80.5/15.6	83.75/14.2

Table 3. Bispectrum results for frame length of 16 and 32 msec window, 20 speakers, various noises.

Note that with additive Gaussian noise, the bispectrum features continue to do well. Smoothing always boosted the overall success rate while averaging helped the overall success rate about half the time. Compared to the cepstrum feature for a cross-condition (see Section 3.3), the bispectrum does extremely well. The success rate of 88.50% for a 32 msec frame length with averaging and smoothing is quite impressive in a cross-condition. The other noise conditions of additive white, additive color, and multiplicative noise also did extremely well yielding results of 82.50%, 80.50%, and 83.75% respectively.

3.2. NTIMIT Results

While the bispectrum does well under the additive Gaussian noise cases, communication lines contain more than just additive noise. The NTIMIT database is a version of TIMIT which has been transmitted over telephone channels. Typical degradations include: bandlimiting, network noise, echoes, and distortions.

Tables 4 gives the results for a 16 msec frame length using NTIMIT for training and testing and Table 5 gives the results for a 32 msec frame length. Unfortunately, the bispectrum does not do as well with the NTIMIT data, possibly due to corrupted phase relations and bandpass filtering. The next subsection explores the cepstrum feature for comparison to the bispectrum.

	No Averaging			
	No Smoothing		Smoothing	
	$\|\|A\|-\|B\|\|$	$\|A-B\|$	$\|\|A\|-\|B\|\|$	$\|A-B\|$
7⇒28	42.5/8.4	37.5/8.7	40.0/8.2	27.5/7.7
9⇒45	47.5/8.9	40.0/8.3	42.5/8.9	27.5/7.5
	Averaging			
	No Smoothing		Smoothing	
	$\|\|A\|-\|B\|\|$	$\|A-B\|$	$\|\|A\|-\|B\|\|$	$\|A-B\|$
7⇒28	45.0/7.9	32.5/8.4	50.0/7.4	25.0/6.7
9⇒45	50.0/9.1	35.0/8.8	55.0/8.5	27.5/7.1

Table 4. Bispectrum results for frame length of 16 msec, 20 speakers, test and train with NTIMIT.

	No Averaging			
	No Smoothing		Smoothing	
	$\|\|A\|-\|B\|\|$	$\|A-B\|$	$\|\|A\|-\|B\|\|$	$\|A-B\|$
7⇒28	25.0/8.0	25.0/7.4	32.5/7.1	15.0/7.3
9⇒45	40.0/8.6	22.5/8.0	35.0/8.2	22.5/7.2
	Averaging			
	No Smoothing		Smoothing	
	$\|\|A\|-\|B\|\|$	$\|A-B\|$	$\|\|A\|-\|B\|\|$	$\|A-B\|$
7⇒28	32.5/8.5	30.0/8.5	40.0/7.5	32.5/7.3
9⇒45	42.5/9.5	30.0/8.6	40.0/9.0	27.5/7.6

Table 5. Bispectrum results for frame length of 32 msec, 20 speakers, test and train with NTIMIT.

3.3. Cepstrum

The cepstrum has been very good with lab quality speech, but has also shown a lack of robustness in noisy environments. This downfall is especially apparent when the training conditions and testing conditions are different. The same test conditions are used again and Table 6 contains the results using the real cepstrum for the various noise cases. A Hamming window is used for smoothing. Not only does the cepstrum suffer a large drop off under similar noise conditions, but cross-conditions are extremely detrimental to the speaker identification success rate. These detrimental results in cross-conditions are confirmed in [1], [2]. For NTIMIT data, the best cepstrum result was 67.5% for a 16 msec window and 32 coefficients.

4. BISPECTRUM DISCUSSION

Although the preceding sections discussed the bispectrum feature and looked at how it performed under varying conditions, it is important to analyze the results. The bispectrum is still relatively simple to compute by using an FFT along

∞	∞	10 dB w	10 dB c	fade
∞	10 dB w	10 dB w	10 dB c	fade
16 msec, Cepstrum, 16 coefficients.				
100/36.9	9.75/5.9	53.0/8.0	49.25/8.0	41.5/6.9
16 msec, Cepstrum, 32 coefficients.				
100/40.4	8.25/5.3	60.0/8.5	57.25/8.3	45.0/7.4
32 msec, Cepstrum, 16 coefficients.				
100/40.2	5.0/5.8	40.25/8.9	42.0/8.8	44.5/6.8
32 msec, Cepstrum, 32 coefficients.				
100/43.9	5.0/5.7	65.75/9.8	66.75/9.8	47.5/8.3

Table 6. Cepstrum results for frame length of 16 and 32 msec window, 20 speakers, various noises.

with computing the triple product $X(\lambda_1)X(\lambda_2)X^*(\lambda_1+\lambda_2)$. Of course, smoothing and averaging the bispectrum would also increase the amount of computations. The difficulty with using HOS for speaker identification is conflicting requirements: HOS need longer data lengths for consistent estimates, yet speech frames need to be short to maintain stationarity assumptions.

In theory, the bispectrum of a Gaussian noise process is zero. Since the bispectrum is estimated with finite data, Gaussian noise does effect the performance although the bispectrum did a very remarkable job for the Gaussian noise case.

The goal of the bispectrum was to capitalize on nonlinearities, phase coupling, and deviations from normality that may be occurring in speech. It is possible that the bispectrum is using all of these and is a continuing point of research. In [7], research indicates that self-phase coupling is not occurring in speech signals.

The method proposed here can be thought of as being based on an ARMA(p,q) model of a non-Gaussian speech signal. In [5], HOS-based methods were proposed for estimating a non-Gaussian ARMA(p,q) process. The authors define a portion of the nonredundant region of the bispectra that can be used to uniquely estimate general (non-) causal and (non-) minimum phase ARMA signals. Thus, the bispectrum method proposed here is similar to a nearest neighbor classifier for ARMA coefficients.

Our results do demonstrate that phase was an important aspect of the feature. For the TIMIT case, using the bispectrum phase information *always* improved the speaker identification performance. For the NTIMIT case, the amplitude and phase information has been corrupted and using the phase information was *always* detrimental.

Although, the bispectrum feature does well in controlled noise environments, using NTIMIT data exposes some weaknesses. The bandpass nature of the telephone removes formants below 300 Hz. Thus, the bispectra of the TIMIT and NTIMIT data may be significantly different.

5. SUMMARY

The stated goal of this research was to investigate the bispectrum as a robust feature to be used for speaker identification. Results were compared to the cepstrum due to its widespread use and success in speaker identification. The goal was not to discover which techniques gave the best results using clean data, but instead to know which features performed the best in varying conditions. While the cepstrum does the best under clean conditions yielding a 100% speaker identification success rate, its performance falls off sharply in varying conditions with cross-conditions being the most difficult for the cepstrum.

The bispectrum feature came from the nonredundant region. By averaging and smoothing, the bispectrum did very well for additive Gaussian noise. Even in the difficult condition of training in clean data and testing in noise, the overall success rate only dropped 4% from 92.5% to 88.5%. The other noise cases of training and testing with white Gaussian, colored Gaussian, or multiplicative noise yielded results of 82.50%, 80.50%, 83.75% respectively. These are all excellent results compared to other published research. However, the bispectrum did not hold up as well when the training and testing were conducted with the NTIMIT database possibly because phase relations were distorted via the communication systems and formants below 300 Hz were removed.

Additional research is planned using speech from other communication systems that do not bandlimit the signal as severely as the telephone channel. Different classifiers should be used in an effort to improve on these features and increase robustness. Plus, other techniques such as channel normalization or frequency warping may be of benefit. A larger speaker set should also be used to verify the usefulness of the bispectrum feature.

REFERENCES

[1] R. P. Ramachandran, M. S. Zilovic, and R. J. Mammone, "A Comparative Study of Robust Linear Prediction Analysis Methods with Applications to Speaker Identification," *IEEE Transactions on Speech and Audio Processing*, pp. 117-125, vol. 3, no. 2, 1995.

[2] L. J. Trent, C. M. Rader, and D. A. Reynolds, "Using Higher Order Statistics to Increase the Noise Robustness of a Speaker Identification System," *ESCA Workshop On Automatic Speaker Recognition, Identification, and Verification*, pp. 221-224, 1994.

[3] C. F. Mullins, and G. B. Giannakis, "Speaker Classification Using Log-Bispectra," *International Symposium on Signal Processing and It's Application*, vol. 1, 1992.

[4] H. M. Teager, and S. M . Teager, "Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract," *Speech Production and Speech Modeling*, (2.J. Hardcastle and a. Marchal, eds.), (the Netherlands), pp. 241-261, Kluwer Academic Publishers, 1990.

[5] G. Giannakis, M. Tsatsanis, "A Unifying Maximum-Likelihood View of Cumulant and Polyspectral Measures for Non-Gaussian Signal Classification and Estimation", *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 386-406, 1992.

[6] C. L. Nikias, M. R. Raghuveer, "Bispectrum Estimation: A Digital Signal Processing Framework," *Proceedings of the IEEE*, vol. 75, no. 7, July 1987.

[7] G. Zhou, G. Giannakis, "Retrieval of Self-Coupled Harmonics", *IEEE Transactions on Signal Processing*, vol. 43, no. 5, pp. 1173-86, 1995.