Sub-banded Reconstructed Phase Spaces for Speech Recognition

By

Kevin Michael Indrebo

A THESIS

SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL, MARQUETTE UNIVERSITY,

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree of

MASTER OF SCIENCE

in Electrical and Computer Engineering

Milwaukee, Wisconsin

May 2004

Preface

A novel method for classification of speech phonemes, based on the combination of dynamical systems theory and filter banks, is introduced. The benefit of this approach is seen in its ability to model nonlinear characteristics of speech, something that traditional methods cannot do. The modeling tool that provides this capability is the reconstructed phase space. This space carries all the dynamical information present in the signal's underlying system. The reconstructed phase spaces used for modeling and classification of the phonemes are built using frequency sub-banded signals that are generated using a set of band-pass filters. This approach is motivated by empirical evidence that suggests humans process and recognize speech in sub-bands. Modeling and classification is performed on the sub-banded reconstructed phase spaces using Gaussian Mixture Models, and the results of the classifications for each sub-band are combined to form an overall classification. Several methods for the combination of the sub-band classifications are examined, and it is found that an un-weighted linear combination produces classification accuracies that are significantly higher than those of a classification system using reconstructed phase spaces of unfiltered signals. Results also demonstrate that the proposed phoneme classification system is competitive with state-of-the-art approaches.

Acknowledgements

I would first like to thank my entire family, especially my parents, for all their support throughout my time as a graduate student. Also deserving of thanks are my committee members: Drs. Richard J. Povinelli, Michael T. Johnson, and George F. Corliss, for all their advice. My advisor, Dr. Richard Povinelli, especially receives many thanks for his guidance regarding many aspects of research and study. I would also like to thank all the members of the Knowledge and Information Discovery and Speech and Signal Processing Labs for all the insightful discussions we have had.

Table of Contents

Table of Figures

Table of Tables

1. Introduction

1.1. Automatic Speech Recognition

1.1.1 Overview

Automatic speech recognition (ASR) is the process of converting human speech signals into text. The speech begins as an acoustic wave and is transformed into an electrical signal by a microphone. The electrical signal is then sampled with an analog-to-digital converter. Signal processing methods, such as linear prediction and cepstral analysis, are used to extract features from the digitized signal [1]. The final step in this process involves pattern recognition of the parameterized signals to produce a hypothesis of the utterance.

Speech recognition has many potential applications, in a variety of settings. Some examples include dictation, hands-free computer control, and machine translation. Each of these relies on accurate translation of speech into text, which may be processed further to extract semantic information.

Dictation systems convert speech into text as part of a letter, report, or some other document. This task is of special significance to members of the medical and legal professions, as they often depend on trained human transcriptionists, who must be familiar with the vast lexicons of the profession.

Hands-free computer control systems allow users to execute basic operations on computers or embedded systems without the use of a mouse, keyboard, or keypad. Booking of flights using telephone airline reservation systems and buy/sell requests for stock trading over the phone are common tasks. Also, hands-free control is important for the physically disabled who cannot use traditional computer terminal interface devices.

Another task that requires effective speech recognition is machine translation, which is the task of automatically translating spoken words from one language into another. The significance of this application is apparent given the increasing globalization of the economy, though a practical machine translation system is years away from realization. This task depends not only on accurate speech-to-text conversion, but also inter-language conversion.

Despite the many benefits that could be gained from implementing speech recognition interfaces in computers and embedded systems, ASR has not yet seen widespread acceptance in the commercial world. The reason for this is that ASR systems are simply not accurate enough to be effective tools for the average user. Several factors make speech recognition a difficult problem. Recognition rates of continuous, natural speech with multiple speakers, large vocabularies, and noisy environments are low, often below 75% word accuracy [2], especially when the conditions present at the time of use do not match the conditions used to train the system, which is often the case.

1.1.2 Traditional Systems

Traditional ASR systems use front-end acoustic parameterization methods that are based on a switched-excitation linear source-filter model of human speech production [1]. This model, which is depicted in Figure 1-1, treats the glottis, the excitation source, as a pulse train for voiced speech and white noise for unvoiced speech. Voiced speech includes vowels and some consonants. Unvoiced speech consists primarily of fricatives such as 'f' and 's'. The signal produced by the glottis enters the vocal tract, which is modeled as a linear time-invariant (LTI) filter. The vocal tract shapes the spectrum of the signal, giving it a distinct sound.

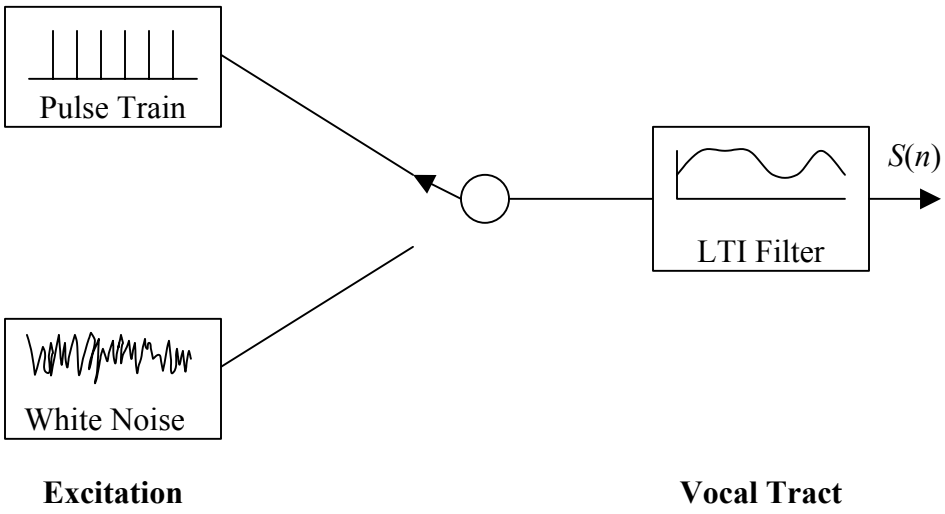

Figure 1-1. Linear model of speech production.

In this model of speech, the excitation is separated from the vocal tract. For English phonemes, the excitation is traditionally seen as not providing useful features for recognition. Rather, it is believed that the relevant information comes from the vocal tract. Cepstral coefficients, which are described in more detail in chapter 2, and specifically Mel-frequency cepstral coefficients (MFCCs), are a standard set of features that are produced by an ASR system's front-end parameterization algorithm [1]. Cepstral analysis is particularly attractive because it allows for separation of the excitation and vocal tract characteristics. These coefficients correspond directly to the general spectral envelope of a segment of speech. Because cepstral coefficients are computed using linear signal processing techniques, they are unable to capture nonlinear or higher-order statistics of the speech signals [3]. Nonlinear processes do not obey the principle of superposition, and cannot be fully represented by power spectra. Traditionally, it has been believed that speech waveforms contain little relevant nonlinear information. It has recently been shown, however, that this is not true, and that significant nonlinear characteristics do exist in human speech signals [4-6].

1.2. Nonlinear Analysis of Speech Signals

Recently, experimental evidence has begun to reveal significant nonlinear characteristics present in speech signals. This work has investigated whether speech signals can be modeled as nonlinear or even chaotic processes. Chaotic systems are very sensitive to initial conditions, and often appear to be random processes, even though they are deterministic. Because of the highly sensitive nature of some chaotic measures to noise [7], the presence of nonlinear and chaotic characteristics in speech is not easily confirmed. Different studies have produced varying conclusions about the question of chaotic components in speech, but the postulate that nonlinear components are present appears more solid [5]. This phenomenon is the basis of the work presented here.

Following the discovery of the presence of nonlinear characteristics, several attempts have been made at incorporating nonlinear-based features into recognizers. Lyapunov exponents[8-10], fractal and other dimensions [11], and polynomial prediction coefficients [12] are among the feature types that have been examined. Though recognition systems based solely on nonlinear features often show substandard performance [13], those that use nonlinear features in combination with the traditional features, such as MFCCs, have shown improvements over systems based exclusively on spectral features [11, 13].

Phase space reconstruction is one of the analysis tools that has been used for nonlinear speech signal analysis [14]. Phase space reconstruction is based in dynamical systems theory, and can be used as a tool for estimating the dynamical invariants of a system, such as Lyapunov exponents and dimension [15]. However, in this thesis, a more direct approach is taken by modeling the natural measure of the structure resulting from phase space reconstruction.

1.3. Contributions of This Thesis

1.3.1 Problem Statement

The goal of this work is to develop an acoustical analysis method that is not bound by the limits of linear techniques and improves the phoneme recognition capabilities of ASR systems. The nonlinear analysis techniques used here are based on reconstructed phase spaces (RPS's), which were originally developed in the field of dynamical systems.

1.3.2 Dynamical Systems

Dynamical systems are often represented with a state structure. This state structure, or state space, determines how the state variables of the system change through time. Mathematically, a dynamical system is represented by a set of differential or difference equations, expressed in terms of a finite set of state variables. The number of state variables determines the dimension of the system. These systems are typically treated as deterministic, i.e. if the state of the system at time t_0 is known, the state of the system at any time t_1 is completely predictable.

Unfortunately, the entire state space of almost all real systems cannot be observed. Often, only one state variable is available. It would seem that accurate characterization of the system is impossible in this case, especially if the dimensionality and nonlinearity of the system are high. However, with the use of a transformation on the observable variable known as a time-delay embedding [16], more information about the system is available than one might expect. This transformation is defined as

$$\Phi_{y,\varphi}(x) = (y(x), y(\varphi(x)), ..., y(\varphi^{2d}(x)),$$

where the vector x is the current system state, φ is the state update function, and y is the function that transforms a state x into an output variable. A structure that is topologically equivalent [16, 17], to the original system state space is created, given certain assumptions [16, 18]. This structure is called a reconstructed phase space (RPS). An example RPS of a vowel is shown in Figure 1-2. This RPS is of dimension two, with a time-lag of six. The embedding dimension and the time-lag of an RPS determine its shape, and affect its representation capability. These parameters will be discussed in further detail in chapter 3.

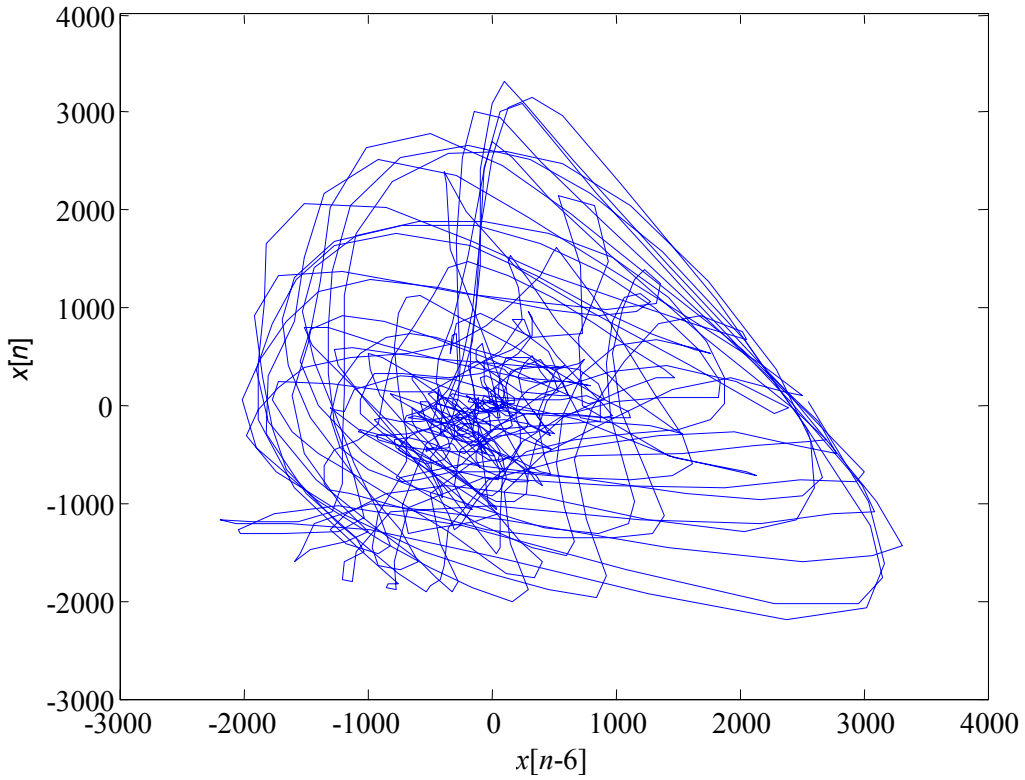

Figure 1-2. Example of a 2-dimensional, lag 6 RPS of the phoneme 'ah'.

As stated previously, this RPS can be used to estimate the dynamical invariants [7] of the system. This work uses the shape or density of the RPS as a basis for modeling and classification of speech phonemes. Only recently has this approach been taken in the realm of speech recognition [19, 20].

This thesis extends previous work in this area by combining this RPS-based method with filter banks. A set of front-end band-pass filters are applied to the speech signals to be modeled before phase space reconstruction. The RPS's in each of these sub-bands are modeled and classified using statistical methods, and the classifications of each sub-band are fused to reach a final decision. The specifics of this process, and the advantages of this approach, are the subject of this thesis.

The remainder of this thesis is as follows. Chapter 2 provides a more thorough background on automatic speech recognition and dynamical systems. In chapter 3, the methods applied here for the classification of speech sounds are presented. Chapter 4 presents and analyzes the results of experiments used to test the proposed methods. Finally, chapter 5 provides a conclusion and discusses possible future work.

2. Background

2.1. History of ASR

Automatic speech recognition is a task that has been studied for many decades. In 1952, a system was built at Bell Labs to recognize spoken digits [1]. It used the long-term spectral moment in two frequency bands as acoustic features, and was built using analog electronic components. Given the level of technology at the time, and the simplicity of the system, it achieved a surprisingly high accuracy rate of 98% for digits uttered by a single speaker separated by pauses.

Following this, speech recognition systems were advanced by improvements in acoustic parameterization and added language models [1]. Short-term spectral estimates became popular acoustic features, computed using filter banks at first, then using signal processing techniques like the Fast Fourier Transform (FFT). In the 1960's, homomorphic or cepstral analysis was developed [3]. Since then, cepstral-based feature sets have become the standard acoustic analysis technique for speech recognition.

Methods for pattern recognition were also developed and adopted for speech recognition, namely dynamic time warping (DTW) and hidden Markov Models (HMM) [1]. The HMM, the current standard used for pattern recognition in automatic speech recognition (ASR) systems, is a statistical analysis tool that works well for continuous speech recognition, where the durations of each word are not known in advance, and word boundaries are ambiguous.

Only recently has ASR entered the commercial world. The first dictation system available to consumers was delivered in 1990 by Dragon Systems [21]. Unlike today's dictation software, it could not handle continuous speech. Instead, users had to pause in

between each word. A few months later, IBM released a competing system, also limited to isolated word recognition [21]. In 1997, Dragon Systems released "Naturally Speaking", the first continuous speech dictation system [22].

In 1996, Charles Schwab and Nuance partnered to develop the first major Interactive Voice Response (IVR) system for the use of investors who wished to process transactions over the telephone [21]. This type of system requires a fixed dialogue, with a limited grammar, which is dictated by the options a user has for control of the dialogue.

Despite the strides made over the past decades, current ASR systems are still inadequate for many desired tasks. Telephone ASR systems are often limited by the number of responses that the application can recognize, and dictation systems require long training times for a single user.

2.2. Why ASR is difficult

Many factors contribute in making automatic speech recognition difficult. Among these are large vocabularies, additive and convolutional noise, coarticulation, and varying speaker characteristics. Research in speech recognition has become extensive enough that much of the current research focuses on solving one of these problems in isolation. Simple tasks such as speaker-independent isolated digit recognition in low noise environments can be handled easily by unsophisticated ASR systems. When the task becomes more difficult, however, even today's state-of-the-art systems do not perform well.

One major issue for ASR is the continuous nature of human speech. Humans do not pause between words when speaking. Nor do they always speak at the same rate. Because of this, recognition systems must be able to determine word boundaries and phoneme durations given nothing but the acoustic waveform of an utterance.

Additionally, a phenomenon known as coarticulation accounts for additional complexity. Coarticulation is caused by the constant motion of vocal tract articulators during speech. The vocal tract does not produce phonemes with a series of stationary configurations, but instead varies its shape smoothly through time. Because of this, the actual sound of a phoneme is dependent on the preceding and following phonemes.

Another complicating factor is that different speakers have varying voice characteristics. The vocal tract length, glottal structure, mouth configuration, and learned speaking patterns all affect the way a person's speech sounds. Since it is desirable for recognition systems to be able to recognize speech from users that have not been previously encountered by the system, a recognition system must be able to perform recognition using features that are not dependent on the speaker.

Large vocabularies are another issue that recognition systems must handle. There are a set number of basic sounds, or phonemes, that are always part of any language. These sounds are fundamental to a language, and are relatively limited in number, around 40 in English [23]. The words in a language are built from these phonemes, but they are not so limited. The English language, for example, contains more than 450,000 words [24]. For most applications, the vocabulary is limited to far fewer than all 450,000, but even 20,000 words, a reasonably modest number, can make for a challenging vocabulary. As the number of words increases, so do the number of words that are very nearly the same acoustically, resulting in an increase of the word error rate.

This thesis addresses the task of speaker independent continuous speech recognition (CSR). This task is appropriate for many applications of speech recognition,

since users often wish to be able to speak in a conversational manner, without strict limits on the vocabulary, or having to spend large amounts of time training a system.

2.3. Traditional Approaches

2.3.1 Speech modeling

Speech recognition can be decomposed into two distinct but not disjoint domains: acoustic modeling and language modeling. Acoustic modeling is the process of analyzing speech signals to generate hypotheses about the sequence of phonemes in an utterance. Language modeling is used to refine these hypotheses using knowledge of the spoken language. This thesis focuses on the acoustic portion of speech recognition.

Traditional acoustic feature extraction approaches are based on a linear source-filter model of speech production, which is illustrated in Figure 1-1. In this domain, speech is modeled as a signal produced by a pulse-train excitation or white noise excitation, which are filtered by a linear time-invariant (LTI) filter. The excitation corresponds to the glottis. During voiced speech, a pulse-train is produced by the glottis, while white noise is generated during unvoiced speech. The resulting excitation signal is then filtered using a LTI filter, which represents the vocal tract.

2.3.2 Cepstral Coefficients

Because speech is nonstationary, meaning the statistics of speech signals change throughout time, it is necessary to segment a speech signal into frames before features can be extracted. These frames are made of a length, usually around 30 milliseconds, such that speech can be considered approximately stationary. Before any spectral-based analysis is performed on these frames, they are typically windowed with a hamming or hanning window [3]. This windowing process smoothes out the spectrum, which would

have high frequency characteristics that would be artificially high, due to the framing process [3].

For English speech, it is a reasonable approximation to consider the excitation as irrelevant to the sounds being produced. Consequently, features that describe the vocal tract filter are used for recognition. Cepstral analysis provides a mechanism for separating the excitation source from the vocal tract response in the acoustic waveform.

The observed speech signal *s(t)* can be written as the convolution of the excitation signal *e(t)* and the vocal tract impulse response, *v(t)*

$$s(t) = e(t) * v(t). \tag{2.1}$$

Blind deconvolution [25] of these two signals is extremely difficult. However, cepstral analysis is able to take advantage of the different spectral characteristics of the excitation and vocal tract. In the frequency domain, equation (2.1) becomes

$$|S(\omega)| = |E(\omega)||V(\omega)|, \tag{2.2}$$

By taking the logarithm of this equation, the convolution is transformed into addition

$$\log|S(\omega)| = \log|E(\omega)| + \log|V(\omega)| \tag{2.3}$$

To obtain the cepstrum, the Inverse Fast Fourier Transform (IFFT) is taken, transforming the log spectrum into the quefrency domain.

$$C(m) = \mathrm{Re}\{IFFT(\log|E(\omega)|) + IFFT(\log|V(\omega)|)\}. \tag{2.4}$$

The first M points are then used as the features for acoustic modeling, as they relate more strongly to the vocal tract, which has more smoothly varying spectral characteristics, resulting in lower quefrency characteristics. The higher-indexed points correspond more to the excitation, as the excitation has more jagged spectral characteristics, and consequently higher quefrency characteristics. This phenomenon can be seen in Figure

2-1. The cepstral values at the beginning represent features of the vocal tract, whereas the humps farther to the right represent the pitch characteristics.

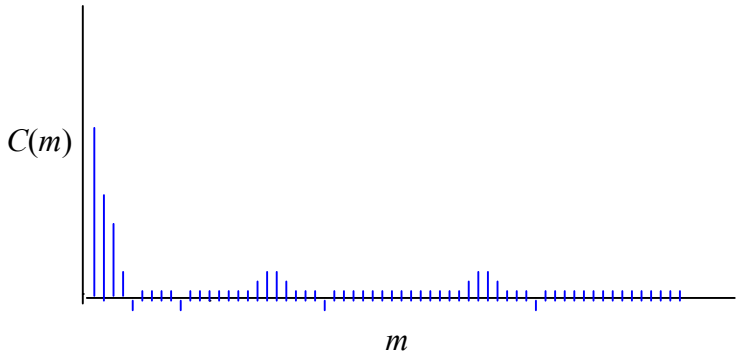

Figure 2-1. Example cepstrum of a voiced speech signal.

An alternate way to compute these cepstral coefficients involves grouping the power spectrum of the windowed frame of a signal into triangular filter channels as shown in Figure 2-2. In this figure, each triangle represents a channel, and the height of the triangle along the frequency axis represents the multiplicative factor for the spectral magnitude coefficient at each frequency for that channel. The sum of the energy in each of these channels becomes a spectral energy coefficient, and the cepstrum can be found using the discrete cosine transform (DCT) by

$$C(m) = \sum_{k=0}^{N} \log\left[e(k)\right] \cos\left(\frac{2\pi k}{N}\right), \tag{2.5}$$

where N is the number of filter channels and $e(k)$ is the energy in the *kth* channel.

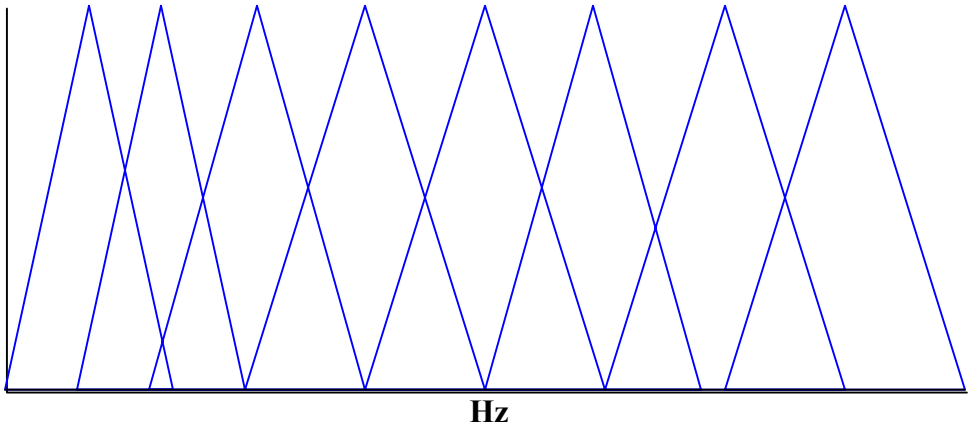

Figure 2-2. Filter channels for cepstral processing.

The Mel-scale is an empirical scale developed by Stanley Smith Stevens, John Volkman, and Edwin Newman. It was created using human listeners who judged the distance between tones, providing a perceptual measure of the difference between tones as a measure of frequency. A function that approximates this scale is given by

$$MELS = 1127 \ln(1 + \frac{f}{700}), \tag{2.6}$$

where f is the frequency in Hz.

2.3.3 Energy and delta coefficients

Because the energy of a speech waveform carries significant information about which sounds are being produced, an energy coefficient is often included along with the Mel-frequency cepstral coefficients (MFCCs) [1]. Typically, this energy coefficient is the log energy of the frame of speech.

Additionally, the trajectory of the MFCC features contains much information about the speech. As the human articulation system changes through time, so does the spectral envelope of the acoustic waveforms produced. Since the articulation mechanics change smoothly, so does the spectrum. Hence, features that track this change, known as

delta coefficients, are used along with the MFCC and energy coefficients. First-order delta coefficients are computed over the static MFCC and energy features using the linear regression equation

$$d_t = \frac{\sum_{\tau=1}^{\mathrm{T}} (c_{t+\tau} - c_{t-\tau})\tau}{2\sum_{\tau=1}^{\mathrm{T}} \tau^2}, \tag{2.7}$$

where c_t is the static cepstral (or energy) coefficient, and T is the maximum lag used. T is typically two or three. Often, second-order delta coefficients are used as well. These features are computed over the first-order delta coefficients using (2.7). As the most common number of cepstral coefficients used is twelve, a feature vector with MFCCs, log energy, first-order deltas, and second-order deltas (referred to as delta-deltas), is typically of size 39.

For a more detailed discussion of cepstral coefficients or speech recognition, [1] and [3] are good resources. More information about speech processing in general can be found in these texts.

2.4. Reconstructed Phase Spaces

2.4.1 RPS Theory

A reconstructed phase space (RPS) of a signal is an embedding that is topologically equivalent to the signal's generating system [18]. It is created by embedding a signal against time delayed versions of itself as

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_{1+(d-1)\tau} \\ \mathbf{x}_{2+(d-1)\tau} \\ \vdots \\ \mathbf{x}_N \end{bmatrix} = \begin{bmatrix} x_{1+(d-1)\tau} & \cdots & x_{1+\tau} & x_1 \\ x_{2+(d-1)\tau} & \cdots & x_{2+\tau} & x_2 \\ \vdots & & \ddots & \\ x_N & \cdots & x_{N-(d-2)\tau} & x_{N-(d-1)\tau} \end{bmatrix}, \tag{2.8}$$

where x_n is the signal value at time-sample n, d is the dimension, and τ is the time lag. A single point in the RPS is given by

$$\mathbf{x}_n = \begin{bmatrix} x_n & x_{n-\tau} & \cdots & x_{n-(d-1)\tau} \end{bmatrix} \quad n = \{(1+(d-1)\tau) \quad \ldots \quad N\}. \tag{2.9}$$

A two-dimensional example of an RPS of the phoneme 'ah' is shown in Figure 2-3.

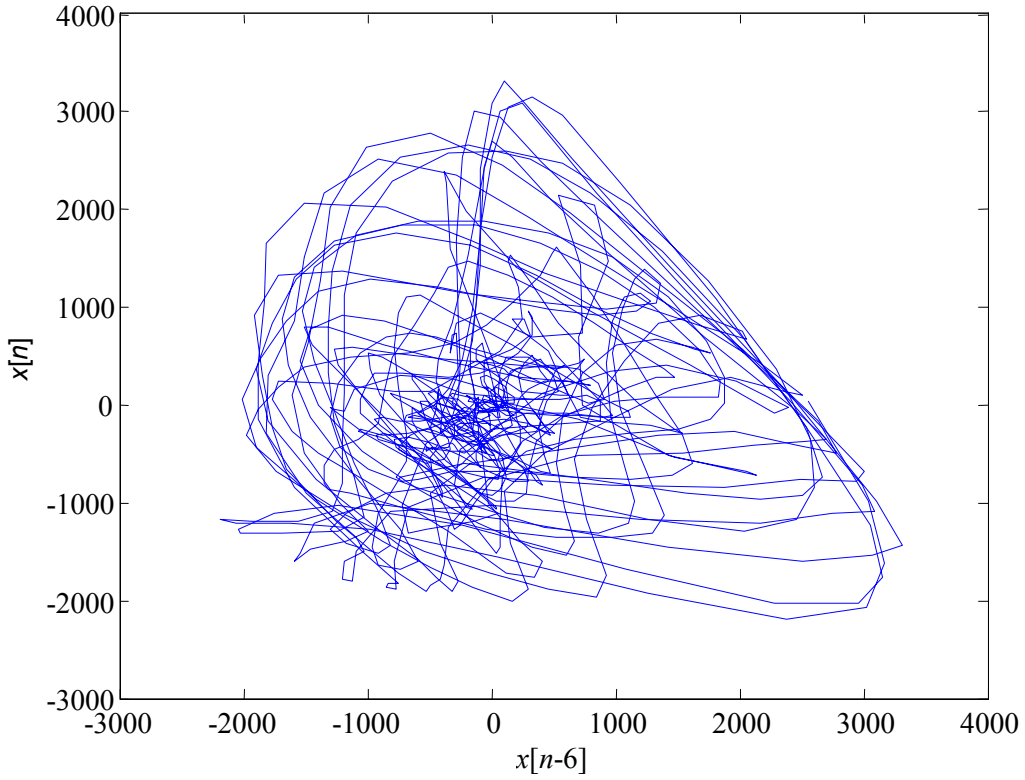

Figure 2-3. An example of a 2-dimensional, lag 6 RPS of the phoneme 'ah'.

Packard et al. [26] first proposed the concept of phase space reconstruction in 1980. Soon after, Takens showed [16] that a delay-coordinate mapping from a generic d-dimensional state space to a space of dimension $2d+1$ preserves topology. The time-delay function from the original state space to the reconstructed phase space is given by

$$\Phi_{\varphi,y}(x) = (y(x), y(\varphi(x)), ..., y(\varphi^{2d}(x)), \tag{2.10}$$

where $\varphi(x)$ is the state-update equation of the system, and $y(x)$ is the map from the state to the system output. Takens proved that, given an uncorrupted signal of infinite length, this

transformation is injective, surjective, and twice-differentiable. These conditions are sufficient for topological equivalence.

Sauer, Yorke, and Casdagli [18] extended Takens' theorem by showing that except in degenerate cases with probability zero, a time-delay embedding can be topologically equivalent to the original dynamical system with looser constraints. In Takens' theorem, the embedding dimension must be greater than twice that of the original system dimension. Instead, Sauer, Yorke, and Casdagli proved that the embedding dimension must only be greater than twice the boxcounting dimension of the system.

These theorems provide important theoretical justification for the use of RPS's for system identification and pattern classification. Because the topology of the RPS is identical to the topology of the underlying system phase space, we can expect the shape and density of the RPS attractor to provide valuable information of the system that generates a signal.

2.4.2 Applications to Signal Processing

Because RPS's preserve the dynamical information in a system, they have advantages over other methods for signal classification. Many classical methods make assumptions about the systems in question, and in doing so can ignore information that does not fit with those assumptions. While data reduction is often desirable for practical purposes, the reduction of data can remove important features. RPS's do not remove any information from the signals; the entire signal can be reconstructed from an RPS.

Unlike traditional linear signal classification methods, RPS's are able to capture nonlinear information that may be present in signals. Classic methods such as LPC

analysis and cepstral analysis work by assuming a linear model and learning parameters to fit the linear model [3]. These approaches work very well if the signals to be analyzed are generated by linear systems, but are fundamentally flawed if that is not the case. These methods are popular for speech signal analysis, based on the linearity assumption. However, recent work has shown that speech may in fact contain nonlinearities [4-6].

Dynamical invariant features extracted from RPS's have been used for classification of speech signals. Lyuponav exponents [8-10] and correlation dimension [27] have been used as features in addition to cepstral coefficients by appending them to the standard feature vectors. While this has been shown to increase accuracy, the improvements have been relatively small.

In addition to invariant features, RPS's can be statistically modeled. This approach has been successfully applied to heart arrhythmia classification [28] as well as motor fault detection [29]. Instead of extracting features from the RPS's, such as Lyuponav exponents and various dimensions, the full attractors are modeled. In this case, there is no data reduction, which is important, as the attractors may contain information not preserved by the dynamical invariants. To accomplish this, a statistical model is built over the attractor, describing the natural dimension of the RPS.

2.4.3 Applications to ASR

While the traditional linear features used in speech recognition systems are unable to capture nonlinear signal characteristics, RPS's can model these nonlinearities. It has been shown that speech signals carry nonlinear information [4-6]. Thus, cepstral coefficients and other linear-based features miss some of the information contained in speech. The potential for benefit from using this information is still under investigation.

One disadvantage of using RPS's for speech recognition is seen in the time complexity for classification of phonemes. Because cepstral coefficients are frame-based, a feature vector is computed and used for statistical analysis every 10 milliseconds, where RPS's have a feature vector at every sample. Under a standard sampling rate of 16 KHz, there are 160 more feature vectors in the RPS method.

Because of this complexity, the experiments conducted for this paper involve isolated phoneme classification as opposed to continuous speech recognition, which is the norm. This is because the time complexity of the Viterbi algorithm [1], which is used for the recognition of speech, is far greater for the RPS approach, due to the amount of data [30]. For RPS based methods to become useful, this issue must be solved. Currently, the best solution appears to be the use of frame-based features.

2.5. Sub-banding Speech Signals

2.5.1 Motivation

Harvey Fletcher, working at Bell Labs, ran experiments testing human speech recognition capabilities using various filters [31]. His intent was to study the capability of humans to understand speech filtered to have energy in different bandwidths and in different levels of background noise, with applications in telephone service.

He performed his experiments by having speakers say nonsense words that followed a consonant-vowel-consonant (CVC) or consonant-vowel (CV) pattern. This removed language cues, isolating acoustic information. He measured the rate at which listeners were able to accurately recognize these words on speech low-pass filtered along range of frequencies, and high-pass filtered along the same range. Using these results he developed an articulation index.

These results indicate that ideal conditions lead to better than 98% phoneme recognition rate. When filters were applied, the accuracy degrades in a systematic way. The total recognition error over a range of frequencies is equal to the product of the errors over the sub-bands in that frequency range. So, if the recognition rate for the frequency range from 300 Hz to 1000 Hz was 75%, and from 1000 Hz to 3000 Hz was also 75%, the accuracy from 300 Hz to 3000 Hz would be 1 – (25% * 25%) = 93.75%. This behavior can be expressed as

$$e_{total} = \prod_i e_i \tag{2.11}$$

where e_{total} is the error rate for the full bandwidth, and each e_i is the error rate for the sub-bands.

One possible interpretation is that humans recognize phonemes in sub-bands independently, and then combine the information across sub-bands [32]. It is unknown, however, how this information is combined. The results suggest that if any of the sub-bands is able to accurately recognize a phoneme, the human brain is able to identify which sub-band is correct.

To understand the mechanism that causes this phenomenon, one must look at the cochlea. The human cochlea has a frequency response that is space-variant [1]. Inside the cochlea resides the basilar membrane. The basilar membrane conducts mechanical vibrations induced by acoustic waves to the hair cells, which rub against the basilar membrane, causing neurons connected to the hair cells to activate.

The basilar membrane is thin on one end and thick on the other. In adult humans it is approximately 32 mm long, and is shaped like the cochlea [1]. Because of the variable width of the basilar membrane, its rigidity changes along its length. The local

rigidity determines the attenuation of the vibrations at each frequency. This way, the basilar membrane acts as a filter bank on incoming audio signals.

2.5.2 Sub-banding in Traditional Systems

ASR systems using traditional features have attempted to improve their robustness to narrowband noise with sub-banding. The work in this area was originally motivated by Fletcher's work, but the main aims are spurred by the need for more robust speech recognition. Many studies have been done using cepstral coefficients derived from sub-banded speech in the presence of many different types of noise [33-37]. 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.

It has been shown that sub-banding speech in certain types of noise (factory, babble, pink, etc.) can lead to better results [35, 37]. In [34], recombination of sub-band recognition hypotheses was used to show significant improvement in artificial narrowband noise, using signal-to-noise ratio estimates for weighting of the hypotheses. Recombination at different levels, including state, phone, and syllable level combination, was examined, showing slightly better accuracy with the syllable level recombination than the other levels. McCourt et al. [36] examined the combination of full-band and sub-banded features in white noise. They found that this combination could significantly

improve the robustness of their recognition system, even though the corrupting noise was not narrowband. However, the accuracy obtained in broadband noise can sometimes actually be degraded by sub-banding [35].

Sub-banding with traditional linear features does not typically improve the recognition accuracy in clean speech significantly [35, 37]. It is generally thought that sub-banding is unimportant for speech to be recognized that is not corrupted by noise. This does not follow from Fletcher's results, as his experiments were not used to study speech in noisy conditions.

2.5.3 Embedding Filtered Signals into RPS's

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 originates in the theorems of Takens [16] and Sauer et al [18]. 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 [38]. Previously, transforms such as principle component analysis (PCA) have been used to reduce the dimension of RPSs [39]. 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.

As a smooth, invertible transform, finite-impulse response (FIR) filters do not destroy the topological equivalence between an RPS and the underlying dynamical system [18]. 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, IIR filters, which require much lower orders to achieve sharp cutoffs, are used. It has been shown that convolution of chaotic signals with infinite-impulse response (IIR) filters can change the dynamical invariants of the systems [40-42]. Specifically, [41] showed that IIR filtering of the logistic map can increase the self-similarity of the map, and thus the fractal dimension. This is a cause for concern if the analysis method involves computing dynamical invariant features such as Lyapunov exponents or dimension. Unfortunately, this also means that the topological equivalence property has been lost. However, that this does not mean effective modeling and classification is impossible. On the contrary, it is shown in this thesis that these filtered RPSs still hold much information for discrimination of phonemes.

2.6. Summary

In the next chapter, the methods developed in this work for speech recognition is presented. This system combines the concept of sub-banding, motivated by Fletcher's work, and used in traditional ASR systems for robust speech recognition, with RPS's. This methodology allows for analysis of nonlinear dynamics to be studied in isolated frequency bands.

3. Methods

3.1. Overview

Shown in Figure 3-1 is a block diagram of the proposed system used for phoneme classification in this paper. It consists of three stages. First, phoneme waveforms are filtered into a set of mutually exclusive frequency sub-bands, which creates a set of waveforms from the original signal. Each new signal is then embedded into a reconstructed phase space (RPS) with parameters that are determined using techniques discussed in section 3.3.1. During the training phase, these RPS's are used to build the phoneme models; during testing, they are used for classification. The RPS's from each sub-band are treated independently; a classification is produced from each sub-band for every example. Finally, these likelihoods are fused to form a final decision.

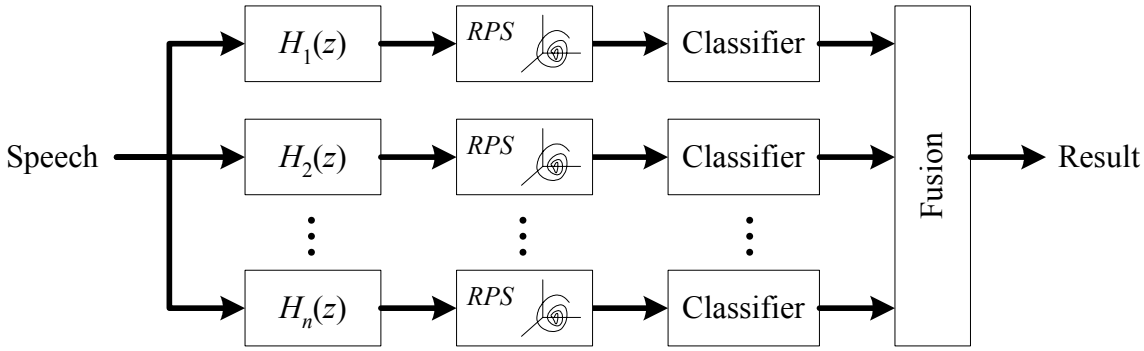

Figure 3-1. Block diagram of the proposed phoneme classification system

The filter characteristics, the RPS parameters, and the fusion strategy each have many possible forms, and the effects of the various methods are discussed in this thesis. Each of these system blocks is discussed in more detail in this chapter.

3.2. Filter Bank

The first stage of the system is a filter bank, which is intended to act similarly to the cochlea in the human auditory system. This front end separates the system dynamics

into separate frequency bands, which may be independent according to Fletcher's studies, but still contain nonlinear information. Because of the nature of the modeling applied, which is based on topological equivalence theorems, it is desired that the transforms applied to the signal do not disturb this topology. This objective can be achieved with finite-impulse response (FIR) filters. Unfortunately, the nature of FIR filters makes their use in this case unworkable. The desired filter bank should have extremely small transition bands and low passband and stopband distortions. These conditions are desired because the sub-banded signals are ideally independent, i.e. there are no common frequency components. FIR filters, though, require lengthy impulse responses, or large orders, to meet these conditions. This results in large distortions in the length and duration of the signal. To counter this, infinite-impulse response (IIR) filters are used with a forward/backward filtering process to eliminate phase distortion. They do not meet the requirements for topological equivalence, but their characteristics, especially the requirement of a lower order for sharp transitions, are superior.

The implemented filter bank is based fundamentally on auditory models, but does not exhibit their specific behavior. In the auditory model view, each channel is shaped as a gammatone filter [1], and these filters are overlapping. However, our approach uses non-overlapping filters that are shaped as closely to ideal filters as possible. This isolates frequency components in separate sub-bands, reduces the number of bands, and simplifies the analysis.

3.2.2 Filter basics

Chebychev type II filters [43] are chosen for analysis because of their control over the stopband. In order to completely isolate the nonlinear dynamics in each sub-band,

filters with very small ripple in the stopband are ideal. The implemented filters are of order 36, with stopband attenuation of 70 dB. Because of the numerical precision issues present when dealing with IIR filters [43], second-order structure (SOS) filters are used. The signals are filtered, time-reversed, filtered again, and finally time-reversed once more. This forward-backward filtering process eliminates any phase distortion of the speech signals, which is important because of the nature of the nonlinear analysis techniques applied here.

This forward-backward filtering can be applied to real-time recognition systems if each frame is filtered individually. Provided the computational requirements of the filtering process are not excessive, a small delay in the recognition process will be introduced, on the order of 20 to 30 milliseconds, depending on the frame-length.

3.2.3 Filter Bank Structure

The first major decision to be made in the design of the filter bank is how many filters to use. While it has been suggested that the human auditory system has between 10 and 30 sub-bands, it may not be best to use the same number of filters in the proposed system. Traditional speech recognition systems using sub-banding typically use two to seven sub-bands [34-36]. To study the effects of the number of sub-bands used on the proposed system, we experiment with a variable filter bank size. Experimental results using sets of two, four, and eight sub-bands are presented in this thesis.

Another factor in the design of the filter bank is the spacing of the filter cutoffs and central frequencies. Given a fixed number of filters in the bank, there are still an infinite number of possible filter combinations. The simplest way to space the filters is to place them linearly along the spectrum. However, human hearing has a logarithmic

nature, in the sense that resolution decreases logarithmically along the spectrum. There are several empirical scales that describe this functionality, including the Mel-scale, the bark-scale, and the equivalent rectangle bandwidth (ERB). To simplify the analysis, experiments are restricted to use of the Mel-scale, which is popular in the field speech processing.

Shown in Figure 3-2 is an example of a phoneme that has been filtered into four Mel-spaced sub-bands, and embedded into a two-dimensional, lag six RPS. In Figure 1-1-a, the RPS of the unfiltered vowel ‘ao’ is depicted. In Figure 3-2-b through Figure 1-1-e, each sub-band RPS is shown. It can be seen that the lower frequencies have a smoother, and apparently more characteristic structure than the higher frequency bands. This is due to the smoothness inherent in low-pass filtered signals, as well as the fact that the characteristic information present in vowels lies predominantly in the lower frequencies.

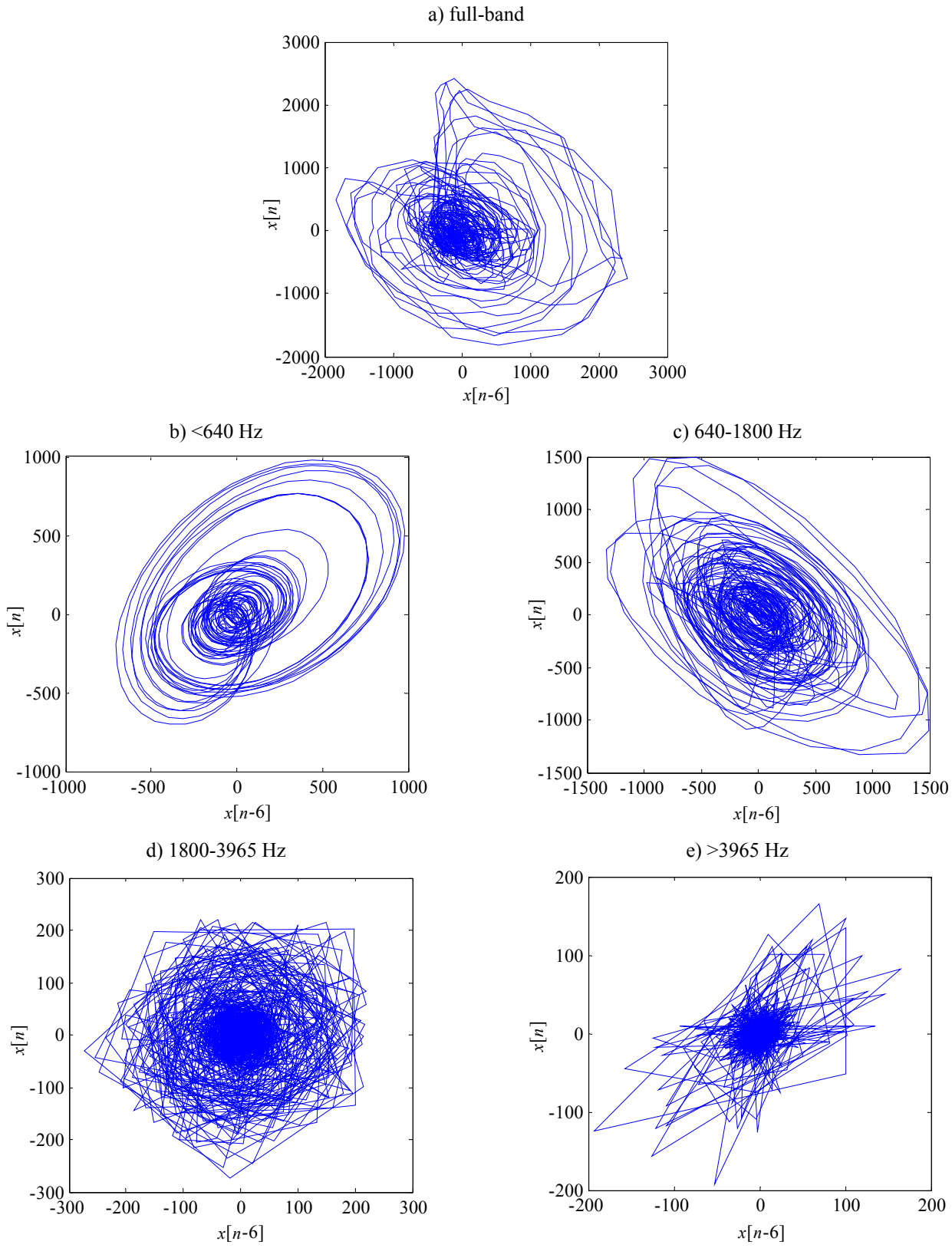

Figure 3-2. The vowel 'ao' filtered into four sub-bands and embedded into 2-dimensional, lag 6 RPS's.

3.3. Reconstructed Phase Space

3.3.1 RPS parameters

Once the speech signals have been filtered, they are embedded into RPS's. Before this step can be completed, though, the embedding dimension, *d*, and time lag, τ, must be determined. Takens' theorem states that the embedding dimension must be greater than twice that of the original state space dimension [16]. Sauer, Yorke, and Casdagli's work [18] showed that *d* must be only twice as large as the box-counting dimension of the system, which is smaller than the requirement given by Takens. There is no theory that gives bounds on the time lag.

Unfortunately, the dimension of the original state space structure is unknown, so some other means of determining an appropriate embedding dimension must be found. Accordingly, heuristics for finding the best dimension and lag have been developed. These methods involve first minimum of auto-mutual information and first zero of autocorrelation for lag identification, and global false nearest neighbors for dimension determination [7]. For the decision of time lag and embedding dimension in this thesis, the auto-mutual information and global false nearest neighbors methods are used, respectively.

The auto-mutual information of a single signal with respect to lag is given by:

$$I_s(\tau) = \sum P(s(n), s(n+\tau)) \frac{\log(P(s(n), s(n+\tau)))}{P(s(n))P(s(n+\tau))}, \tag{3.1}$$

where $P(s(n))$ is the probability of finding the value $s(n)$, and $P(s(n), s(n+\tau))$ is joint probability of measuring $s(n)$ and $s(n+\tau)$. This gives a measure of the redundancy of information between a signal and lagged versions of itself. To minimize the amount of

redundant information, a lag is chosen with the first minimum value from this function for the embedding time lag, τ.

With the lag determined, the global false nearest neighbors [7] method determines the embedding dimension. As the dimension of the RPS is increased, points in the RPS that are close together can be pulled farther apart. The close proximity of these points in the lower dimension is sometimes the result of 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. At this step, adding more dimensions is unnecessary. Global false nearest neighbors uses this strategy to estimate the appropriate embedding dimension. 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

$$D_n(d)^2 = \left\| \mathbf{x}_n(d) - \mathbf{x}_n^{NN}(d) \right\|^2 = \sum_{i=0}^{d-1} \left[x_{n-i\tau}(d) - x_{n-i\tau}^{NN}(d) \right]^2. \tag{3.2}$$

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

$$D_n(d+1)^2 - D_n(d)^2 = \sum_{i=0}^{d} \left[x_{n-i\tau}(d) - x_{n-i\tau}^{NN}(d) \right]^2 - \sum_{i=0}^{d-1} \left[x_{n-i\tau}(d) - x_{n-i\tau}^{NN}(d) \right]^2. \tag{3.3}$$

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. If this threshold is exceeded, these points are considered false neighbors in dimension *d*. If the number of false nearest neighbors exceeds a specified percentage, it is assumed that the attractor has not yet unfolded. The dimension is then increased, and the testing process repeated until the attractor is assumed to be fully unfolded. Using this method, the embedding dimension is chosen to be five.

3.3.2 Normalization

Depending on the speaker, the microphone, and the words spoken, the attractor for each phoneme can be transformed in the RPS in ways such as scaling and translation. To deal with this, a normalization procedure is implemented [19]. First, to deal with potential translation issues, each RPS is zero-meaned, causing the centroid of the RPS to reside at the origin. Secondly, each RPS is radially normalized. This is accomplished by computing the standard deviation of the radius of the RPS by

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i^{\,2}}. \tag{3.4}$$

Each point in the RPS is then divided by σ, causing every RPS to have the same radial standard deviation.

3.4. Gaussian Mixture Models

Gaussian Mixture Models (GMMs) are a generalization of Gaussian distribution functions. Gaussian probability density functions (pdf) are ideal because mathematical analysis of Gaussian's is simple, and because of the central limit theorem, which states that in the limit, the sum of random variables with identical distributions of any type has a Gaussian distribution. A Gaussian pdf of an independent random variable x is given by:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{(x-\mu)^2}{2\sigma^2}}, \tag{3.5}$$

where μ is the mean, and σ^2 is the variance of the Gaussian. Often, distribution functions of many correlated random variables are needed. The multivariate Gaussian function of the random vector $\mathbf{x}$ is given by:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}\left|\sum\right|^{1/2}} e^{\frac{(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\sum^{-1}(\mathbf{x}-\boldsymbol{\mu})}{2}}, \tag{3.6}$$

where **μ** is the mean vector, Σ is the covariance matrix, and n is the dimension. Instead of a single variance, the covariance matrix defines the variance $E[x_i^2]$ of each element of the random vector **x** and the covariances $E[x_i x_j]$ for each pair of elements in **x**. Σ is always symmetric and positive semi-definite. The elements of x can be treated as uncorrelated, and in this case Σ becomes a diagonal matrix. This allows for simplified computation and analysis.

Gaussian pdfs are limited in their ability to representative capability, however, as they are unimodal and are completely characterized by their first and second order moments. Hence, more complex probability models are often required. If a number of Gaussians are used, multi-modal distributions can be effectively modeled. This leads to the definition of a GMM, which is a sum of weighted Gaussian distributions. Provided the weights sum to unity, a GMM is a valid probability density function. Given an arbitrarily high number of mixtures, a GMM can approximate any distribution. In order to model a set of data with a GMM, the weights of each Gaussian distribution and their means and covariance matrices must be learned. This is accomplished via a well-known iterative process called expectation maximization (EM) [44].

3.5. Expectation Maximization

Given an initial hypothesis of the means and covariances of the Gaussian distributions that compose a GMM, EM can be used to iteratively improve the hypothesis [44]. The quality of a particular hypothesis is defined by the likelihood that the data used for training was generated by the proposed GMM. This likelihood is given as

$$P(\mathbf{x} \mid \Theta) = \sum_{k=1}^{M} c_k \, p(\mathbf{x} \mid m_k, \Theta), \tag{3.7}$$

where Θ represents the GMM parameters, **x** is the observation, each m_k is a Gaussian pdf,

and c_k are the weights of the Gaussian pdfs. The maximizer of this function can be found by setting the partial derivatives with respect to the means and vectors to zero and solving. This leads to the update equations for the parameters:

$$\boldsymbol{\mu}_k = \frac{\sum_{n=1}^{N} P(m_k \mid \mathbf{x}, \Theta)\mathbf{x}}{\sum_{n=1}^{N} P(m_k \mid \mathbf{x}, \Theta)}, \tag{3.8}$$

$$\sigma_k^{\,2} = \frac{\sum_{n=1}^{N} P(k \mid \mathbf{x}, \Theta)(\mathbf{x} - \boldsymbol{\mu}_k)^2}{\sum_{n=1}^{N} P(k \mid \mathbf{x}, \Theta)}. \tag{3.9}$$

3.6. Bayesian Classification

Once all the models are trained, classification is performed on each individual sub-band. A Bayes' classifier approach is taken, producing a log-likelihood score for each class on each example test phoneme. The probability of each class given the data is denoted $\hat{p}(c_i \mid \mathbf{x})$. The class that maximizes this quantity is

$$\hat{c} = \arg\max_{i=1\ldots C} \hat{p}(c_i \mid x). \tag{3.10}$$

Because $\hat{p}(c_i \mid \mathbf{x})$ is unavailable, Bayes' theorem is used, yielding

$$\hat{c} = \arg\max_{i=1\ldots C} \frac{\hat{p}(\mathbf{x} \mid c_i)\hat{p}(c_i)}{\hat{p}(\mathbf{x})}. \tag{3.11}$$

As the probability of the data $\hat{p}(x)$ does not affect the maximization operation, it can be removed, giving

$$\hat{c} = \arg\max_{i=1\ldots C} \hat{p}(\mathbf{x} \mid c_i)\hat{p}(c_i). \tag{3.12}$$

Assuming uniform distribution of the class prior probabilities, this equation becomes

$$\hat{c} = \arg\max_{i=1\ldots C} \hat{p}(\mathbf{x} \mid c_i), \tag{3.13}$$

which is known as the maximum likelihood equation. This quantity is not a true probability, and is called the class likelihood. It is denoted as $l_c(x)$. Because likelihood values are often very small, numerical computation issues can become problematic when dealing with large amounts of data, as the product of several likelihoods can quickly approach zero. To handle this, log-likelihoods are used in place of likelihoods.

For each phoneme, there are hundreds or thousands of points in the RPS. A log-likelihood value of each point for each GMM is computed. The overall log-likelihood of the phoneme for each class is

$$ll_c(\mathbf{X}) = \sum_{i=1}^{N} \log l_c(\mathbf{x}_i), \tag{3.14}$$

where x_i is the *ith* point in the RPS attractor, and N is the number of RPS points in the phoneme. The likelihood of each point for a GMM is

$$l_c(\mathbf{x}_i) = \sum_{k=1}^{M} c_k \frac{1}{(2\pi)^{n/2} |\sum_k|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{u}_k)^T \sum_k^{-1} (\mathbf{x}-\mathbf{u}_k)}, \tag{3.15}$$

where c_k is the weight of mixture k, $\mathbf{x}$ is the RPS point, $\mathbf{u}_k$ is the mean vector of the *kth* mixture, and Σ_k is the covariance matrix of the *kth* mixture. To avoid exponentiation and subsequent log calculation, the likelihood values are often calculated directly in the log domain.

3.7. Fusion

The final step of the proposed recognition system is the fusion operation. In this step, the class log-likelihoods for each sub-band are combined to yield a single log-likelihood for each class. Many strategies are available for fusion, including linear combinations [45], nonlinear combinations [46], and hierarchical schemes [47]. There is a large body of research pertaining to data fusion or classification combination. Image

processing [45], multi-sensor systems [48], and speech recognition [33, 36, 37] are some example applications that have seen interest in data fusion.

Several approaches can be taken in regard to the treatment of uncertainty for data fusion. Among these frameworks are Bayesian (probabilistic), Dempster-Schafer theory [49], possibility theory, and fuzzy logic. Each of these approaches makes different commitments to the ontology of uncertainty. In this thesis, a Bayesian approach is adopted. All fusion strategies presented are in the form of a linear combination of the sub-band log-likelihoods, given by

$$\hat{p}(\mathbf{X}) = \sum_{j=1}^{K} w_{j,c}\hat{p}_j(\mathbf{x}), \tag{3.16}$$

where the $w_{j,c}$ coefficients are the weights given to each band, **x** is the data for one sub-band, and **X** is the data for all bands. Several weighting strategies are examined.

3.7.1 Equal weights

The probability of a class given the set of features from all sub-bands is computed as

$$\hat{c} = \arg\max_{i=1\ldots C} p(c_i \mid \mathbf{x}_1\mathbf{x}_2\ldots\mathbf{x}_K) = \arg\max_{i=1\ldots C} \frac{p(\mathbf{x}_1\mathbf{x}_2\ldots\mathbf{x}_K \mid c_i)p(c_i)}{p(\mathbf{x}_1\mathbf{x}_2\ldots\mathbf{x}_K)}. \tag{3.17}$$

Because the $p(\mathbf{x}_1\mathbf{x}_2\ldots\mathbf{x}_R)$ term does not affect the maximization, and the class priors are treated as uniformly distributed, equation (3.17) becomes

$$\hat{c} = \arg\max_{i=1\ldots C} p(\mathbf{x}_1\mathbf{x}_2\ldots\mathbf{x}_K \mid c_i). \tag{3.18}$$

This quantity is difficult to analyze, so the class log-likelihoods produced by each sub-band are assumed to be independent. Given Fletcher's findings, this appears to be a reasonable approximation. Using this assumption, equation (3.18) becomes

$$\hat{c} = \arg\max \prod_{j=1}^{K} p(\mathbf{x}_j \mid c_i). \tag{3.19}$$

Since log-likelihoods are used, (3.19) becomes

$$\hat{c} = \arg\max_{i=1\ldots C} \sum_{j=1}^{K} \ln p(\mathbf{x}_j \mid c_i). \tag{3.20}$$

Thus, according to Bayes' theorem and assuming conditional independence of the classifiers, the log-likelihood for each class is given by the un-weighted sum of the log-likelihoods for each sub-band RPS classifier.

3.7.2 Sub-band accuracy-based weights

Though equation (3.20) gives an un-weighted sum of log-likelihoods as the correct likelihood for each class, intuitively this does not seem the best possible approach. Instead, it would seem appropriate that classifiers that carry more information, i.e. those that provide higher individual accuracies, would be given more weight towards the overall classification decision. Accordingly, a second fusion strategy is adopted, in which the weights in equation (3.16) are computed in proportion to the accuracy of each individual sub-band, computed over a development set that is discussed in section 4.1. The weight for sub-band *j* with development set accuracy $p_j(c \mid \mathbf{x} = c)$ is given by

$$w_j = \frac{p_j(c \mid \mathbf{x} = c)}{\sum_{k=0}^{K} p_k(c \mid \mathbf{x} = c)}. \tag{3.21}$$

3.7.3 Sub-band accuracy by class-based weights

As presented in [50], the sub-band RPS classifiers perform differently on various types of phonemes. For example, the lower frequency sub-band RPS classifiers produce

much higher accuracy on vowels than do the higher frequency bands, and the higher frequency sub-band RPS classifiers produce higher accuracy on fricatives than lower frequency bands. Therefore, a strategy that weights each log-likelihood based on the sub-band and class is used.

Because each sub-band produces log-likelihoods with varying scales, these numbers must be normalized, or weighting by class would produce invalid results. Without normalization, each class would be scaled according to its own weight, and the comparison of the fused log-likelihoods would not be valid. Normalization is executed by exponentiating each log-likelihood, then rescaling each of these values so that they sum to unity. The weight for the *ith* class of the *jth* sub-band is given by

$$w_{i,j} = \frac{p_j(c_i \mid \mathbf{x} = c_i)}{\sum_{k=0}^{K} p_j(c_i \mid \mathbf{x} = c_i)}. \tag{3.22}$$

3.7.4 Variance of sub-band log-likelihood-based weights

In [34], multi-classifier combination is implemented using entropy of log-likelihoods as a basis for weighting. This scheme makes the assumption that classifiers that have higher entropy, or more uniformly distributed log-likelihood values, are less reliable than those with lower entropy. The weight given to each classifier is inversely proportional to its entropy.

A similar approach is taken here, but variance is used instead of entropy. Sub-bands that have a higher variance of the log-likelihood values, and accordingly a larger spread between the highest class log-likelihoods and the lowest class log-likelihoods, are assumed more reliable than other bands. The sub-band weights are then distributed in proportion to the variance of the log-likelihood values of each band. Unlike the three

previous weighting strategies, the weights in this scheme are not the same for every test example, as they are computed during the testing phase for each phoneme. The weight of the *jth* sub-band is given by

$$w_j = \frac{\sum_{i=1}^{C} p_j(\mathbf{x} \mid c_i)^2}{\sum_{k=1}^{K}\sum_{i=1}^{C} p_k(\mathbf{x} \mid c_i)^2}, \tag{3.23}$$

where $p_j(\mathbf{x} \mid c_i)$ is the normalized probability of the data given class *i* in sub-band *j*.

3.7.5 Optimized Weights

A fifth approach is examined in which each sub-band weight is learned using an optimization technique known as the Nelder-Mead simplex method [51]. This algorithm finds the weights that give the greatest classification accuracy over the development set that was previously mentioned and is detailed in section 4.1.

3.7.6 Energy

In section 3.3.2, the radial normalization procedure used to deal with possible scaling effects was explained. This technique removes the energy associated with the RPS, which can be a valuable feature for discrimination of phonemes. By fusing energy with the RPS sub-bands, the phoneme classification accuracy could potentially be significantly increased. This is performed by fusing the sub-band RPS log-likelihoods with log-likelihoods obtained from a classifier that uses an energy vector created using Mel-spaced triangular filter channel log energies, along with the log energy of the full signal.

3.8. Summary

In this chapter, the proposed methodology for classification of phonemes based on

sub-banded RPS's has been introduced. Five techniques for fusion of the individual sub-band classifications are discussed. In chapter 4, the experiments studying the proposed system and comparing the fusion techniques are explained, and the results are presented.

4. Experiments

4.1. Data set

The proposed system is tested using the well-known TIMIT speech corpus [52]. Different from many speech databases, TIMIT has expertly-labeled phonetic boundaries. Most speech databases include word-level transcriptions of the recorded utterances, but the phonetic transcriptions must be obtained by substitution using a pronunciation dictionary. For TIMIT, which was developed by Texas Instruments and the Massachusetts Institute of Technology, time-stamped phonetic boundaries were determined by linguists, allowing for precise extraction of the phonemes in each utterance.

TIMIT contains utterances from 630 speakers, both male and female, from eight dialect regions in the United States. Each person read ten predefined sentences into a headset microphone, used to ensure the quality of the speech signals. Two of the sentences spoken by each individual are common across all speakers, and are not typically used for most speech recognition evaluation tasks. The test and training sets contain 168 and 462 speakers respectively, with no speakers contributing utterances to both sets. Thus, the most common use of TIMIT is for speaker-independent recognition.

Because the experiments presented in this thesis involve isolated phoneme classification, all phonemes are extracted from each utterance that is used. There are approximately 132,000 phonemes in the training set, and just over 48,000 in the test set. Though the phonetic labels in TIMIT were created using a set of 64 phonemes, not all of these are used as classes for modeling and recognition. Instead, a folding procedure

outlined in [53] is used. The set of 64 phonemes is reduced to 48 classes by re-labeling some of the phonemes, creating equivalence classes. This folding is defined as

Vowels	{ih ix} {ax ah} {ao aa} iy eh ey ae aw ay ox ow uh uw er
Semivowels	{el l} r w y hh
Stops	b d g p t k dx
Nasals	{n en} m ng
Fricatives	{sh zh} jh ch s z f th v dh

Additionally, though forty-eight models are learned and used for classification, errors between specified classes are not counted as errors. These between-class error reductions are indicated by the brackets in the above folding definition.

For the sub-band RPS approach, several types of fusion strategies are examined. Because some of these fusion techniques require learning parameters, and it is undesirable to use the test set to learn these parameters, a development set is created. To do this, the training set is randomly partitioned into two new sets: a new, smaller training set, which consists of 90% of the entire training data, and a development set, which contains the other 10% of the original training data. All models, both for the baselines and the proposed system, are trained using this 90% training partition. Any fusion parameters used for the sub-band RPS are learned using the development set.

4.2. Baselines

4.2.1 MFCC

The standard feature type used in state-of-the-art automatic speech recognition (ASR) systems is the Mel-frequency cepstral coefficient (MFCC). To establish a baseline with which to compare the proposed system, isolated phoneme classification experiments are performed for two typical feature sets. These include a set of 12 MFCCs, appended

with log energy, totaling 13 features, as well as a 39 feature baseline of 12 MFCCs, log energy, and first and second order deltas calculated on those 13 base features. Deltas are a linear regression that measures the general trajectory of the base features.

Gaussian Mixture Models (GMMs) are learned over these feature sets using the binary-split expectation maximization (EM) algorithm detailed in section 3.5. The models are trained to have 16 mixtures. A maximum likelihood classifier is used for testing. The results are given in Table 1.

12 MFCC, log energy	52.33%
12 MFCC, log energy, delta, delta-delta	56.94%

Table 1. Phoneme classification accuracies for MFCC baselines.

4.2.2 Fullband RPS

The sub-band reconstructed phase space (RPS) approach is also compared to the unfiltered RPS method to examine the effects of sub-banding RPS's. The parameters for embedding dimension and lag are found using the heuristics introduced in section 3-3. These parameters are $d = 10$ (five base dimensions plus five delta dimensions), and $\tau = 6$. Again, GMMs are used to model the RPS features, and are learned using binary-split EM. The number of mixtures used, which was determined empirically in [30], is 128. The classification accuracy for an RPS of dimension 10, with delta dimensions is 38.81%.

4.3. Sub-band RPS

The sub-band RPS system introduced in chapter 3 is tested on the same data and with the same methodology outlined in the previous baseline experiments section. In order to study the behavior of the system with respect to the number of sub-bands, three

sets of experiments are run, each using a different filter bank size. Sub-banding and fusion experiments are performed using two, four, and eight sub-bands. As stated previously in chapter 3, the filters are non-overlapping and equally spaced along the Mel-space.

4.3.1 Individual Band Results

The classification results for the individual RPS sub-bands are given in Table 2, Table 3, and Table 4. It can be seen from these numbers that classification of phonemes in filtered RPS's is possible; the sub-bands still carry discriminatory information.

Band	< 1800 Hz	> 1800 Hz
Dev set	36.26%	23.49%
Test set	34.57%	23.25%

Table 2. Phoneme classification accuracies for sub-banded RPS in two bands.

Band	< 640 Hz	640 – 1800 Hz	1800 – 3965 Hz	> 3965 Hz
Dev set	26.31%	25.36%	20.57%	14.89%
Test set	25.22%	24.47%	20.75%	14.77%

Table 3. Phoneme classification accuracies for sub-banded RPS in four bands.

Band (Hz)	<285	285-640	640-1130	1130-1800	1800-2715	2715-3965	3965-5670	>5670
Dev set	17.76%	20.42%	16.93%	19.11%	15.64%	14.47%	15.08%	14.15%
Test set	17.22%	19.92%	16.74%	18.43%	16.11%	14.30%	14.34%	14.58%

Table 4. Phoneme classification accuracies for sub-banded RPS in eight bands.

As was discussed in chapter 2, IIR filters do not preserve the topology of spaces as do FIR filters. However, because of the need for extensively long windows in FIR

filters to meet the small transition band requirements desired, IIR filters are instead chosen for analysis in this work. Hence, though IIR filters are the clear choice practically, the theoretical justification of reconstructed phase spaces becomes questionable. It is important, then, to verify that the classification of sub-banded phoneme RPS's is feasible.

These results show that the proposition that sub-banded RPS's can effectively model and classify phonemes is valid. This verification is an important first step in the evaluation of the proposed recognition system. It is interesting to note that the accuracies do not decrease at the same rate as the bandwidths in each band, especially in the lower frequencies.

4.3.2 Fusion Experiments

As discussed in section 3.7, the individual sub-band class log-likelihoods are fused to obtain an overall classification for each phoneme. The fusion strategies presented here are all forms of linear combination. The overall class log-likelihoods are a weighted sum of each sub-band class likelihood. This is defined by

$$\hat{\omega} = \arg\max_{c} \sum_{j=1}^{K} w_{j,c}\, ll_{j,c}(x), \tag{4.1}$$

where $ll_{c,m}(x)$ is the log-likelihood of class c given the data x in sub-band m, and w coefficients are the weights. Several weighting schemes are examined here.

An additional fused feature is based on energy. Energy can be an important feature for phoneme discrimination. Because of the radial normalization performed on the RPS's, though, it is removed. Therefore, adding likelihoods based on energy, or log energy, can potentially improve the sub-band RPS classifications. In order to accomplish this, an energy feature set is fused in the same manner as the sub-band features. Energy

feature class likelihoods are weighted and summed along with the sub-band class likelihoods to give an overall classification. The energy feature vector is created using Mel-spaced triangular filter channel log energies, along with the log energy of the full signal. The number of channels matches the number of bands for each fusion experiment, so for the two, four, and eight sub-band fusion experiments, the energy feature vectors are of length three, five, and nine, respectively.

4.3.2.1 Equal Weights

The simplest weighting scheme assigns equal importance to each sub-band class likelihood. All *w* coefficients are assigned a value of 1 for all classes and bands. As shown in section 3.7, this method is optimal if the sub-bands are truly independent. The phoneme accuracies for this strategy are shown in Table 5.

# 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 5. Phoneme classification of sub-band RPS with equal-weight based fusion.

From these results, it is clear that the sub-banded RPS approach improves the accuracy over an unfiltered RPS approach for classification of phonemes. The best accuracy with no energy is seen in the four-band case, where an absolute improvement of 5.4% is seen over the full-band approach. The eight-band case produces an accuracy that is lower than the two-band classifier when energy is not used, but the superior classification of the eight-channel energy features causes the eight-band classifier to outperform all other classifiers for the RPS-plus-energy case. This accuracy of 54.86% is greater than the accuracy produced by the MFCC feature set without delta coefficients.

This is significant because the proposed sub-band RPS classifier does not make use of comparable long-term deltas.

4.3.2.2 Accuracy-based weights by sub-band

As stated in section 3.7.2, weighting the log-likelihoods in each sub-band by the development set accuracy for that band seems intuitively more appropriate than weighting each band equally. The results for this fusion method are shown in Table 6.

# Sub-bands	2	4	8
Accuracy w/o energy	42.27%	43.62%	43.53%
Accuracy w/ energy	49.63%	50.11%	52.81%

Table 6. Phoneme classification of sub-band RPS with sub-band-accuracy-weight based fusion.

Comparing these results to those shown in Table 5, which represent the equal-weight fusion results, it is observed that the sub-band-accuracy-based weighting scheme is inferior to the equal-weight scheme. This is a bit surprising, as the individual sub-band accuracies are not equal. One might expect the lower frequency bands, which produce higher individual accuracies, to contribute more reliable information, and therefore deserve a larger weight.

4.3.2.3 Accuracy-based weights by sub-band and class

Based on the results of [50], a fusion weighting strategy based not only on the particular sub-band but the class seems suitable. However, the classification accuracies, shown in Table 7, are lower than both the equal-weight and sub-band based weight schemes.

# Sub-bands	2	4	8
Accuracy w/o energy	40.53%	37.17%	29.47%
Accuracy w/ energy	43.98%	46.65%	51.96%

Table 7. Phoneme classification of sub-band RPS with sub-band-by-class-weight based fusion

As can be seen by comparing Table 7 with Table 5 and Table 6, this method is consistently inferior to the strategies that use weights that are consistent across class. This is somewhat surprising, given the results in [50], which demonstrate that each sub-band has different strengths and weaknesses based on phonetic class. In contrast to the expectations produced from the results in this paper, weight determination using the likelihood of a sub-band classifier correctly classifying a particular phoneme class degrades the classification accuracy in comparison to a naïve weighting scheme.

4.3.2.4 Variance-based weights by sub-band

As discussed in section 3.7.4, a weighting scheme based on the variance of each sub-band RPS classifiers log-likelihood values is implemented. Classifiers with greater variances are expected to be more reliable, as they have a greater spread between the highest and lowest class log-likelihoods. As seen in Table 8, this strategy performs worse than all other strategies for the set of two sub-bands, and only outperforms the class-weighted scheme on the set of four and eight sub-bands when energy is not used.

# Sub-bands	2	4	8
Accuracy w/o energy	38.90%	38.54%	37.67%
Accuracy w/ energy	43.74%	44.33%	48.79%

Table 8. Phoneme classification of sub-band RPS with variance weight based fusion

The particular distribution of class log-likelihoods for each sub-band RPS classifier does not seem to provide especially useful information for fusion weighting, at least not with the strategy implemented here. However, one might expect that if the task was to classify phonemes corrupted by noise, this fusion strategy might be more successful. In [54], the recognition likelihood entropy was exploited for computation of a reliability measure in noisy speech, and this measure improved the robustness of the recognition system. It is possible that a similar approach using a reliability measure based on variance or entropy would provide better results for the experiments examined in this thesis if the task was robust phoneme classification.

4.3.2.5 Optimized sub-band weights

Parameter optimization is a common task in machine learning and pattern recognition. Artificial neural networks, support vector machines, and decision trees all learn a set of parameters over training data that will minimize the classification error. Here, the development set previously described is used to learn the *w* coefficients that maximize the classification accuracy. These weights are then used for classification over the test set. The results are given in Table 9.

# 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 9. Phoneme classification of sub-band RPS with optimized weight based fusion.

Comparison of these results to those of the equal-weighting scheme, shown in Table 5, reveals that the optimized weight classification accuracies are greater for each experiment. This is not surprising, as weights learned to give optimal classification

accuracy could only be inferior to any other set of weights if the optimal weights for classification on the development set and test set are significantly different. The difference in classification accuracies between the optimized weights and the equal weights is not large, however, demonstrating the efficacy of the equal-weighting scheme.

4.3.2.6 Fusion with MFCC features

As discussed in section 3.7.6, MFCC features are used to produce classifications that are fused with the sub-band RPS classifiers. Because the equal-weighting scheme produces the highest accuracies for the sub-band RPS fusion, that strategy is used for these experiments. The fusion results for the classifier using twelve MFCCs, log energy, deltas, and delta-deltas, and the sub-band RPS classifiers are presented in Table 10.

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

Table 10. Phoneme classification accuracies for fusion of sub-band RPS and MFCC classifiers.

The classification accuracies of the fusion of sub-band RPS and MFCC classifiers outperform the MFCC-only classifier for every configuration. The best fusion accuracy is seen with the eight sub-band RPS plus MFCC fusion using optimized weights. This accuracy produces a 2.38% absolute improvement. This suggests the presence of significant discriminatory information contained in the sub-band RPS representation that is not found in the MFCC representation.

4.4. Summary of Results

Examination of the experimental results, which are summarized in Table 11, shows that the optimized-weight strategy outperforms all other strategies for every set of

sub-bands. It is interesting to note, however, that the accuracies produced by this method are not significantly higher than the equal-weighting scheme. This is not necessarily the expected result, because the sub-band classifiers do not perform equally, especially in the case of only two sub-bands. Also the equal-weighting scheme does not make use of the information about the relative strengths and weaknesses of each sub-band classifier. However, as was derived in section 3.7.1, the equal-weighting scheme is optimal assuming independence among the bands. The best sub-band RPS fusion accuracy is seen for the eight sub-band RPS plus energy classifier. This accuracy is 54.95%, an absolute improvement of 16.14% over the full-band RPS classifier. It also outperforms the MFCC feature set that does not contain delta coefficients.

Fullband RPS baseline = 38.81%	2 bands	4 bands	8 bands
Optimized weights	50.65%	52.18%	54.95%
Equal weights	50.14%	51.96%	54.84%
Sub-band accuracy-based weights	49.63%	50.11%	52.81%
Sub-band & class accuracy-based weights	43.98%	46.65%	51.96%
Variance-based weights	43.74%	44.33%	48.79%

Table 11. Summary of fusion results with energy.

Fusion of the sub-band RPS classifiers with MFCC features improves the classification accuracy over the MFCC-only baseline. The best fusion accuracy is 59.32% for the eight sub-band RPS plus MFCC classifier, an absolute improvement of 2.38%. This result demonstrates that RPS's can potentially be used to improve speech recognition systems.

5. Conclusion

5.1. Comparison of Sub-banded RPS with MFCCs

The best accuracy reached by the proposed system is realized by the set of eight sub-bands with energy. Though it does not match the greatest cepstral baseline, that of the twelve Mel-frequency cepstral coefficients (MFCCs), log energy, deltas, and delta-deltas, it does outperform the set of MFCCs that includes energy but not delta coefficients. This is significant because the nature of the proposed system does not allow for inclusion of comparable delta coefficients. If similar features could be built for the system developed in this work, it may be able to outperform the top MFCC baseline.

5.2. Combination with MFCCs

As can be seen in Table 10, the fusion of the MFCC classifier and the sub-band reconstructed phase space (RPS) classifiers improved the accuracy over the MFCC-only baseline. This appears to suggest that there is information present in the sub-band RPS's that is not captured by the MFCCs. It seems likely, however, that there is some overlap in the information captured by the two methods. It would then be desirable to eliminate this overlap, and develop a feature set based on the sub-band RPS approach that models only the nonlinear or higher-order statistical information. These features could then be combined with the MFCC features in a more efficient way.

5.3. Future Directions

The performance of the proposed methodology for the classification of isolated phonemes suggests that this approach may be valuable for use in practical automatic speech recognition systems. Before that can be accomplished, this methodology must be applied to continuous speech recognition. The time complexity of the current approach,

however, does not allow for real-time recognition of continuous speech. Since each point in the RPS is treated as a frame, with likelihoods calculated for each model every 6.25 us for a sampling rate of 16,000 Hz, there are many more frames in this approach than in the standard spectral feature-based methodology. This leads to a recognition time complexity that is on the order of 100 times greater than automatic speech recognition systems based on spectral features [30].

One way to reduce the computational time is to develop a set of features based on the sub-banded RPS's. The computational complexity of cepstral features is $O(n \log n)$, where n is the frame length. If an RPS-based feature extraction technique with a similar complexity could be developed, the recognition time of the RPS approach would be reduced, making the approach practical. Additionally, using features computed from the sub-band RPS's would allow for delta and delta-delta features, important pieces of information, to be incorporated.

Sub-banding has been used in spectral-based for recognition of noisy speech. Though all experiments presented in this thesis are performed on clean speech, it would be interesting to apply the methodology presented here to speech corrupted by various types of noise. The use of sub-banded RPS's may benefit ASR systems that must be robust to noise.

5.4. Conclusion

This thesis has introduced a full phoneme classification system using sub-banded reconstructed phase spaces for classification of phonemes. Experiments studying the effects of filter bank size and fusion methods have been presented and discussed. The RPS modeling approach used 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 introduced 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 proposed in this paper produces classification accuracies inferior to that of the full MFCC feature set with log energy and regression coefficients, it does outperform the MFCC feature set that includes log energy but no delta coefficients. Combination of the sub-band RPS features and MFCCs shows a 2.38% absolute improvement over the MFCC-only features with equal-weighted log-likelihood fusion.

The results presented show that further research into sub-banded RPS's as an alternative for front-end parameterization in speech recognition systems is warranted. This approach has the potential to benefit real ASR systems, including those that must perform recognition on noisy speech.

6. References

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

[2] B. Pellom and K. Hacioglu, “Recent improvements in the CU SONIC ASR system for noisy speech: The SPINE task,” proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Hong Kong, 2003.

[3] J. R. Deller, J. H. L. Hansen, and J. G. Proakis, *Discrete-time processing of speech signals*, Second ed. New York: IEEE Press, 2000.

[4] M. Banbrook and S. McLaughlin, “Is speech chaotic?,” proceedings of IEE Colloquium on Exploiting Chaos in Signal Processing, 1994, pp. 8/1-8/8.

[5] 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.

[6] H. M. Teager and S. M. Teager, “Evidence for nonlinear sound production mechanisms in the vocal tract,” proceedings of NATO ASI on Speech Production and Speech Modelling, 1990, pp. 241-261.

[7] H. D. I. Abarbanel, *Analysis of observed chaotic data*. New York: Springer, 1996.

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

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

[10] 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.

[11] 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.

[12] X. Liu, R. J. Povinelli, and M. T. Johnson, “Vowel classification by global dynamic modeling,” proceedings of ISCA Tutorial and Research Workshop on Non-linear Speech Processing (NOLISP), Le Croisic, France, 2003, pp. 111-114.

[13] 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.

[14] R. J. Povinelli, M. T. Johnson, A. C. Lindgren, and J. Ye, "Time series classification using Gaussian mixture models of reconstructed phase spaces," *IEEE Transactions on Knowledge & Data Engineering*, in press.

[15] H. D. I. Abarbanel, T. A. Carroll, L. M. Pecora, J. J. Sidorowich, and L. S. Tsimring, "Predicting physical variables in time-delay embedding," *Physical Review E*, vol. 49, pp. 1840-1853, 1994.

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

[17] J. R. Munkres, *Topology*, 2nd ed. Upper Saddle River, NJ: Prentice Hall Inc., 2000.

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

[19] 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.

[20] 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.

[21] “Speech recognition software and medical transcription history,” available at http://www.dragon-medical-transcription.com/historyspeechrecognitiontimeline.html.

[22] “History of speech recognition and transcription software,” available at http://www.dragon-medical-transcription.com/historyspeechrecognition.html.

[23] “Phoneme,” available at http://en.wikipedia.org/wiki/Phoneme.

[24] *Webster's third new international dictionary*, third ed: Merriam-Webster, 2000.

[25] S. Haykin, *Adaptive filter theory*, 3rd ed. Upper Saddle River, New Jersey: Prentice Hall, 1996.

[26] 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.

[27] V. Pitsikalis and P. Maragos, “Some advances on speech analysis using chaotic models,” proceedings of ISCA Tutorial and Research Workshop on Non-linear Speech Processing (NOLISP), La Croisic, France, 2003.

[28] F. M. Roberts, R. J. Povinelli, and K. M. Ropella, “Identification of ECG arrhythmias using phase space reconstruction,” proceedings of Principles and Practice of Knowledge Discovery in Databases (PKDD'01), Freiburg, Germany, 2001, pp. 411-423.

[29] R. J. Povinelli, J. F. Bangura, N. A. O. Demerdash, and R. H. Brown, "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*, vol. 17, pp. 39-46, 2002.

[30] M. T. Johnson, R. J. Povinelli, A. C. Lindgren, J. Ye, X. Liu, and K. M. Indrebo, "Time-domain isolated phoneme classification using reconstructed phase spaces," *IEEE Transactions on Speech and Audio Processing*, in press.

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

[32] J. B. Allen, "How do humans process and recognize speech?," *IEEE Trans. on Speech and Audio Processing*, vol. 2, pp. 567-577, 1994.

[33] H. Bourlard and S. Dupont, “Subband-based speech recognition,” proceedings of International Conference on Acoustics, Speech, and Signal Processing, 1997, pp. 21-24.

[34] H. Bourlard and S. Dupont, “A new ASR approach based on independent processing and recombination of partial frequency bands,” proceedings of Fourth International Conference on Spoken Language Processing, 1996, pp. 426-429.

[35] 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.

[36] 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.

[37] 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.

[38] J. F. Gibson, J. D. Farmer, M. Casdagli, and S. Eubank, "An analytic approach to practical state space reconstruction," *Physica D*, vol. 57, pp. 1-30, 1992.

[39] J. Ye, M. T. Johnson, and R. J. Povinelli, “Phoneme classification over reconstructed phase space using principal component analysis,” proceedings of ISCA Tutorial and Research Workshop on Non-linear Speech Processing (NOLISP), Le Croisic, France, 2003, pp. 11-16.

[40] R. Badii, G. Broggi, B. Derighetti, and M. Ravini, "Dimension increase in filtered chaotic signals," *Physical Review Letters*, vol. 60, pp. 979-982, 1988.

[41] A. Chennaoui, K. Pawelzik, W. Liebert, H. G. Schuster, and G. Pfister, "Attractor reconstruction from filtered chaotic signals," *Physical Review A*, vol. 41, pp. 4151-4159, 1990.

[42] S. H. Isabelle, A. V. Oppenheim, and G. W. Wornell, “Effects of convolution on chaotic signals,” proceedings of ICASSP, 1992, pp. 133-136.

[43] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-time signal processing*, 2nd ed. Upper Saddle River, New Jersey: Prentice-Hall, 1999.

[44] A. Webb, *Statistical pattern recognition*. New York: Oxford University Press, 1999.

[45] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas, "On combining classifiers," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 226-239, 1998.

[46] S. Baglio, S. Graziani, G. Manganaro, and N. Pitrone, “Cellular neural networks: A new paradigm for multisensor data fusion,” proceedings of Electrotechnical Conference, 1996. MELECON '96., 8th Mediterranean, 1996, pp. 509-512 vol.1.

[47] M. Petrakos, J. A. Benediktsson, and I. Kanellopoulos, "The effect of classifier agreement on the accuracy of the combined classifier in decision level fusion," *IEEE Trans. on Geoscience and Remote Sensing*, vol. 39, pp. 2539-2546, 2001.

[48] M. Dietl, J.-S. Gutmann, and B. Nebel, “Cooperative sensing in dynamic environments,” proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS'01), Maui, Hawaii, 2001.

[49] G. Schafer, *A mathematical theory of evidence*: Princeton University Press, 1976.

[50] 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.

[51] J. A. Nelder and R. Mead, "A simplex method for function minimization," *Computer Journal*, vol. 7, pp. 308-313, 1965.

[52] 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.

[53] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, pp. 1641-1648, 1989.

[54] H. Misra, H. Bourlard, and V. Tyagi, “New entropy based combination rules in HMM/ann multi-stream ASR,” proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003, pp. 741-744.