A COMPARISON OF RECONSTRUCTED PHASE SPACES AND CEPSTRAL COEFFICIENTS FOR MULTI-BAND PHONEME CLASSIFICATION

Kevin M. Indrebo Richard J. Povinelli Michael T. Johnson

Department of Electrical and Computer Engineering
Marquette University, Milwaukee, WI USA
Email: {Kevin.Indrebo, Richard.Povinelli, Mike.Johnson}@Marquette.edu

Abstract: This paper examines the use of multi-band reconstructed phase spaces as models for phoneme classification. Sub-banding reconstructed phase spaces combines linear, frequency-based techniques with a nonlinear modeling approach to speech recognition. Experiments comparing the effects of filtering speech signals for both reconstructed phase space and traditional speech recognition approaches are presented. These experiments study the use of two non-overlapping sub-bands for isolated phoneme classification on the TIMIT corpus. It is shown that while classification accuracy using Mel frequency cepstral coefficients as features does not improve with sub-banding, the accuracy increases from 36.1% to 42.0% using sub-banded reconstructed phase spaces to model the phonemes.

1. INTRODUCTION

Nonlinear acoustic modeling of human speech is an emerging area of research [1-3]. Unlike the analysis performed by typical automatic speech recognition (ASR) systems, nonlinear methods primarily use time-domain analysis rather than frequency-domain analysis. Examples of nonlinear features that can be extracted from speech waveforms include fractal and correlation dimension, modulation features, and Lyapunov exponents [4-6]. Investigation into nonlinear modeling of speech is motivated by research that suggests human speech contains nonlinear components [1]. Unfortunately, nonlinear techniques have drawbacks as well, including added complexity, and lack of a standard fundamental nonlinear speech production model.

The nonlinear approach proposed here is based on phase space reconstruction. Phase space reconstruction provides a mechanism to recover the dynamics of a system from a sampled signal generated by that system [7-9]. Recent work has shown that reconstructed phase spaces (RPS) combined with Gaussian Mixture Models (GMM) can be effectively used for modeling and classifying isolated phonemes [10]. While the RPS approach has not outperformed state-of-the-art modeling techniques, it has approached the ability of cepstral analysis on isolated phoneme classification.

In previous work [11], the authors have shown that RPSs created from sub-banded signals can discriminate amongst phonemes. In this work, we extend the investigation to show how Bayesian combination of sub-band phoneme classification log likelihoods improves overall classification accuracies.

In the rest of the paper, background on reconstructed phase spaces and sub-bands is given, followed by the methodology and experimental setup used for this paper, and the experimental results.

2. BACKGROUND

2.1. Reconstructed Phase Spaces

An RPS is a model of the dynamics of a system that can be created using a time series of one of the state-variables of the system. To generate an RPS, a trajectory matrix is established as

$$\mathbf{X}=\begin{bmatrix}\mathbf{x}_{1+(d-1)\tau}\\ \mathbf{x}_{2+(d-1)\tau}\\ \vdots\\ \mathbf{x}_{N}\end{bmatrix}=\begin{bmatrix}x_{1+(d-1)\tau} & \cdots & x_{1+\tau} & x_1\\ x_{2+(d-1)\tau} & \cdots & x_{2+\tau} & x_2\\ \vdots & & \ddots & \\ x_N & \cdots & x_{N-(d-2)\tau} & x_{N-(d-1)\tau}\end{bmatrix}_{(N-(d-1)\tau)\times d},$$

where d is the embedding dimension, τ is the time lag, and x_n is the *nth* point in the signal.

The concept of phase space reconstruction, also known as a delay-coordinate mapping or time-delay embedding, was first introduced by Packard [9]. Using Whitney's theorem [12], Takens showed that a delay-coordinate mapping from a generic n dimensional state space to a space of dimension $2n+1$ preserves topology [7]. Later, Sauer and Yorke extended this by showing that, with probability 1, a time-delay embedding is topologically equivalent to the original dynamical system, provided that the embedding dimension is greater than twice the box-counting dimension of the original system [8].

2.2. Sub-bands

There are reasons to believe that performing speech recognition in frequency sub-bands may be beneficial. Experiments performed on human speech recognition have suggested that humans recognize sounds in individual sub-bands, and then integrate the resulting information [13]. This allows for robust recognition that can function even if some frequency bands are distorted or eliminated.

Multi-band acoustic features have been used in an attempt to replicate the robustness of human speech recognition in ASR systems. It has been shown that recognizers using traditional linear features derived from frequency sub-bands can outperform fullband systems when narrowband noise has been added to test signals [14, 15]. Typically, though, little or no improvement in recognition accuracy is seen on clean speech when using a simple Bayesian combination [14, 16].

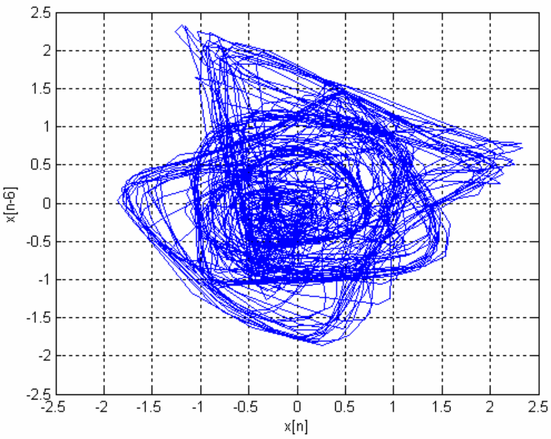

Figure 1. RPS of phoneme '/ae/'.

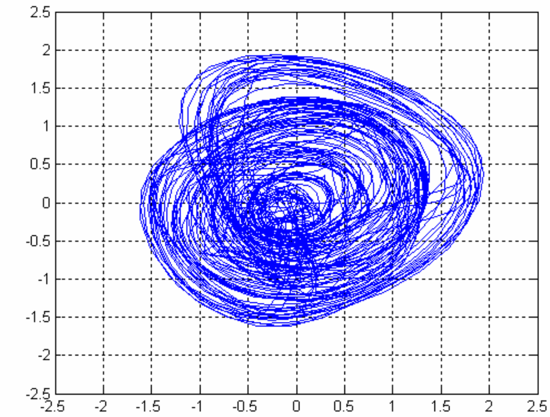

Figure 2. RPS of phoneme '/ae/' lowpass filtered at 1500 Hz.

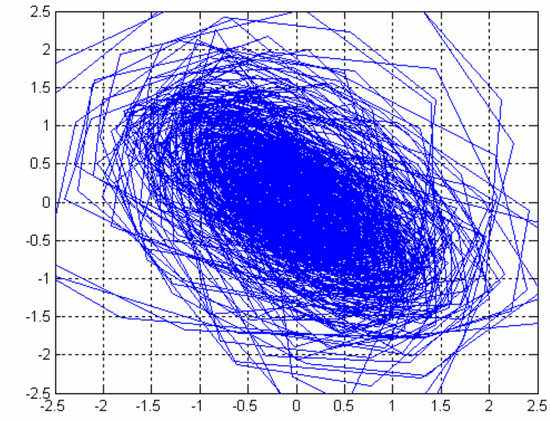

Figure 3. RPS of phoneme '/ae/' highpass filtered at 1500 Hz.

Unlike Mel frequency cepstral coefficients (MFCC) and other traditional speech recognition features, RPSs are not based on the frequency spectrum of speech signals. RPSs model the time-domain characteristics of the signal, but not explicitly the spectral characteristics. Therefore, sub-banding speech signals before embedding them into RPSs could potentially improve their discriminatory power by integrating frequency information with the dynamics.

A two dimensional RPS of the vowel '/ae/' is shown in Figure 1. Exploring the characteristic structure of this plot is our objective. Figures 2 and 3 display a sub-banded version of this phoneme, which was filtered using a highpass and lowpass filter, each with a cutoff at 1500 Hz. The lowband RPS appears similar in structure to the unfiltered version, but is clearly smoother, displaying the slower moving dynamics of the vowel. The highband RPS appears less structured, and could be of a higher dimension, or could perhaps contain less information.

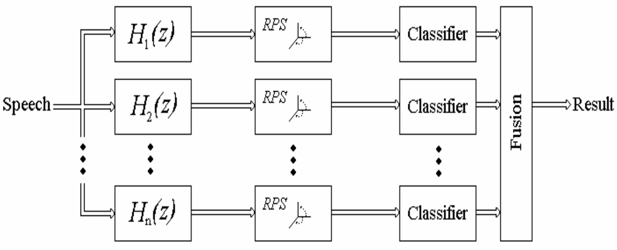

Figure 4. RPS sub-band classifier.

3. METHODOLOGY

The process of sub-band RPS classification is illustrated in Figure 2. Each speech signal is separated into a number of frequency band-limited signals. The filters are designed to completely isolate the dynamics of the desired sub-band, so the attenuation in the stop band is greater than the loudest frequencies found in human speech. Infinite impulse response filters are used because of the need for narrow transition bands.

These filtered signals are then embedded into an RPS, which is zero-meaned and radial normalized. A GMM models the RPS of each phoneme class. The GMM probability distribution of the *ith* class, with M mixtures is defined as

$$\hat{p}_i(\mathrm{x}) = \sum_{m=1}^{M} w_{im}\mathcal{N}(\mathrm{x};\mathrm{u}_{im},\Sigma_{im}),$$

where μ_{im} is the mean, and Σ_{im} is the covariance matrix of mixture m. Each mixture has an associated weight w_{im}. The sum of the M weights must be equal to one. An example GMM is illustrated in Figure 5, which shows a 16 mixture GMM learned over a two dimensional RPS. The ellipses represent one standard deviation from the centroid of each mixture. For classification of a given exemplar, each sub-banded RPS signal is assigned a log likelihood $\hat{\omega}$ for the *ith* class according to

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

and these sub-band likelihoods are then fused to produce one final likelihood for each class.

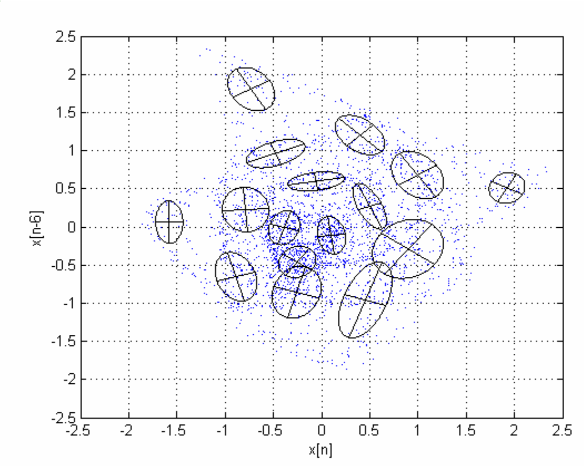

Figure 5. 16 mixtures learned over phoneme '/ae/'.

4. EXPERIMENTAL SETUP

4.1. Data Set

To test the proposed sub-banding approach, we ran two sets of experiments over the TIMIT database [17]. TIMIT is a speaker independent database with read speech. There are 462 speakers in the training set and 24 speakers in the core test set. The task is isolated phoneme classification, so the expertly labeled phonemes in TIMIT are extracted for modeling and classification.

4.2. Setup

The experiments presented here examine the classification ability of RPSs and MFCCs in pairs of sub-bands, a lowband and a highband, each filtered with the same cutoff. Six cutoffs are chosen based on the approximate Mel-scale:

$$Mels = 1127 * \ln(1 + \frac{f}{700}).$$

The classification accuracies of each method on the lowband data, the highband data, and the Bayesian fusion of the two bands are examined. These results are compared against a fullband baseline.

After all individual phonemes are extracted using the time stamps provided, each signal is filtered with a lowpass or a highpass Chebychev type II filter. Each filter is of order 36, with a stop band attenuation of 70 dB. These parameters are chosen to ensure that the dynamics outside of the selected sub-band are completely removed, and the transition band is as narrow as possible. All signals are filtered twice, once forward in time and once backward in time to eliminate phase distortion.

HTK [18] is used for training and testing of both experiment sets. Each sub-band is represented as a stream in HTK, and the combination is done with equal stream weights.

4.3. RPS Experiments

The filtered signals are embedded into a ten dimension RPS with five axes as the time points of the signal at each lag, and five axes of deltas computed over each of the lagged signals. Each RPS point is defined as

$$\mathrm{x}_n = [x_{n-4\tau} \cdots x_{n-\tau} \; x_n \; \Delta_{n-4\tau} \cdots \Delta_{n-\tau} \; \Delta_n],$$

and the delta coefficients are given by

$$\Delta_n = \frac{\sum_{\Theta}^{\Theta} \theta(x_{n+\theta} - x_{n-\theta})}{2\sum_{\theta=1}^{\Theta} \theta^2},$$

with Θ = 5. The delta function is equivalent to an FIR filter, which, as a smooth linear transformation on the space preserves topological equivalence of the RPS [8]. The dimension, lag, and number of mixtures in the GMM, which are 10, 6, and 128 respectively, were determined empirically [10].

4.4. Cepstral Experiments

Training and testing for the Mel frequency cepstral coefficient (MFCC) features is done in a manner similar to that of the RPS. To have a comparable number of features as in the RPS experiments, 10 MFCCs are computed from each sub-band using twelve Mel frequency channels. The GMMs used to model the MFCC features contain 16 mixtures.

5. RESULTS

The results for the RPS sub-band experiments are shown in Table 1, along with the baseline fullband RPS accuracy. The greatest accuracy is found in the fusion of the two sub-bands filtered at 1380 Hz, and is approximately 6% higher than the fullband RPS accuracy.

It is interesting to note that at the third, fourth, and fifth cutoffs the lowband accuracy is greater than that of the fullband. As can be seen from Table 1, the accuracies of the low and high bands are equivalent somewhere below 750 Hz. Unlike the lowband, the highband performance is severely degraded when only the lowest 10% of the frequency spectrum is filtered out. Clearly, the RPS method appears to be able to model signals with slow moving dynamics better than those with fast moving dynamics.

Fullband baseline: 36.10%			
Cutoff Freq	Lowband	Highband	Fusion
320 Hz	18.84%	29.01%	29.76%
750 Hz	32.06%	22.34%	39.06%
1380 Hz	37.61%	19.68%	42.01%
2280 Hz	37.74%	18.30%	40.39%
3560 Hz	37.48%	14.70%	38.52%
5380 Hz	36.09%	13.09%	36.96%

Table 1. RPS accuracies for each filter cutoff.

Table 2 shows the results for the sub-banded MFCC experiments. The highband cepstrals provided for better classification than the highband RPSs across the board, but the lowband cepstrals could not match their RPS counterparts for the lower three cutoffs. More importantly, the greatest MFCC fusion accuracy was nearly equal to the fullband MFCC baseline.

Fullband baseline: 47.78%			
Cutoff Freq	Lowband	Highband	Fusion
320 Hz	15.26%	42.61%	43.09%
750 Hz	27.44%	37.94%	47.12%
1380 Hz	33.69%	32.28%	47.58%
2280 Hz	40.78%	25.07%	47.14%
3560 Hz	43.49%	20.30%	47.14%
5380 Hz	46.68%	15.59%	47.83%

Table 2. MFCC accuracies for each filter cutoff.

6. CONCLUSION

We have shown that sub-band decomposition of speech signals can substantially improve the classification performance of the reconstructed phase space approach. Filtering phonemes into two sub-bands and fusing the RPS classification likelihoods with a naïve Bayesian combination yielded an increase in accuracy from 36.1% to 42.0%. In contrast, there appears to be no benefit to sub-banding MFCCs in clean speech.

In future work, we will investigate why RPSs are able to model lowpass filtered signals more effectively than highpass filtered signals, and also look at increasing the number of sub-bands. Additionally more sophisticated fusion strategies for combining subband likelihoods will be considered.

7. REFERENCES

[1] 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.

[2] N. D. Warakagoda and M. H. Johnsen, “Nonlinear dynamical system based acoustic modeling for asr,” proceedings of Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01). 2001 IEEE International Conference on, 2001, pp. 525-528 vol.1.

[3] P. Maragos, A. G. Dimakis, and I. Kokkinos, “Some advances in nonlinear speech modeling using modulations, fractals, and chaos,” proceedings of Digital Signal Processing, 2002. DSP 2002. 2002 14th International Conference on, 2002, pp. 325-332 vol.1.

[4] D. Dimitriadis, P. Maragos, and A. Potamianos, “Modulation features for speech recognition,” proceedings of Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on, 2002, pp. I-377-I-380 vol.1.

[5] V. Pitsikalis and P. Maragos, “Speech analysis and feature extraction using chaotic models,” proceedings of Acoustics, Speech, and Signal Processing, 2002. Proceedings. (ICASSP '02). IEEE International Conference on, 2002, pp. I-533-I-536 vol.1.

[6] S. S. Narayanan and A. A. Alwan, "A nonlinear dynamical systems analysis of fricative consonants," *Journal of the Acoustical Society of America*, vol. 97, pp. 2511-2524, 1995.

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

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

[9] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, "Geometry from a time series," *Physical Review Letters*, vol. 45, pp. 712-716, 1980.

[10] 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.

[11] K. M. Indrebo, R. J. Povinelli, and M. T. Johnson, “A combined sub-band and reconstructed phase space approach to phoneme classification,” proceedings of ISCA Tutorial and Research Workshop on Non-linear Speech Processing (NOLISP), Le Croisic, France, 2003, pp. 107-110.

[12] H. Whitney, "Differentiable manifolds," *The Annals of Mathematics, 2nd Series*, vol. 37, pp. 645-680, 1936.

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

[14] H. Hermansky, S. Tibrewala, and M. Pavel, “Towards asr on partially corrupted speech,” proceedings of Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, 1996, pp. 462-465 vol.1.

[15] P. McCourt, S. Vaseght, and N. Harte, “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.

[16] S. Tibrewala and H. Hermansky, “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.

[17] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "Timit acoustic-phonetic continuous speech corpus," Linguistic Data Consortium, 1993.

[18] “Htk,” Version 2.1: Entropic Cambridge Research Laboratory Ltd., 1997.