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

This paper presents a method of classifying phonemes by combining a dynamical system approach with sub-band decomposition of speech signals. The ability of reconstructed phase spaces to effectively model sub-bands of phonemes in different phonological classes is studied. The current results are taken from a small speaker-independent set. For the final version of this paper, the entire TIMIT database will be used for experimentation.

Introduction

Standard automatic speech recognition (ASR) systems use acoustic features that are based on linear models, which have proven to be successful in many applications [1]. However, ASR systems often fail when presented with conditions that differ even slightly from those with which the system was trained. Current systems are far inferior to humans, and there are many factors such as noise and speaker variability that severely degrade recognition performance.

Because of the deficiencies in today's ASR systems, interest in nonlinear analysis of speech signals has emerged [2-4]. It is now believed that human speech production includes some nonlinear processes [2].

Background and Motivation

Today, the most popular acoustic features used for speech recognition are cepstral coefficients, which come from a linear model of speech production [1]. This model describes human speech production as an excitation source 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. Because of the linearity assumption, current ASR systems may be ignoring important information contained in speech signals.

Reconstructed phase space (RPS) methods offer an alternative, nonlinear approach to phoneme modeling and classification. Here, analysis is performed in the time domain rather than the frequency domain. A RPS is created by generating *d*-dimensional vectors with the signal and *d-1* time-delayed versions of itself. It has been shown [5, 6] that if *d* is large enough, a RPS is topologically equivalent to the original system that created the signal, and therefore has all of the original information. An example of a two dimension RPS of the phoneme '/ao/', which has been zero-meaned and radial normalized, is shown in Figure 1.

There has been recent work applying a sub-band approach to analyzing speech signals [7-9]. The goal of such work is to improve recognition of noisy speech by combining recognition results from individual sub-bands. This approach is motivated in part by experimental work done by Harvey Fletcher at Bell Labs in the 1920's [10]. His results suggest that humans recognize speech in independent frequency bands.

There is substantial evidence that the human cochlea acts as a filter bank, possibly splitting the speech waveform into several sub-bands for recognition [11]. The basilar membrane (BM), 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 BM reacts most strongly to a particular

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

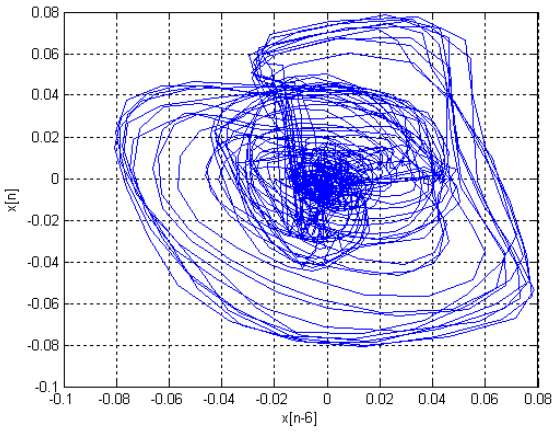

Figure 1. Reconstructed phase space of '/ao/' phoneme.

Sub-band RPS approach

Previous studies have shown that recognition of speech in sub-bands can make ASR systems more robust to narrowband noise [7-9]. If the noise is located in one frequency band, it can be isolated by performing recognition on multiple sub-bands independently. Combining the sub-band recognitions can then minimize noise effects. In some cases, using this sub-band approach has shown small improvements even on uncorrupted speech.

We take a similar approach using reconstructed phase spaces. Before embedding the speech signals into RPS, they are passed through a filter bank. The filters are sets of FIR filters with Kaiser windows of length 255, spaced logarithmically according to the approximate Mel-scale. Figure 2 shows the RPS of two sub-bands of one phoneme. The sub-bands shown are created with a lowpass filter with a cutoff of 1800 Hz, and a highpass filter with the same cutoff. Linear-phase filters are used to avoid phase distortion of the signal.

After filtering, a RPS is created from each filtered signal as a sequence of *d*-dimensional vectors, with the *n*th vector defined by

$$\mathbf{x}_n = \left[x_{n-(d-1)\tau} \quad \cdots \quad x_{n-\tau} \quad x_n \right] \quad n = \left(1 + \left(d-1\right)\tau\right) \ldots N \tag{1}$$

where τ is the time lag. Gaussian mixture models (GMM) of the phase space points are then built for each class using the expectation maximization algorithm. The GMM for a class describes the distribution of the RPS points over all examples of that class. We use 38 mixtures, which maximizes the descriptive power of the GMM with reasonable computational requirements. Each test phoneme is classified with a Bayes' classifier that finds the likelihood of each class for that test example. The likelihood is computed as

$$\hat{\omega} = \underset{i=1\ldots C}{\arg\max} \left\{ \hat{p}_i \left(x \right) \right\} \tag{2}$$

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.

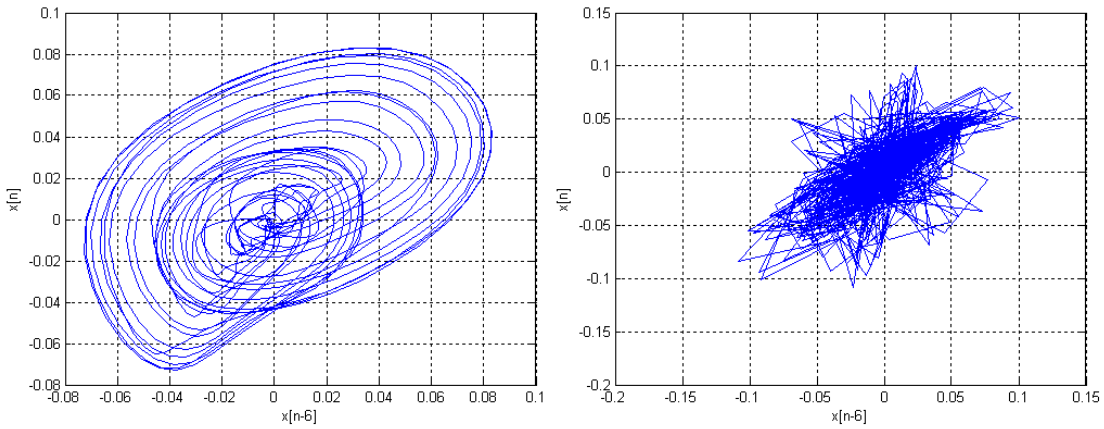

Figure 2. RPS of '/iy/' phoneme low pass filtered (left), and high pass filtered (right) at 1800 Hz.

Experiments

The experiments are performed over a small speaker independent data set (20 speakers) from the 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 training and testing set. Results for the final paper will include experiments run over all of Timit. In addition, the same experiments will be run using Mel-frequency cepstral coefficients for comparison.

	Vowels	Fricatives	Nasals	Stops
Training Set	1,336	572	368	581
Testing Set	874	354	259	371

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) [12]. Then the signals from each data set are filtered into four independent sub-bands, and classification is performed on each sub-band individually. The FIR filters are implemented with a Kaiser window with $\beta = 8$ and window length $M = 255$.

Results

The results for the RPS experiments are shown in Table 2. For two of the phonological classes, there is at least one sub-band that had better accuracy than the fullband. The other two phonological classes have at least one sub-band with nearly the same accuracy as the fullband. Also, the relative performance of the sub-bands is not uniform across the four classes. Vowels and stops are better recognized in the low to mid frequency bands, whereas fricatives are recognized more accurately in the higher bands.

Class	Fullband	< 630 Hz	630–1790 Hz	1790–3955 Hz	> 3955 Hz
Vowels	29.38%	22.32%	28.47%	20.27%	15.49%
Fricatives	49.72%	37.99%	28.49%	44.69%	48.04%
Nasals	38.29%	36.49%	45.05%	33.33%	46.40%
Stops	37.97%	42.03%	34.94%	33.42%	32.91%

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

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 RPS could yield significant improvements over the fullband approach. In future work, combination of sub-band classifications will need to be investigated thoroughly. Specifically, there are several questions that need to be addressed:

1. How many sub-bands should be used?
2. What are the appropriate center frequencies and bandwidths?
3. How should the individual classifications be combined?

Further experimentation with larger data sets will help us understand the nature of RPS methods in sub-bands of speech. The experiments discussed will be run on the entire TIMIT database. Also, Mel-frequency cepstral coefficients will be used as features to provide comparisons to the RPS techniques.

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

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

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

[6] T. Sauer, J. A. Yorke, and M. Casdagli, "Embedology," *Journal of Statistical Physics*, vol. 65, pp. 579-616, 1991.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] A. C. Lindgren, M. T. Johnson, and R. J. Povinelli, “Speech recognition using reconstructed phase space features,” proceedings of Acoustics, Speech, and Signal Processing, 2003. ICASSP-03., 2003 IEEE International Conference on, in press.