A Combined Sub-band and Reconstructed Phase Space Approach to Phoneme Classification

Kevin M. Indrebo, Richard J. Povinelli, Michael T. Johnson
Department of Electrical and Computer Engineering, Marquette University, Milwaukee, WI USA
{kevin.indrebo, richard.povinelli, mike.johnson}@marquette.edu

***Abstract* - In this paper a method of classifying phonemes by combining a dynamical systems approach with sub-band decomposition of speech signals is presented. The ability of reconstructed phase spaces to effectively model sub-bands of phonemes in different phonological classes is demonstrated. Experiments performed over the TIMIT database show how well phonemes from different phonological classes can be recognized in different frequency bands. It is hypothesized that given these results, filtering signals before embedding has the potential to improve classification accuracy.**

I. Introduction

Standard automatic speech recognition (ASR) systems use acoustic features based on linear models [1]. The most common of these linearly based acoustic features are cepstral coefficients [1]. The underlying model, upon which cepstral coefficients are based, describes human speech production as an excitation source representing the glottis and a linear time-invariant filter representing the vocal tract. Cepstral analysis allows the excitation source energy to be separated from the frequency response characteristics of the vocal tract.

Such linear assumptions have resulted in many successful speech applications [1]. However, approaches that can capture potential nonlinearities of the vocal tract and coupling of the glottis and vocal tract systems without dramatically increasing the time and space complexity of the corresponding models have the potential to improve ASR performance. Recent studies of nonlinear acoustic features show that exploiting nonlinearities in the speech production system can result in ASR system performance improvements [2-5].

II. Background and Motivation

Our approach to capturing the nonlinearities of the speech production system is based on a dynamical system method called phase space reconstruction. Takens has shown [6] that given proper assumptions a reconstructed phase space (RPS) can be constructed that is topologically equivalent to an original system. In this work the original system is the speech production system. Using the speech signal generated by the speech production system a RPS is formed as follows:

$$\mathbf{x}_n = \begin{bmatrix} x_n & x_{n-\tau} & \cdots & x_{n-(d-1)\tau} \end{bmatrix} \quad n = \left(1+(d-1)\tau\right)\ldots N,$$

where τ is the time lag and d is the dimension of the RPS. A two dimensional RPS is illustrated in Figure 1 for the phoneme '/ao/'.

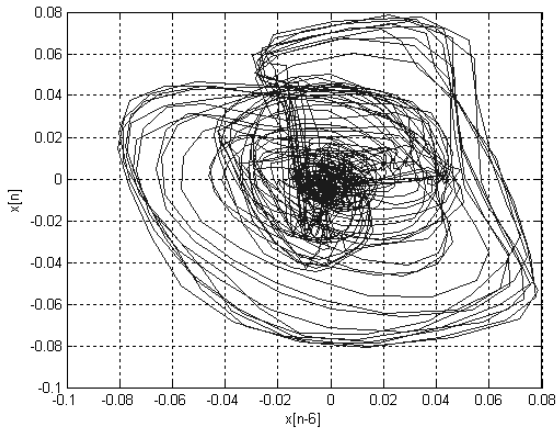

Figure 1 – Reconstructed phase space of '/ao/' phoneme.

Because the RPS is topologically equivalent when d is large enough, the nonlinearities present in the original system are still present in the RPS. Hence, an RPS based approach can capture nonlinear characteristics of the speech production system. We have studied various models of the resulting patterns visible in RPSs with initial promising results [3, 7]. In studying the RPS patterns, an apparent slow/fast dynamic can be observed. This is also seen in traditional speech processing.

One mechanism for exploiting the apparent slow/fast dynamic is through sub-banding the speech signal. Such sub-banding has been previously applied to speech signals with the goal of improving recognition of noisy speech [8-10]. Sub-banding work is motivated in part by experimental work done by Harvey Fletcher at Bell Labs in the 1920's [11]. His results suggest that humans recognize speech in independent frequency bands. There is also substantial evidence that the human cochlea acts as a filter bank, possibly splitting the speech waveform into several sub-bands for recognition [12]. The basilar membrane, which conducts energy received from the outer and middle ears to the hair cells in the inner ear, is shaped in such a way that high frequencies cause large amounts of vibration on one end, and low frequencies cause strong vibrations on the opposite end. Because of this, each location on the basilar membrane reacts most strongly to a

particular frequency, passing the signal components with that frequency on, and attenuating the other frequency components.

III. Sub-band RPS approach

Previous studies have shown that recognition of speech in sub-bands yields ASR systems more robust to narrowband noise [8-10]. If the noise is concentrated in one frequency band, it can be isolated by performing recognition on multiple sub-bands independently, therefore combining the sub-band recognitions can minimize deleterious noise effects on overall speech recognition. In some cases, using sub-band approaches has shown small improvements on uncorrupted speech.

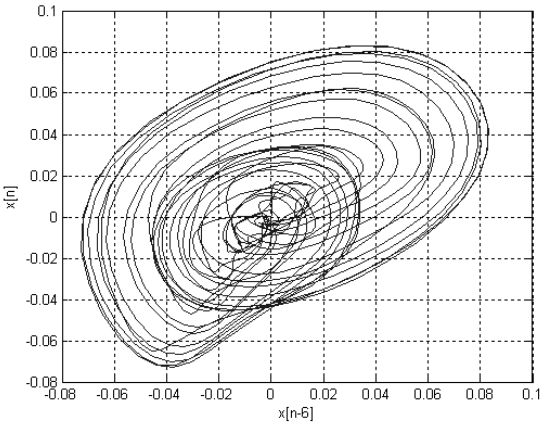

Figure 2 – RPS of '/iy/' phoneme low pass filtered at 1800 Hz.

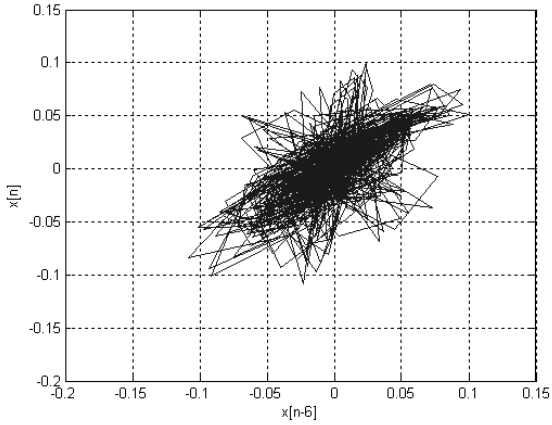

Figure 3 – RPS of '/iy/' phoneme high pass filtered at 1800 Hz.

We take a similar approach using RPSs. Before creating an RPS of the speech signal, the speech signal is passed through a bank of filters, which are Chebychev type II, spaced logarithmically according to the approximate Mel-scale. Figures 2 and 3 show the RPSs for two sub-bands of the '/iy/' phoneme. These sub-bands are created with a lowpass filter with a cutoff of 1800 Hz, and a highpass filter with the same cutoff. The signals are passed through the filters forward and backward to avoid phase distortion of the signal.

After filtering, an RPS is created from each filtered signal. Gaussian mixture models (GMM) of the phase space are learned for each class using the EM algorithm. The GMM for each class uses 128 Gaussian mixtures to describe the distribution of the RPS points over all examples of that class. Each test phoneme is classified with a Bayes' classifier that determines the likelihood of each class for that test example. The likelihood is computed as

$$\hat{\omega} = \arg\max_{i=1\ldots C} \left\{ \hat{p}_i(x) \right\},$$

where x is the test data vector, and C is the number of classes. The class with the greatest likelihood is selected by the classifier.

IV. Experiments

Experiments are performed over the entire TIMIT database. The phoneme set is split into four phonological categories: vowels, fricatives, nasals, and stops. Table 1 shows the number of examples in each testing set.

	Vowels	Fricatives	Nasals	Stops
Testing Set	20,914	7,724	5,105	7,932

Table 1 – Number of examples in training and testing sets for all four categories.

As a baseline, fullband (unfiltered waveform) signals are classified using the RPS/GMM approach with $\tau = 6$ (time lag) and $d = 5$ (dimension) [3]. Then the signals from each data set are filtered into four independent sub-bands, and classification is performed on each sub-band individually. The Chebychev II filters are of order 36, with a stopband attenuation of 70 dB, and are implemented using a second-order section structure.

V. Results

The results for the RPS experiments are shown in Table 2. For the nasal class, one of the sub-bands gave better accuracy than the fullband case, and the stop class had a sub-band with nearly the same accuracy as the fullband. The relative performance of the sub-bands is not uniform across the four classes. Stops are best recognized in the highest and lowest frequency bands, while vowels are better recognized in the middle frequency bands. Fricatives and nasals, though, have the best accuracies in the first and third bands.

Class	Fullband	< 630 Hz	630–1790 Hz	1790–4000 Hz	> 4000 Hz
Vowels	29.59%	17.97%	25.76%	19.03%	7.79%
Fricatives	36.68%	28.11%	22.33%	30.31%	21.48%
Nasals	31.48%	26.68%	27.56%	34.16%	26.15%
Stops	36.28%	35.26%	26.34%	25.29%	31.61%

Table 2 – Classification accuracies of phonemes in four phonological categories in various sub-bands.

VI. Discussion and Conclusions

It was shown that individual RPS sub-bands of phonemes can be used for classification, and that different phoneme classes are classified more successfully in different frequency ranges. Clearly, recognition accuracy could be improved if the recognizer can decide which band(s) to regard as more reliable on an individual phoneme basis.

Developing a system that uses sub-band decomposition and RPSs could yield improvements over the fullband approach. In future work, combination of sub-band classifications will be investigated. Specifically, we will study the number of sub-bands to use, appropriate center frequencies and bandwidths, and methods for combining individual sub-band classifications.

Acknowledgements

This material is based on work supported by the National Science Foundation under Grant No. IIS-0113508 and the Department of Education GAANN Fellowship. The authors would like to thank Andrew Lindgren, Jinjin Ye, and Felice Roberts for portions of the code used in the experiments.

References

[1] B. Gold and N. Morgan, *Speech and audio signal processing*. New York, New York: John Wiley and Sons, 2000.

[2] M. Banbrook, S. McLaughlin, and I. Mann, "Speech characterization and synthesis by nonlinear methods," *IEEE Transactions on Speech and Audio Processing*, vol. 7, pp. 1 -17, 1999.

[3] A. C. Lindgren, M. T. Johnson, and R. J. Povinelli, "Speech recognition using reconstructed phase space features," proceedings of International Conference on Acoustics, Speech and Signal Processing, Hong Kong, 2003, pp. 61-63.

[4] G. Kubin, "Nonlinear speech processing," in *Speech coding and synthesis*, W. B. Kleijn and K. K. Paliwal, Eds.: Elsevier Science, 1995.

[5] A. Kumar and S. K. Mullick, "Nonlinear dynamical analysis of speech," *Journal of the Acoustical Society of America*, vol. 100, pp. 615-629, 1996.

[6] F. Takens, "Detecting strange attractors in turbulence," proceedings of Dynamical Systems and Turbulence, Warwick, 1980, pp. 366-381.

[7] J. Ye, R. J. Povinelli, and M. T. Johnson, "Phoneme classification using naive bayes classifier in reconstructed phase space," proceedings of IEEE Signal Processing Society 10th Digital Signal Processing Workshop, 2002, pp. 2.2.

[8] H. T. Hermansky, S.; Pavel, M., "Towards asr on partially corrupted speech," proceedings of Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, 1996, pp. 462-465 vol.1.

[9] P. V. McCourt, S.; Harte, N., "Multi-resolution cepstral features for phoneme recognition across speech sub-bands," proceedings of Acoustics, Speech, and Signal Processing, 1998. ICASSP '98. Proceedings of the 1998 IEEE International Conference on, 1998, pp. 557-560 vol.1.

[10] S. H. Tibrewala, H., "Sub-band based recognition of noisy speech," proceedings of Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on, 1997, pp. 1255-1258 vol.2.

[11] H. Fletcher, *Speech and hearing in communication*, [2d ed. New York,: Van Nostrand, 1953.

[12] F. Baumgarte, "A computationally efficient cochlear filter bank for perceptual audio coding," proceedings of Acoustics, Speech, and Signal Processing, 2001. Proceedings. 2001 IEEE International Conference on, 2001, pp. 3265-3268 vol.5.