AN SVM FRONT-END LANDMARK SPEECH RECOGNITION SYSTEM

BY

SARAH E. BORYS

B.S., University of Illinois at Urbana-Champaign, 2003

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Master of Science in Electrical and Computer Engineering
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2008

Urbana, Illinois

Adviser:

Associate Professor Mark Hasegawa-Johnson

ABSTRACT

Support vector machines (SVMs) can be trained to detect manner transitions between phones and to identify the manner and place of articulation of any given phone. The SVMs can perform these tasks with high accuracy using a variety of acoustic representations. The SVMs generalize well to unseen test data if these data were created under identical conditions to the training corpus. Unseen acoustic data from different corpora present a problem for the SVM, even if these acoustic data were generated under similar conditions. The discriminant outputs of these SVMs are used to create both a hybrid SVM/HMM (hidden Markov model) phone recognition system and a hybrid SVM/HMM word recognition system. There is a significant improvement in both phone and word recognition accuracy when these SVM discriminant features are used instead of mel frequency cepstral coefficients (MFCCs).

ACKNOWLEDGMENTS

This thesis would not have been possible without the guidance, encouragement, patience, and understanding of my adviser, Mark Hasegawa-Johnson.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Since at least 1955, psychophysical experiments in human speech perception have demonstrated that speech perception is multiscale and structured: coarse-scale information (prosody, syllable structure, sonorancy) can be perceived independently of fine-grained information (place of articulation) [2, 3, 4, 5, 6, 7, 8, 9]. Human ability to generalize quickly and effortlessly from one speaking style, signal-to-noise ratio (SNR), or channel condition to another has been attributed to this multiscale characteristic of speech perception [10, 11, 12]. Despite the importance of multiscale perception in human speech perception, psychologically realistic multiscale models have failed to outperform single-scale models such as the hidden Markov model (HMM). The apparent cause of the success of the HMM is the property of simultaneously optimal parameters: it is possible to simultaneously adjust every parameter in an HMM in order to optimize a global recognition performance metric (maximum likelihood, maximum mutual information, or minimum classification error). Until the 1990s, the HMM was the only large vocabulary speech recognition model with the characteristic of simultaneously optimal parameters; therefore, psychologically realistic hierarchical multiscale models were not competitive.

Current-generation automatic speech recognition (ASR) systems are based on an architecture (HMMs) that is both time-consuming to train, and extremely vulnerable to acoustic interference and variation in speaking style. The conventional methods for enhancing ASR performance often require enormous amounts of data

Text in this chapter has previously been published in [1].

collection and annotation, as well as extensive training on representative material. This dependence on training materials shapes the entire fabric of ASR methodology and makes it exceedingly difficult (and expensive) to introduce innovative concepts into speech recognition. As a consequence, the pace of innovation and refinement is considerably slower than it might otherwise be.

Current-generation ASR systems represent words as sequences of context-dependent phonemes. In order to train acoustic models proficient in classifying phonemic units, vast amounts of training material are required. Even with such material, state-of-the-art recognition systems generally misclassify 30-40% of the phonetic constituents [13]. Performance improves only slightly when a word transcript is provided. And yet, phonetic classification is critical for ASR performance; the word error rate (WER) is highly correlated with phonetic classification error [14, 15]. Substantial improvement of phonetic classification would likely yield a significant gain in ASR performance. Moreover, if phonetic classification were extremely accurate and pronunciation models in the lexicon precisely matched the phonetic classification data, ASR performance would improve dramatically [16]. Unfortunately, ASR systems are nowhere close to achieving such goals. An entirely different approach is required - one that melds state-of-the-art phonetic classifiers with realistic pronunciation models representative of the speaking styles and conditions associated with the recognition task.

Early work in automatic speech recognition included relatively sophisticated linguistic representations of phonology [17, 18], syntax [19, 20], and semantics [21]. By contrast, it was often assumed that expert knowledge of acoustic phonetics could add little to the knowledge automatically acquired by dynamic programming [22], finite state automaton [23], or HMM [24] algorithm; the success of these algorithms was so great that Klatt proposed a model of human speech perception based on frame-based finite state automata [25]. It was frequently argued that the acoustic correlates of a phoneme are so variable and context-dependent that context-independent phoneme classification is impossible; thus, human speech perception must integrate a tremendous amount of context for even simple phoneme perception tasks [26]. The possibility

of achieving very low phoneme classification error rates with limited context was first demonstrated in two quite different sets of experiments: spectrogram reading experiments [27], and experiments with neural networks [28]. Later experiments with hybrid neural-network/HMM systems hinted at the strong correlation between phoneme error rate and word error rate of an automatic speech recognizer [29, 30, 31, 32], leading to a renewed engineering focus on the linguistic discipline of acoustic phonetics [33, 34].

The "landmark-based speech recognition" approach described in this thesis draws on ideas initially proposed by Stevens et al. [35, 36]. In 1992, Stevens and his colleagues proposed a framework for automatic speech recognition based on his theory of human speech perception [36]. The algorithm described by Stevens begins with the detection of perceptually salient phonetic landmarks. These landmarks are of different types, including obstruent and nasal closures and releases, glide extrema, and the "steady state" center regions of vowels and syllabic consonants. Because landmarks are of different types, the detection of a landmark also specifies the values of distinctive features which define the landmark type. Stevens calls distinctive features which define a landmark type "articulator-free features," because they can be implemented by any articulator; in his 2000 proposal, the articulator-free features are [vowel, glide, consonant, sonorant, continuant, strident]. Using knowledge-based algorithms, Liu was able to detect closure and release of [-sonorant] consonants with an accuracy of approximately 95% [37, 38]. Liu detected closure and release of [+sonorant, -continuant] consonants (nasal consonants) with an accuracy of about 89%, and Chen [39] was able to detect nasalization in vowels adjacent to 94% of all nasal consonants. Howitt [40] used a multilayer perceptron to detect vowel landmarks with 93% accuracy. Espy-Wilson developed semivowel detectors with similarly high accuracy [41].

Glass and Zue [42] proposed the use of a simple spectral-change metric to detect phoneme segment boundaries in the SUMMIT system, and Halberstadt and Glass [43, 44] used the SUMMIT segment boundaries to anchor phoneme classification in a landmark-based system. Both papers propose that the landmark detector should

be allowed to generate a large number of false landmarks, in order to avoid the false rejection of any true landmarks. In the system proposed by Halberstadt and Glass, a lexical alignment program finds the best match between each sentence candidate and the proposed list of landmarks. As a by-product of lexical alignment, the program determines which landmarks are true segment-boundary landmarks, and which are segment-internal landmarks.

Landmark-based and segment-based speech recognition methods have been incorporated into hidden Markov models in a number of ways. Ostendorf et al. described a large family of methods for modeling variability in the duration and temporal sequencing of phonetic events; both segment-based and hidden Markov models were shown to be special cases of the general family of methods [45]. Bilmes et al. [46] used an HMM with models of phonetic auditory events (avents) separated by phoneme-independent steady state models, and achieved a 1.2% word error rate on the DIGITS+ database (ten digits plus "oh," "no," and "yes"). Word error rate did not increase as much in noise as a standard speech recognizer; at 10 dB SNR, word error rate was 8.1%. Omar, Hasegawa-Johnson, and Levinson created an HMM with special observation probability density models of phoneme boundaries; stop consonant recognition error rate was reduced by a factor of three, but overall phoneme recognition error rate was unchanged because of degraded recognition performance for vowels and glides [47]. The inappropriateness of standard MFCC features for a landmark-based speech recognizer motivated Omar and Hasegawa-Johnson to develop a generalized maximum likelihood nonlinear acoustic feature transformation for use in mixture Gaussian HMMs [48, 49].

Niyogi and Ramesh trained radial basis function support vector machines (RBF SVMs) to detect stop release segments in the TIMIT database [50, 51]. For the same level of false acceptances (about 7%), the RBF stop detector incurred fewer false stop rejections (21% vs. 30%) than an HMM phoneme recognizer. Niyogi and Burges [51] have shown that the nonlinear discriminant functions $g(\vec{x})$ computed using an RBF SVM have the property of imitating the perceptual magnet effect. Specifically, the distance $|g(\vec{x}_1) - g(\vec{x}_2)|$ decreases as vectors $\vec{x}_1$ and $\vec{x}_2$ are moved away

from the $g(\vec{x}) = 0$ separatrix. Equivalently, the sensitivity $|\nabla g(\vec{x})|$ is a monotonically decreasing function of $|g(\vec{x})|$. In the few cases that we have carefully observed, $g(\vec{x})$ as learned by an RBF SVM, tends to resemble an arctangent nonlinearity along the direction orthogonal to the separatrix, and therefore, we can specify that the perceptual magnet effect learned by an RBF SVM seems to resemble the following form:

$$|\nabla g(\vec{x})| \sim 1 - g^2(\vec{x}) \tag{1.1}$$

Juneja and Espy-Wilson combined the approaches of Stevens et al. and of Niyogi et al. in order to create an automatic speech recognition algorithm that combines SVM-based landmark detectors with a dynamic programming algorithm for the temporal alignment and classification of phoneme boundaries [52, 53, 54, 55]. SVM-based landmark detectors were trained for onsets and offsets of the distinctive features [silence] (94% recognition accuracy), [syllabic] (79% accuracy), [sonorant] (93%), and [continuant] (94%). Six-manner-class recognition accuracy on TIMIT was 80%, using a total of 160 trainable parameters.

All results listed above were obtained using clean speech recorded with a 16 kHz sampling frequency. In 2004, Hasegawa-Johnson et al. [1, 56] developed SVMs for landmark detection and classification in telephone speech; SVM discriminant outputs were observed by distinctive-feature-based lexical access systems based on either maximum entropy or dynamic Bayesian network probability models. The goal of this thesis is to report substantial improvement in the accuracy and computational complexity of the SVMs in [1], as well as their integration with an HMM back end. Part of this work has previously been reported in [57].

CHAPTER 2

BACKGROUND

2.1 Support Vector Machines

2.1.1 Problem formulation

We are given two sets of data. The members in each set are represented in pairs $(\vec{x}_i, y_i)$, where $\vec{x}_i$ is a vector of predetermined features, $y_i = -1$ or 1 is the class label, and there are $i = 1 \ldots N$ members in the universe D. The distribution of the data, $P(\vec{x}, y)$, is unknown. The vector $\vec{x}_i$ contains d elements.

We wish to find an optimal separatrix, $g(\vec{x}) = \vec{w} \cdot \vec{x} + b$, such that $h(\vec{x}) = \text{sign}\ (g(\vec{x})) = y$. We define the optimal separatrix as the hyperplane with the Vapnik-Chervonenkis (VC) dimension d_{VC} that minimizes the expected risk $R(\vec{w}, b)$. The VC dimension of a hyperplane is the logarithm of the maximum number of points in a training set that can be shattered by that hyperplane. Since there are N points with binary labels in the training set, it follows that there are at most 2^N different ways to label the dataset. The number of points *shattered* by $g(\vec{x})$ is the number of points correctly classified by $g(\vec{x})$ for a given labeling permutation.

The expected risk is written as

$$R(\vec{w}, b) = \frac{1}{2} \int \int |y - g(\vec{x})| p(\vec{x}, y) \partial \vec{x} \partial y \tag{2.1}$$

Because the expected risk depends on the probability density $p(\vec{x}, y)$, it is impossible to minimize this equation without prior knowledge of the distribution of the data.

Text in this chapter has been previously published in [1] and [57].

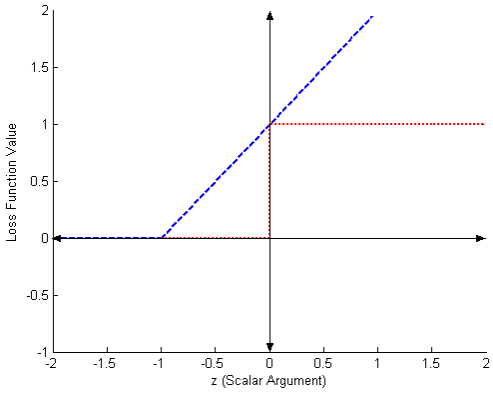

Figure 2.1 The ideal loss function (red) is discontinuous at 0 and is replaced with the "hinge" loss (blue) by the SVM.

Since this knowledge is usually not available, it is easier to bound $R(\vec{w}, b)$ with the empirical risk $R_{emp}(w, b, D)$ and a generalization bound $G(H, N)$, and minimize the right-hand side of the following inequality:

$$R(\vec{w}, b) \leq R_{emp}(\vec{w}, b, D) + G(H, N) \tag{2.2}$$

The set H contains a given class of mapping functions with adjustable parameters.

The empirical risk is defined as

$$R_{emp}(\vec{w}, b, D) = \frac{1}{N} \sum_{i=1}^{N} u(-y_i(\vec{w} \cdot \vec{x}_i + b)) \tag{2.3}$$

where $u(\cdot)$ is the "unit step" function. The unit-step function is shown in Figure 2.1 and is called the loss ℓ. Note that the unit step function is discontinuous at 0. Gradient-based optimization methods are not guaranteed to converge over discontinuous functions so the actual loss is approximated with a "hinge" loss ℓ_h, where

$$\ell_h = \max(0, z + 1) \tag{2.4}$$

The hinge loss function is shown in Figure 2.1. The use of ℓ_h instead of ℓ creates a "margin" around the line $g(\vec{x}) = 0$. Any member of D that lies between the lines $|\vec{w} \cdot \vec{x}_i + b| = 1$ is called a partial error. The width of the margin is $\frac{2}{\|\vec{w}\|}$.

How can we train a classifier using partial errors? One thing to do would be to say that all partial errors are "correct." Classifiers that use this philosophy

are called "gap-tolerant" classifiers. Another approach would be to treat all partial errors as if they were true errors. Classifiers that make this assertion are said to be "gap-intolerant."

Recall that we wish to minimize the right-hand side of Equation (2.2). The mapping functions in H are determined by whether a gap-tolerant or gap-intolerant approach is used. These two different kinds of mapping functions will be referred to as H_1 and H_2, respectively. The functions in H_1 are equivalent to classifiers with d-dimensional hyperplanes. Members in the set H_2 depend on the ratio of R, the radius of a sphere that encircles the data, to the width of the space between the margins. With probability $1 - \varepsilon$, the generalization error can be written as

$$\begin{aligned} G(H, N) &\leq \frac{\log(\frac{1}{\varepsilon})}{N} \times d_{VC}(H) \\ &= \frac{\log(\frac{1}{\varepsilon})}{N} \times \min(d, \frac{2}{\|\vec{w}\|}) \end{aligned} \tag{2.5}$$

For SVMs, we must have $d_{VC} \geq \frac{R^2}{\frac{2}{\|\vec{w}\|}}$.

The SVM problem formulation is depicted in Figure 2.2.

2.1.2 Optimization

To solve the SVM problem, we must solve the primal Lagrangian

$$\mathcal{L}_p = \min_{\vec{w},b} \ \frac{1}{2}\|\vec{w}\|_2^2 + C\sum_i \xi_i - \sum_i \alpha_i(y_i(\vec{w}\cdot\vec{x}_i + b - 1 + \xi_i)) - \sum_i \mu_i\xi_i \tag{2.6}$$

subject to

$$\alpha_i(y_i(\vec{w}\cdot\vec{x}_i + b - 1 + \xi_i)) = 0 \tag{2.7}$$

$$\mu_i\xi_i = 0 \tag{2.8}$$

where α_i and μ_i are Lagrange multipliers. The quantity ξ_i is called the error distance. The error distance is defined as the distance of a point from its corresponding class-margin boundary, i.e., the boundary a point must cross to become an error.

The constraints in Equations (2.7) and (2.8) are part of the Karush-Kuhn-Tucker (KKT) conditions. The KKT conditions are necessary and sufficient for $\vec{w}$, b,

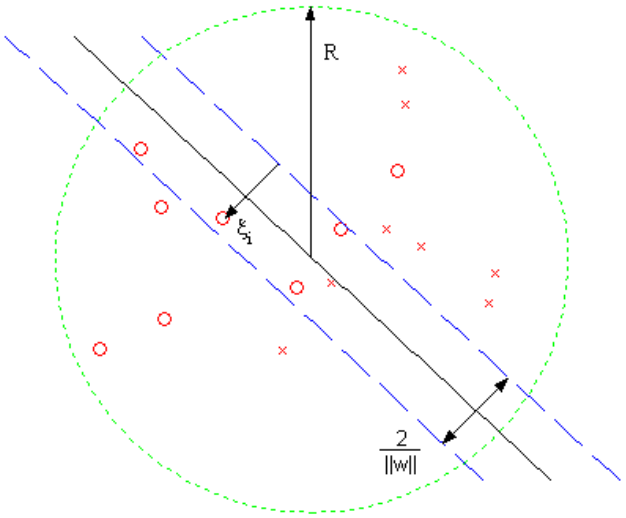

Figure 2.2 The SVM finds the optimal separating hyperplane by minimizing the ratio of R, the radius of a data-encircling sphere, to $\|\frac{2}{\vec{w}}\|$, the width between the separating margins. Tokens located between the margins are called "partial" errors and are located at a distance of ξ_i away from its corresponding class margin.

and $\vec{\alpha}$ to be a solution to the SVM problem. In fact, solving the KKT conditions is equivalent to solving the SVM problem! Maximizing Equation (2.6) gives

$$\sum_i \alpha_i y_i \vec{x}_i = \vec{w} \tag{2.9}$$

$$\sum_i \alpha_i y_i = 0 \tag{2.10}$$

$$C - \alpha_i - \mu_i = 0 \tag{2.11}$$

$$\vec{w} \cdot \vec{x}_i + b - 1 + \xi_i \geq 0 \tag{2.12}$$

$$\alpha_i \geq 0 \tag{2.13}$$

$$\mu_i \geq 0 \tag{2.14}$$

$$\xi_i \geq 0 \tag{2.15}$$

It is worth making a few observations about what the KKT conditions imply about the Lagrange multipliers and the error distance. When $\vec{x}_i$ is an error, then $\xi_i > 0$ by definition. Therefore, μ_i must be zero-valued to satisfy the condition in Equation (2.8). Equation (2.11) then requires that $\alpha_i = C$. If $\vec{x}_i$ has been labeled

Table 2.1 A summary of error distance values and the Lagrange multiplication coefficient values for the three different possible classifications of $\vec{x}_i$, correctly labeled, located on the margin, or incorrectly labeled. Tokens on the margin are technically errors according to the derivation in Section 2.1.1.

	ξ_i	α_i	μ_i
Correctly labeled	$= 0$	$= 0$	$= C$
On the margin	$= 0$	> 0	> 0
Error	> 0	$= C$	$= 0$

correctly, then $\xi_i = 0$. Because Equation (2.7) tells us that $\alpha_i = 0$ for any correctly labeled point, then it follows from Equation (2.8) that $\mu_i = C$. If $\vec{x}_i$ is actually located on the margin, then its distance from the margin is still zero. In this case, both $\alpha_i > 0$ and $\mu_i > 0$. Table 2.1 summarizes these results.

Because both the objective function and its constraints are convex, this quadratic programming problem is also a convex optimization problem. To solve $\mathcal{L}_p$, we must simultaneously minimize over $\vec{w}$ and b while requiring that all derivatives with respect to the α_i's and the μ_i's are equal to zero. However, because we know our optimization problem is convex, we can solve the dual problem by maximizing $\mathcal{L}_p$ over α_i and μ_i, and requiring that all derivatives with respect to $\vec{w}$ and b are equal to zero. Note that in the primal problem given by Equation (2.6), μ_i ultimately disappears from the SVM dual problem formulation because of the constraint in Equation (2.8). Therefore, we can ignore μ_i and optimize only over α_i.

Using the KKT conditions, the dual problem $\mathcal{L}_d$ is found to be

$$\mathcal{L}_d = \max_{\alpha_i} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_y y_j \vec{x}_i \cdot \vec{x}_j \tag{2.16}$$

$$= \max_{\vec{\alpha}} \vec{\alpha} \cdot \vec{e} - \frac{1}{2} \vec{\alpha}' \mathbf{Q} \vec{\alpha} \tag{2.17}$$

where $\vec{e}$ is a vector of 1's and $\mathbf{Q}$ is a Gram matrix. The element of $\mathbf{Q}$ in the ith row and the jth column, $q_{i,j}$, is $y_i y_j \vec{x}_i \cdot \vec{x}_j$.

Suppose, in the hopes of increasing the accuracy of our dichotomizer, we wish to divide the data using a nonlinear discriminant function. We can do so by first substituting Equation (2.9) into $g(\vec{x})$ to get

$$g(\vec{x}) = \sum_j \alpha_j y_j \vec{x}_j \cdot \vec{x} \tag{2.18}$$

Table 2.2 The four required properties a metric must satisfy. The vectors $\vec{a}$, $\vec{b}$, and $\vec{c}$ are arbitrary.

Nonnegative	$d(\vec{a},\vec{b}) \geq 0$
Reflex	$d(\vec{a},\vec{a}) = 0$
Symmetric	$d(\vec{a},\vec{b}) = d(\vec{b},\vec{a})$
Triangle Inequality	$d(\vec{a},\vec{b}) + d(\vec{a},\vec{c}) \leq d(\vec{b},\vec{c})$

Equation (2.18) shows us that $g(\vec{x})$ ultimately depends on only the dot product between a vector $\vec{x}$ and $\vec{x}_j$. $\mathcal{L}_d$ also depends on dot products - the dot products between the vector $\vec{x}_i$ and $\vec{x}_j$. Any function that has the form

$$d(\vec{x}_i,\vec{x}_j) = K(\vec{x}_i,\vec{x}_i) + K(\vec{x}_j,\vec{x}_j) - 2K(\vec{x}_i,\vec{x}_j) \tag{2.19}$$

and satisfies the properties listed in Table 2.2 is a metric, and the function $K(\cdot,\cdot)$ is called the kernel function. The kernel function can be decomposed into the product

$$K(\vec{x}_i,\vec{x}_j) = \Phi(\vec{x}_i)'\Phi(\vec{x}_j) \tag{2.20}$$

where $\Phi(\cdot)$ is an implicit projection into another feature space.

The properties of a dot product allow us to let $g(\vec{x})$ and $\mathcal{L}_d$ depend on the product in Equation (2.20). The equations for $g(\vec{x})$ and $\mathcal{L}_d$ can be rewritten as

$$g(\Phi(\vec{x})) = \sum_j \alpha_j y_j K(\vec{x}_j,\vec{x}) + b \tag{2.21}$$

$$\mathcal{L}_d = \vec{\alpha}\cdot\vec{e} - \frac{1}{2}\vec{\alpha}'\tilde{\mathbf{Q}}\vec{\alpha} \tag{2.22}$$

The $\tilde{q}_{i,j}$'s are now $y_i y_j K(\vec{x}_i,\vec{x}_j)$. It is useful that Equations (2.21) and (2.22) depend only on $K(\vec{x}_i,\vec{x}_j)$ because it means we do not need to estimate $\Phi(\cdot)$ directly. Direct estimation of $\Phi(\cdot)$ can be difficult and $\Phi(\cdot)$ may be infinite dimensional.

In addition to the linear kernel, which we used in the above derivation, there are three other commonly used kernels. They are the radial basis function

$$K_{RBF}(\vec{x}_i,\vec{x}_j) = \alpha_i y_i \exp^{-\gamma|\vec{x}_i-\vec{x}_j|^2} \tag{2.23}$$

the polynomial kernel

$$K_{poly}(\vec{x}_i,\vec{x}_j) = (\vec{x}_i\cdot\vec{x}_j + 1)^P \tag{2.24}$$

and the hyperbolic tangent

$$K_{tanh}(\vec{x}_i, \vec{x}_j) = \tanh(\kappa \vec{x}_i \cdot \vec{x}_j - \delta) \tag{2.25}$$

where γ, P, κ, and δ are predetermined parameters of their respective kernels.

Recall that when the optimal solution to $\mathcal{L}_d$ is found, some data tokens will have $\alpha_i \neq 0$. Such tokens can only be either errors or located on the margin (Table 2.1). Such tokens are called "support vectors."

2.1.3 Implementation

The convexity of the SVM objective function makes it an ideal candidate to be solved with standard quadratic programming (QP) methods. However, SVM training is computationally expensive. Some alternative quadratic programming methods have been derived and implemented in SVM software packages such as LibSVM [58] and SVM^{Light} [59].

LibSVM uses a modified version of the sequential minimal optimization (SMO) algorithm developed by Platt [60]. SMO solves the smallest possible QP problem during every iteration until the stopping criteria are met. In general, the smallest possible QP SVM-problem has two tokens. This is due to the linear equality constraints obeyed by the Lagrange multipliers. SMO chooses which set of tokens to optimize over (referred to as the working set) by selecting the maximal violating pair (MVP), the set of tokens that maximally violate the KKT conditions. LibSVM finds the working set using MVP selection methods described in [61]. SVM^{Light} uses a similar SMO-like training algorithm, but the working set consists of q tokens instead of two, where $q < N$.

In addition to the basic SMO training algorithm, the current version of SVM^{Light} implements additional algorithms to assess SVM generalization, train transductive SVMs, and learn ranking functions.

The $\xi\alpha$ algorithm [62] is one algorithm used by SVM^{Light} that can be used to obtain a biased measure of the generalization performance. The algorithm assumes that any token that satisfies the inequality $\rho\alpha_i R^2 + \xi_i \geq 1$ is an error. The total error

is $\frac{d}{N}$, where d is the number of tokens that satisfy the inequality. The parameter R is related to the kernel type and the value of ρ may be task dependent. For example, $\rho = 1$ provides a better generalization error estimate than $\rho = 2$ for the task of text classification.

SVM^{Light} also uses a second algorithm that provides a fairly unbiased estimate of the generalization error. In [63], the $\xi\alpha$ algorithm is adapted to estimate the generalization error using a leave-one-out error estimate. The algorithm assumes the data are processed in batches and that the data change over time.

Algorithms also exist to reduce the generalization error. Hastie et al. [64] propose finding the optimal error weight and kernel parameters by examining every possible regularization path by reformulating the SVM problem in terms of a loss-penalty formulation. Their proposed method fixes all variable parameters and finds the set of support vectors that minimizes the error for those parameters. In contrast to this guess and check method proposed by Hastie et al., Keerthi [65] proposes estimating the parameters from the Radius-Margin bound $f(\vec{w}) = \frac{1}{N}R^2\|\vec{w}\|$ using gradient methods.

Burges and Schölkopf [66] propose the use of virtual support vectors (VSVs) to reduce generalization error. VSVs exploit the fact that vectors in the training set are ignored ($\alpha_i = 0$) if they are not chosen to be support vectors ($\alpha_i > 0$). If all other vectors are removed from the training set and the SVM is retrained on only the support vectors (with ξ_i fixed), then the solution will be the same. VSVs take this concept a step further by generating "virtual" training examples based on estimated properties from the original training set. Handwritten digit classification experiments preformed in [66] confirm that VSVs do decrease the generalization error; however, they also dramatically increase the number of support vectors in the solution for large data sets.

An increased number of vectors in the supporting set increases the time it takes to process any test data. Burges and Schölkopf [66, 67] counter this adverse increase with a reduced set (RS) method. Recall from Section 2.1.2 that the normal vector $\vec{w}$ can be written $\sum_{i=1}^{N} \alpha_i y_i \Phi(\vec{x}_i)$. The vector $\vec{w}$ can be equivalently defined as

$\sum_{j=1}^{M} \sigma_j \Psi(\vec{z}_j)$. The set $\mathbf{Z} = \vec{z}_1, \vec{z}_2, \ldots, \vec{z}_M$ is defined as the reduced set with $M < N$ and can be found directly if so desired; however, gradient methods tend to provide more powerful and flexible solutions. In [67], experiments showed that the number of support vectors in a machine with a polynomial kernel can be reduced by a factor of 10 without a significant increase in the number of errors. The results in [66] confirm this and imply that the number of support vectors needed to maintain generalization accuracy can be decreased even further.

2.2 Speech Perception and Landmarks

In 1952, Jakobson, Fant and Halle suggested encoding each phoneme as a vector of binary "distinctive features": voiced vs. unvoiced, lowpass vs. highpass, spectrally compact vs. spectrally diffuse, etc. [68]. The idea that a phoneme can be decomposed into independently manipulable dimensions is quite old: classical Greek, Hebrew, Arabic, and Japanese, for example, mark secondary distinctions such as vowel length and consonant gemination (Arabic), voicing (Japanese), and syllable-initial aspiration or glottalization (classical Greek) by means of diacritics. The Hangul writing system, published by King Sejong of Korea in 1446 [69], independently encodes the place, manner, and voicing of every consonant: each consonant is composed of a fundamental symbol encoding place (labial, dental, alveolar, velar, or pharyngeal), modified by diacritics encoding manner and voicing. In 1876, the phonetician Alexander Bell proposed an international phonetic alphabet, capable of representing any place or manner distinction specified by any of the world's languages [70]. Bell's initial notation was based on a symbol encoding the place of the consonant, annotated by diacritics encoding manner and voicing, much like the Hangul system; because of the high cost of typesetting Bell's symbols, his notation was eventually replaced by an international consensus system called the International Phonetic Alphabet (IPA) [71]. Given the very long history of place-manner notation, the binary distinctive feature notation of Jakobson, Fant, and Halle was significant primarily for two reasons. First, their notation was the first to declare that all phonemic distinctions can be encoded in

a binary notation, as opposed to the N-ary place and manner distinctions proposed by Sejong and Bell. Second, their notation was important in part because, within three years after Jakobson's paper, Miller and Nicely were able to prove the psychological reality of a nearly binary distinctive feature notation similar to Jakobson's [2].

Miller and Nicely [2] asked listeners to transcribe noisy recordings of consonant-vowel syllables. Miller and Nicely compiled their results into confusion matrices, in which element (i, j) of the matrix shows the number of times that phoneme i was misrecognized as phone j. They found that human listeners rarely misidentified nonsense syllables under quiet listening conditions, but with enough noise, it is possible to get listeners to make mistakes, and the mistakes they make are revealing. First, some distinctive features are more susceptible to noise than others: place of articulation is reliably communicated only at SNR above -6 dB, while sonorancy is reliably communicated even at -12 dB SNR. Second, errors in the perception of distinctive features are approximately independent, in the following sense: given that the true values of the N distinctive features are $F = [f_1, \ldots, f_N]^T$, the SNR-dependent probability that a listener will perceive the vector $\hat{F} = [\hat{f}_1, \ldots, \hat{f}_N]^T$ is given by

$$p(\hat{F}|F, \text{SNR}) \approx \prod_{i=1}^{N} p(\hat{f}_i|f_i, SNR) \tag{2.26}$$

Equation (2.26) does not specify the dependence of distinctive feature errors on any particular acoustic signal. Several authors have suggested an implementation of Equation (2.26) that makes signal-dependence explicit in the following way, where X is the particular acoustic signal used to transmit feature vector F:

$$p(\hat{F}|X) = \prod_{i=1}^{N} p(\hat{f}_i|X) \tag{2.27}$$

Equation (2.27) is motivated by training considerations. Each feature has two possible settings ($f_i = 1$ and $f_i = -1$), so the feature vector F has 2^N possible settings. A classifier trained to represent $p(\hat{F}|X)$ must distinguish 2^N different labels, while a classifier trained to represent $p(\hat{f}_i|X)$ only distinguishes two labels; the former therefore typically requires 2^{N-1} times as much training data as the latter. Unfortunately, Equation (2.27) is incorrect in three ways. First, it is neither a necessary nor sufficient

condition for Equation (2.26). Second, it is suboptimal as an engineering system: a classifier trained to model $p(\hat{F}|X)$ directly, without factoring as shown in Equation (2.27), usually results in fewer errors than a bank of classifiers trained as in Equation (2.27). Third, it is not a correct model of human speech perception. Volaitis and Miller [72], for example, have demonstrated that a voice onset time (VOT) of 40 ms is sufficient to turn a synthesized /b/ into /p/, but that /g/ only becomes /k/ when the VOT passes 50 ms, i.e., $p(\text{voiced}|X, \text{labial}) \neq p(\text{voiced}|X, \text{palatal})$.

A somewhat better approximation of Equation (2.26) may be created by assuming that the perceived feature vector $\hat{F}$ is a deterministic function of the signal X; that is, assume that any given listener will always hear the same sequence of phonemes in response to a given acoustic signal. Specifically, choose any continuous function $G(X) = [g_1(X), \ldots, g_N(X)]^T$ that specifies the response pattern of listeners by the constraint $\hat{f}_i = \text{sgn}(g_i(X))$. If $G(X)$ is assumed to be a deterministic function, then Equation (2.26) is equivalent to

$$p(\hat{F}|F, \text{SNR}) \approx \prod_{i=1}^{N} \int_{\hat{f}_i g_i(X)>0} p(X|f_i, \text{SNR}) dX \tag{2.28}$$

The function $G(X)$ is, thus far, completely unconstrained, except that $\hat{f}_i = \text{sgn}(g_i(X))$ and Equation (2.28) holds. Given these constraints, it is possible to choose $G(X)$ such that the dimensions of $G(X)$ are conditionally independent, i.e.,

$$\int_{\hat{f}_i g_i(X)>0} p(X|f_i, \text{SNR}) dX = \int_0^\infty p(g_i(X)|f_i, \text{SNR}) dg_i \tag{2.29}$$

where the limits of the right-hand integral are $(0, \infty)$ as shown if $\hat{f}_i = 1$, and $(-\infty, 0)$ if $\hat{f}_i = -1$.

By combining Equations (2.28) and (2.29), a parsimonious speech sound classifier is produced. The classifier consists of two functions: a class-independent multi-dimensional transform $G(X)$, and a set of class-dependent scalar PDFs $\hat{p}(g_i(X)|f_i)$. The task of a human learner, or of a mathematical model of human speech perception, is to learn functions $G(X)$ and $\hat{p}(g_i(X)|f_i)$ that optimally approximate the unknown PDF $p(X, F)$.

Equation (2.29) suggests that the problem of speech sound classification is really, in some sense, a problem of acoustic-to-perceptual speech sound transformation.

But what is the transformation? Is it linear, or nonlinear? Is it learned or innate? Again, the answers to both questions are provided by the speech perception literature.

The ability of listeners to discriminate two nearly identical synthesized speech waveforms (e.g., identical except for a 50 Hz difference in the second formant) is highest if the two waveforms straddle a phoneme boundary (e.g., if one waveform is classified as /iy/ while the other is classified as /ih/). Kuhl et al. [73] have demonstrated that the phoneme boundary does not need to lie between the two waveforms in order to increase their discriminability: two waveforms that are both classified as /iy/, but that are both close to the /iy/-/ih/ boundary, are more discriminable than are two waveforms that are both close to the center of the /i/ region in acoustic space. They explain their results by positing a continuous-valued "perceptual space" computed by the listener as a nonlinear transformation of the acoustic space, $G(X) = [g_1(X), \ldots, g_N(X)]$, such that the magnitude of the Jacobian of the transform is smaller near the center of a phoneme region than it is near the border between phoneme regions [74]. These variations in the value of the Jacobian they term the "perceptual magnet effect." The proposed perceptual space $G(X)$ is controversial, but continues to serve as an organizing paradigm for new experiments, e.g., [75].

Listeners do not need to hear all of the acoustic evidence for a distinctive feature in order to correctly recognize the feature setting. Phoneticians have catalogued a handful of primary acoustic correlates (characteristic spectrotemporal patterns) that may be used to signal the setting of each distinctive feature. A signal synthesized with any one of these acoustic correlates will be heard to have the target distinctive feature. Consider, for example, the word "backed." This word contains three stop consonants; because of their relative positions in the word, the places of articulation of these three stops are communicated by three very different types of acoustic information. The place of the final /d/ is communicated by a turbulent burst spectrum. The place of the /k/ is communicated by formant transitions during the last 70 ms of the vowel. The place of the initial /b/ is communicated by both a turbulent burst and by formant transitions during the first 70 ms of the vowel, but experiments with synthetic speech [76] and digitally modified natural speech [77] have shown that ei-

ther of these cues may be excised without impairing listeners' ability to understand the stop. The closure transition, burst spectrum, and release transition of a stop are thus redundant acoustic correlates; unambiguous presence of any one of these three acoustic patterns is enough to force listeners to hear the desired distinctive feature.

The redundancy principle operates under at least two circumstances. First, one or more acoustic correlates may be missing because of syllable position, as in the example word "backed." Second, one or more acoustic correlates may be inaudible because of noise. When all acoustic correlates are masked by noise, listeners forced to guess the identity of a stop will choose a place of articulation at random. When the noise is lowered sufficiently to unmask either the burst peak or the formant transition, recognition accuracy rapidly approaches 100% [78].

The three sample acoustic correlates discussed above—closure transition, burst spectrum, and release transition—share an important characteristic. All three can only be correctly recognized using a signal representation precisely synchronized with an acoustic-phonetic "landmark": an instant of sudden signal change, e.g., a consonant closure or consonant release. The mammalian auditory system is uniquely sensitive to sudden onsets and sudden offsets of signal energy [79, 80]. Stevens [35] and Stevens et al. [36] have proposed a "landmark-based" model of speech perception and recognition, according to which acoustic phonetic landmarks proposed by a preprocessor are then classified by a set of distinctive feature classifiers. Redundancy of asynchronous acoustic observations occurs because landmarks are only classified if they are first detected by the preprocessor; thus, if X_1 is a sequence of spectra covering a 140 ms period centered at the instant of stop closure, X_2 is a sequence of spectra centered at the stop release, and $\mathcal{X} = [X_1, X_2]$ is their union, then

$$p(\mathcal{X}|F) = \begin{cases} p(X_1|F) & \text{if only closure exists} \\ p(X_2|F) & \text{if only release exists} \\ p(X_1|F)p(X_2|F) & \text{if both exist} \end{cases} \tag{2.30}$$

Humans and machines recognize consonants on the basis of acoustic cues present just after consonant release, and just before consonant closure; acoustic spectra during the closure interval itself provide little phonetic information [81]. Stevens

et al. have proposed [36] that consonant closures and releases, as well as syllable peaks and dips, compose a series of "acoustic landmarks" around which human and automatic speech recognition may be organized. Detection of these landmarks provides two sets of cues to a human or automatic speech recognizer: (1) detected manner-change landmarks specify the manner of articulation (stop, nasal, fricative, glide, vowel) of the phonemes, and (2) manner-change landmarks can be used to synchronize classifiers that seek to identify place and voicing.

Stevens proposed four types of landmarks: consonant releases (release of a nasal, stop, or fricative consonant into a vowel or glide), consonant closures, syllable nuclei, and intersyllabic energy dips. The four landmarks proposed by Stevens can be interpreted as the four synchronization points in a typical syllable: the onset, the nucleus, the offset, and the dip. A number of speech perception and neurological studies have shown that syllable counting is a perceptual skill that is distinct from and perhaps a necessary prerequisite for speech perception. Siok et al. demonstrated, using fMRI, that syllable counting and phoneme recognition are performed using different brain regions [82]. Jusczyk et al. [6] have shown that, within the first 24 hours of life, infants are capable of discriminating their native language from other languages on the basis of syllabic prosody, apparently because they have learned the prosody of their native language while still in the womb. By about 6-8 months of age, infants begin to segment and recognize individual words in their native language, but only if the words are produced using characteristic prosody (trochaic for English, iambic for French); by 10 months of age, infants become capable of segmenting words using other cues such as phonotactics [5]. Finally, there is some evidence that human speech perception may employ a coarse-to-fine recognition algorithm, in which mistakes in syllable-counting sometimes preclude correct recognition of the fine phonetic detail. Warren et al. have demonstrated a "vowel sequence illusion" suggesting that listeners are unable to correctly recognize the phonemes in an utterance unless they are also able to correctly syllabify the utterance [83]. Steady-state vowels, spliced together into a repeating sequence, are easily recognized if each vowel segment is long enough to be a naturally spoken syllable. If the vowel segments are too short to be natural

syllables (e.g., 70 ms), listeners fail to hear the correct vowels. Instead, listeners hear the signal as a recording of two talkers speaking simultaneously, each talking at a plausible English syllable rate, with phoneme content suggesting that listeners are attributing energy in the high band (above 1500 Hz) to one talker, and are attributing energy in the low band (below 1500 Hz) to the second talker.

2.3 Distinctive Features

Distinctive features [84] allow for an economical way of classifying phone segments and also allow for a better understanding of allophonic variation. Each phone can be classified by a unique set of binary valued (either positive (+) or negative (-)) distinctive features. There are two categories of distinctive features, articulator free and articulator bound.

An articulator free (manner) feature is a parameter of phonological structure that encodes a perceptually salient aspect of speech production. The five manner features we are primarily concerned with are [silence, continuant, sonorant, syllabic, consonantal]. The feature [silence] specfies whether a sound was created by the human vocal apparatus ([-silence]) or whether it is silence or other ambient noise ([+silence]). [Continuant] describes the airflow through the oral cavity. A phone that is [+continuant] is made with air flowing through the mouth. [Sonorant] determines how resonant a phone is. [+Sonorant] denotes loud, continuous voicing. A [-sonorant] sound is produced with an oral obstruction that raises the air pressure in the vocal tract, impeding vocal fold vibration. [+Syllabic] sounds are those that can occur in the nucleus of a syllable. [Consonantal] determines if there is a narrow constriction in the oral cavity ([+consonantal]). Manner features allow phones to be grouped into broad class categories such as vowels, glides, nasals, stops, fricatives, and others. Table 2.3 lists the value of the manner features for each phonetic class considered in experiments described in Chapter 4.

An articulator bound (place) feature is a parameter that describes a physical, articulator-dependent aspect of human speech production. Place features that can

Table 2.3 Every phonetic category (listed in column 1) can be described in terms of a set of binary (+ or -) valued manner features. A blank space indicates that a manner feature is undefined for a given class of sounds.

	Silence	*Continuant*	*Sonorant*	*Syllabic*	*Consonantal*
Nonspeech	+				
Flap	-	-	+	+	+
Fricatives	-	+	-	-	+
Liquids	-	+	+	-	-
Nasals	-	-	+	-	+
Closures	-	-	-	-	+
Releases	-	+	+	-	+
Syllabic Liquids	-	+	+	+	-
Syllabic Nasals	-	-	+	+	+
Vowels	-	+	+	+	-

be sensibly defined for a phone are manner dependent; i.e., (most) different manner classes will have different place features.

Nasals ([+sonorant, -continuant]) can be characterized by the features [alveolar, labial, palatal]. [Alveolar] sounds are those that are made by pressing the tongue blade to the back of the alveolar ridge, as in the nasal /n/. [Labial] sounds are created by pressing the lips together. The sound /m/ is the labial nasal. Finally, [palatal] sounds are made by pressing the tongue body to the hard or soft palate. Many languages distinguish between hard palate and soft palate constrictions, calling the latter "velar." English /ng/,/g/, and /k/ constrictions may be produced in either place. The palatal nasal is /ng/.

Syllabic nasals are nasals that occur in the nucleus of a syllable, just as a vowel normally does. Syllabic nasals are defined by the same place features as nasals. The /en/ sound at the end of the word "button" is a syllabic nasal. The other syllabic nasals are /em/ and /eng/.

Like nasals, stop closures and stop releases ([-sonorant, -continuant] and [-sonorant,+continuant], respectively) are also characterized by the features [alveolar] (/tcl/, /dcl/, /t/, /d/), [labial] (/pcl/, /bcl/, /p/, /b/), [palatal] (/kcl/, /gcl/, /k/, /g/). In addition, stop closures and releases are also classified by the feature [voice]. Voiced sounds ([+voice]) are those that are made with the vibration of the vocal folds.

The sounds /p/, /t/, and /k/ are the unvoiced stop releases whereas /b/, /d/, and /g/ are the voiced stop releases. Unvoiced and voice stop releases are preceded by unvoiced and voiced closures, respectively. The unvoiced closures are /pcl/, /tcl/, and /kcl/. The voiced closures are /bcl/, /dcl/, and /gcl/. Stop closures and stop releases will be referred to as closures and releases, respectively.

There is some uncertainty as to how to deal with closures. Closures are produced by completely obstructing the oral and nasal cavities and some researchers, such as Juneja [55], consider the stop closure to be in the class of sounds that would fall into the [+silence] category. However, the vocal tract is a lossy system and audible sounds can sometimes radiate through the throat and face during the production of the closure. Therefore, the experiments reported in this thesis make the distinction between the stop closure and silence.

Fricatives ([-sonorant, +continuant]) can be described by the features [anterior, dental, labial, strident, voice]. [Anterior] fricatives are created with a constriction anterior to the alveolar ridge, such as /s/ or /th/. A phone with the feature [+dental] is realized by pressing the tongue against teeth. The phone /th/ is an example of a dental fricative. [Strident] fricatives are those that have an obstacle placed in front of the constriction in the vocal tract, thereby increasing the amplitude of the turbulent noise, as in the phone /z/. An example of a labial fricative is the sound /f/ and an example of a voiced fricative would be the sound /zh/.

The glides ([+continuant,+sonorant]) /w/, and /y/, and the liquids ([+continuant, +sonorant]) /r/ and /l/, are unique in that each is articulated with a different region of the oral cavity. Their place features are [+labial], [+palatal], [+rhotic], and [+lateral], respectively. Liquids and glides have similar acoustic characteristics; the difference between the two kinds of sounds is that liquids can be substituted for vowels in the nucleus of a syllable and glides can form diphthongs. We differentiate between liquids that occur in the nucleus of a syllable and those that do not by referring to the former as syllabic liquids. An example of a syllabic liquid is the /er/ sound in the word "bird." The other syllabic liquid is /el/. Because we make an explicit distinction between syllabic and nonsyllabic liquids, the nonsyllabic liquids and glides

are collectively referred to as liquids unless otherwise specified. The sound /h/ is sometimes considered to be a glottal fricative by linguists because its production is similar in nature to fricative sounds; however, the acoustic characteristics of /h/ are more similar to the acoustic characteristics of liquids than they are to the acoustic characteristics of the fricatives. We therefore consider /h/ to be a liquid. Despite its acoustic similarity to other liquid sounds, /h/ can neither be syllabic nor form diphthongs in English. The sound /h/ is defined by the feature [+glottal].

English flaps are made by quickly tapping the tongue against the alveolar ridge. There are three flaps in English. One is the sound /dx/ in the word "butter." The nasal flap /nx/ occurs in words such as "banner." Some English speakers also flap the rhotic sound in words such as "three." The rhotic flap is not considered in this work and is merged with the phone /r/. The flaps are [-continuant, +sonorant].

Vowels ([+continuant, +sonorant]) are defined by the features [advanced tongue root (ATR), constricted pharynx (CP), front, high, low, reduced, round, tense]. The features [front, low, high] describe the tongue body position during production of the vowel. A vowel with the feature [+ATR] is produced with a widened pharynx (/ey/ vs. /ih/). The sound /ae/ is produced with the pharynx constricted ([+CP]), whereas the vowel /ah/ is not ([-CP]). [Round] denotes lip rounding during vowel production. The vowel /uw/ is made with the lips rounded and is therefore considered to be [+round]. [Tense] vowels, like /aa/, are usually longer in duration, have a higher pitch and higher tongue position than lax ([-tense]) vowels, such as /uh/. Vowels that are [reduced] are generally unstressed, such as the schwa /ax/.

Diphthongs, or vowels merged with glides, are separated into their phonetic components. The diphthongs of English are /oy/ as in the word "boy," /ay/ as in the word "eye," and /aw/ as in the word "house." The diphthongs /oy/ and /ay/ can be divided into the vowel-glide pairs /ow/ and /y/ and /aa/ and /y/, respectively. The /aw/ sound can be decomposed into the phones /ae/ and /w/.

The place features for vowels are listed in Table 2.4. Consonant place features are given in Table 2.5. Syllabic and nonsyllabic liquids are listed in Table 2.6.

Table 2.4 The vowels in the dataset and their place feature values. A blank space indicates that the place feature is undefined for a given phone.

	ATR	CP	Front	High	Low	Reduced	Round	Tense
aa		+	-	-	+	-	-	+
ae		+	+	-	+	-	-	+
ah		-	-	-	+	-	-	-
ao		+	-	-	+	-	+	+
ax			-			+		
ax-h			-			+		
eh		-	+	-	+	-	-	-
ey	+		+	-	-	-	-	+
ih	-		+	-	-	-	-	-
ix			+			+		
iy	+		+	+	-	-	-	+
ow	+		-	-	-	-	+	+
uh	-		-	-	-	-	+	-
uw	+		-	+	-	-	+	+
ux	+		+	+	-	-	+	+

Table 2.5 The consonants in the dataset and their place feature values. A blank space indicates that a feature is undefined for a given phone.

	Dental	Labial	Anterior	Alveolar	Palatal	Nasal	Strident	Voice
pcl		+		-	-			-
tcl		-		+	-			-
kcl		-		-	+			-
bcl		+		-	-			+
dcl		-		+	-			+
gcl		-		-	+			+
p		+		-	-			-
t		-		+	-			-
k		-		-	+			-
b		+		-	-			+
d		-		+	-			+
g		-		-	+			+
f		+					-	-
v		+					-	+
th	+	-	+				-	-
dh	+	-	+				-	+
s	-	-	+				+	-
z	-	-	+				+	+
sh	-	-	-				+	-
zh	-	-	-				+	+
ch	-	-	-				+	-
jh	-	-	-				+	-
m		+		-	-	+		+
n		-		+	-	+		+
nx		-		+	-	+		+
ng		-		-	+	+		+
em		+		-	-	+		+
en		-		+	-	+		+
eng		-		-	+	+		+
dx		-		+	-	-		

Table 2.6 The liquids in the dataset and their place feature values.

	glottal	lateral	rhotic	palatal	labial
h	+	-	-	-	-
l	-	+	-	-	-
el	-	+	-	-	-
r	-	-	+	-	-
axr	-	-	+	-	-
er	-	-	+	-	-
w	-	-	-	+	-
y	-	-	-	-	+

CHAPTER 3

RELATED WORK

The power of artificial neural networks (ANNs) and SVMs is in their ability to learn complex nonlinear functions and achieve good generalization performance on unseen data. Neural networks are mentioned here in addition to SVMs because any neural network has an equivalent SVM formulation [85]. Both pattern classifiers have successfully been applied to speech.

3.1 Distinctive Feature Recognition

Many researchers have had success detecting various kinds of distinctive features using a variety of methods.

Esposito et al. [86] used time delay neural networks (TDNNs) to classify the place and voicing of stop consonants from the TIMIT corpus. The TDNNs were able to achieve >90% accuracy for every stop consonant with the exception of /p/ which was only classified correctly 80% of the time.

Niyogi et al. [87] compared the ability of linear and nonlinear SVMs and an HMM to detect stop consonants in continuous speech on the TIMIT corpus. The SVM input features were the log of the total signal energy, the log of the signal energy above 3 kHz, and spectral flatness. These acoustic features were calculated every millisecond. The HMM used MFCCs as its input features. ROC diagrams in [87] show that SVMs appear to be more suited to the task of stop detection than HMMs. Nonlinear SVMs outperform linear SVMs.

In [51] Niyogi and Burges determine what factors allow linear and nonlinear SVMs to excel at stop detection. The authors discuss how lower values of C (see Section 2.1.1) emphasizes model complexity whereas a larger C value places more emphasis on the training errors. The normalization of the SVM discriminant function is discussed for probability density estimation in lower dimensional spaces. The behavior of nonlinear kernels is likened to the perceptual magnet effect.

Ali et al. [88] built a system to classify fricatives from the TIMIT corpus. The system can distinguish between voiced and unvoiced fricatives with 95% accuracy. The distinction between silibant and nonsilibant fricatives can be made with 94% accuracy. Palatal fricatives are distinguished from nonpalatal fricatives with 99% accuracy. The three-way distinction between alveolar, dental and palatal fricatives can be made with 97% accuracy. Overall, the system can classify all eight individual fricatives with 90% accuracy.

Glass and Zue [89] detected nasals from confusable phones such as glides and weakly voiced fricatives with the goal of incorporating nasal detection into a speech recognizer. The authors were able to classify nasal regions in the acoustic wave with 83.5% accuracy when phonetic boundary labels are available. When trying to detect nasal regions with unknown boundaries, the classifier was able to identify 95% of the nasal regions; however, the ratio of false positives to correctly labeled regions was 2:1. The mistakes made by the classifier are highly systematic and context dependent.

Using SVMs, Pruthi and Espy-Wilson [90] classified nasals and semivowels from the TIMIT corpus. The distinction between prevocalic, postvocalic, and intervocalic sounds was made. The SVM used formant, spectral energy, and other acoustic measures as input. Prevocalic nasals and semivowels are classified with 90% and 88% accuracy, respectively. Postvocalic nasals and semivowels are both classified with around 95% accuracy. Intervocalic nasals and semivowels are classified with 89% and 82% accuracy, respectively.

King and Taylor [91] proposed building a system to detect phonological features and using that information in a speech recognition system. In [91], the first part of this goal is accomplished using TDNNs. Different sets of phonological fea-

tures are chosen and TDNNs are used for detection. All TDNNs are able to detect their respective feature with accuracy that is well above chance.

3.2 ANN and SVM Systems for Speech Processing

ANN-only and SVM-only systems have been applied to some speech recognition tasks.

Harrison [92] proposed a network architecture for continuous phone recognition in isolated nonsense words. First, a three-layer network classifies and transforms sequences of phones. These output sequences are input to a second network that determines the spoken phone sequence. Of the 15 distinct sounds in the dataset (14 phones and silence) recognition error was greater than 25% for only word-initial /b/, word-final /p/, and word-final /t/.

Robinson [93] trained a recurrent ANN to classify phones from the TIMIT corpus. Using all 61 phonetic distinctions presented in TIMIT, the recurrent network achieves an error rate of 30.7%. Robinson compares his results to other recognition systems based on an HMM paradigm, but some of these systems have fewer than 61 tokens in their respective datasets because they have combined similar phones.

ANN systems have been applied to small isolated word recognition tasks such as the task of distinguishing the word "yes" from the word "no" for dysarthric speakers [94]. In this task, the ANN achieves a 0% error rate when the test data and training data are spoken by the same subject. Due to the variability across speakers with dysarthria, the ANNs performance degrades significantly for some unseen dysarthric speakers, but not others.

Iso and Watanabe [95] used a sequence of multilayer perceptrons (MLPs) to perform speaker-independent word recognition. Each MLP was trained to distinguish a single word from every other word in the data. The data consisted of a 10-digit vocabulary spoken by 107 different speakers. The sequential MLP system achieved a 0.2% error rate.

In [96], Yousafzai et al. use the acoustic wave and PLP features in combination with SVMs to examine the robustness of these two acoustic representations to Gaussian noise corruption. They find that the SVM has a higher phone recognition rate for clean speech than it does for noisy speech when using PLPs. Conversely, the SVM performs poorly on clean speech when using just the acoustic wave, but it does relatively better in noisy speech.

ANNs and SVMs have also been applied to other speech and language related tasks such as speaker recognition [97, 98], part of speech tagging [99, 100], text-categorization [101, 102, 103], and speech emotion recognition [104, 105].

3.3 Hybrid Speech Recognition Systems

Many researchers have proposed and implemented various methods that combine ANNs with HMMs. These systems have been used successfully to recognize words, strings, and phones. Like the SVM, the ANN can be used to approximate accurate representations of acoustic (or other) data of overdetermined problems, but both the ANN and SVM by themselves are unable to represent temporally dynamic systems. The HMM, however, does provide a good model of temporally dynamic systems, such as speech. The goal of hybrid systems is to combine the ANNs power of mathematical function representation with the HMM's ability to accurately portray events from systems that change as a function of time. An overview of ANN/HMM hybrid systems for speech along with several examples of different architectures and implementations used to recognize words and strings can be found in [106]. A brief discussion of the advantages and disadvantages of traditional HMM-only speech recognizers is also provided.

Like the SVM/HMM hybrid system described in Section 4.4, the system designed and implemented by Bengio et al. [30, 33] is used for phone recognition. In [30], the authors developed an ANN training algorithm that finds the global, rather than local, optimum of an objective function using maximum likelihood estimation (MLE). In experiments using their ANN/HMM hybrid [33], Bengio et al. trained

three individual ANNs that could either determine the manner of articulation of a phone, distinguish between stop consonants and nasal consonants, or identify fricatives. These three ANNs had average error rates of 17.7%, 25.4%, and 25.2% for their respective tasks. Using the global optimization algorithm from [30], Bengio et al. combined the three ANNs into a single ANN/HMM phone recognition system for read speech. The ANN/HMM phone recognition system improved 9% relative to an HMM-only system that the authors trained to perform the same task. In [30], the authors provide experimental results that support the claim that an ANN by itself provides for a poor model of speech dynamics.

Schwenk [107] built several baseline hybrid systems using ANNs to calculate the posterior probabilities of confusable sets of numbers and an HMM to decode the ANN outputs. Several kinds of features, including RASTA-PLP and modulation-spectrogram features, were used as the initial input to the ANNs. Schwenk also built ANN/HMM hybrids using syllable and frame level representations. The system that achieved the lowest word error rate (WER) of 5.3% on the TIMIT database combined Boosting [108, 109] with the ANN classifiers. A discussion about the use of Boosting on neural networks can be found in [110].

Traditionally, ANN/HMM hybrid systems use the ANN to calculate posterior probabilities of words, syllables, phones, or subphone units. These probabilities are then interpreted and decoded by the HMM. Like traditional hybrids, the Tandem System [111] also uses ANNs to calculate the posterior probabilities of the acoustic data. Unlike traditional hybrid systems, the Tandem System transforms the ANN output and uses the warped probabilities as input features for a Gaussian mixture model (GMM) based recognizer. The Tandem system is more robust than traditional hybrid systems in both clean speaking conditions and in noise.

Kirchhoff and Bilmes [112] deviate from traditional hybrid architecture by training several different ANNs on multiple acoustic representations of the same data. The ANN outputs are then combined and decoded. The system achieves a 5.4% error rate for telephone-band continuously spoken digits.

Robinson et al. [113] applied a recurrent ANN/HMM hybrid system for both the task of phone recognition on TIMIT and the task of word recognition on the Wall Street Journal (WSJ) Corpus. In [113], Robinson et al. describe three improvements they make to their baseline system. First, by adding context information to the recurrent network, the authors reduce the word recognition error of their system by 10%; the phone error rate (PER) improvement is insignificant. Secondly, by combining multiple models that use different acoustic representations of the input data, the authors were able to reduce their baseline WER by 25% and their baseline PER by 8%. Finally, by applying duration modeling, Robinson et al. are able to reduce the WER on WSJ by 25% and find no significant improvement in PER on TIMIT.

ANNs have been used in other kinds of hybrid systems. The isolated phone recognition system described in [114] combines ANNs with decision trees. This combination reduces the computational complexity of the ANN and reduces the error rate when the trees are allowed to share phonetic information.

ANNs have also been used in combination with dynamic Bayesian networks (DBNs). In [115], Kirchhoff used DBNs to detect phonological features. The information provided by both articulatory and acoustic features is analyzed to determine an optimal way to utilize both kinds of knowledge for speech recognition. Kirchhoff builds large vocabulary speech recognition systems for both the German and English languages and examines the usefulness of articulatory features when recognizing noisy speech.

Some researchers have previously used TDNNs instead of the hybrid architecture. TDNN-only speech recognition systems, such as those in [116, 117, 118], have shown an ability to accurately represent timing information and delays in unaligned data just as the HMM does. Like ANNs, TDNNs also provide accurate estimations of complex decision surfaces. Ahmadi et al. [119] compared the performance of TDNNs and ANNs for the task of voiced stop recognition on the TIMIT database. Both networks used input derived from a spectrogram image. The relative difference in recognition accuracy between the TDNN and the ANN on this task is 5.1%.

SVM/HMM hybrid systems have also been built. Smith and Gales [120] use HMMs to extract score-space features from a speech signal and then train an SVM to perform isolated word recognition. On utterances containing isolated letters from the ISOLET corpus, the hybrid system tends to outperform an HMM-only system for both ML and MMI training. In contrast, the system described in Chapter 4 uses SVMs to generate input features for the HMM.

CHAPTER 4

EXPERIMENTS

Section 4.1 describes the corpora used in the experiments discussed in this thesis. Section 4.3 describes the SVM training, and Section 4.4 describes the SVM/HMM hybrid system.

4.1 Corpora

The NTIMIT [121] corpus was used for most experiments described in this chapter. NTIMIT was constructed from the TIMIT [122] database by first filtering the original utterances through telephone channels and then high-pass filtering them. The original TIMIT database contains 6300 sentences that were collected from 630 different speakers, both male and female, from 8 different dialect regions of the United States. Each utterance is phonetically rich. The original TIMIT corpus contains detailed phonetic transcriptions. NTIMIT has been time aligned with TIMIT so the original phonetic transcriptions can be used with the NTIMIT data as well.

The NTIMIT phone transcriptions are converted into landmark transcriptions using a Perl script. The landmark transcriptions specify the instants when a phone of one manner class ends and a phone of a different manner class begins, i.e. the landmark times. The transcription also labeled the type of each landmark, i.e., the manner-change features of the landmark, the manner of articulation features, and

Text and some experimental results presented in this chapter have previously been published in [1] and [57].

the place of articulation features of the phones on both sides of the landmark. Vowel center landmarks were estimated and labeled as well. These modified transcriptions were used to extract landmarks for SVM training and testing. Figure 4.1 shows an example of how different landmarks correspond to the NTIMIT phonetic transcription, the spectrogram, and the acoustic waveform. The figure shows three example landmarks: a [-+consonantal] landmark (closure from the [-consonantal] phoneme /ow/ into the [+consonantal] phoneme /kcl/), a vowel center landmark (annotated with one of its place features: [+front]), and a [+-sonorant] landmark (closure from [+sonorant] /ih/ into [-sonorant] /tcl/).

A subset of the Switchboard conversational telephone speech corpus [123], WS97 [124], was used to test the SVMs generalization ability to new, but similar, data. Switchboard is a corpus of recorded spontaneous telephone conversations between two strangers. Though the speech was spontaneous, the topic of conversation was predetermined. Switchboard is interesting with regard to landmarks because landmarks and phonetic features often become altered or can even be deleted in conversational speech due to factors such as the speaker's rate of speech or less-careful articulation of words and individual sounds. Switchboard contains no phone-level transcriptions. WS97 was used because it contains manually labeled phonetic transcriptions. Unfortunately, WS97 is neither as large as nor as phonetically rich as the NTIMIT corpus.

4.2 Periodic Vector Toolkit

The Periodic Vector Toolkit (PVTK) [125] was used to extract the SVM training data from NTIMIT. It contains three tools: VTransform, VExtract, and VApplySVMs. The tools use the Hidden Markov Toolkit (HTK) [126] libraries, so many of the PVTK input files are similar in format to files that would be used with HTK.

VTransform is used to manipulate acoustic data. VTransform input data files are stored in the HTK HParm binary format; i.e., VTransform manipulates acoustic

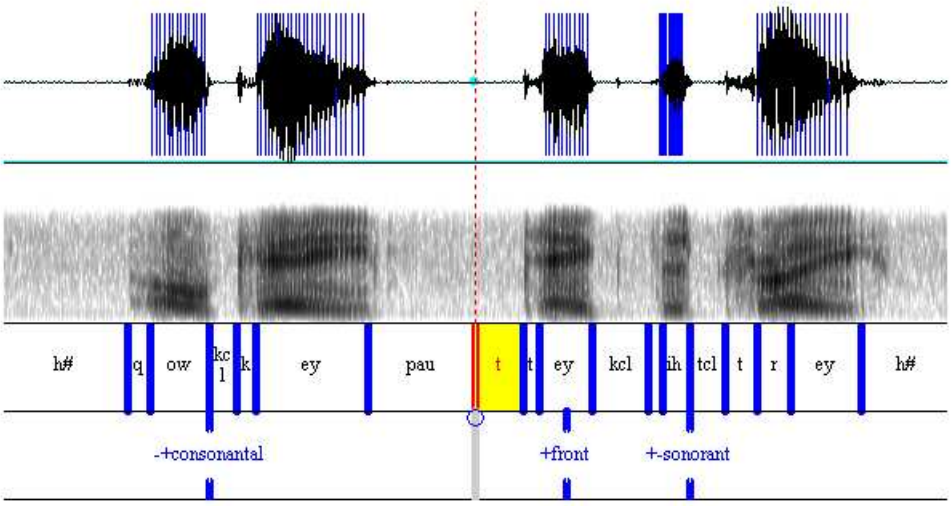

Figure 4.1 An example of how the acoustic wave, the spectrogram, the phonetic transcription, and landmarks are related. The utterance is "Okay, take the tray." Three example landmarks are shown: two stop closure landmarks and one vowel center landmark.

feature files identical to those generated using the HTK tool HCopy. VTransform can be used to easily combine files containing different acoustic features. VTransform also can apply mathematical transformations to frames of data. The output files generated by VTransform are also written in HParm format for use with the other PVTK tools or the HTK tools. In the experiments reported in this thesis, VTransform was used to concatenate corresponding frames in the different feature files; it was not used to alter the data any other way.

VExtract is the tool used to the extract the frames of data used for the SVM training input. VExtract uses both binary acoustic feature files and PVTK format transcription files to do this. The general format of a PVTK transcription is

$$t_{start} \quad t_{end} \quad description \quad token \tag{4.1}$$

If the *description* field contains more than one entry, then those entries must be separated by a comma with no spaces between any of the entries.

HTK uses transcription files of the format shown in Figure 4.2. VExtract uses a format similar to that in Figure 4.3. In both figures, the first two columns of

the transcription specify the start and end times, respectively. In the PVTK format transcription, the start and end times of the landmark are the same because landmarks occur at instantaneous points in the speech signal; however, the information provided at the landmark is not instantaneous. Acoustic cues for landmark classification may be found even as far away as 200 ms on either side of the landmark. VExtract does not necessarily need the start and end times to be the same. If the two times are different, VExtract will extract multiple frames between the two times. The number of frames extracted depends on the length of time associated with a transcribed symbol and the "Sample Period" parameter specified in the HTK binary data files. This is the same parameter specified by the TARGETRATE variable in an HTK configuration file.

The third column of an HTK format transcription and the fourth column of a PVTK format transcription specify what event happens at and in between the start and end times.[1] The third column of the PVTK format transcription provides additional information about the event transcribed in column four. In the case of acoustic-phonetic landmarks, the third column specifies any changes in manner features, manner features of both phones associated with the landmark, and place features of both phones associated with the landmark. In Figure 4.3, the third column contains only a small number of the phonetic features described in Section 2.3. In the transcriptions used in the experiments described below, all the phonetic features listed in Tables 2.4, 2.5, and 2.6 are contained in the transcription. Due to the lack of space, all but a few have been eliminated from the figure.

In Figure 4.2, there are two kinds of entries in the fourth column. The first kind are in the form $phn1:$phn2. Both $phn1 and $phn2 can represent any phone in the dataset. The ":" represents a transition between $phn1 and $phn2. The second kind of entry in column four is of the form "$phn1." This time, $phn1 represents a landmark located in the nucleus of a phone; i.e., this kind of entry represents a vowel

[1]HTK format transcription labels can have multiple tiers that can specify phones, syllables, words, and other segment labels for a given start and end time. More details can be found in [126], but this thesis briefly mentions HTK transcriptions to compare and contrast with the similar PVTK format.

```
"*/MFGK0SX214.lab"
0          925000     h#
925000     1375000    dh
1375000    1827500    ax
1827500    2663125    m
2663125    3846250    ao
3846250    4438125    r
                ⋮
```

Figure 4.2 An example of a phonetic transcription used by HTK. Time is in units of 100 ns. The first column contains start times. The second column contains stop times. The third column contains the phone uttered by the speaker at and between those times. The transcription is of the beginning of the utterance, "The morning dew on the spider web glistened in the sun."

```
"*/MFGK0SX214.lab"
925000     925000     +-silence,+consonantal,+continuant     h#:dh
1375000    1375000    +-consonantal,+anterior,-lips          dh:ax
1601250    1601250    +syllabic,-front,+reduced              ax
1827500    1827500    +-continuant,+sonorant,+labial         ax:m
2663125    2663125    -silence,+-consonantal,+nasal          m:ao
3254687    3254687    +syllabic,+CP,-high                    ao
3846250    3846250    +continuant,+sonorant,+-syllabic       ao:r
                                ⋮
```

Figure 4.3 An example landmark transcription for use with the PVTK tool VExtract. The first and second columns are the landmark start and stop times, respectively. The third column of the transcription provides information about the landmark. The fourth column specifies the actual phones that make up the landmark. The transcription is of the utterance, "The morning dew on the spider web glistened in the sun."

center landmark. In the figure, the transitions between phones are lines 1, 2, 4, 5, and 7 of the transcription. The vowel center landmarks are in lines 3 and 6. Notice that phone transition times can be taken directly from the phonetic transcription in Figure 4.2. The vowel center landmark times are estimated.

VExtract searches both the *description* field and the *label* field for one or more user provided class-specific patterns. Patterns can represent necessary or sufficient conditions for a token to be included in a given class. As an example, it is necessary for a transition from a fricative or a stop closure to any other class of sound listed in Table 2.3 to be [-+sonorant]. A "-+" indicates a transition from a sound that has the

property [-$feature] to a sound that has the property [+$feature]. Any vowel phone is sufficient to cause the transition to be marked +continuant.

VExtract can be used to extract tokens from multiple classes at the same time. The output produced by VExtract is formatted to be used with LibSVM, SVM^{Light}, and any other SVM package that uses the same input format. VExtract also provides the option to write its output in HParm format.

VapplySvms can read in one or more SVM definition files in the format generated by LibSVM or SVM^{Light}. VApplySvms applies each SVM to each frame in a user-specified list of data files. The input data files must be in HTK binary data format. The output of VApplySvms will also be in HTK binary format where each frame is a vector of SVM discriminant values. All HParm file header information is preserved, with the exception that the data code (specified by the TARGETKIND variable in the HTK configuration file) is changed to "USER." VApplySVMs also saves its output as an HTK compressed file (specified by "SAVECOMPRESSED = T" in the HTK config file).

4.3 SVM Training

The systems tested in this thesis make use of both landmark detection and landmark classification SVMs.

Landmark detection SVMs are trained to compute a univariate discriminant function $g_f(\vec{x}_t)$ where $\vec{x}_t$ is the observation cepstrogram centered at time t, such that $g_f(\vec{x}_t)$ optimally discriminates between the cases that a landmark of type f exists ($\delta(f,t) = 1$) or does not exist ($\delta(f,t) = -1$) at time t.

Landmarks are places of sudden signal change that correspond to a change in manner features, such as a consonant closure or a consonant release. Landmark detection SVMs are trained for the landmarks f that correspond to changes in one or more of the manner transition features listed in Table 4.4 (p. 44). Phrased another way, $g(\vec{x}_t)$ is trained to be a minimum structural risk univariate summary of $\vec{x}_t$; i.e., the minimum structural risk classifier, given $\vec{x}_t$, is $\hat{\delta}(f_i, t|\vec{x}_t) = \text{sign}(g_f(\vec{x}_t))$.

Landmark classification SVMs are trained to be conditionally discriminative for a given feature f_i, given the existence of other context features f_j; i.e., $\hat{\delta}(f_i, t|\vec{x}_t, f_j) = \text{sign}(g_i(\vec{x}_t))$. Landmark classifiers label place, voicing, and vowel quality features given the presence of a landmark. A list of our landmark class classification SVMs, together with their context features, is given in Tables 2.4, 2.5, and 2.6.

4.3.1 SVM training and acoustic feature selection

Each landmark detector was trained on a total of 13 000 training tokens: 6500 positive examples and 6500 negative examples. A positive example is a frame that includes the desired landmark, whereas a negative example would then be any other frame that does not include the landmark. Each manner detector was trained on acoustic feature vectors $\vec{x}_t$ containing 11 concatenated frames of acoustic information. The first frame was sampled at 50 ms before the landmark, the 6th frame was sampled at the landmark time, and the 11th frame was sampled at 50 ms after the landmark; i.e., $\vec{x}_t \equiv [\vec{y}_{t-50}, \ldots, \vec{y}_t, \ldots, \vec{y}_{t+50}]$ where $\vec{y}_t$ included a concatenation of different acoustic features that were calculated at each frame.

Place classification SVMs were trained using the maximum number of positive and negative samples available. While NTIMIT is a phonetically rich corpus, some phonetic distinctions are better represented than others. Positive place information was taken from the release or closure of consonants or from the beginning of vowels exhibiting the desired positive place features. Negative place information is taken from the release or closure of phones with the desired negative value for a given place feature.

Different acoustic features provide different information. For example, PLPs provide perceptual information, whereas MFCCs provide a more noise-robust frequency representation using the discrete cosine transform of log filterbank amplitudes. Some information is redundant between different acoustic feature representations. Each set of acoustic features added to $\vec{x}_t$ also increases computation time. For these reasons, manner detection and place classification SVMs were trained using a variety of acoustic feature combinations. Acoustic features were selected from

PLPs, MFCCs, delta coefficients (for both PLPs and MFCCs), acceleration coefficients (for both MFCCs and PLPs), formants from Zheng and Hasegawa-Johnson's formant tracking system [127], and knowledge-based acoustic parameters (APs) [128]. Energy and zero-crossing rate were also used as features for fricatives [129]. Not all combinations of acoustic features were tried. The "best" combination of features is the MFCCs, formants, and APs. The results of the feature selection experiments are given in Tables 4.1 and 4.2. Feature selection experiments were not performed for the landmark detectors.

The ability of the SVMs to utilize the spectral information provided by both PLPs and MFCCs is also explored. Table 4.3 compares the accuracy of SVMs trained using MFCCs and PLPs for the task of place detection. Both MFCCs and PLPs include their delta and acceleration coefficients. The performances of linear and RBF SVMs are also compared for each task.

Manner information is available to humans up to 200 ms before and after the landmark, depending on the characteristics of the sound. Experiments were done for some sounds to determine the best number of frames to concatenate. These experiments were not tried for all feature combinations. The acoustic features were determined first and then the number and location-in-time of the frames to be concatenated was determined. The number of frames concatenated in the SVM input vector was determined on the TIMIT corpus in previous experiments that are not reported here.

In general, most place classification SVM observation vectors could contain 7 frames ($\vec{x}_t = [\vec{y}_{t-60}, \vec{y}_{t-50}, \ldots, \vec{y}_t]$) or 11 frames ($\vec{x}_t = [\vec{y}_{t-50}, \ldots, \vec{y}_t, \ldots, \vec{y}_{t+50}]$). Stop closure place detection SVMs achieved the highest accuracy using vectors of 20 concatenated frames ($\vec{x}_t = [\vec{y}_{t-100}], \vec{y}_{t-90}, \ldots, \vec{y}_t$). As mentioned earlier, the SVM observation vector for manner classification also has the 11 frame format. The total number of features per frame is 90. The 11-frame vector has a total dimension of 990 features, the 7-frame vector has a total of 630 features, and the 20-frame vector has a total of 1800 features.

Table 4.1 The SVMs were trained on NTIMIT using different sets of acoustic features. The accuracies of the place classification SVMs for this experiment are listed. MFCC (MFC) or PLP features also included their first and second derivatives. An F means that formants were included. ZCR is the zero-crossing rate and energy. AP indicates that the APs were included.

	MFC F AP	MFC F	MFC ZCR	MFC F ZCR	PLP F
Release					
Alveolar	85.27	84.46	85.10	84.42	84.44
Labial	87.60	87.75	88.66	87.70	88.70
Velar	90.13	88.83	89.80	88.85	88.46
Voice	86.71	85.29	85.78	85.41	84.61
Nasal					
Alveolar	83.75	84.34	84.90	83.93	86.85
Labial	88.32	86.34	86.61	86.43	87.91
Nasal	97.25	93.42	93.08	93.32	93.28
Velar	97.80	98.29	98.52	98.29	98.56
Fricative					
Anterior	95.40	94.10	94.35	94.13	94.04
Dental	94.53	94.77	94.87	94.84	94.98
Labial	87.03	85.70	86.89	85.86	86.24
Strident	89.89	88.64	89.21	89.15	88.85
Voice	89.59	89.04	89.46	89.34	87.49
Liquid					
Glottal	96.22	94.84	95.28	95.04	95.36
Lateral	90.24	88.31	89.23	88.35	87.54
Rhotic	92.99	92.78	93.46	92.67	93.02
Labial	91.67	91.53	91.79	91.72	92.08
Palatal	96.76	96.55	96.90	96.50	96.86
Closure					
Alveolar	74.53	73.28	74.12	73.66	79.45
Labial	82.71	84.99	85.34	85.01	88.85
Velar	80.01	77.73	78.75	78.24	82.78
Voice	69.05	74.17	75.13	74.28	72.95
Vowel					
ATR	85.60	82.57	82.72	83.05	82.06
CP	85.62	84.70	84.98	84.52	84.72
Front	87.47	88.13	88.18	88.23	88.13
High	92.33	90.88	91.18	90.98	91.40
Low	88.63	88.40	88.80	88.35	88.20
Reduced	86.07	85.18	85.08	85.08	85.15
Round	97.55	94.13	94.00	94.22	93.06
Tense	83.15	81.63	81.90	81.73	79.70

Table 4.2 The SVMs were trained on NTIMIT using different sets of acoustic features. The accuracies of the manner classification SVMs for this experiment are listed. MFCC (MFC) or PLP features also included their first and second derivatives. An F means that formants were included. ZCR is the zero-crossing rate and energy. AP indicates that the APs were included.

	MFC F AP	MFC F	MFC ZCR	MFC F ZCR	PLP F
Closure	94.75	94.30	94.75	94.38	94.75
Flap	96.66	94.79	95.27	94.66	95.34
Fricative	94.18	93.18	93.58	93.35	93.85
Liquid	87.10	84.15	84.78	83.75	84.88
Nasal	92.83	90.93	91.68	91.13	91.60
Release	95.92	94.40	95.15	94.43	94.53
Syllabic Liquid	89.93	88.35	89.13	88.68	89.03
Syllabic Nasal	98.15	97.57	98.00	97.52	97.47
Vowel	94.13	92.28	93.30	92.83	93.23

4.3.2 Landmark detection and classification using SVMs

The accuracy of the 10 SVMs trained to classify different landmarks is listed in Table 4.4. In addition to the landmark classifiers, a set of manner classifiers were also trained. The accuracies of the 9 manner classifiers are listed in Table 4.5. Place of articulation detection accuracies are listed in Table 4.6.

Because the majority of the SVMs had several hundred, even several thousand, support vectors, the reduced set (RS) method described in Section 2.1.3 was used to reduce the number of support vectors by 90%. For some of the SVMs, as much as 99% of the support vectors were able to be eliminated without a significant change in accuracy. RS SVM accuracies are also given in Tables 4.4, 4.5, and 4.6. Unless otherwise specified, the RS SVMs were used for all experiments.

The ROC curves of the landmark detectors and the manner classifiers are given in Figures 4.4 – 4.8. These ROC curves plot percentage insertion vs. percentage deletion of frames of the given manner class (Figures 4.4 and 4.5 and the top left of Figure 4.6) or landmarks of a specified manner transition type (the top right, bottom left, and bottom right of Figure 4.6 and Figures 4.7 and 4.8), when a landmark is declared every time that $g_f(x_t)$ exceeds some fixed threshold. The balance between insertions and deletions is changed by adjusting the threshold. As shown, all SVMs

Table 4.3 In general, SVMs seem to detect place information more accurately using MFCCs (MFC) instead of PLPs. RBF SVMs are more accurate than linear SVMs at classifying acoustic input. The MFCCs and PLPs also included their first and second derivatives.

	MFC		PLP	
	RBF	Linear	RBF	Linear
Stop				
Alveolar	85.2	76.9	84.6	77.4
Labial	88.5	83.5	87.0	83.7
Palatal	88.9	81.5	87.0	80.3
Voice	85.3	82.9	84.4	83.0
Nasal				
Alveolar	79.4	74.8	77.3	73.5
Labial	85.4	81.3	82.7	83.5
Flapped	97.8	96.6	98.0	96.5
Palatal	95.6	94.8	95.5	94.8
Fricative				
Anterior	94.2	93.8	94.3	92.8
Dental	89.0	87.1	88.5	87.5
Labial	84.2	79.0	83.7	78.6
Strident	88.6	86.7	88.7	86.9
Voice	83.5	82.0	83.6	83.3
Liquid				
Glottal	94.1	91.3	93.4	91.0
Labial	87.5	83.7	86.9	83.4
Lateral	90.3	87.8	90.4	88.1
Palatal	92.0	87.4	91.6	87.8
Rhotic	89.3	85.9	89.9	85.4
Vowel				
ATR	85.6	80.8	84.4	79.2
CP	85.6	82.6	85.3	82.3
Front	87.5	85.9	87.0	85.8
High	92.3	91.4	92.2	91.6
Low	88.6	87.5	88.9	87.0
Reduced	86.1	85.1	85.3	85.3
Round	97.6	97.0	97.2	96.8
Tense	83.2	77.5	81.8	77.0

Table 4.4 Accuracies of the landmark detection SVMs and RS SVMs. Chance is 50%.

SVM	ACC	ACC (RS)	SVM	ACC	ACC (RS)
-+Silence	91.5	90.5	+-Silence	92.2	91.1
-+Continuant	81.1	80.8	+-Continuant	81.5	81.7
-+Sonorant	79.8	79.5	+-Sonorant	82.2	82.0
-+Syllabic	88.0	79.65	+-Syllabic	78.5	78.4
-+Consonantal	78.8	78.5	+-Consonantal	74.7	74.1

Table 4.5 Accuracies of the manner of articulation classification SVMs and RS SVMs. Chance is 50%.

SVM	ACC	ACC (RS)	SVM	ACC	ACC (RS)
Closure	94.8	94.3	Release	95.9	95.3
Flap	96.7	96.2	Syllabic Liquid	89.3	89.3
Fricative	94.2	93.1	Syllabic Nasal	98.2	98.1
Liquid	87.1	85.4	Vowel	94.1	92.4
Nasal	92.8	91.8			

can achieve very low insertion rates, but few can achieve a deletion rate below a classifier-dependent minimum, often as high as 20% or 30%. The high number of deletions may be caused by sparsity of the training data; i.e., deleted landmarks are landmarks that fail to adequately resemble any of the support vectors. Of particular interest is the ROC curve of the syllabic nasal shown in the bottom right plot of Figure 4.5. The deletion rate is almost constant and very small. However, the insertion rate is high. A high insertion rate may indicate that a given phonetic feature may not be distinct given the acoustic features or number of training samples.

The SVMs are clearly able to make some distinctions in contexts in which they were not trained. The discriminant values are output in such a way that they are consistent given these new contexts. Information of this sort may be of use to the HMM to distinguish between phones.

4.3.3 Cross corpus generalization of the SVM

The SVM dichotomizer is designed to minimize the generalization error on unseen data.

Table 4.6 Accuracies of the place of articulation classification SVMs and RS SVMs. Chance is 50%.

SVM	ACC	ACC (RS))
Release		
Alveolar	85.3	85.0
Labial	87.6	85.8
Palatal	90.1	87.9
Voice	86.7	86.1
Nasal		
Alveolar	83.8	82.9
Flapped	97.3	95.4
Labial	88.3	87.5
Palatal	97.8	96.5
Fricative		
Anterior	96.4	95.7
Dental	94.5	94.3
Labial	87.0	86.4
Strident	89.9	89.1
Voice	89.6	89.1
Liquid		
Glottal	96.2	95.5
Labial	91.7	89.9
Lateral	90.2	88.7
Palatal	96.8	96.3
Rhotic	93.0	91.2
Closure		
Alveolar	74.5	73.4
Labial	82.7	81.4
Palatal	80.0	79.1
Voice	69.0	69.1
Vowel		
ATR	85.6	83.8
CP	85.6	84.3
Front	87.5	86.8
High	92.3	91.9
Low	88.6	87.6
Reduced	86.1	86.0
Round	97.6	96.6
Tense	83.2	81.9

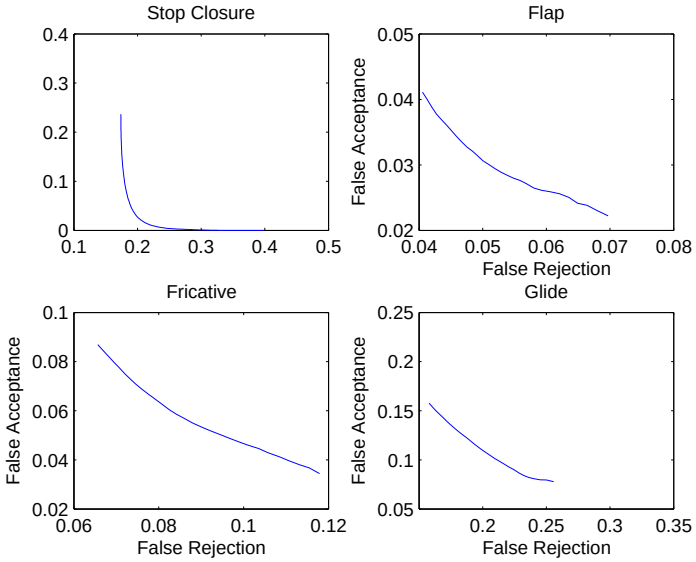

Figure 4.4 (Top left) The ROC curve of the stop closure classifier. (Top right) The ROC curve of the flap classifier. (Bottom left) The ROC curve of the fricative classifier. (Bottom right) The ROC curve of the glide classifier.

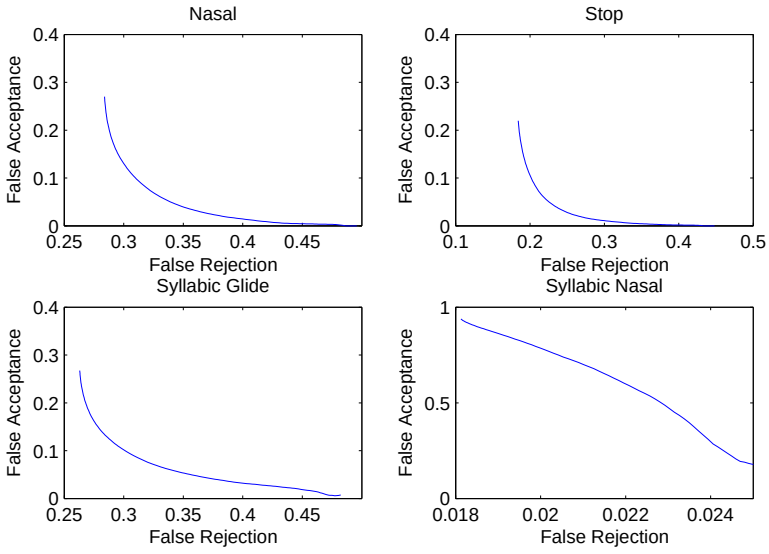

Figure 4.5 (Top left) The ROC curve of the nasal classifier. (Top right) The ROC curve of the stop release classifier. (Bottom left) The ROC curve of the syllabic glide classifier. (Bottom right) The ROC curve of the syllabic nasal classifier.

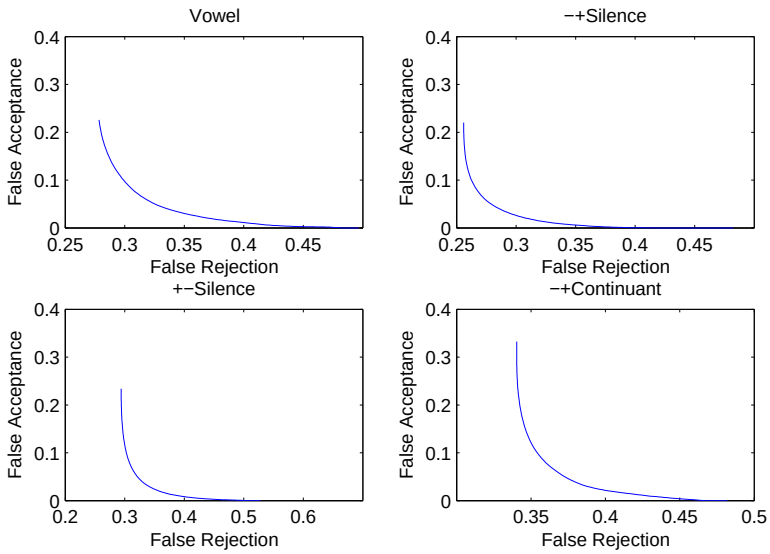

Figure 4.6 (Top left) The ROC curve of the vowel classifier. (Top right) The ROC curve of the -+silence classifier. (Bottom left) The ROC curve of the +-silence classifier. (Bottom right) The ROC curve of the -+continuant classifier.

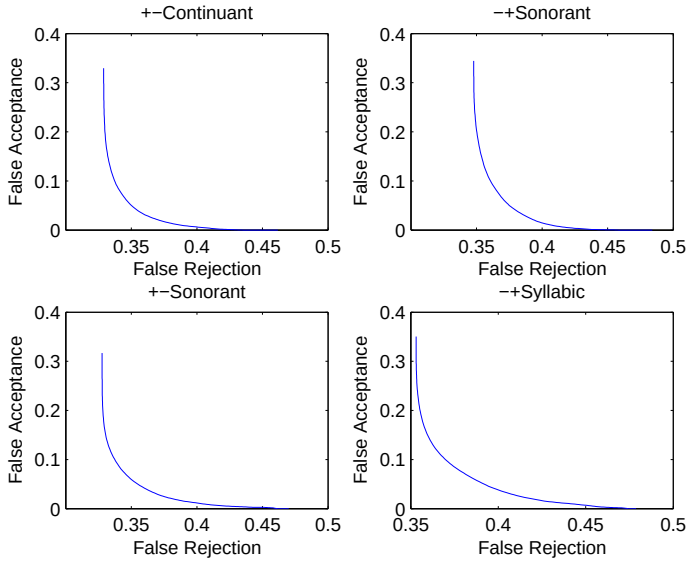

Figure 4.7 (Top left) The ROC curve of the +-continuant classifier. (Top right) The ROC curve of the -+sonorant classifier. (Bottom left) The ROC curve of the +-sonorant classifier. (Bottom right) The ROC curve of the -+syllabic classifier.

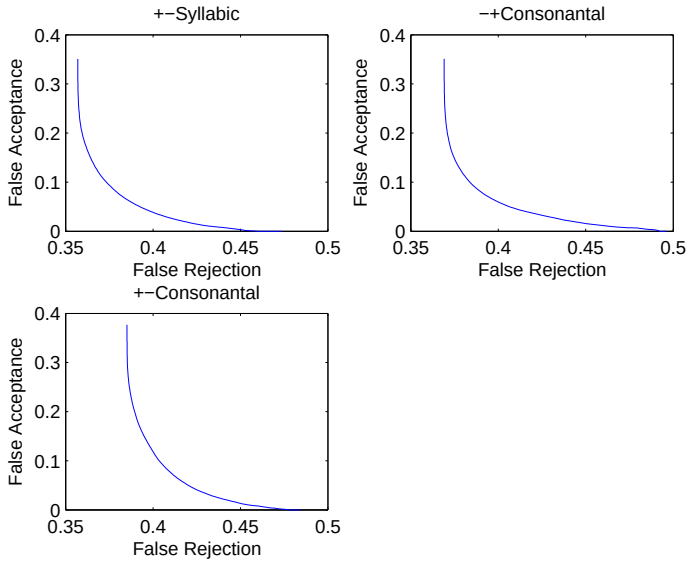

Figure 4.8 (Top left) The ROC curve of the +-syllabic classifier. (Top right) The ROC curve of the -+consonantal classifier. (Bottom left) The ROC curve of the +-consonantal classifier.

Given that we know the location of the landmark frame, we can classify manner and place information quite accurately because the landmark detectors were trained only on positive and negative examples of different landmarks. If we do not know the location of the landmark frame, can we still achieve this same accuracy? What happens when the landmark detectors are forced to output a discriminant value for all the frames in the NTIMIT test set, i.e., nonlandmark frames? To answer these questions, we processed every frame in the test corpus using the manner classification SVMs. The results of this experiment are shown in Table 4.7.

Place classification SVMs are trained to be meaningful only in a desired context; e.g., the alveolar (nasal) SVM is trained to discriminate between [+alveolar] and [-alveolar] nasal consonant release landmarks. The HMM, however, will be forced to observe discriminant outputs of these SVMs in every frame - even frames for which the SVM target output is undefined. The question arises as to what the place classification SVMs will do when used out of context. Figures 4.9 and 4.10 show the discriminant plotted as a function of time for the [front] and [alveolar stop] classifiers, respectively. Also shown in the figures are the spectrograms of the utterances. While

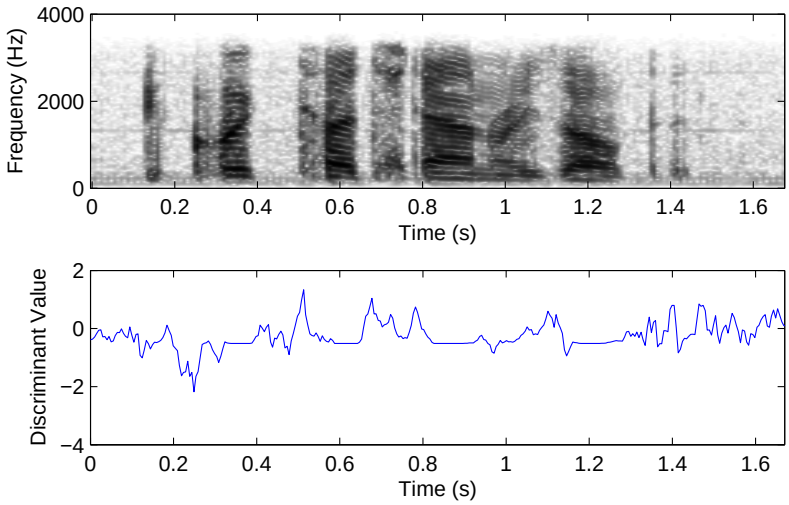

Figure 4.9 (Top) A spectrogram of the utterance "Quick touchdown result." (Bottom) The discriminant as a function of time. The discriminant was obtained from the [alveolar] stop classifier.

it is interesting to observe the behavior of the discriminant function of each place classification SVM, it is unclear what the SVM output signifies when a place feature is undefined for a particular phone; therefore, the place classification SVMs cannot be tested on every frame in the test corpus.

In Figure 4.9, the utterance is "Quick touchdown result." The [alveolar stop] release classifier discriminant is largest at the alveolar consonants /t/, /ch/ and /d/ in the word "touchdown" and at the /t/ in the word "result". It is also fairly large at the /z/ in "result" despite having been trained for stop classification. Again, we find the discriminant is smaller in magnitude or negative at other places in the signal, including the /k/ release of "quick" (nonalveolar stop) and the /n/ of "touchdown" (nonstop alveolar).

The utterance in Figure 4.10 is "Okay, take the tray." As can be seen in the figure, the discriminant of the [front] classifier is large and positive-valued in the context of the front vowel /ey/ and of the fronted schwa in "the," but is smaller in magnitude or negative at other places in the utterance.

SVM generalization was also tested on the WS96 corpus. Landmarks were extracted using the manually labeled phonetic transcriptions. The results of this

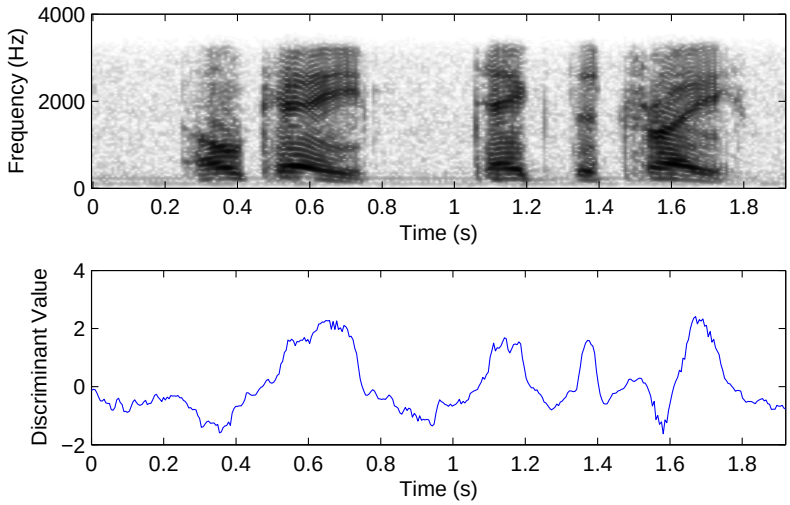

Figure 4.10 (Top) A spectrogram of the utterance "Okay, take the tray." (Bottom) The discriminant as a function of time. The discriminant was obtained from the [front] classifier.

Table 4.7 Accuracies of the RS manner classifiers when used to classify all of the NTIMIT test corpus.

SVM	ACC	SVM	ACC
Stop Closure	74.8	Stop Release	86.4
Flap	98.4	Syllabic Glide	80.6
Fricative	77.4	Syllabic Nasal	99.2
Glide	65.5	Vowel	84.9
Nasal	84.7		

experiment are given in Tables (4.8) and (4.9). The results from SVM experiments on the NTIMIT corpus are given again for convenience. This experiment was only run for place and manner classifiers. Stop release place-of-articulation features are not included in Table 4.8 and the stop release manner feature is not included in Table 4.9. This is because the stop closure and release are merged in the WS96 labeling. This makes it impossible to know the exact location of the release because the length of the stop closure is extremely variable.

Table 4.8 The generalization of the NTIMIT-trained place classification RS SVMs on the Switchboard corpus. The accuracies of the SVMs from Table 4.6 are given again for convenience.

	Switchboard	NTIMIT
Nasal		
Alveolar	57.1	82.9
Flapped	77.0	95.4
Labial	60.3	87.5
Palatal	95.7	96.5
Fricative		
Anterior	85.4	95.7
Dental	75.1	94.3
Labial	74.8	86.4
Strident	51.4	89.1
Voice	57.0	89.1
Liquid		
Glottal	88.9	95.5
Lateral	82.2	88.7
Rhotic	66.2	91.2
Labial	73.6	89.9
Palatal	83.6	96.3
Closure		
Alveolar	52.9	73.4
Labial	77.1	81.4
Palatal	65.2	79.1
Voice	55.60	69.1
Vowel		
ATR	46.4	83.8
CP	56.0	84.3
Front	63.2	86.8
High	70.0	91.93
Low	53.9	87.7
Reduced	48.2	86.0
Round	96.0	96.6
Tense	49.7	81.9

Table 4.9 The generalization of the NTIMIT-trained manner classification RS SVMs on the Switchboard corpus. The accuracies of the SVMs from Table (4.5) are given again for convenience.

	Switchboard	NTIMIT
Closure	50.1	74.8
Flap	80.0	97.4
Fricative	53.4	77.4
Liquid	56.33	65.5
Nasal	53.65	84.7
Syllabic Liquid	37.93	80.6
Syllabic Nasal	89.53	99.2
Vowel	64.88	84.9

4.4 Automatic Speech Recognition on NTIMIT

We wish to integrate the SVMs and the HMM into the system shown in Figure 4.11. The SVMs each receive a vector of acoustic data once per frame. Each SVM will output a discriminant value. These values are concatenated together to form the observation vector for the HMM. The HMM uses this new information to output a phonetic transcription.

We define

$$p(\mathbf{G}|\Lambda) = \sum_{Q} \sum_{K} p(\mathbf{G}, \mathbf{Q}, \mathbf{K}|\Lambda) \tag{4.2}$$

where $\mathbf{G} = [\vec{g}_1 \dots \vec{g}_T]$ is the sequence of observations, $\mathbf{Q} = [q_1 \dots q_T]$ is the sequence of states, and $\mathbf{K} = [k_1 \dots k_T]$ is the sequence of mixtures. $\vec{g}_t = [g_{t1} \dots g_{tD}]$ is the set of SVM discriminant outputs at time t, where g_{td} is the output of the dth SVM at time t. In other words, if the support vectors of the dth SVM are $\vec{s}_{1d} \dots \vec{s}_{jd}$, and the test acoustic observation vector at time t is $\vec{x}_t$, then

$$g_{td} = \sum_{j=1}^{M} a_{jd} e^{-\gamma|\vec{x}_t - \vec{s}_{jd}|^2} + b_d \tag{4.3}$$

where a_{jd} is the signed weight of the jth support vector in the dth SVM multiplied by the data label (either ± 1), γ is a parameter of the RBF kernel, and b_d is the bias of the dth SVM. The dth SVM has M support vectors.

The likelihood is given by

$$p(\mathbf{G}, \mathbf{Q}, \mathbf{K}|\Lambda) \quad = \quad \pi_{q_1} c_{q_1 k_1} \frac{1}{(2\pi|\Sigma_{k_1}|)^{\frac{d}{2}}} e^{\frac{-1}{2}(\vec{g}_1 - \vec{\mu}_{k_1})' \Sigma_{k_1}^{-1} (\vec{g}_1 - \vec{\mu}_{k_1})} \tag{4.4}$$

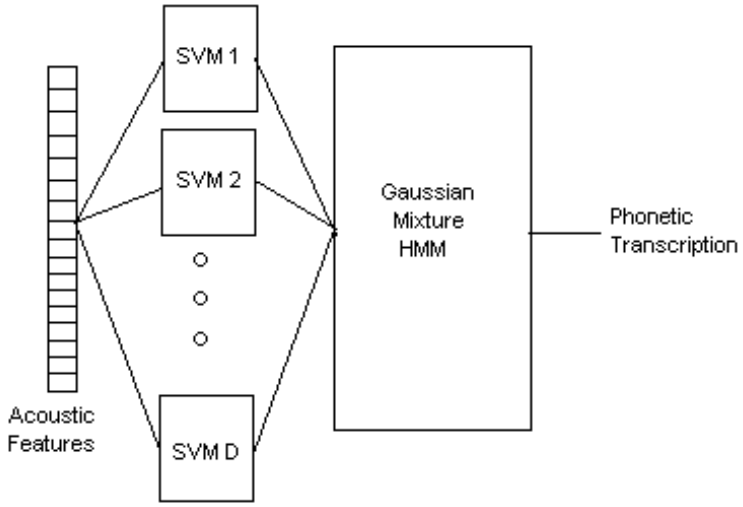

Figure 4.11 The SVM/HMM hybrid system.

$$\times a_{q_1 q_2} c_{q_2 k_2} \frac{1}{(2\pi|\Sigma_{k_2}|)^{\frac{d}{2}}} e^{\frac{-1}{2}(\vec{g}_2 - \vec{\mu}_{k_2})' \Sigma_{k_2}^{-1} (\vec{g}_2 - \vec{\mu}_{k_2})}$$
$$\ldots$$

where π_{q_1} is the initial state distribution, $c_{q_t k_t}$ are the mixture weights, $\vec{\mu}_{k_t}$ is the mean vector, Σ_{k_t} is the covariance matrix, $a_{q_{t-1} q_j}$ are the state transition probabilities, and $\vec{g}_t$ are the observations.

The baseline system was an HMM-based recognizer trained using HTK [126]. Each phone model consisted of a 5-state HMM (3 emitting states) with 1 to 33 Gaussian mixtures per state. Each mixture contained 12 MFCCs, 12 delta coefficients, 12 acceleration coefficients, and three energy coefficients. This observation vector size was chosen so that the number of parameters in the baseline system and in the SVM/HMM hybrid system are equal.

Two SVM/HMM hybrid systems were constructed. One utilized the landmark features from Table 4.4 and the other utilized the manner features from Table 4.5. Both also contained some subset of the place features given in Table 4.6. Like the baseline system, the SVM/HMM hybrid systems were HMM-based recognizers that modeled phonemes with a 5-state HMM (3 emitting states). Each emitting state contained 1 to 33 Gaussian mixtures.

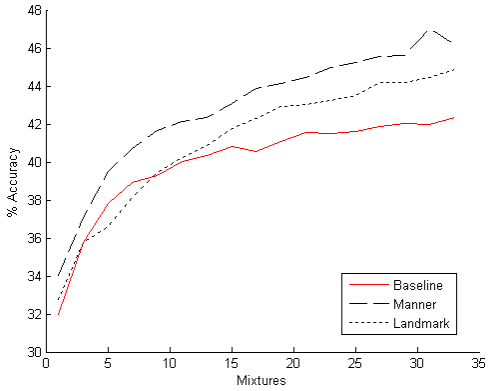

Figure 4.12 Phoneme recognition results for the baseline system, the SVM/HMM landmark feature based system, and the SVM/HMM manner feature based system as a function of the number of mixtures.

The performances of our baseline and SVM/HMM hybrid systems are shown in Figure 4.12. Our baseline system is roughly comparable with other results [130, 131] on telephone band speech.

Which kind of acoustic-phonetic information (manner changes at landmarks, manner class, or place of articulation) is most useful for phone recognition? To determine this, three additional phone recognizers were trained. One was trained on only the 10 landmark-change discriminant features. The second used only the 9 manner class discriminant features. The third used only place of articulation discriminant features. These recognizers will be referred to as LM, MC, and PA, respectively. Each of these recognizers modeled phones using a 5-state (3 emitting states) HMM. The number of mixtures in each HMM state was varied from 1 to 33. Each recognizer was used to perform phone recognition on NTIMIT. Results are shown in Figure 4.13. As seen in the figure, the majority of the information to the recognizer is provided by the place features.

In Chapter 1, we speculated that improved phone recognition accuracy should mean improved word recognition accuracy. To confirm this, tree-clustered triphone models were trained. The results of these experiments are shown in Figure 4.14. Word recognition accuracy is plotted as a function of the number of mixtures. When comparing Figures 4.12 and 4.14, we can confirm that the system with the lowest phone

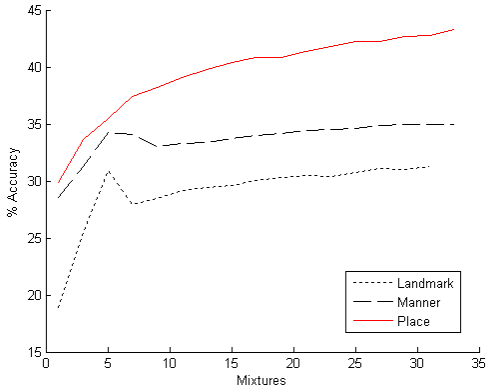

Figure 4.13 NTIMIT phone recognition accuracy using place features only, manner features only, and landmark features only. The accuracy of each recognizer is plotted as a function of the number of mixtures.

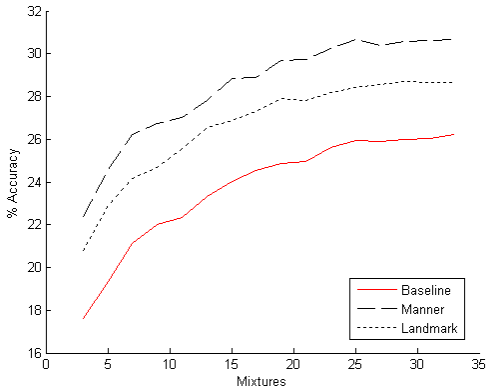

Figure 4.14 Word recognition accuracy as a function of the number of mixtures for the MFCC baseline, the SVM/HMM landmark feature-based system, and the SVM/HMM manner feature-based system.

recognition accuracy, the baseline system, has the lowest word recognition accuracy. The system with the highest phone recognition accuracy, the SVM/HMM system that uses manner and place feature information, has the highest word recognition accuracy. The SVM/HMM landmark feature-based system outperforms the baseline system at the task of phone recognition, but does worse than the SVM/HMM manner feature-based system at this same task. The results are similar for word recognition. The baseline system is comparable to the MFCC-based baseline system trained by Selouani and O'Shaughnessy in [132].

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

Recognition of 8 kHz speech tends to be more difficult than recognition of 16 kHz speech. This may be due in part to the loss of information given by the higher frequencies. Gaussian classifiers trained on the TIMIT corpus (16 kHz speech) [44] perform rather well when compared to systems trained on the NTIMIT corpus (8 kHz speech) [130, 131]. This drop in performance between 16 kHz speech and 8 kHz speech is not limited to Gaussian classifiers but is evident in other classification systems as well, such as the SVM [1].

SVMs are computationally complex. The calculation of one discriminant value requires (N-1)+N(D+1) multiplications (where N is the number of support vectors and D is the length of the support vector) and 2DN+(N-1) additions. Before we reduced the number of support vectors using the RS method (Section 2.1.3), the average number of support vectors for any given SVM was around 4620 with 39 SVMs, and so the total computational complexity of the SVMs was 18 941 900 multiplications per second of speech. Reducing the number of support vectors by a factor of 100 reduces the multiplications and additions by a factor of 100. Therefore, the SVMs in our final hybrid system required 188 500 multiplications per second of speech.

It is unclear how SVM accuracy affects phone recognition accuracy. If the SVMs were 100% accurate, would phone recognition be 100% accurate? Furthermore, how does a small change in the classification accuracy of the SVM (such as the change in accuracy between standard and RS SVMs in Tables 4.4, 4.5, and 4.6) affect phone recognition accuracy? Unfortunately, this experiment could not be run; the

computational complexity of the RBF SVM combined with a large number of support vectors makes this experiment infeasible.

Data and information are not synonyms. Perhaps at one point in time, when recorded speech data were rare or when only a few different acoustic feature extraction methods existed to represent information such as frequency, amplitude, energy, phase, perceptual, or other content of a speech signal, data and information were equivalent. We currently attempt to extract something useful from the data using all kinds of signal processing techniques and mathematical transformations that are based on ideas from multiple disciplines such as psychology, engineering, pattern recognition, and linguistics. Yet, even these different ideas and formulations of this "something useful" that we say exists in the speech signal are redundant and often highly correlated with each other. The improvement in accuracy between SVMs trained on MFCC, delta, and acceleration coefficient input and MFCC, delta, acceleration, formant, and AP input described in Section 4.3.1 is disappointingly small for the additional work that was put into generating the extra two signal representations. Intuitively, one would expect a much higher increase between the two SVMs trained on those respective feature sets if the different kinds of signal representations were completely independent of each other and provided new, previously unseen information. Speech recognizers are not performing poorly due to lack of data, but due to lack of useful nonredundant information.

The solution to the SVM problem provides a sparse representation of the data if the data are noiseless. Noisy data, however, can present a problem for the SVM. The SVM finds an optimally sparse subset of the training dataset, a subset called the support vectors. Though the solution in the data space may be sparse, the same solution may not be sparse in the feature space. In fact, the SVM solution may actually be extremely dense in the feature space because every feature makes a weighted contribution to the solution. This is why the SVM is, in many cases, unable to ignore noise or redundancy in the data. Even if two different sets of data are drawn from the same distribution, they will provide for two distinct solutions to the

overdetermined SVM problem. These different solutions have the potential to greatly affect the generalization ability of the classifier, despite both solutions' optimality.

The use of sparse data representations is growing in popularity. Obtaining the sparsest solution to a problem is desirable because it not only saves disk space and computation time, but also because sparsity of the feature representation can bound generalization error, in much the same way that sparsity of the support vector set can bound generalization error . Semidefinite programming methods [133] can be used to find the sparsest solution. The SVM can be reformulated into a semidefinite programming problem as shown in [134, 135] where the SVM is rederived using quadratically constrained quadratic programming (QCQP) and semidefinite relaxation (SDR) methods. Currently, many tools exist to solve semidefinite programming problems. The SVM front-end system presented in this thesis could benefit from the use of sparse representations.

Integration with a distinctive feature based back end (e.g., the system of Livescu et al. [136]) seems to be a desirable goal. This thesis has demonstrated that a landmark based front end provides information useful to an HMM. It may be possible to get further benefit by optimizing the back end to better model the information provided by the SVMs; such optimization is a topic for future research.

REFERENCES

[1] M. Hasegawa-Johnson, J. Baker, S. Greenberg, K. Kirchhoff, J. Muller, K. Sommez, S. Borys, K. Chen, A. Juneja, K. Livescu, S. Mohan, E. Coogan, and T. Wong, "Landmark-Based speech recognition: Report of the 2004 Johns Hopkins summer workshop," Johns Hopkins University, Center for Speech and Language Processing, Baltimore, MD, Tech. Rep., 2004.

[2] G. A. Miller and P. E. Nicely, "Analysis of perceptual confusions among some English consonants," *Journal of the Acoustical Society of America*, vol. 27, pp. 338–352, 1955.

[3] R. Shannon, F. Zeng, V. Kamath, J. Wygonski, and M. Ekelid, "Speech recognition with primarily temporal cues," *Science*, vol. 270, pp. 303–304, October 1995.

[4] J. Mehler, P. W. Jusczyk, G. Lambertz, N. Halstead, J. Bertoncini, and C. Amiel-Tison, "A precursor of language acquisition in young infants," *Cognition*, vol. 29, pp. 143–178, 1988.

[5] P. Jusczyk, "Picking up regularities in the sound structure of the native language," in *Speech Perception and Linguistic Experience: Issues in Cross-Language Speech Research*, W. Strange, Ed. Timonium, MD: York, 1995.

[6] P. W. Jusczyk, A. D. Friederici, J. M. I. Wessels, V. Y. Svenkerud, and A. M. Jusczyk, "Infants' sensitivity to the sound patterns of native language words," *Journal of Memory and Language*, vol. 32, pp. 402–420, 1993.

[7] K. Saberi and D. R. Perrott, "Cognitive restoration of reversed speech," *Nature*, vol. 398, no. 6730, p. 760, April 1996.

[8] A. Boothroyd, B. Mulhearn, J. Gong, and J. Ostroff, "Effects of spectral smearing on phoneme and word recognition," *Journal of the Acoustical Society of America*, vol. 100, no. 3, pp. 1807–1818, 1996.

[9] Q.-J. Fu and R. V. Shannon, "Recognition of spectrally degraded speech in noise with nonlinear amplitude mapping," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Phoenix, AZ, 1999, pp. 369–372.

[10] J. B. Allen, "How do humans process and recognize speech?" *IEEE Transactions on Speech and Audio Processing*, vol. 2, no. 4, pp. 567–577, October 1994.

[11] H. Hermansky and S. Sharma, "Temporal patterns (TRAPS) in ASR of noisy speech," in *International Conference on Acoustics, Speech, and Signal Processing*, Phoenix, AZ, 1999, pp. 289–292.

[12] L. K. Saul, M. G. Rahim, and J. B. Allen, "A statistical model for robust integration of narrowband cues in speech," *Computer Speech and Language*, vol. 15, no. 2, pp. 175–194, 2001.

[13] S. Greenberg, S. Chang, and J. Hollenback, "An introduction to the diagnostic evaluation of the switchboard-corpus automatic speech recognition systems," in *Proceedings of the NIST Speech Transcription Workshop*, College Park, MD, 2000.

[14] S. Greenberg, "Recognition in a new key — towards a science of spoken language," in *International Conference on Acoustics, Speech, and Signal Processing*, Seattle, WA, 1998, pp. 1041–1045.

[15] S. Greenberg, "Speaking in shorthand - a syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, pp. 159–176, 1999.

[16] D. McAllaster, L. Gillick, F. Scattone, and M. Newman, "Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch," in *International Conference on Speech and Language Processing*, 1998, pp. 1847–1850.

[17] P. S. Cohen and R. L. Mercer, *Phonological Component of an Automatic Speech Recognizer*. New York, NY: Academic Press, 1975.

[18] B. T. Oshika, V. W. Zue, R. V. Weeks, H. Neu, and J. Aurbach, "The role of phonological rules in speech understanding research," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 104–112, February 1975.

[19] M. Bates, "The use of syntax in a speech understanding system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 112–117, 1975.

[20] L.-W. Fung and K.-S. Fu, "Stochastic syntactic decoding for pattern classification," *IEEE Transactions on Computers*, vol. 23, no. 1, pp. 662–667, 1975.

[21] B. Nash-Webber, "Semantic support for a speech understanding system," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 124–129, 1975.

[22] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, no. 1, pp. 67–72, 1975.

[23] F. Jelinek, "Continuous speech recognition by statistical methods," *Procedings of the IEEE*, vol. 64, no. 4, pp. 532–556, 1976.

[24] J. Baker, "The Dragon system — an overview," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 23, pp. 24–29, 1975.

[25] D. H. Klatt, "Speech perception: A model of acoustic-phonetic analysis and lexical access," *Journal of Phonetics*, vol. 7, pp. 279–312, 1979.

[26] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revisited," *Cognition*, vol. 21, pp. 1–36, 1985.

[27] V. Zue, "The use of speech knowledge in automatic speech recognition," *Proceedings of the. IEEE*, vol. 73, no. 11, pp. 1602–1615, November 1985.

[28] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, pp. 328–339, 1989.

[29] E. McDermott, H. Iwamida, S. Katagiri, and Y. Tohkura, *Shift-Tolerant LVQ and Hybrid LVQ-HMM for Phoneme Recognition.* San Mateo, CA: Morgan Kaufmann, 1990.

[30] Y. Bengio, R. D. Mori, G. Flammia, and R. Kompe, "Global optimization of a neural network-hidden Markov model hybrid," *IEEE Transactions on Neural Networks*, vol. 3, no. 2, pp. 252–259, 1992.

[31] H. Bourlard, N. Morgan, C. Wooters, and S. Renals, "CDNN: A context-dependent neural network for continuous speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, San Francisco, CA, 1992, pp. 349–352.

[32] H. Bourlard and N. Morgan, *Connectionist Speech Recognition: A Hybrid Approach.* Boston, MA: Kluwer Academic Publishers, 1994.

[33] Y. Bengio, R. D. Mori, G. Flammia, and R. Kompe, "Phonetically motivated acoustic parameters for continuous speech recognition using artificial neural networks," *Speech Communication*, vol. 111, no. 2-3, pp. 261–271, 1992.

[34] F. Bimbot, G. Chollet, and J.-P. Tubach, "TDNNs for phonetic features extraction: A visual exploration," in *International Conference on Acoustics, Speech, and Signal Processing*, 1991, pp. 73–76.

[35] K. N. Stevens, "Evidence for the role of acoustic boundaries in the perception of speech sounds," in *Phonetic Linguistics: Essays in Honor of Peter Ladefoged*, V. A. Fromkin, Ed. Orlando, FL: Academic Press, 1985, pp. 243–255.

[36] K. N. Stevens, S. Y. Manuel, S. Shattuck-Hufnagel, and S. Liu, "Implementation of a model for lexical access based on features," in *International Conference on Speech and Language Processing*, vol. 1, Banff, Alberta, Canada, 1992, pp. 499–502.

[37] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, May 1995.

[38] S. A. Liu, "Landmark detection for distinctive feature-based speech recognition," *Journal of the Acoustical Society of America*, vol. 100, no. 5, pp. 3417–3430, November 1996.

[39] M. Chen, "Nasal landmark detection," in *International Conference on Speech and Language Processing*, 2000, pp. 636–639.

[40] A. W. Howitt, "Vowel landmark detection," in *International Conference on Speech and Language Processing*, 2000.

[41] C. Espy-Wilson, "A feature-based semi-vowel recognition system," *Journal of the Acoustical Society of America*, vol. 96, no. 1, pp. 65–72, July 1994.

[42] J. R. Glass and V. W. Zue, "Multi-level acoustic segmentation of continuous speech," in *International Conference on Acoustics, Speech, and Signal Processing*, New York, NY, April 1988, pp. 429–432.

[43] A. K. Halberstadt and J. R. Glass, "Heterogeneous acoustic measurements for phonetic classification," in *EUROSPEECH*, 1997, pp. 401–404.

[44] A. K. Halberstadt and J. R. Glass, "Heterogeneous measurements and multiple classifiers for speech recognition," in *International Conference on Speech and Language Processing*, Sydney, Australia, November 1998.

[45] M. Ostendorf, V. V. Digalakis, and O. A. Kimball, "From HMM's to segment models: A unified view of stochastic modeling for speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 4, no. 5, pp. 360–378, 1996.

[46] J. Bilmes, N. Morgan, S.-L. Wu, and H. Bourlard, "Stochastic perceptual speech models with durational dependence," in *International Conference on Acoustics, Speech, and Signal Processing*, 1996, pp. 1301–1304.

[47] M. K. Omar, M. Hasegawa-Johnson, and S. E. Levinson, "Gaussian mixture models of phonetic boundaries for speech recognition," in *IEEE Workshop on Automatic Speech Recognition and Understanding*, 2001, pp. 33–36.

[48] M. K. Omar and M. Hasegawa-Johnson, "Approximately independent factors of speech using non-linear symplectic transformation," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 660–671, 2003.

[49] M. K. Omar and M. Hasegawa-Johnson, "Model enforcement: A unified feature transformation framework for classification and recognition," *IEEE Transactions on Signal Processing*, vol. 52, no. 10, pp. 2701–2710, 2004.

[50] P. Niyogi and P. Ramesh, "Incorporating voice onset time to improve letter recognition accuracies," in *International Conference on Acoustics, Speech, and Signal Processing*, 1998, pp. 13–16.

[51] P. Niyogi and C. Burges, "Detecting and interpreting acoustic features by support vector machine," University of Chicago, Tech. Rep. 2002-02, 2002.

[52] A. Juneja and C. Espy-Wilson, "Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines," in *International Joint Conference on Neural Networks*, Portland, OR, 2003.

[53] A. Juneja and C. Espy-Wilson, "A novel probabilistic framework for event-based speech recognition," *Journal of the Acoustical Society of America*, vol. 114, no. 4(A), p. 2395, 2003.

[54] A. Juneja and C. Espy-Wilson, "Significance of invariant acoustic cues in a probabilistic framework for landmark-based speech recognition," in *From Sound to Sense: 50+ Years of Discoveries in Speech Communication*. Cambridge, MA: MIT, 2004, pp. C–151–C–156.

[55] A. Juneja, "Speech recognition based on phonetic features and acoustic landmarks," Ph.D. dissertation, University of Maryland, College Park, MD, 2004.

[56] M. Hasegawa-Johnson, J. Baker, S. Borys, K. Chen, E. Coogan, S. Greenberg, A. Juneja, K. Kirchhoff, K. Livescu, S. Mohan, J. Muller, K. Sommez, and T. Wong, "Landmark-based speech recognition: Report of the 2004 Johns Hopkins summer workshop," in *International Conference on Acoustics, Speech, and Signal Processing*, 2005, pp. 213–216.

[57] S. Borys and M. Hasegawa-Johnson, "Distinctive feature based discriminant features for improvements to phone recognition on telephone band speech," in *Eurospeech*, 2005, pp. 679–700.

[58] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[59] T. Joachims, "Making large-scale support vector machine learning practical," in *Advances in kernel methods: support vector learning*, B. Schökopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1999, pp. 169–184.

[60] J. C. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," Microsoft Research, Redmond, WA, Tech. Rep. MSR-TR-98-14, 1998.

[61] R.-E. Fan, P.-H. Chen, and C.-J. Li, "Working set selection using second order information for training support vector machines," *Journal of Machine Learning Research*, vol. 6, pp. 1889–1918, November 2005.

[62] T. Joachims, "Estimating the generalization performance of the SVM efficiently," in *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, pp. 431–438.

[63] R. Klinkenburg and T. Joachims, "Detecting concept drift with support vector machines," in *Proceedings of the Seventeenth International Conference on Machine Learning*, 2000.

[64] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu, "The entire regularization path for the support vector machine," *Journal of Machine Learning Research*, no. 5, pp. 1391–1415, 2004.

[65] S. Keerthi, "Efficient tuning of SVM hyperparameters using radius/margin bound and iterative algorithms," *IEEE Transactions on Neural Networks*, vol. 13, no. 5, pp. 1225–1229, 2002.

[66] C. Burges and B. Schölkopf, "Improving the accuracy and speed of support vector machines," in *Advances in Neural Information Processing Systems*, vol. 9, 1997, pp. 375–381.

[67] C. Burges, "Simplified support vector decision rules," in *Proceedings of the 13th International Conference on Machine Learning*, 1996, pp. 71–77.

[68] R. Jakobson, G. Fant, and M. Halle, "Preliminaries to speech analysis," MIT Acoustics Laboratory, Cambridge, MA, Tech. Rep. 13, 1952.

[69] King Sejong, *Hunmin Jeongeum*. Kingdom of Joseon, 1446.

[70] A. Bell, *Visible Speech: The Science of the Universal Alphabetic*. London, England: Simpkin, Marshal, and Co., 1876.

[71] International Phonetic Association (IPA), "International phonetic alphabet," 1993. [Online]. Available: http://www.arts.gla.ac.uk/IPA/ipa.html.

[72] L. E. Volaitis and J. L. Miller, "Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories," *Journal of the Acoustical Society of America*, vol. 92, no. 2, pp. 723–735, August 1992.

[73] P. K. Kuhl, K. A. Williams, F. Lacerda, K. N. Stevens, and B. Linblom, "Linguistic experience alters phonetic perception in infants by 6 months of age," *Science*, vol. 255, pp. 606–608, 1992.

[74] F. H. Guenther and M. N. Gjaja, "The perceptual magnet effect as an emergent property of neural map formation," *Journal of the Acoustical Society of America*, vol. 100, pp. 1111–1121, 1996.

[75] A. Sharma and M. F. Dorman, "Exploration of the perceptual magnet effect using the mismatch negativity auditory evoked potential," *Journal of the Acoustical Society of America*, vol. 104, pp. 511–517, 1998.

[76] P. C. Delattre, A. M. Liberman, and F. S. Cooper, "Acoustic loci and transitional cues for consonants," *Journal of the Acoustical Society of America*, vol. 27, no. 4, pp. 769–773, July 1955.

[77] Z. B. Nossair and S. A. Zahorian, "Dynamic spectral shape features as acoustic correlates for initial stop consonants," *Journal of the Acoustical Society of America*, vol. 89, no. 6, pp. 2978–2991, June 1991.

[78] A. Alwan, "Modeling speech perception in noise: The stop consonants as a case study," Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, MA, February 1992.

[79] M. Miller and M. Sachs, "Representation of stop consonants in the discharge patterns of auditory-nerve fibers," *Journal of the Acoustical Society of America*, vol. 74, pp. 502–517, 1983.

[80] A. S. Bregman, *Auditory Scene Analysis*. Cambridge, MA: MIT Press, 1990.

[81] S. Furui, "On the role of spectral transition for speech perception," *Journal of the Acoustical Society of America*, vol. 80, no. 4, pp. 1016–1025, 1983.

[82] W. T. Siok, Z. Jin, P. Fletcher, and L. H. Tan, "Distinct brain regions associated with syllable and phoneme," *Human Brain Mapping*, vol. 18, no. 3, pp. 201–207, 2003.

[83] R. M. Warren, E. W. Healy, and M. H. Chalikia, "The vowel-sequency illusion: Intrasubject stability and intersubject agreement of syllabic forms," *Journal of the Acoustical Society of America*, vol. 100, pp. 2452–2461, 1996.

[84] K. Stevens, *Acoustic Phonetics*. Cambridge, MA: MIT Press, 1999.

[85] P. Andres, "The equivalence of support vector machine and regularization neural networks," in *Neural Processing Letters*, 2002, pp. 97–104.

[86] A. Esposito, C. E. Ezin, and M. Ceccarelli, "Preprocessing and neural classification of english stop consonants [b, d, g, p, t, k]," in *International Conference on Speech and Language Processing*, 1996.

[87] P. Niyogi, C. Burges, and P. Ramesh, "Distinctive feature detection using support vector machines," in *International Conference on Acoustics, Speech, and Signal Processing*, 1999, pp. 961–964.

[88] A. A. Ali, J. van der Spiegel, and P. Meuller, "An acoustic-phonetic feature based system for the automatic recognition of fricative consonants," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1998.

[89] J. Glass and V. Zue, "Detection and recognition of nasal consonants in American English," in *International Conference on Acoustics, Speech, and Signal Processing*, Tokyo, Japan, 1996, pp. 2767–2770.

[90] T. Pruthi and C. Espy-Wilson, "Automatic classification of nasals and semivowels," in *15th International Congress of Phonetic Sciences*, Barcelona, Spain, 2003, pp. 3061–3064.

[91] S. King and P. Taylor, "Detection of phonological features in continuous speech using neural networks," *Computer Speech and Language*, vol. 14, no. 4, pp. 333–353, 2000.

[92] T. D. Harrison and F. Fallside, "A connectionist model for phoneme recognition in continuous speech," in *Internatioal Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1989, pp. 417–420.

[93] A. Robinson, "An application of recurrent nets to phone probability estimation," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 298–305, March 1994.

[94] P. Polur, R. Zhou, J. Yang, F. Adnani, and R. Hobson, "Isolated speech recognition using artificial neural networks," in *Proceedings of the 23rd Annual EMBS International Conference*, Istanbul, Turkey, 2001.

[95] K. ichi Iso and T. Watanabe, "Speaker independent word recognition using a neural prediction tool," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, April 1990, pp. 441–444.

[96] J. Yousafzai, Z. Cvetkovic, P. Sollich, and B. Yu, "Combined plp-acoustic waveform classification for robust phoneme recognition using support vector machines," in *Proceedings of European Signal Processing Conference*, 2008. [Online]. Available: http://www.eurasip.org/Proceedings/Eusipco/Eusipco2008/.

[97] A. Hatch, A. Stolke, and B. Peskin, "Combining feature sets with support vector machines: Application to speaker recognition," in *Automatic Speech Recognition and Understanding*, 2005, pp. 75–79.

[98] S. Kajarekar, "Four weightings and a fusion: A cepstral SVM system for speaker recognition," in *Automatic Speech Recognition and Understanding*, 2005, pp. 17–22.

[99] H. Schmid, "Part-of-speech tagging using neural networks," in *Proceedings of the 15th International Conference on Computational Linguistics*, August 1994, pp. 172–176.

[100] J. Gimenex and L. Marquez, "Fast and accurate part-of-speech tagging: The SVM approach revisited," in *Recent Advances in Natural Language Processing*, 2003, pp. 153–162.

[101] T. Joachims, *Text Categorization with Support Vector Machines: Learning from many Relevant Features.* Heidelburg, Germany: Springer, 1997.

[102] S. Dumais, J. Platt, D. Heckerman, and M. Sehami, "Inductive learning algorithms and representations for text categorization," in *Proceedings of the 7th International Conference on Information and Knowledge Management*, 1998, pp. 148–155.

[103] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter, "Distributional word clusters vs. text catergorization," *Journal of Machine Learning*, vol. 3, pp. 1183–1208, March 2003.

[104] J. Nicholson, K. Takahashi, and R. Nakatsu, "Emotion recognition in speech using neural networks," *Neural Computing and Applications*, vol. 9, no. 4, pp. 290–296, December 2000.

[105] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in *EUROSPEECH*, 2003, pp. 125–128.

[106] N. Morgan and H. Boulard, "Continuous speech recognition," *IEEE Signal Processing Magazine*, vol. 12, no. 3, pp. 24–42, May 1995.

[107] H. Schwenk, "Using boosting to improve a hybrid HMM/neural network speech recognizer," in *International Conference on Acoustics, Speech, and Signal Processing*, 1999, pp. 1009–1012.

[108] Y. Freund and R. Schapire, "A short introduction to boosting," *Journal of Japanese Society for Artificial Intelligence*, vol. 14, no. 5, pp. 771–780, September 1999.

[109] R. Schapire, "A brief introduction to boosting," in *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence*, 1999, pp. 1401–1406.

[110] H. Schwenk and Y. Bengio, "Boosting neural networks," *Neural Computing*, vol. 12, no. 8, pp. 1869–1887, August 2000.

[111] H. Hermansky, D. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, Istanbul, Turkey, 2000, pp. 1635–1638.

[112] K. Kirchhoff and J. A. Bilmes, "Dynamic classifier combination in hybrid speech recognition systems using utterance-level confidence values," in *International Conference on Acoustics, Speech, and Signal Processing*, 1999, pp. 693–696.

[113] T. Robinson, M. Hockberg, and S. Renals, "IPA: Improved phone modeling with recurrent neural networks," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, March 1994, pp. 137–140.

[114] M. Rahim, "A neural tree network for phoneme classification with experiments on the TIMIT database," in *International Conference on Acoustics, Speech, and Signal Processing*, 1992, pp. 345–348.

[115] K. Kirchhoff, "Robust speech recognition using articulatory information," Ph.D. dissertation, University of Bielefeld, Bielefeld, Germany, 1999.

[116] P. Haffner, M. Franzini, and A. Waibel, "Integrating time alignment and neural networks for high performance continuous speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, 1991, pp. 105–108.

[117] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, March 1989.

[118] A. Waibel, H. Sawai, and K. Shikano, "Consonant recognition by modular construction of large phonemic time-delay neural networks," in *International Conference on Acoustics, Speech, and Signal Processing*, May 1989, pp. 215–223.

[119] N. Ahmadi, N. J. Bailey, and B. S. Hoyle, "Phoneme recognition using speech image (spectrogram)," in *Proceedings of the 3rd International Conference on Signal Processing*, 1996, pp. 675–677.

[120] N. D. Smith and M. J. F. Gales, "Using SVMs and discriminative models for speech recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2002, pp. 77–80.

[121] C. Jankowski, J. Kalyanswamy, S. Basson, and J. Spritz, "NTIMIT: A phonetically blaanced, continuous speech, telephone bandwidth speech database," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1990, pp. 109–112.

[122] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren, "The DARPA TIMIT acoustic phonetic speech corpus," National Institute of Standards and Technology, Gaithersburg, MD, Tech. Rep., 1993.

[123] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *International Conference on Acoustics, Speech, and Signal Processing*, 1992, pp. 517–520.

[124] S. Greenberg, "Strategies for automatic multi-tier annotation of spoken language corpora," in *Eurospeech*, 2003.

[125] M. Hasegawa-Johnson, "The periodic vector toolkit (PVTK)," 2004. [Online]. Available: http://www.ifp.uiuc.edu/speech/software.

[126] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book*. Cambridge, UK: Cambridge University Engineering Department, 2002.

[127] Y. Zheng, M. Hasegawa-Johnson, and S. Borys, "Stop consonant classification by dynamic formant trajectory," in *International Conference on Speech and Language Processing*, Jeju Island, Korea, 2004, pp. 2481–2484.

[128] N. Bitar and C. Espy-Wilson, "A knowledge-based signal representation for speech recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, Atlanta, GA, 1996, pp. 29–32.

[129] K. Stevens, S. Blurmstein, L. Glicksman, M. Burton, and K. Kurowski, "Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters," *Journal of the Acoustical Society of America*, vol. 91, no. 5, May 1992.

[130] B. Chigier, "Phonetic classification on wide-band and telephone quality speech," in *International Conference on Acoustics, Speech, and Signal Processing*, 1991.

[131] P. Moreno and R. Stern, "Sources of degradation of speech recognition in the telephone network," in *IEEE Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1994, pp. 109–112.

[132] S. A. Selouani and D. O'Shaughnessy, "Robustness of speech recognition using genetic algorithms and a mel-cepstral subspace approach," in *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, May 2004, pp. 201–204.

[133] L. Vandenberghe and S. Boyd, "Semidefinite programming," *Siam Review*, vol. 38, no. 1, pp. 49–95, March 1996.

[134] A. Chan, N. Vasconcelos, and G. Lanckriet, "Direct convex relaxations of sparse SVM," in *International Conference on Machine Learning*, 2007, pp. 145–153.

[135] A. Chan, N. Vasconcelos, and G. Lanckriet, "Duals of the QCQP and SDP sparse SVM," University of California, San Diego, Tech. Rep. SVCL-TR-2007-02, 2007.

[136] K. Livescu, J. Glass, and J. Bilmes, "Hidden feature models for speech recognition using dynamic Bayesian networks," in *EUROSPEECH*, 2003, pp. 173–180.