Sub-banded Reconstructed Phase Spaces for Speech Recognition

Kevin M Indrebo [a,*], Richard J. Povinelli [a], Michael T. Johnson [a]

[a] Department of Electrical and Computer Engineering
Marquette University
1515 W. Wisconsin Ave.
Milwaukee, WI 53233, USA
{kevin.indrebo, richard.povinelli, mike.johnson}@marquette.edu

Abstract – A novel method combining filter banks and reconstructed phase spaces is proposed for the modeling and classification of speech. Reconstructed phase spaces, which are based on dynamical systems theory, have advantages over spectral-based analysis methods in that they can capture nonlinear or higher-order statistics. Recent work has shown that the natural measure of a reconstructed phase space can be used for modeling and classification of phonemes. In this work, sub-banding of speech, which has been examined for recognition of noise-corrupted speech, is studied in combination with phase space reconstruction. This sub-banding, which is motivated by empirical psychoacoustical studies, is shown to dramatically improve the phoneme classification accuracy of reconstructed phase space-based approaches. Experiments that examine the performance of fused sub-banded reconstructed phase spaces for phoneme classification are presented. Comparisons against a cepstral-based classifier show that the proposed approach is competitive with state-of-the-art methods for modeling and classification of phonemes. Combination of cepstral-based features and the sub-band RPS features shows improvement over a cepstral-only baseline.

Keywords: Speech Recognition, Dynamical Systems, Nonlinear signal processing, Sub-bands

[*] Corresponding Author

1. Introduction

1.1 Traditional ASR systems

Today's state-of-the-art speech recognition systems use Hidden Markov Models (HMM) to model speech signals using features extracted from the short-term power spectrum of the waveforms. Fourier analysis is used to compute the magnitude spectrum of the signals, and features are extracted for modeling. The most common spectral feature type is the Mel-frequency cepstral coefficient (MFCC), which is based on the linear time-invariant model of the human vocal tract separating the excitation from the vocal-tract characteristics. For English speech, this is considered a good base model, as the excitation is considered to have minor discriminatory power for individual phonemes. Cepstral coefficients are a good match for this model because they capture primarily vocal tract characteristics, however because they are derived from the power spectrum of speech, they are unable to capture nonlinear or phase characteristics. Traditionally, it was thought that there was little relevant information outside of the speech magnitude spectrum, but recent studies have shown significant nonlinear components in speech (Banbrook and McLaughlin, 1994; Banbrook et al., 1999; Teager and Teager, 1990).

Due to these findings, research in analyzing and modeling speech signals as nonlinear processes has been increasing. This work is typically based on higher-order statistics (Moreno and Rutllan, 1996), dynamical systems (Kubin, 1995), nonlinear models of speech production (Dimitriadis et al., 2002), or chaos theory (Pitsikalis and Maragos, 2002). The work presented here is based on dynamical systems theory, namely the embedding theorems for reconstructed phase spaces (RPSs) (Sauer et al., 1991;

Takens, 1980). In this paper, we integrate a filter bank, using ideas from Fletcher's work (Fletcher, 1953), with a modeling and classification system based on RPSs.

1.2 Reconstructed Phase Spaces

The justification for the use of reconstructed phase spaces as a tool for signal classification lies in theorems that show topological equivalence between the original state space of the system and the reconstructed phase space. Takens (Takens, 1980) showed that, given a system of dimension d described by a state vector $\mathbf{x}$, a state evolution function $\varphi(\mathbf{x})$, and a smooth map y from the system state to an output variable, the transformation

$$\Phi_{(y,\varphi)}(\mathbf{x}) = \left(y(\mathbf{x}), y(\varphi(\mathbf{x})), \ldots, y\left(\varphi^{2n}(\mathbf{x})\right)\right) \qquad (1)$$

is an embedding, meaning it is bijective and differentiable. Sauer, Yorke and Casdagli (Sauer et al., 1991) proved that almost every time delay map $\Phi_{(y,\varphi)}$ is an embedding, showing that, except for a set of degenerate cases with measure zero, the topological equivalence property is guaranteed. Together, these theorems guarantee that for almost every time delay embedding, the reconstructed dynamics of the map are topologically identical to the true dynamics of the system.

This theorem leads directly to the definition of an RPS, which is created by embedding a signal against time-lagged versions of itself. The trajectory matrix $\mathbf{X}$ generated by phase space reconstruction of a signal $\mathbf{x}$ is given by

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

where x_n is the signal value at sample n, d is the embedding dimension, and τ is the time-lag. There are $(d-1)\tau$ fewer points in the RPS trajectory matrix than in the original signal. Though RPSs can be used to estimate dynamical invariants such as Lyapunov exponents and correlation dimension, we model these spaces directly by estimating the natural measure with Gaussian Mixture Models (GMMs).

Examples of two-dimensional RPSs of four phonemes are shown in Figure 1. The phonemes shown here represent four major phonetic classes, namely vowels, fricatives, plosives, and nasals. These RPSs show that the trajectories of vowels show the smoothest structure, while fricatives appear to have a much less distinct structure. The other phonemes have RPS shapes that tend to fall somewhere in between these two extremes.

1.3 RPSs for phoneme classification

Previously, statistical modeling of RPSs for classification of phonemes has been studied. In (Lindgren et al., 2003), experiments comparing Gaussian Mixture Modeling of RPS features and MFCCs for speaker-independent isolated phoneme classification on the TIMIT corpus were presented. Sub-banding of RPSs for classification of phonemes in four major phonological classes was studied in (Indrebo et al., 2003). In (Johnson et al., in press), the effect of embedding dimension and lag on phoneme classification accuracy was examined.

1.4 The contributions of this paper

In this paper, we extend previous work by combining the RPS approach with filter banks to study the nonlinear dynamics of speech signals in frequency sub-bands. Sub-banding of speech signals for speech recognition has been studied, with potential advantages for robust recognition. Those approaches already use frequency-based analysis, so that the benefits of sub-banding for recognition of clean speech signals are minimal. Here, the results show that combining frequency sub-banding with the RPS approach improves the classification of phonemes even for uncorrupted speech signals.

The remainder of this paper is as follows. Section 2 describes the motivation for sub-banding, followed by a presentation of the methodology in section 3. In section 4, the experimental results are given. Section 5 provides an analysis of the results, and a discussion of possible future work. A conclusion is then presented in section 6.

2. Sub-banding/Filtering

2.1 Motivation for sub-banding

Harvey Fletcher, working at Bell Labs in the 1910's and 1920's, performed experiments examining human speech recognition in sub-bands (Fletcher, 1953). He utilized low-pass and high-pass filters with varying cutoffs to study the ability of humans to correctly recognized phonemes with limited frequency bandwidths. The subjects listened to nonsense syllables to remove any language information. He found that the phoneme recognition error rate over a particular band was equal to the product of the error rates for the component sub-bands. This relationship is expressed as

$$e_{total} = \prod_{i=0}^{N-1} e_i \tag{3}$$

where e_{total} is the error rate of a band, N is the number of sub-bands composing the band, and e_i are the errors in each sub-band. Using these findings, he developed his well-known articulation index, a function that defines sub-bands of equal articulation, or phoneme recognition rate. His results lend strong support for the theory that humans recognize phonemes in sub-bands independently.

The structure and behavior of the human cochlea fit well with Fletcher's findings. The basilar membrane, which conducts acoustic vibrations to the inner hair cells that in turn excite neurons in the auditory nerve, has a spatially variant rigidity. This causes the basilar to act as set of band pass filters, with a frequency response dependent on the location on the basilar membrane (Gold and Morgan, 2000). This results in the appearance of tuning curves in the auditory nerve fibers, which isolate bands of frequency information.

In addition to the physiological and psychoacoustic reasons for sub-banding, practical issues in RPS analysis motivate this approach. Because of the nature of human speech, the dynamics of speech signals residing primarily in sub-bands with comparatively low energy can become hidden by dynamics in higher energy bands. Since statistical models of the RPS density are used for classification here, small perturbations in data points lying along a trajectory in the RPS are not captured. This is often seen in the form of low energy high frequency dynamics conveyed along a higher energy low frequency carrier wave. Sub-banding allows for modeling of lower energy dynamics with better resolution.

An example of an RPS, a vowel, filtered into four sub-bands is depicted in Figure 2, where it is observed that the lower frequency band RPSs have a smoother structure, and appear to be more structured than the higher frequency band RPSs. This is likely the product of not only that low frequency signals appear smoother, but also that, as a vowel, this phoneme carries less information in the higher frequencies.

2.2 Sub-banding in Traditional ASR systems

Sub-banding has been studied for robust recognition of noisy speech (Bourlard and Dupont, 1996; Hermansky et al., 1996; McCourt et al., 1998; Tibrewala and Hermansky, 1997). Traditional systems, i.e. those using MFCCs as acoustic features, can benefit from this approach in situations where speech is corrupted by narrowband noise. As each cepstral coefficient contains information from the entire spectrum, noise in one region of the spectrum will corrupt every feature. If the cepstral coefficients are computed in multiple bands that are isolated from each other, only a portion of the coefficients is distorted. Suppression of the significance of these corrupted features on pattern recognition can result in recognizers that are more robust.

Though sub-banding has been shown to increase the robustness of ASR systems in several types of noise (Hagen et al., 2001; Hermansky et al., 1996; McCourt et al., 1998), it generally has not proven to significantly improve spectral-based recognition systems on uncorrupted speech (McCourt et al., 1998; Tibrewala and Hermansky, 1997). This is not surprising, considering that these traditional features are already based on analysis of narrow regions in the power spectrum. A major point of emphasis in this paper lies in the contrast of the proposed system with spectrum-based systems, with respect to classification accuracy enhancement in clean speech through sub-banding.

2.3 Embedding filtered signals into RPS

Given the theory behind RPSs, the question of preserved topology arises when filtered signals are embedded into RPSs. The justification for statistical modeling of RPSs originated in the theorems of Takens (Takens, 1980) and Sauer et al. (Sauer et al., 1991), followed by the empirical success of this method for classification of heart arrhythmias (Roberts et al., 2001), motor faults (Povinelli et al., 2002), and speech phonemes (Johnson et al., in press; Povinelli et al., 2004). With the addition of the filter bank front-end proposed in this paper, though, a re-examination of topology equivalence is warranted.

Any linear transformation on a space preserves the topology of that space (Gibson et al., 1992). Previously, transforms such as principle component analysis (PCA) have been used to reduce the dimension of RPSs (Ye et al., 2003). A PCA projects an RPS into a lower dimension, where each new dimension is an orthogonal linear combination of the original dimensions. This operation has shown the ability to improve phoneme classification accuracy in some cases.

Augmenting RPSs with trajectory information can also improve their discrimination capabilities. In (Lindgren et al., 2003), short-term delta coefficients were used to augment five-dimensional RPSs, creating ten-dimensional RPSs. This operation improved the classification accuracy significantly.

As a smooth, invertible transform, finite-impulse response (FIR) filters do not destroy the topological equivalence between an RPS and the underlying dynamical system (Sauer et al., 1991). This property can be exploited for RPS analysis of signals that have noise components that must be removed. FIR filters could also be implemented for the front-end filter bank in the proposed system. However, to isolate the dynamics in each band, small transition bands are desirable, requiring unfortunately long impulse responses.

Instead, we have chosen to use IIR filters, which require much lower orders to achieve sharp cutoffs. It has been shown that convolution of chaotic signals with infiniteimpulse response (IIR) filters can change the dynamical invariants of the systems (Badii et al., 1988; Chennaoui et al., 1990; Isabelle et al., 1992). This is a cause for concern if the analysis method involves computing dynamical invariant features such as Lyapunov exponents or fractal dimensions. Unfortunately, this also means that the topological equivalence property has been lost. It is the authors' belief, however, that this does not mean effective modeling and classification is impossible. On the contrary, it will be shown that these filtered RPSs still hold much information for discrimination of phonemes.

3. Methodology

3.1 System overview

The proposed system for modeling and classification of phonemes is depicted in Figure 3. The speech signal to be analyzed is first filtered with the front-end filter bank, creating a set of sub-banded signals. The signal is then embedded into an RPS, which is modeled using a Gaussian Mixture Models (GMM). Classification is performed on the RPSs using Bayesian classifiers. The likelihoods produced by these classifiers are fused to form an overall classification. Each of these modules is discussed in this section.

3.2 Filter Bank structure

The number of filters utilized in the filter bank is the first variable that defines the performance of the proposed system. It has been suggested that the human auditory system has at least ten sub-bands, and likely no more than thirty (Allen, 1994). While the proposed methodology is originally based on models of human hearing, it may not be best to completely mimic all facets of the human ear. Traditional ASR systems using sub-banding typically have two to seven bands (Hermansky et al., 1996; McCourt et al., 1998;

Tibrewala and Hermansky, 1997). These systems are constrained by the resolution of the spectral analysis. With too many sub-bands, the features extracted from each band will contain minimal information. Thus, this range may be inappropriate for our analysis. However, because of the complexity of the RPS approach, we are limited to the number of experiments that can be carried out in a reasonable amount of time. We have chosen to study the effect of the size of the filter bank on the system performance, by varying the number of filters, using three different sizes: sets of two, four, and eight sub-bands.

Another design parameter of this system concerns the central frequencies and bandwidths of the filters. The simplest method would be to simply divide the bandwidth equally. However, this does not seem appropriate, as it does not correspond to the human ear. The spectral resolution of the human ear varies logarithmically along the frequency axis, with better resolution at lower frequencies (Gold and Morgan, 2000). Therefore, it seems reasonable to space the filters logarithmically, according to some scale that resembles human audition. Potential choices include the Mel-scale and the Bark-scale, both of which are empirically determined using human subjects. Due to its popularity in the speech recognition community, the Mel-scale is utilized as the basis for spacing the filter channels. To simplify the analysis, the filter banks are designed with no overlapping filters.

Though the filters in the human auditory model are approximately gammatone shaped (Gold and Morgan, 2000), we used approximately brickwall-shaped filters to simplify our analysis. Chebychev type II filters are chosen for analysis because of their control over the stopband. As it is preferred that the dynamics in each band are isolated, and a frame of speech can contain frequency components with very different energies, up to seventy or eighty decibels, it is desirable to have control over the ripple in the stopband. Therefore, Chevychev type II filters, which are designed using the δ for the stopband, which specifies the maximum ripple, are used for implementation of the filter bank.

To avoid phase distortion, the speech signals are filtered forward and backward. It is very difficult to design an IIR filter with a linear phase response. Because the RPS approach can model phase characteristics of signals, distortion of the signal's phase may be detrimental to the proposed system. To combat this, a forward-backward filtering technique is used. Any phase distortion introduced in the forward filtering operation is reversed in the backward filtering, eliminating this problem.

3.3 RPS parameters

Though there are theoretical requirements for the embedding dimension, these require knowledge of the underlying system dimension (Takens, 1980). Typically, this quantity is not known, necessitating alternative methods for selecting the best embedding dimension. Additionally, these theorems provide no conditions on the time lag. Though any lag provides for topological equivalence, lag affects the shape of the RPS, and so can affect the performance of the respective modeling and classification algorithm.

Popular heuristics have been developed to find appropriate embedding dimension and time-lag (Abarbanel, 1996). These techniques are not directly derived from the theory behind RPSs, but are based rather on common-sense approaches for maximizing the information captured by the RPS.

Two common heuristics for determining the time lag involve the autocorrelation and auto-mutual information functions of the signal. These functions determine the amount of redundant information in higher-dimensional representations in a signal as a function of lag. As it is slightly more common in the literature, the method of selecting the first minimum of the auto-mutual information is chosen for determining the best time-lag in this paper. Auto-mutual information is defined by

$$I(\tau)=\sum_{i,j} p_{ij}(\tau)\ln\frac{p_{ij}(\tau)}{p_i(\tau)p_j(\tau)}, \tag{4}$$

where τ is the lag, and p_{ij} is the joint probability of the signal with values i and j. This method has led to the selection of six as the time lag. A more detailed analysis can be found in (Johnson et al., in press).

Once the time lag is chosen, global false nearest neighbors is used for selecting the embedding dimension. As the dimension of the RPS is increased, points in the space that are close together are sometimes pulled farther apart. The close proximity of these points in the lower dimension may be due to projection rather than geometry. Once a sufficient dimension is reached, and the attractor is completely unfolded, all points that lie in the same neighborhood remain together as the dimension increases. This signals that adding more dimensions is unnecessary, and will not be beneficial. This process can be carried out algorithmically, using the global false nearest neighbors approach. A measure of distance between a point $\mathbf{x}_n(d)$ in a phase space of dimension d and its nearest neighbor $\mathbf{x}_n^{NN}(d)$ is defined by

$$D_n(d)^2=\left\|\mathbf{x}_n(d)-\mathbf{x}_n^{NN}(d)\right\|^2=\sum_{i=0}^{d-1}\left[x_{n-i\tau}(d)-x_{n-i\tau}^{NN}(d)\right]^2, \tag{5}$$

The difference between the distances of the two points in dimension d and d+1 is then defined as

$$D_n(d+1)^2 - D_n(d)^2 = \sum_{i=0}^{d}\left[x_{n-i\tau}(d) - x_{n-i\tau}^{NN}(d)\right]^2 - \sum_{i=0}^{d-1}\left[x_{n-i\tau}(d) - x_{n-i\tau}^{NN}(d)\right]^2. \qquad (6)$$

This measure can be used to determine if a point and its nearest neighbor were proximal in dimension *d* because of projection rather than geometry by comparing this difference to a threshold. These points are then considered false neighbors in dimension *d*. If enough of the point-pairs move apart as the dimension in increased, it can be inferred that the attractor has not yet unfolded. Once the percentage of false nearest neighbors for a dimension drops below a threshold, that dimension is then considered a sufficient embedding dimension. Using this method, the embedding dimension is chosen to be five. Again, a more detailed analysis can be seen in (Johnson et al., in press).

Though the RPS of dimension five with lag six is considered to contain all the dynamics of the underlying speech production system, direct statistical modeling may not capture all the relevant information. An augmentation is made to the RPS by adding five more dimensions, each a linear regression on the first five dimensions. These dimensions, computed in the same manner as delta coefficients in traditional cepstral analysis, carry information about the trajectory of the RPS attractor. This augmentation has been shown to significantly increase the classification accuracy of the RPS method (Lindgren, 2003).

3.4 GMM modeling/classification

A GMM is a weighted sum of Gaussians, where the sum of the weights is required to equal unity (Duda et al., 2001). The equation for a GMM pdf is defined as

$$\hat{p}(\mathbf{x}) = \sum_{k=1}^{M} c_k \frac{1}{(2\pi)^{n/2} \,|\, \Sigma_k \,|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathrm{T}} \Sigma_k^{-1} (\mathbf{x}-\boldsymbol{\mu}_k)}, \qquad (7)$$

where **x** is the data vector, c_k is the weight of the *kth* mixture, **μ** is the mean vector of the *kth* mixture, and Σ_k is the covariance matrix of the *kth* mixture. In our methodology, the covariance matrices are diagonal. Given enough mixtures, a GMM can approximate any continuous pdf. Here, GMMs are used to directly model the natural measure of the RPSs. A model for each phoneme class is learned over all the points in the composite class RPSs.

Previous empirical study has shown that normalizing the phoneme RPSs prior to modeling can improve the classification accuracy (Ye et al., 2002). This operation is performed computing the standard deviation of the radius of the attractor by

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left|\mathbf{x}_i - \boldsymbol{\mu}\right|^2}, \tag{8}$$

where $\boldsymbol{\mu}$ is the centroid of the attractor, and then dividing each point in the RPS by σ.

The expectation-maximization (EM) algorithm is used to learn the GMM parameters for each class. Each GMM is initialized with a single Gaussian, and a sequence of binary-split operations is performed to build the models up to 128 mixtures. The number of mixtures is determined empirically, and the analysis is discussed in (Lindgren et al., 2003).

Bayesian classification is used for assigning log-likelihoods to each class for a given example phoneme. Using Bayes' theorem, the posterior probability of a class c_i given the data $\mathbf{x}$ is defined as

$$p(c_i \mid \mathbf{x}) = \frac{p(\mathbf{x} \mid c_i)p(c_i)}{p(\mathbf{x})}. \tag{9}$$

As the classification is determined by the c_i that maximizes (9), the $p(\mathbf{x})$ term is irrelevant. Also, the prior class likelihoods are treated as uniformly distributed, making this a Maximum Likelihood (ML) classifier. Each class log-likelihood is computed by summing the log-likelihoods for each point in the phoneme, given as

$$\hat{l}_i(\mathbf{X}) = \sum_{n=1}^{N} \ln \hat{p}_i(\mathbf{x}_n), \tag{10}$$

where the $\hat{p}_i(\mathbf{x}_n)$ values are computed from equation (7). For each individual sub-band, a classification can be obtained using these log-likelihoods by

$$\hat{c} = \underset{i=1\ldots C}{\arg\max} \left\{ \hat{l}_i(X) \right\}, \tag{11}$$

where C is the number of phoneme classes. The proposed system fuses the class log-likelihoods computed in equation (10) to reach a final decision for $\hat{c}$.

3.5 Fusion

Data fusion can take many forms, and can be used for a wide range of applications (see, for example (Kittler et al., 1998)). Fusion of information produced by multiple sources, e.g. sensors, classifiers, etc., can be used to reach a decision in a situation where any degree of uncertainty about observations prevails. This methodology can be beneficial when multiple observations provide non-redundant information. A priori knowledge of the reliability of the observers and the correlations among them may or may not be available.

Image processing and multi-sensor systems have seen extensive research in the application of data fusion methodology. Classifier fusion is common for the task of object identification, including facial recognition with implications to security (Kittler et al., 1998). Classifications based on different viewpoints, or with other forms of media such as speech, can be fused, often yielding better classification than any single classifier (Bourlard and Dupont, 1997; Misra et al., 2003). Fusion is found in multi-sensor systems such as toxic gas detection, where sensors react in different ways with chemicals that must be detected (Wei et al., 2002).

Data fusion has also been studied in the context of speech recognition. Hermansky et al. (Hermansky et al., 1996), for example, have examined fusion of sub-band features for recognition of noisy speech. We take a similar approach, though our experiments are based on sub-bands of clean speech.

Examples of data fusion frameworks include combination of hypotheses based on Bayesian methods or Dempster-Schafer (Schafer, 1976) theory. Bayesian approaches use probabilistic analysis to determine the probability of a hypothesis conditioned on a set of observations. Independence assumptions are often made to simplify analysis when full joint-probability distributions are unknown. Dempster-Schafer theory measures the probability that an observation supports a hypothesis. Given a lack of evidence, it makes no commitment to the probability of a hypothesis, and can be interpreted as defining a range of probabilities as opposed to one probability.

In this paper, a Bayesian framework is adopted. The classification is produced using Bayes' theorem as

$$\hat{c} = \underset{i=1\ldots C}{\arg\max}\; p(c_i \mid x_1 x_2 \ldots x_R) = \underset{i=1\ldots C}{\arg\max}\, \frac{p(x_1 x_2 \ldots x_R \mid c_i)\, p(c_i)}{p(x_1 x_2 \ldots x_R)}, \tag{12}$$

Again, the denominator term does not affect the maximization, and the class priors are assumed uniformly distributed, giving

$$\hat{c} = \underset{i=1\ldots C}{\arg\max}\; p(\mathbf{x}_1 \mathbf{x}_2 \ldots \mathbf{x}_R \mid c_i). \tag{13}$$

To simplify the analysis, the class log-likelihoods produced by each sub-band are assumed independent. Given Fletcher's findings, this appears to be a reasonable approximation. This assumption gives

$$\hat{c} = \arg\max \prod_{j=1}^{R} p(\mathbf{x}_j \mid c_i), \tag{14}$$

and since log-likelihoods are used, this becomes

$$\hat{c} = \arg\max_{i=1\ldots C} \sum_{j=1}^{R} \ln p(\mathbf{x}_j \mid c_i). \tag{15}$$

3.5.1 Optimized Weights

Though the assumption of independence may be a good approximation, equal weights may not be truly optimal. Therefore, a second weighting strategy is implemented, in which the weights are learned using the Nelder-Mead simplex (Nelder and Mead, 1965) method. This is implemented by minimizing the classification error over a development set that is described in section 4.1.

$$\hat{c} = \arg\max_{i=1\ldots C} \sum_{j=1}^{R} w_j \ln p(x_j \mid c_i), \tag{16}$$

where w_j is the weight of the *jth* sub-band.

3.6 Energy

Because energy is removed from the RPSs via radial normalization, fusing the sub-band RPS likelihoods with likelihoods based on energy features has the potential to improve the classification accuracy. Without the fusion methodology proposed here, the question of how to include information based on energy is a difficult one. The setup of our system, however, provides a convenient method with which to accomplish this.

Energy is added to the proposed system using a feature vector composed of mel-spaced filter channel log energies and full signal energy. The energy features are used to produce another set of class log-likelihoods, which are then fused along with the sub-band RPS log-likelihoods.

In (Lindgren et al., 2004), the RPS features were combined with MFCCs, improving classification accuracy over the MFCC-only baseline. It should be reasonable, then, to combine the sub-banded RPS method with the MFCC features as well. This is done fusing the log-likelihoods generated by the MFCC classifier along with the sub-band RPS log-likelihoods.

4. Experiments

4.1 Data set

The data set used for experimentation is the well-known speech database TIMIT (Garofolo et al., 1993). TIMIT is used because it has expertly-labeled phonetic boundaries on all utterances. This allows for extraction of the phonemes for use in isolated phoneme classification experiments. As is common in the literature, all SI and SX sentences are used, and the SA sentences are not used. Because some of the fusion experiments use parameters that are learned over data, the training set is randomly partitioned into a training set, which contains 90% of the training examples, and a development set containing the remaining 10%. The new training set contains 119 549 examples, and the development set has 13 283. The entire TIMIT test set, which contains 48 072 examples, is used for testing.

TIMIT has 64 distinct phoneme labels. For modeling, this set is reduced to 48 phonemes as defined in (Lee and Hon, 1989). Each test example is classified as one of these 48 phonemes, but errors among certain classes are not counted due to considerable linguistic similarities, (Lee and Hon, 1989), yielding a final set of 39 phoneme classes on which accuracy is determined.

4.2 Baselines

A set of baseline experiments is run for comparison to the proposed system. These baselines use traditional spectral-based features, as well as full-band RPS-based features. Twelve MFCCs plus log energy form the feature vector for the first baseline. Additionally, an experiment using the twelve MFCCs and log energy, plus deltas and delta-deltas is run. Both feature sets are modeled with GMMs of 16 mixtures. The accuracies from these experiments are 52.33% for the set of 13 features, and 56.94% for the full set including delta coefficients.

To examine the effects of sub-banding on classification accuracy of the RPS approach, a full-band RPS baseline accuracy is obtained. As mentioned previously, the dimension for this RPS is ten, with five base dimensions, and five delta dimensions. The lag is six, and 128 mixture GMMs are used. These parameters are the same for the sub-band RPS experiments. The resulting RPS full-band baseline is 38.81%.

4.3 Setup of 2,4,8 sub-bands

Three sets of sub-band experiments are presented in this paper. As stated in section 3, we wish to study the effect of the number of sub-bands on the performance of the system. Accordingly, experiments using 2, 4, and 8 sub-bands are performed. These are similar to traditional systems using sub-banding, but likely lower than the number of sub-bands in the human auditory system.

4.4 Individual band results

The phoneme accuracies for the individual sub-band RPSs in each sub-band set are shown in Tables 1, 2, and 3. The accuracies for each band degrade as the bandwidth is decreased, but not at such a rate that would prohibit the use of sub-banding for classification of phonemes.

4.5 Fusion

The log-likelihoods produced by the classifications of each sub-band are fused to give final classifications. As stated before, the first fusion scheme is an un-weighted sum of the class log-likelihoods. As shown in section 3, if all classifiers are assumed independent, this is the optimal linear combination method. Weights optimized over a development set are used for the second fusion strategy. Tables 4 and 5 show the results for these experiments. In each table, the first row gives the accuracies for the sub-band RPS fusions. In the second row, the accuracies for the sub-band RPS plus energy feature fusions are shown. It can be seen that, for both weighting schemes, the four-band set gives the highest accuracy if energy is not used, but an increase over the four-band set is seen with eight bands, solely due to the energy features. Previous studies have shown that the dimension and mixture size used in these experiments are near the asymptotes of the classification accuracy curves, indicating that the improvement seen is likely primarily a result of the sub-banding technique, rather than an artifact of the increase in number of parameters (Johnson et al., in press; Lindgren et al., 2003).

4.6 Fusion with MFCC features

In Table 6, classification accuracies produced by the fusion of sub-band RPS log-likelihoods and full, 39 coefficient MFCC feature set log-likelihoods. For each number of sub-bands, the accuracy outperforms the MFCC-only baseline of 56.94%, with the greatest increase seen with four RPS sub-bands.

5. Discussion/Future work

5.1 Examination of individual band results

Given the results of convolution on RPSs discussed in section 2.2, it is important to observe that sub-banded RPSs do provide discriminatory information about phonemes. It is not surprising that the classification accuracy of each band decreases as the bandwidth decreases. Not only does the amount of linear information become less at smaller bandwidths, but the amount of nonlinear information, which is spread across frequencies, also degrades. This important trade-off must be considered with regard to the design parameters of the system.

It can be seen from Tables 1, 2, and 3 that the lower frequency sub-bands generally outperform the higher frequency bands. This could be because there is simply less information in the higher bands, but this would contradict Fletcher's experimental results on human speech recognition. It could also be due to the nature of the RPS methodology. RPSs are based on theory pertaining to deterministic systems, which often have smooth attractors. Signals that have been high-pass filtered, however, tend to have RPSs that have a much less smooth structure.

5.2 Discussion of Fusion results

Clearly, sub-banding and likelihood fusion can improve the classification ability of RPSs. Because the dimension and number of Gaussian mixtures chosen give near-maximum classification accuracy for the full band RPS approach, this improvement can be attributed to the new sub-banding technique, rather than the increase in model size (Johnson et al., in press; Lindgren et al., 2003). A greater than 4% absolute increase is seen when two sub-bands are used. Further gains are made with four sub-bands. It appears that, at least for the configuration presented here, eight sub-bands may be near the upper limit for classification. However, the accuracy of the eight-band RPS plus energy experiment exceeds that of the four-band RPS, this is due entirely to the energy features. It seems likely that, as the number of bands continues to increase over eight, the amount of nonlinear information captured by each band will be minimal, and the fusion will suffer.

The most appropriate number of bands for the sub-band RPS approach appears to be below the suggested number of bands in the human ear. This could be due to several factors. The filter bank used here is implemented with non-overlapping approximate brick wall filters, in contrast to the overlapping gammatone shaped filters of the human auditory system. In addition, it is unknown exactly how the partial phoneme recognition and fusion mechanisms work in human hearing. They are likely quite different from the methodology presented here.

With energy added, the sub-band RPS method classification accuracy outperforms the 13 coefficient MFCC plus energy feature set. Unfortunately, there is no easy way to incorporate the long-term trajectory features that are part of the 39 coefficient MFCC feature set into the proposed system given the current modeling approach. Along with the time complexity, this is an issue that must be dealt with for this approach to be feasible for continuous speech recognition.

5.3 Further possibility for improvement

Because the RPS parameters are chosen heuristically, further investigation into the appropriate lag and dimension may yield additional improvements in accuracy. A discussion of this can be found in (Johnson et al., in press). Analysis of phoneme classification accuracy as a function of RPS dimension and lag has shown that, given a sufficiently high embedding dimension, a lag of one consistently provides for better classification than other lags. Incorporation of this knowledge could possibly result in further improvements to the sub-band RPS system.

As the MFCC experiments are based on feature vectors extracted from the signals, clustering of these features is possible. Pronunciation patterns can vary because of speaker variation, different dialects, etc. Because of this, speech features may not cluster in a single region of the feature space, but many regions. Consequently, the use of GMMs for modeling of speech features improves the performance of recognition systems significantly over the use of single Gaussians.

Because of the nature of our RPS modeling scheme, we are unable to implement a similar clustering mechanism. Though GMMs are implemented for modeling, they represent the distribution of a collection of RPS points, which in themselves do not provide much discrimination power. Only by computing likelihoods over a set of points is discrimination feasible. This methodology does not allow for clustering, as there is no mechanism by which to identify if two proximal RPS points should be placed together or in separate speech pattern clusters. One possible solution is to compute features over the RPS of a frame of speech, allowing for analysis similar to MFCC-based methods.

Though all experiments presented in this paper are based on clean speech, the proposed methodology has the potential to be beneficial for recognition of speech corrupted by narrowband noise. Sub-banding has already been shown to improve the robustness of traditional ASR systems in certain types of noise, and one would expect to see the same benefits for the sub-band RPS approach.

6. Conclusion

This paper has shown the advantage of sub-banding speech signals for analysis with reconstructed phase spaces. A full phoneme classification system using this sub-banding method is introduced, and experiments studying the effects of filter bank size and fusion methods have been presented. The RPS modeling approach has important theoretical advantages over spectral-based approaches in that nonlinear or higher-order statistical characteristics present in speech signals can be captured. The filter bank front-end clearly improves the classification ability over the standard full-band RPS approach.

The sub-band RPS approach is competitive with the standard Mel-frequency cepstral coefficient features for classification of isolated phonemes. Though the approach adopted in this paper still produces classification accuracies below that of the full MFCC feature set with log energy and regression coefficients, it does outperform the MFCC feature set using log energy but no deltas. Combination of the sub-band RPS features and MFCCs shows a 2.38% absolute improvement over the MFCC-only features with fusion of log-likelihoods using optimized weights.

The results presented show that further research into sub-banded RPSs in alternative front-end parameterization methodologies for speech recognition is warranted. This approach has the potential to benefit real ASR systems, including those that must perform recognition on noisy speech. Computational complexity issues that prevent the current modeling approach from being adopted for continuous speech recognition must be addressed, and efficient computation of features from framed RPSs is the emphasis of current and future research.

Acknowledgement

This material is based on work supported by the National Science Foundation under Grant No. IIS-0113508 and the Department of Education GAANN Fellowship.

References

Abarbanel, H. D. I., 1996, Analysis of Observed Chaotic Data, Springer, New York.

Allen, J. B., 1994, How Do Humans Process and Recognize Speech?, IEEE Transactions on Speech and Audio Processing, 2, 567-577.

Badii, R., Broggi, G., Derighetti, B., and Ravini, M., 1988, Dimension Increase in Filtered Chaotic Signals, Physical Review Letters, 60, 979-982.

Banbrook, M., and McLaughlin, S., 1994, "Is Speech Chaotic?," *IEE Colloquium on Exploiting Chaos in Signal Processing*, 8/1-8/8.

Banbrook, M., McLaughlin, S., and Mann, I., 1999, Speech Characterization and Synthesis by Nonlinear Methods, IEEE Transactions on Speech and Audio Processing, 7, 1 -17.

Bourlard, H., and Dupont, S., 1996, "A New ASR Approach Based on Independent Processing and Recombination of Partial Frequency Bands," *International Conference on Spoken Language Processing (ICSLP)*, 426-429.

Bourlard, H., and Dupont, S., 1997, "Subband-Based Speech Recognition," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 21-24.

Chennaoui, A., Pawelzik, K., Liebert, W., Schuster, H. G., and Pfister, G., 1990, Attractor Reconstruction from Filtered Chaotic Signals, Physical Review A, 41, 4151-4159.

Dimitriadis, D., Maragos, P., and Potamianos, A., 2002, "Modulation Features for Speech Recognition," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, I-377-I-380.

Duda, R. O., Hart, P. E., and Stork, D. G., 2001, Pattern Classification, John Wiley & Sons, New York, New York.

Fletcher, H., 1953, Speech and Hearing in Communication, Van Nostrand, New York,.

Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., Dahlgren, N., and Zue, V., 1993, Timit Acoustic-Phonetic Continuous Speech Corpus, Linguistic Data Consortium.

Gibson, J. F., Farmer, J. D., Casdagli, M., and Eubank, S., 1992, An Analytic Approach to Practical State Space Reconstruction, Physica D, 57, 1-30.

Gold, B., and Morgan, N., 2000, Speech and Audio Signal Processing, John Wiley and Sons, New York, New York.

Hagen, A., Bourlard, H., and Morris, A., 2001, "Adaptive Ml-Weighting in Multi-Band Recombination of Gaussian Mixture ASR," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 257-260.

Hermansky, H., Tibrewala, S., and Pavel, M., 1996, "Towards ASR on Partially Corrupted Speech," *Fourth International Conference on Spoken Language (ICSLP)*, 462-465 vol.461.

Indrebo, K. M., Povinelli, R. J., and Johnson, M. T., 2003, "A Combined Sub-Band and Reconstructed Phase Space Approach to Phoneme Classification," *ISCA Tutorial and Research Workshop on Non-linear Speech Processing (NOLISP)*, Le Croisic, France, 107-110.

Isabelle, S. H., Oppenheim, A. V., and Wornell, G. W., 1992, "Effects of Convolution on Chaotic Signals," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 133-136.

Johnson, M. T., Povinelli, R. J., Lindgren, A. C., Ye, J., Liu, X., and Indrebo, K. M., in press, Time-Domain Isolated Phoneme Classification Using Reconstructed Phase Spaces, IEEE Transactions on Speech and Audio Processing.

Kittler, J., Hatef, M., Duin, R. P. W., and Matas, J., 1998, On Combining Classifiers, IEEE Transactions on Pattern Analysis and Machine Intelligence, 20, 226-239.

Kubin, G., 1995, Nonlinear Speech Processing, Speech Coding and Synthesis, W. B. Kleijn and K. K. Paliwal, eds., Elsevier Science.

Lee, K.-F., and Hon, H.-W., 1989, Speaker-Independent Phone Recognition Using Hidden Markov Models, IEEE Transactions on Acoustics, Speech and Signal Processing, 37, 1641-1648.

Lindgren, A. C., 2003, Speech Recognition Using Features Extracted from Phase Space Reconstructions, M.S. Thesis, Marquette University, Milwaukee, Wisconsin.

Lindgren, A. C., Johnson, M. T., and Povinelli, R. J., 2003, "Speech Recognition Using Reconstructed Phase Space Features," *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Hong Kong, 61-63.

Lindgren, A. C., Johnson, M. T., and Povinelli, R. J., 2004, "Joint Frequency Domain and Reconstructed Phase Space Features for Speech Recognition," *International Conference on Acoustics, Speech and Signal Processing*, Montreal, Canada, I-533-I-536.

McCourt, P., Vaseght, S., and Harte, N., 1998, "Multi-Resolution Cepstral Features for Phoneme Recognition across Speech Sub-Bands," *Acoustics, Speech, and Signal Processing, 1998. ICASSP '98. Proceedings of the 1998 IEEE International Conference on*, 557-560 vol.551.

Misra, H., Bourlard, H., and Tyagi, V., 2003, "New Entropy Based Combination Rules in Hmm/Ann Multi-Stream ASR," *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 741-744.

Moreno, A., and Rutllan, M., 1996, "Integrate Polispectrum on Speech Recognition," *International Conference on Spoken Language Processing*, Philadelphia, 1281-1284.

Nelder, J. A., and Mead, R., 1965, A Simplex Method for Function Minimization, Computer Journal, 7, 308-313.

Pitsikalis, V., and Maragos, P., 2002, "Speech Analysis and Feature Extraction Using Chaotic Models," *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, I-533-I-536 vol.531.

Povinelli, R. J., Bangura, J. F., Demerdash, N. A. O., and Brown, R. H., 2002, Diagnostics of Bar and End-Ring Connector Breakage Faults in Polyphase Induction Motors through a Novel Dual Track of Time-Series Data Mining and Time-Stepping Coupled Fe-State Space Modeling, IEEE Transactions on Energy Conversion, 17, 39-46.

Povinelli, R. J., Johnson, M. T., Lindgren, A. C., and Ye, J., 2004, Time Series Classification Using Gaussian Mixture Models of Reconstructed Phase Spaces, IEEE Transactions on Knowledge and Data Engineering, 16, 779-783.

Roberts, F. M., Povinelli, R. J., and Ropella, K. M., 2001, "Identification of ECG Arrhythmias Using Phase Space Reconstruction," *Principles and Practice of Knowledge Discovery in Databases (PKDD'01)*, Freiburg, Germany, 411-423.

Sauer, T., Yorke, J. A., and Casdagli, M., 1991, Embedology, Journal of Statistical Physics, 65, 579-616.

Schafer, G., 1976, A Mathematical Theory of Evidence, Princeton University Press.

Takens, F., 1980, "Detecting Strange Attractors in Turbulence," *Dynamical Systems and Turbulence*, Warwick, 366-381.

Teager, H. M., and Teager, S. M., 1990, "Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract," *NATO ASI on Speech Production and Speech Modelling*, 241-261.

Tibrewala, S., and Hermansky, H., 1997, "Sub-Band Based Recognition of Noisy Speech," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1255-1258 vol.1252.

Wei, G., Chan, P. C. H., Tang, Z., Yu, J., and Wang, L., 2002, "Gas Mixture Analysis with Dynamic Gas Sensor Array Signals," *Second International Symposium on Instrumentation Science and Technology*, 637-642.

Ye, J., Johnson, M. T., and Povinelli, R. J., 2003, "Phoneme Classification over Reconstructed Phase Space Using Principal Component Analysis," *ISCA Tutorial and Research Workshop on Non-linear Speech Processing (NOLISP)*, Le Croisic, France, 11-16.

Ye, J., Povinelli, R. J., and Johnson, M. T., 2002, "Phoneme Classification Using Naive Bayes Classifier in Reconstructed Phase Space," *IEEE Signal Processing Society 10th Digital Signal Processing Workshop*, 2.2.

Vitae

Kevin Indrebo

Kevin Indrebo received a B.S. in computer engineering and a M.S. in electrical and computer engineering in 2002 and 2004, respectively, from Marquette University, Milwaukee, Wisconsin.

He has been with the Knowledge and Information Discovery Lab in the Electrical and Computer Engineering department at Marquette University since May 2002. His primary research interests include speech recognition and nonlinear signal processing.

From 2001 to 2002, he served as the vice president of the Marquette chapter of the Association for Computing Machinery (ACM), and is currently a member of the ACM, and the International Speech Communication Association.

Richard Povinelli

Richard J. Povinelli received the B.S. degree in electrical engineering and the B.A. degree in psychology from the University of Illinois, Champaign-Urbana, in 1987, the M.S. degree in computer and systems engineering from Rensselaer Polytechnic Institute, Troy, NY, in 1989, and the Ph.D. degree in electrical and computer engineering from Marquette University, Milwaukee, WI, in 1999.

From 1987 to 1990, he was a Software Engineer with General Electric Corporate Research and Development. From 1990 to 1994, he was with GE Medical Systems, where he served as a Program Manager and then as a Global Project Leader. From 1995 to 1998, he held the positions of Lecturer and Adjunct Assistant Professor with the Department of Electrical and Computer Engineering, Marquette University, where, since 1999, he has been an Assistant Professor. His research interests include data mining of time series, chaos and dynamical systems, computational intelligence, and financial engineering.

Dr. Povinelli is a senior member of the Institute of Electrical and Electronics Engineers, Association for Computing Machinery, American Society of Engineering Education, Tau Beta Pi, Phi Beta Kappa, Sigma Xi, Eta Kappa Nu, Upsilon Pi Epsilon, and Golden Key. He was voted the Young Engineering of the Year for 2003 by the Engineers and Scientists of Milwaukee, Inc.

Mike Johnson

Michael T. Johnson received a B.S. degree in Computer Science and a B.S. degree in Engineering, Electrical Concentration from LeTourneau University in 1989 and 1990, respectively. He received an M.S.E.E. degree in Electrical Engineering from the

University of Texas at San Antonio in 1994 and a Ph.D. degree from Purdue University in 2000.

Dr. Johnson worked as a design engineer from 1990-1991 for Micronyx, Inc. and Microtechnology Services, Inc., in Dallas, TX, and from 1991-1993 for Datapoint Corp. in San Antonio TX. He was Senior Engineer and Engineering manager for SNC Manufacturing in Oshkosh, WI from 1993-1996. Since 2000 he has been an Assistant Professor of Electrical and Computer Engineering at Marquette University in Milwaukee, WI. His primary research area is speech and signal processing, with an emphasis in speech recognition systems. Other areas of research include machine learning, statistical pattern recognition, and non-linear signal processing.

Dr. Johnson is a registered as a professional engineer in the state of Wisconsin. He is a senior member of the Institute of Electrical and Electronics Engineers, and is also a member of the Association for Computing Machinery, the Acoustical Society of America, and the International Speech Communications Association. Honor society memberships include Sigma Xi, Eta Kappa Nu, and Upsilon Pi Epsilon.

Tables

< 1800 Hz	> 1800 Hz
34.57%	23.25%

Table 1. Individual sub-band classification accuracies for 2 band set.

< 640 Hz	640 – 1800 Hz	1800 – 3965 Hz	> 3965 Hz
25.22%	24.47%	20.75%	14.77%

Table 2. Individual sub-band classification accuracies for 4 band set.

<285 Hz	285-640 Hz	640-1130 Hz	1130-1800 Hz	1800-2715 Hz	2715-3965 Hz	3965-5670 Hz	>5670 Hz
17.76%	20.42%	16.93%	19.11%	15.64%	19.97%	15.08%	14.15%

Table 3. Individual sub-band classification accuracies for 8 band set.

# Sub-bands	2	4	8
Accuracy w/o energy	42.91%	44.21%	43.99%
Accuracy w/ energy	50.14%	51.96%	54.84%

Table 4. Classification accuracies of equal-weight fusion of sub-band RPS.

# Sub-bands	2	4	8
Accuracy w/o energy	43.05%	44.51%	44.28%
Accuracy w/ energy	50.65%	52.18%	54.95%

Table 5. Classification accuracies of optimized-weight fusion of sub-band RPS.

# Sub-bands	2	4	8
Equal Weights	58.85%	59.19%	58.99%
Optimized	58.95%	59.22%	59.32%

Table 6. Classification accuracies of fusion of sub-band RPS & MFCC features.

Figures

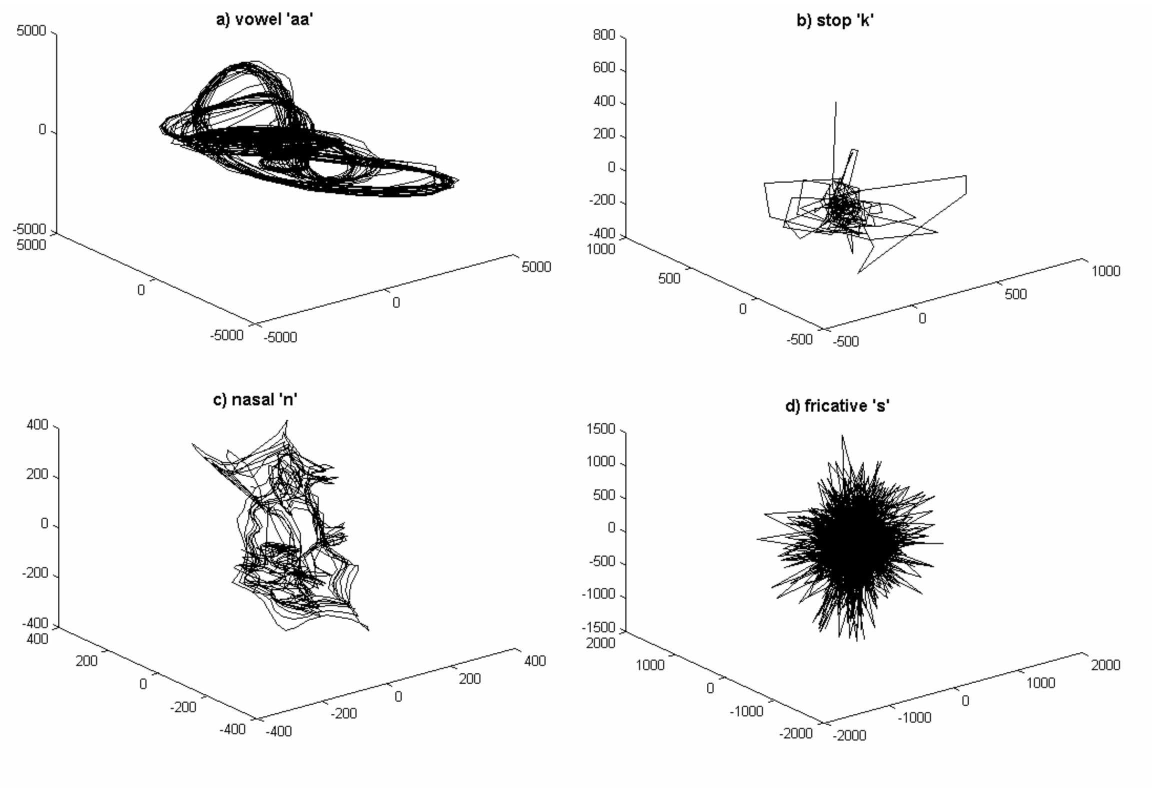

Figure 1. Four examples of phonemes in 3-dimensional reconstructed phase space.

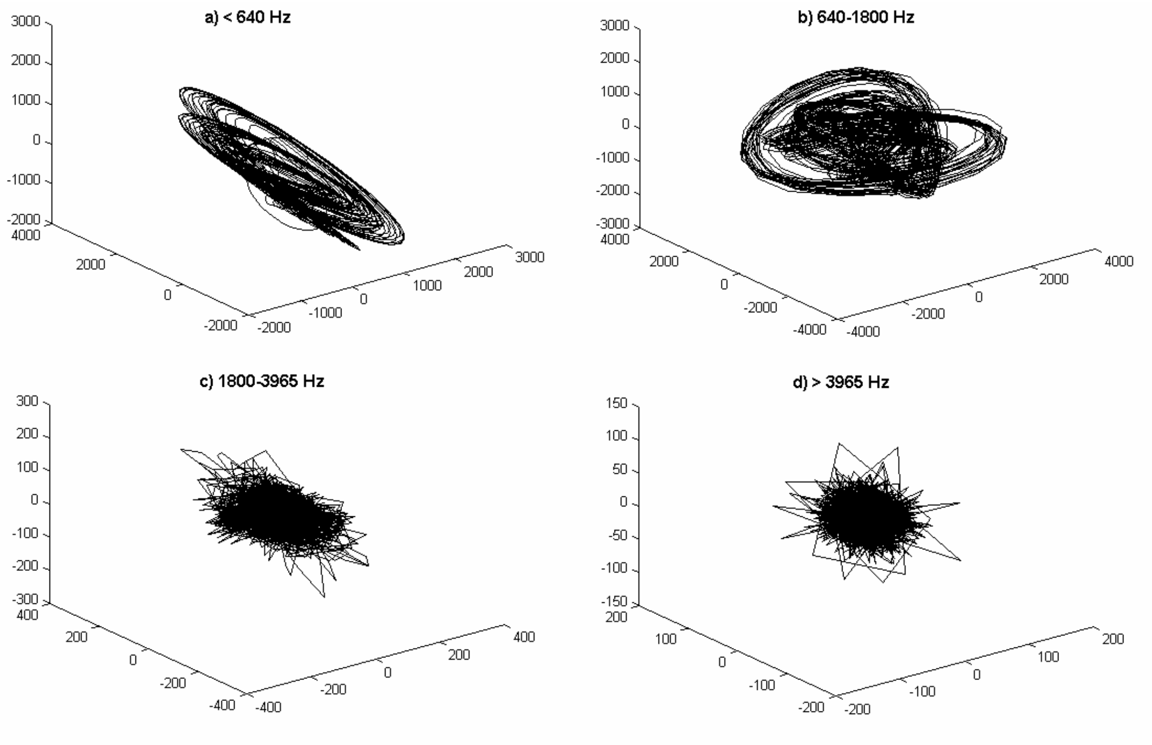

Figure 2. An RPS of the phoneme 'aa' in four sub-bands.

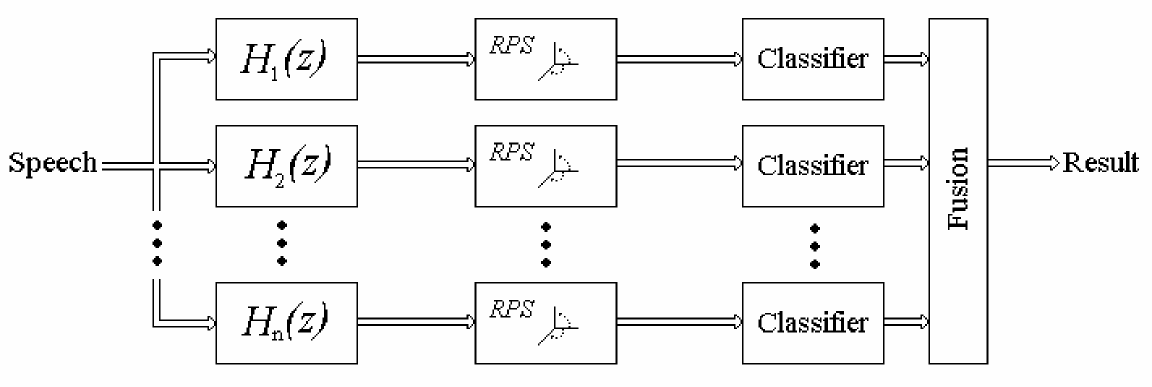

Figure 3. System diagram of the proposed phoneme classification system.