Detecting and Interpreting Acoustic Features by Support Vector Machines

Partha Niyogi[1] and Chris Burges[2]
(1) The University of Chicago
and
(2) Microsoft Research.

Abstract

An important aspect of distinctive feature based approaches to automatic speech recognition is the formulation of a framework for the robust detection of these features. We discuss the application of the Support Vector Machines (SVM) that arise when the structural risk minimization principle is applied to such feature detection problems. In particular, we consider in some detail the problem of detecting stop consonants in continuous speech. We track dynamic acoustic properties and discuss an SVM framework for detecting these sounds. In this paper, we use both linear and non-linear SVMs and present experimental results to illustrate the factors upon which superior performance depends. We also relate the detectors to perceptual phenomena like categorical perception and the perceptual magnet effect. We show how the detectors operate by comparing sounds in a transformed space leading to many different distance metrics that may then be defined. Only one of these is compatible with the perceptual magnet effect.

1 INTRODUCTION

Any approach to speech perception or recognition will have to specify a mechanism by means of which the acoustic input is mapped to discrete linguistic objects or symbols. In most conventional speech recognition systems, the primitive linguistic objects are taken to be phonemes. We are pursuing an approach that considers the primary linguistic objects to be distinctive features (Jakobson et al , 1952) that will then need to be recovered from the speech signal. We proceed by developing detectors for various distinctive

features from their acoustic correlates in the speech signal. Crucial to the success of such an approach are the following:

(1) the determination of acoustic signatures for different sound classes and the development of signal representations in which those acoustic signatures best express themselves;

(2) the construction of statistical learning paradigms that operate on the above representations and separate the positive instances of a distinctive feature from the negative instances of the same feature.

Traditionally, (1) is regarded as the front-end and (2) as the back end of a speech recognition system. In most traditional speech recognition designs, the *same* representation is used for all sound classes. For example, cepstral coefficients, auditory filterbank outputs, short time fourier transforms are common representations computed with a fixed analysis window and stepping rate. The front-end is thus a vector time series. The back-end is typically an HMM of one form or another (Rabiner and Juang, 1993; Jelinek, 1997).

In contrast, we consider using different representations for the different sound classes. Acoustic-phonetic knowledge (Stevens, 1998) guides us in choosing appropriate representations. However, another important question that needs to be addressed is: given a particular representation, what sort of a statistical learning machine should be used to optimally separate the positive exemplars from the negative ones? In this paper, we investigate the application of Support Vector Machines for the kinds of detection problems that would naturally arise in our feature based approach to speech recognition.

It is worthwhile to note the following:

1. Previous attempts at speech recognition using either distinctive features or acoustic-phonetic approaches have often been cast within the framework of knowledge based systems where the primary structures of control are logical rather than statistical. Such logic-based systems tend to be brittle and often preclude the possibility of using inductive methods that are data-driven. In contrast, a statistical approach to feature based recognition attempts to create a suitable statistical framework around the notions of features and their acoustic correlates. In this paper, we explore Support Vector Machines (Vapnik, 1998) for such a purpose.

2. In the construction of a statistical learning paradigm, an important

issue that needs to be addressed is the ability of the machine to generalize from its finite training set to its test set. While it has been recognized in the practical pattern recognition community that this generalization ability depends upon the complexity of the learning machine, the Vapnik-Chervonenkis (Vapnik, 1995) theory provides a more precise characterization in terms of its VC dimension (also see Niyogi and Girosi, 1996 for a discussion of generalization error). The Support Vector Machine is one instantiation of an algorithm derived within the framework of the VC theory that potentially allows us to control the size (capacity) of the machine to achieve good generalization. In contrast, traditional neural networks or Hidden Markov Models both of which have been deployed in practical speech pattern recognition problems usually have inadequate methods of capacity control at present. We examine a variety of Support Vector Machines with different architectures and capacity control bounds to understand and emphasize its importance in the successful utilization of this technology for speech recognition.

3. All results with the Support Vector Machine are presented on a stop detection problem that we have worked on. We provide a very brief description of the problem and direct the reader to Niyogi and Sondhi (2000,2001a) for a more detailed discussion of stop consonants, their acoustic-phonetic attributes and their common confusions. In Niyogi and Sondhi (2000,2001a), we have a wide ranging discussion on stop consonants and present results for a number of different detectors. As it turns out, the non-linear detector based on Support Vector Machines yielded the best performance. The emphasis in Niyogi and Sondhi (2000,2001a) was on understanding acoustic phonetic properties, patterns of errors and robustness of detectors to environmental noise of various sorts. An analysis of the underlying principles of the SVM detector was beyond the scope of that paper. Here, we consider the theory and practice of SVM based detectors in considerable detail. The detector operates on dynamic acoustic features in the speech signal. It is also important to emphasize that various feature detection problems (like vowel detection or nasal detection or obstruent detection and so on) would give rise to conceptually similar learning problems that can in principle be treated within the Support Vector formalism. Interestingly, the distinctive feature theory often focuses on binary oppositions and therefore results in two class pattern recognition problems that are

particularly well handled by current formulations of Support Vector Machines.

4. The problem of detecting acoustic features also arises in the context of speech perception and acoustic phonetics. A substantial part of this paper attempts to relate the support vectors and the detectors based on them to perceptual phenomena like categorical perception (Harnad, 1987) and the perceptual magnet effect (Kuhl, 1991). The perceptual magnet effect characterizes the relationship between the topologies implied by the acoustic distance and the perceptual distance in the space of speech sounds. We discuss how SVMs map the speech data from the acoustic space to a manifold embedded in an infinite dimensional Hilbert space using a kernel function. For kernels defined by radial basis functions, we show that the manifold admits a Riemannian structure giving rise to three new distance measures (in addition to the distance in the original acoustic space) between speech sounds. Only one of these obeys the perceptual magnet effect and in fact it is this measure that is ultimately used by SVMs in comparing and classifying sounds into stop consonants and non-stops. It is also worth mentioning here that the acoustic space considered captures the dynamic transition between closure and burst in stop consonants.

The rest of the paper is structured as follows. In section 2, we provide a description of the stop detection problem and motivate the signal representation. In section 3, we introduce the support vector machine framework and its associated theoretical underpinnings. The treatment is partly tutorial for the benefit of those working in speech recognition who might not have been exposed to it previously. In section 4, we investigate issues that are involved in the successful deployment of SVMs for stop detection. In section 5, we explore the detection mechanism based on SVMs in the context of acoustic phonetics and speech perception. Finally, we conclude by reiterating our major results.

2 A Brief Description of the Stop Detection Problem

Stop consonants are produced by causing a complete closure of the vocal tract followed by a sharp, sudden release. Hence they are signalled in continuous speech by a period of extremely low energy (corresponding to the

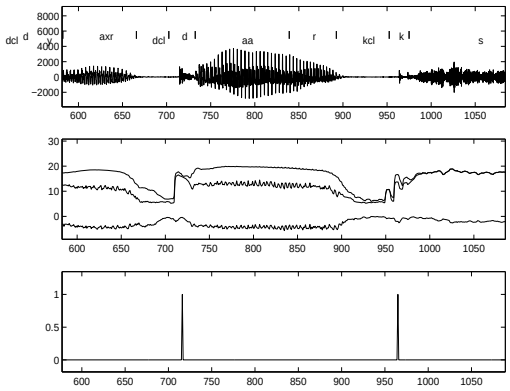

Figure 1: Portion of the speech waveform $s(n)$,(top panel), the associated three-dimensional feature vector, $\mathbf{x}(n)$ (middle panel), and the desired output $y(n)$ bottom panel marking the times of the closure-burst transition, x axis is time in msec.

period of closure) followed by a sharp, broad band signal (corresponding to the release). As a result, stops consonants are highly transient (dynamic) sounds that have a varying duration lasting anywhere from 5 to 100 ms. In American English, the class of stops consists of the sounds { p,t,k,b,d,g}.

In order to build a detector for stop consonants in running speech, the speech signal, $s(t)$, is characterized by a vector time series with three dimensions: (i) log(total Energy) (ii) log(Energy above 3kHz) (iii) spectral flatness measure based on Wiener Entropy defined as $\int log(S(f,t))df - log(\int S(f,t)df)$ where $S(f,t)$ is spectral energy at frequency f and time t. All quantities are computed using 5 ms windows moved every 1 ms. Thus, we have $\mathbf{x}(n) = [x_1(n)\ x_2(n)\ x_3(n)]'$ where n represents time (discretized in units of milliseconds) and x_1 through x_3 are the three acoustic quantities that are measured every 1 ms. Energies at 1 ms intervals potentially allow us to track rapid transitions that would otherwise be smoothed out by a coarser temporal resolution. For more on stop consonants in this context, see Niyogi, Mitra, Sondhi (1998) and Niyogi and Sondhi (2000,2001a).

We need to find an operator on the feature vector time series that will return a single dimensional time series that takes on large values around the times that stops occur and small values otherwise. The most natural points in time that mark the presence of stops are the transition from closure to burst release. Shown in fig. 1 is an example of a speech waveform $s(n)$, the associated feature vector time series $\mathbf{x}(n)$ and a desired output $y(n)$.

The technical goal is to find an operator g on the time series $\mathbf{x}(n)$ that produces an output $y_g = g \circ \mathbf{x}$ with values in {0,1} at each point in time, such that $\| y - y_g \|$ is small in some sense (norm). Specifically, we choose the optimal operator (from some class $\mathcal{G}$ of operators) according to the criterion

$$g_{opt} = \arg\min_{g \in \mathcal{G}} R(g) = \arg\min_{g \in \mathcal{G}} E[(y - y_g)^2] \tag{1}$$

Since both y and y_g have values in $\{0, 1\}$, this results in a two class, pattern classification problem[1] at each point in time n. In particular, we consider operators given by

$$y_h(n) = h(\mathbf{x}(n - W), \ldots, \mathbf{x}(n + W))$$

where h is a function from $R^{3(2W+1)} \longrightarrow \{0, 1\}$. In other words, we consider $2W + 1$-long blocks of the time series $\mathbf{x}$ at a time and attempt to classify them as stops or non-stops. Since $\mathbf{x}$ is itself a three-dimensional vector time series, this results effectively in a pattern classification problem in a $3(2W + 1)$-dimensional space. If h is a linear hyperplane in this space, the operator y_h becomes a linear filter and the problem reduces to one of optimal linear filtering (Wiener filtering of processes). If h is a non-linear function, the operator is a non-linear filter. We consider both situations in this paper. In the experiments that we report, $W = 5$.

In our approach to speech recognition, we decompose the phoneme recognition problem into a collection of feature detection problems that have a structure very similar to that of the stop detection problem described above. For example, a vowel detector might be built with a representation $\mathbf{x}$ consisting of dimensions like degree of periodicity in the signal, ratio of high frequency to low frequency energy, and so on. The same is true of other feature detectors like a nasal detector or a fricative detector. In all of these detection problems, the support vector machine framework might provide the basis for constructing optimal detectors that are learned from the data.

[1] It should be noted that the expectation $E[(y-y_g)^2]$ may be defined as the time average or the space average of the processes y and y_g. One may consider each speech utterance to yield a particular realization of the process $\mathbf{x}$ and therefore y and y_g are two 0-1 valued discrete time processes that lie in l_2 – the space of square summable sequences. Then $E[(y - y_g)^2]$ may be interpreted as the space average (over all possible realizations in the ensemble) of the quantity $||y - y_g||^2 = \sum_n (y(n) - y_g(n))^2$. This will be our interpretation throughout the rest of this paper.

3 Support Vector Machines

Thus the need arises to solve two-class pattern recognition problems. We discuss support vector machines for this purpose.

3.1 Theoretical Development

Consider a two-class pattern recognition problem with labelled examples, i.e., $(\mathbf{x}_i, y_i)$ pairs drawn according to some unknown distribution $P(\mathbf{x}, y)$ on the space $X \times Y$. The goal is to construct a function h (a mapping from X to Y or a decision rule) that is able to classify unknown patterns $\mathbf{x}$ into the appropriate class with minimum misclassification error. Such a function h is in principle chosen from some class $\mathcal{H}$ according to some predetermined criterion. The most commonly used criterion is the principle of empirical risk minimization, i.e.,

$$\hat{h}_l = \arg\min_{f \in \mathcal{H}} R_{emp}(f) = \arg\min_{f \in \mathcal{H}} \frac{1}{l}(y_i - f(\mathbf{x}_i))^2$$

Note that if $Y = \{0, 1\}$ and $\mathcal{H}$ is a set of indicator functions, then the empirical risk, $R_{emp}(f)$ is precisely the error rate on the training set of examples. This is the straightforward approach commonly pursued in statistical pattern recognition[2]. The hope is that the empirically chosen function $\hat{h}_l$ would generalize successfully to novel unlabelled examples.

If $\mathcal{H}$ were chosen to be large enough, it is always possible to reduce the training error to zero — unfortunately, such solutions are known to have poor generalization properties leading to the phenomenon of overfitting. On the other hand, if $\mathcal{H}$ is too small, then the function might perform suboptimally as it might be unable to express a potentially complex decision surface.

In the rest of this section, we will assume without loss of generality that $Y = \{-1, 1\}$ as this will be convenient for the discussion that follows.

3.2 Structural Risk Minimization

The straightforward approach of minimizing the empirical risk turns out not to guarantee a small actual risk on the test set, particularly, if the number l of training examples are limited. As a result of the work of Vapnik and

[2] In speech recognition, for example, this approach has been developed as the minimum error rate training procedure [4] to choose an opimal classifier (function; model) from a class of available classifiers.

Chervonenkis [23] (see also [20, 21]), a new induction principle has emerged, the principle of structural risk minimization (SRM). This is based on the fact that the true goal of the learner should be to minimize the *expected risk*, i.e.,

$$h_{opt} = \arg\min_{f \in \mathcal{H}} R(f) = \arg\min_{f \in \mathcal{H}} E[(y - f(\mathbf{x}))^2]$$

where the expectation is taken according to the true distribution P. h_{opt} is thus the function that provides the minimum expected misclassification error. Since P is unknown, one approximates the above functional $R(f)$ by the following large deviation bound (that holds with probability greater than $1 - \eta$)

$$R(f) \leq R_{emp}(f) + \Phi(\frac{d}{l}, \frac{\log(\eta)}{l}) \tag{2}$$

where Φ is defined as

$$\Phi(\frac{d}{l}, \frac{\log(\eta)}{l}) = \sqrt{\frac{d\left(\log\frac{2l}{d} + 1\right) - \log(\eta/4)}{l}}$$

The parameter d is called the VC–dimension [23] of the set of functions, $\mathcal{H}$ and is a measure of its complexity[3]. Some aspects of eq. 2 are worth highlighting. First, to guarantee minimization of the true risk, $R(f)$, one has to ensure that both terms, $R_{emp}(f)$ and $\Phi(\frac{d}{l}, \frac{\log(\eta)}{l})$ are made sufficiently small — significantly, it is noted that minimizing R_{emp} alone is not enough. Second, for a fixed amount of training data, l, the two components represent a complexity regularization trade-off. As the class $\mathcal{H}$ becomes larger, the minimum empirical risk becomes smaller but the VC term typically becomes larger. Generalization is controlled by controlling each of these two terms.

Conceptually this is done by imposing a structure $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \mathcal{H}_3 \ldots \mathcal{H}$ of nested subsets of $\mathcal{H}$. One then searches within this hierarchy for the class $\mathcal{H}_j$ that minimizes the bound, i.e.,

$$\hat{h} = \arg\min_{\mathcal{H}_j \in \mathcal{H}} (R_{emp}(f) + \Phi()) \tag{3}$$

[3]The VC dimension, d describes the capacity of a set of functions implementable by the learning machine. For binary classification, d is the maximal number of points in X that can be separated into two classes in all possible 2^d ways by using functions drawn from $\mathcal{H}$. Clearly, for two arbitrary classes, $\mathcal{H}_1$ and $\mathcal{H}_2$ such that $\mathcal{H}_1 \subset \mathcal{H}_2$, we have $d_1 \leq d_2$. See Vapnik, 1982 for more details.

3.3 Hyperplanes

Here we discuss the application of the SRM principle to the case where the class $\mathcal{H}$ is the class of linear hyperplanes in X, i.e.,

$$\mathcal{H} = \{h : X \to \{-1, 1\} | h(\mathbf{x}) = \theta(\mathbf{w}.\mathbf{x} + b)\}$$

where $\theta(z) = 1$ if $z \geq 0$, and $\theta(z) = -1$ otherwise. Thus each pair $(\mathbf{w}, b)$ corresponds to a unique hyperplane Note, however, that each hyperplane does not correspond to a unique $(\mathbf{w}, b)$ because $(\mathbf{w}, b)$ and $(a\mathbf{w}, ab)$ correspond to the same hyperplane for any arbitrary a. Therefore, a canonical form is required for hyperplanes to normalize out this scaling factor. This is done by representing each hyperplane by the pair $(\mathbf{w}, b)$ such that $\min_i |\mathbf{w}.\mathbf{x}_i + b| = 1$ where $(\mathbf{x}_i, y_i)$ are an arbitrary finite set of points in (X, Y) (taken later in this paper to be the set of training examples).

Now we use the following theorem [23]:

Theorem 1 *For an arbitrary set of points, $\mathbf{x}_1, \ldots, \mathbf{x}_r$ in X, let $B_{\mathbf{x}_1,\ldots,\mathbf{x}_r} = \{\mathbf{x} \in \mathbf{X} : \|\mathbf{x} - \mathbf{a}\| < V\}$ $(\mathbf{a} \in \mathbf{X})$ be the smallest ball containing those points and*

$$\mathcal{H}_A = \{f_{\mathbf{w},b} = \theta\left((\mathbf{w}.\mathbf{x}) + b\right) | \quad ||\mathbf{w}|| \leq A\} \tag{4}$$

be a subclass of hyperplanes in canonical form with respect to $\{\mathbf{x}_1, \ldots, \mathbf{x}_r\}$. Then, $\mathcal{H}_A$ has a VC-dimension d satisfying

$$d \leq V^2 A^2. \tag{5}$$

Thus, if we fix an arbitrary set of points in X with respect to which we normalize our hyperplanes to get their canonical forms, then the theorem suggests that by structuring hyperplanes according to the norm of their weights, $||\mathbf{w}||$, we can obtain a structured sequence of hyperplanes over which we can perform structural risk minimization. Clearly, $\mathcal{H} = \cup_{A>0}\mathcal{H}_A$. It is possible to show that utilizing such a structure transforms the problem posed in eq. 3 to

$$\hat{h} = \arg\min_{\mathbf{w},b} \left(R_{emp}(\mathbf{w}, b) + \lambda \mathbf{w}.\mathbf{w}\right)$$

λ trades-off the fit to data (measured by R_{emp}) with model complexity (measured by the VC-dimension).

In the rest of this paper, following the treatment of Vapnik (1995), the hyperplanes are made canonical with respect to the data points[4] in the

[4]Strictly, according to the classical structural risk minimization principle, the structure on the space of functions $\mathcal{H}$ has to be picked *a-priori* before any data has arrived. For a further discussion of this issue and modifications for the data-dependent case, see [20, 21].

training set, $\{\mathbf{x}_1, \ldots, \mathbf{x}_l\}$.

3.3.1 Separable Case

For the case, where the data is separable by hyperplanes, one way of implementing the structural risk minimization principle is to minimize the VC-dimension of $\mathcal{H}_A$ while keeping the empirical risk R_{emp} fixed at 0. This is equivalent to

$$\min \frac{1}{2}\mathbf{w}.\mathbf{w} \tag{6}$$

subject to

$$y_i(\mathbf{w}.\mathbf{x}_i + b) \geq 1 \quad \forall i \tag{7}$$

The ith constraint is satisfied only if the ith data point is correctly classified by the hyperplane classifier. Introducing Lagarange multipliers for each of the constraints, the Lagrangian is formed as

$$\mathcal{L}(\mathbf{w}, b, \{\alpha_i\}) = \frac{1}{2}\mathbf{w}.\mathbf{w} - \sum_i \alpha_i(y_i(\mathbf{w}.\mathbf{x}_i + b) - 1)$$

For such a quadratic programming problem with inequality constraints, the optimal solution lies at the saddle point of the Lagrangian and satisfies the following conditions [1]:

1. differentiating with respect to $\mathbf{w}$ and setting to 0 yields $\mathbf{w}^* = \sum_{i=1}^{l} y_i\alpha_i\mathbf{x}_i$, i.e., the optimal hyperplane is a linear combination of the training vectors,

2. differentiating with respect to b and setting to 0 yields $\sum_i \alpha_i y_i = 0$,

3. first order Kuhn Tucker conditions yield $\alpha_i(y_i(\mathbf{w}.\mathbf{x}_i + b) - 1) = 0$. From this we see that all points $\mathbf{x}_i$ for which α_i is non-zero lie on the *margin hyperplanes* $\mathbf{w}^*.\mathbf{x} + b = 1$ or $\mathbf{w}^*.\mathbf{x} + b = -1$. Such points are called *support vectors*. All other points are exterior points and do not enter in the expansion of the optimal hyperplane $\mathbf{w}^*$ since they have $\alpha_i = 0$. As a result of these properties, the learning machine is called the *support vector machine*.

Substituting for $\mathbf{w}$ in the Lagrangian yields the quadratic programming problem,

$$\max \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i y_i \alpha_j y_j (\mathbf{x_i}.\mathbf{x}_j)$$

subject to

$$\sum_i \alpha_i y_i = 0.$$

3.3.2 Non-separable Case

In many practical applications, a perfectly separating hyperplane does not exist. To allow for the possibility of examples violating (7), [3] introduce slack variables

$$\xi_i \geq 0, \quad i = 1, \ldots, l, \tag{8}$$

to get

$$y_i((\mathbf{w}.\mathbf{x}_i) + b) \geq 1 - \xi_i, \quad i = 1, \ldots, l. \tag{9}$$

The structural risk minimization approach to minimizing the guaranteed risk bound (2) consists of the following:

$$\min \Phi(\mathbf{w}, \xi) = \frac{1}{2}\mathbf{w}.\mathbf{w} + U \sum_{i=1}^{l} (\xi_i)^\epsilon \tag{10}$$

subject to the constraints (8) and (9).

In eq. 10, the $\mathbf{w}.\mathbf{w}$ corresponds to the VC-dimension of the learning machine as before. The term $\sum_{i=1}^{l}(\xi_i)^\epsilon$ (for small ϵ) is equivalent to the number of misclassifications on the training set and therefore a measure of empirical risk. The constant U therefore controls the trade-off between the empirical risk $\sum_{i=1}^{l}(\xi_i)^\epsilon$ and the VC-dimension of the learning machine. In actual practice, one has to make choices both for U and ϵ. For computational reasons, ϵ is chosen to be 1 because that translates the optimization problem of eq. 10 into a quadratic programming problem like that of the previous section (in fact, $\epsilon = 2$ does also). Introducing Lagrange multipliers for each of the constraints in eqs. 8 (λ_i's) and 9 (α_i's), we form the Lagrangian $\mathcal{L}(\mathbf{w}, \{\xi_i\}, b, \{\alpha_i\}, \{\lambda_j\})$ as before

$$\mathcal{L} = \frac{1}{2}\mathbf{w}.\mathbf{w} + U \sum_{i=1}^{l}(\xi_i) - \sum_{j=1}^{l} \lambda_j \xi_j - \sum_{i=1}^{l} \alpha_i (y_i((\mathbf{w}.\mathbf{x}_i) + b) - 1)$$

First order conditions show that the optimal hyperplane is given by $\mathbf{w}^* = \sum_i y_i \alpha_i \mathbf{x}_i$, an expansion in terms of the training vectors. Additionally taking into account the Kuhn-Tucker conditions, eq. 10 is transformed into

$$\max \sum_i \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y_i y_j (\mathbf{x}_i.\mathbf{x}_j)$$

subject to

$$\sum_i \alpha_i y_i = 0; 0 \leq \alpha_i \leq U.$$

Once the α_i's are obtained by solving the above problem, the optimal hyperplane is easily found by substitution and is easily seen to be a linear expansion in terms of the *support vectors*.

3.4 Non-linear Extensions: Kernels

So far we have discussed how linear hyperplanes can be optimally chosen within the framework of structural risk minimization. The final conceptual idea in support vector machines involves the construction of non-linear decision boundaries. This is done by transforming the data non-linearly (by a fixed non-linear transformation) and using linear techniques in the transformed space. We now describe this below.

Let us consider a transformation $\psi : \mathbf{X} \longrightarrow \mathbf{Z}$ which maps points $\mathbf{x}_i \in \mathbf{X}$ to corresponding $\mathbf{z}_i \in \mathbf{Z}$. We wish to construct hyperplanes in $\mathbf{Z}$ according to the principles outlined in the earlier sections. This ultimately reduces to solving optimization problems of the sort described earlier with $\mathbf{x}_i$'s replaced by $\mathbf{z}_i$'s. Significantly, we note that the only form in which $\mathbf{z}_i$'s appear in the optimization problem is inner products, i.e., $(\mathbf{z}_i, \mathbf{z}_j)$. Therefore, it is not really necessary to know what the $\mathbf{z}_i$'s actually are —- it is enough to know the inner product between pairs of them. Consequently, we characterize the transformation ψ by the innerproduct it imposes on the space $\mathbf{Z}$. Specifically, we consider mappings of the form $\psi_K : \mathbf{X} \longrightarrow \mathbf{Z}$ such that for any $\mathbf{z}_1, \mathbf{z}_2 \in \mathbf{Z}$, we have

$$(\mathbf{z}_1, \mathbf{z}_2) = K(\mathbf{x}_1, \mathbf{x}_2)$$

where K is a fixed, chosen kernel. Rather than choosing a transformation ψ, we choose a kernel K. The kernel K must be "well behaved" in the sense that it must ensure that the space $\mathbf{Z}$ created as a result is truly an inner product space (Hilbert space). It is enough if K is symmetric, positive semi-definite since we can then take recourse to Mercer's theorem [5] stated below:

Theorem 2 *If $K(\mathbf{x}, \mathbf{y})$ is a symmetric, positive semi-definite kernel on a compact space, then it has an eigenfunction expansion that can be written as*

$$K(\mathbf{x}, \mathbf{y}) = \sum_i \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{y})$$

where λ_i's are the eigenvalues that are guaranteed to be non-negative and the ϕ_i's are the corresponding eigenfunctions.

Each different choice of the kernel K defines a different choice of the transformation ψ_K. In fact, given Mercer's theorem, we can now get a feel for the transformation ψ_K. Let us consider a particular K with its eigenfunction decomposition as $K = \sum_i \lambda_i \phi_i(\mathbf{x})\phi_j(\mathbf{y})$. Then, any $\mathbf{x} \in X$ is mapped to its image $\mathbf{z} = [\sqrt{(}\lambda_1)\phi_1(\mathbf{x})\sqrt{(}\lambda_2)\phi_2(\mathbf{x}) \ldots \sqrt{(}\lambda_i)\phi_i(\mathbf{x}) \ldots]^T$. One can see that $\mathbf{z}$ lies in a potentially infinite-dimensional Hilbert space with inner product given by

$$(\mathbf{z}_1, \mathbf{z}_2) = \sum_i \lambda_i \phi_i(\mathbf{x}_1)\phi_i(\mathbf{x}_2) = K(\mathbf{x}_1, \mathbf{x}_2)$$

where $\mathbf{z}_1$ and $\mathbf{z}_2$ are the images of $\mathbf{x}_1$ and $\mathbf{x}_2$ respectively. The dimensionality of the transformed space Z is equal to the number of non-zero eigenvalues of the kernel K and could easily be infinite. The functions ϕ_i are in general non-linear.

If we set up the optimization problem appropriately, we see that the optimal hyperplane has the form

$$\mathbf{w}^* = \sum_i y_i \alpha_i \mathbf{z}_i$$

For a new data point $\mathbf{x}$ that needs to be classified, one proceeds by conceptually mapping $\mathbf{x}$ to $\mathbf{z}$ and computing inner products with $\mathbf{w}^*$. The corresponding decision rule given by:

$$I(\mathbf{x}) = \theta((\mathbf{w}^*, \mathbf{z}) + b) = \theta(\sum_i y_i \alpha_i (\mathbf{z}, \mathbf{z}_i) + b) = \theta(\sum_i y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b)$$

We see that the form of this decision rule is like a feed-forward neural network, or a kernel-based non-parametric scheme with two significant differences (a) the parameters are estimated by the principle of structural risk minimization (b) the number of "hidden nodes" or "basis functions" is chosen automatically by the procedure. Thus an important problem of model selection is resolved.

4 Experimental Results

As discussed in section 2, our particular way of formulating the stop detection problem has led to the need for solving two class pattern classification

problems for which the support vector technology described above is eminently suitable. We discuss in this section results of the various experiments we have conducted with such support vector machines.

Detection experiments were conducted on dialect region 4 of the TIMIT test set that consists of 32 speakers, 16 male and 16 female uttering 10 sentences each making for a total of 320 sentences in all. The representation consists of the three dimensional vector time series described earlier. Training was performed on 4 randomly chosen speakers from the TIMIT training set from different dialect regions. Each training speaker had 10 sentences making for a total of 40 sentences in all. The forty training sentences yielded 133 positive examples (closure-burst transitions corresponding to stops) and 10760 negative examples.

4.1 Hyperplanes

In this section we consider results obtained when the class of hyperplanes is used as a decision boundary between stops and non-stops. Since the data is non-separable, this corresponds exactly to the formulation of section 3.3.2.

4.1.1 A First Experiment

There were 10893 training data points in R^{33} that were separated using the SVM formalism of section 3.3.2. The value of U was set to $U = 1.$ 163 (1.5% of the total training vectors) support vectors were generated. The error rate on the training set consisted of 25 false positives and 37 false negatives. Values for the hyperplane $\mathbf{w}$ and the threshold b were automatically chosen by the SVM formalism.

Shown in fig. 2 is a distribution of $d(\mathbf{x}) = \mathbf{w}.\mathbf{x} + b$ for the positive and negative examples. Recall that $\frac{d(\mathbf{x})}{||\mathbf{w}||}$ is precisely the distance of each datapoint $\mathbf{x}$ to the separating hyperplane — hence $d(\mathbf{x})$ is proportional to the distance from the separating hyperplane. All positive examples that have $d < 0$ and negative examples that have $d > 0$ are misclassified. The region $-1 \leq d \leq 1$ corresponds to the *margin*, i.e., points that lie within the strip given by the hyperplanes $\mathbf{w}.\mathbf{x} + b = 1$ and $\mathbf{w}.\mathbf{x} + b = -1$.

Histograms have been used to characterize the distributions. The curves have been normalized to add up to 1. The histogram for non-stops is constructed from 10760 examples and is therefore considerably smoother than the one for stops which has been constructed from only 133 examples. The separation on the training set appears to be quite good. Interestingly, how-

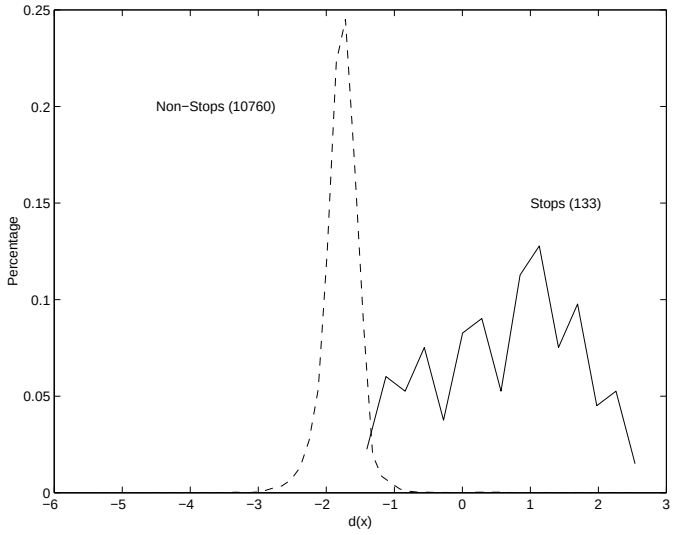

Figure 2: Histogram of $d(x)$ for stops and non-stops on the training set. There are 10760 non-stops and only 133 stops on which the support vectors were trained.

ever, there seems to be a greater spread among stops than non-stops. Additionally, a larger proportion of the stops are close to the boundary and appear as support vectors in the solution.

Shown in fig. 3 are the first twenty-two dimensions of the optimal hyperplane $\mathbf{w}$. Recall that the output of the operator on the time series $\mathbf{x}$ is given by $o(n) = \sum_{i=1}^{11} x_1(n-i-5)w_i + x_2(n-i-5)w_{i+11} + x_3(n-i-5)w_{i+22}$ where x_1 and x_2 correspond to the total energy and the high frequency energy respectively. w_j is the jth dimension of the hyperplane $\mathbf{w}$. Fig. 3 plots w_1–w_{11} and w_{12}–w_{22}. Clearly, the w_i's can be interpreted as the coefficients of a filter and from the shape of the w_i's it is clear that the first twenty two dimensions of the hyperplane essentially act like a differential operator by taking smoothed differences in total energy and high-frequency energy respectively.

For the problem that we consider here, the issue is really one of accurate *detection* of stops. Consequently, by changing the threshold (currently set at $d = 0$) one can obtain a trade-off between type I and type II errors for the stop detection problem. Shown in fig. 4 is the ROC curve generated by varying such a threshold of acceptance for the stops.

The '*' plotted in the figure is worth noting. It denotes the performance

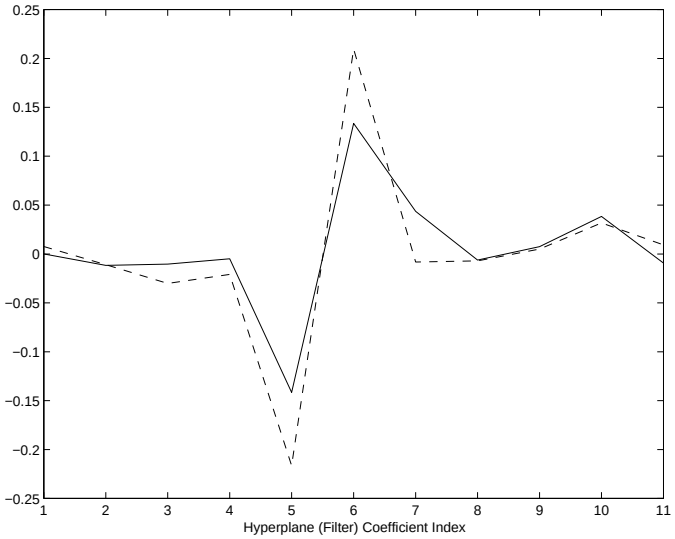

Figure 3: The first twenty two dimensions of the hyperplane $\mathbf{w}$ plotted. The solid line corresponds to w_1 through $w_1 1$ that act upon $x_1(n)$–the total energy. The dotted line corresponds to $w_1 2$ through $w_2 2$ that act upon $x_2(n)$–the high frequency energy. Both serve essentially as differential operators on the time series $x_i(n)$.

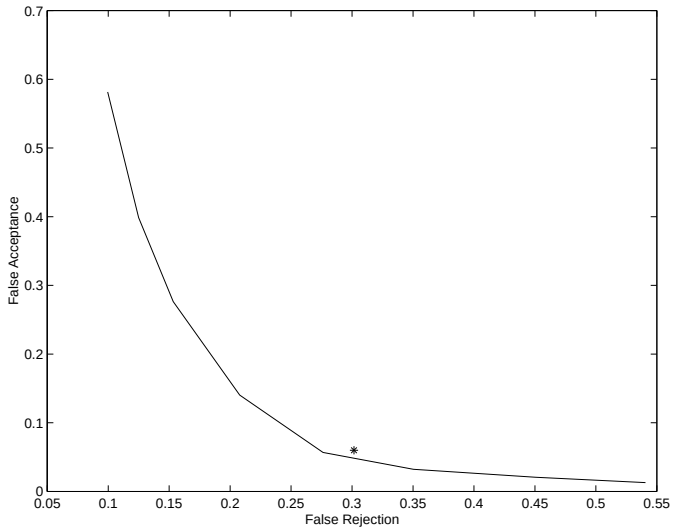

Figure 4: ROC curves for the support vector machine in our first experiment.

of a full blown HMM running with free grammar on the test sentences. The HMM based system consisted of 47 phones with 3 state left to right HMMs per phone. The output probabilites were mixtures of Gaussians (16 mixtures per state) and the front-end was a 39 dimensional vector time series obtained from the first 12 cepstral coefficients and total energy and their first and second differences (delta and delta-delta). A stop in a test sentence was considered to be correctly detected if the closure-burst transition was *anywhere* within the segment postulated as a stop by the HMM recognizer. If a particular segment postulated by the HMM recognizer as a stop did not contain a closure-burst transition, it was deemed a false insert.

4.1.2 Choice of U and its effect

In the earlier section, we had arbitrarily picked a value for $U = 1$. As we discussed earlier, U controls the trade-off between empirical fit to the data and capacity of the learning machine. Here, we consider a variety of choices for U and show that an appropriate choice of U can have a significant effect on the generalizing power of the solution obtained.

Shown in fig. 5 are the ROC curves obtained for each of the values of $U = 1000, 100, 10, 1, 0.2$ and 0.05. A significant trend that is noticed is that test performance improves as a function of U. Shown in table 1 are the values of U, the value of $\mathbf{w}.\mathbf{w}$ for the optimal hyperplane, the total number of support vectors, the training misclassifications, equal error rate (type I error rate = type II error rate from ROC curve) on the test set for each of the experiments. Notice that as U decreases, the value of $\mathbf{w}.\mathbf{w}$ (inversely proportional to the width of the margin and directly proportional to the capacity of the learning machine) decreases, the number of support vectors increase and the ratio of false positives to false negatives in the training set decreases. Correspondingly, over this range, the equal error rate of the detector on the test set becomes progressively less suggesting better generalization at lower values of U. This is not unreasonable since lower values of U de-emphasize the importance of training error and emphasize the importance of model complexity. If U can be decreased without significantly altering training error performance — and it can be argued that over the range of U considered here, this is the case — the test error (generalization ability) of the resulting SVM is likely to be favorable.

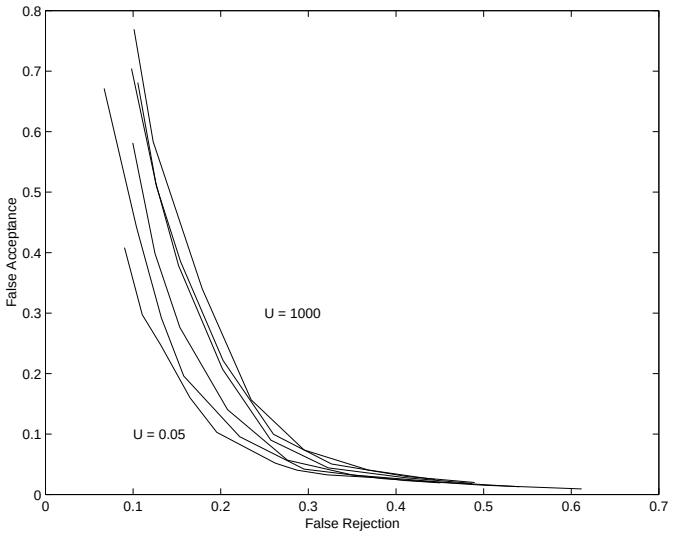

Figure 5: Performance of linear support vector machines as a function of U. ROC curves have been generated for values of U ranging from 1000 to 0.05. Performance on the test set (a measure of generalization) gets progressively better as U decreases over this range.

U	$\mathbf{w}.\mathbf{w}$	S.V. (+ve,−ve)	Training Errors	Equal Error
1000		156 (70,86)	56 (25,31)	22%
100	72.72	154	57 (25,32)	21.5 %
10	32.83	157	57 (24,33)	20.8 %
1	15.08	163	62 (25,37)	19 %
0.2	7.01	194	69 (27,42)	17.8 %
0.05	3.00	237	69 (20,49)	16.5 %

Table 1: Variation of SVM characteristics as a function of U (column 1). Shown are the maximum value of the objective function of the quadratic programming problem that is solved to get support vectors (column 1), total number of support vectors with break up between positive and negative labels (column 3), number of misclassifications in the training set (column 4) and equal error rate (column 5).

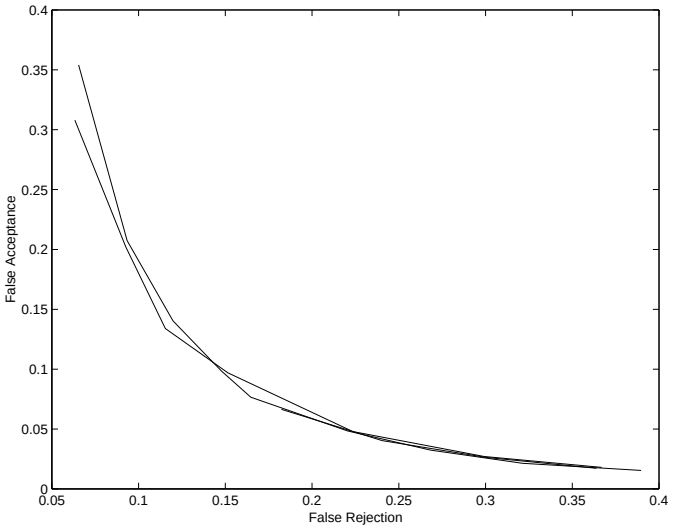

Figure 6: ROC curves for RBF support vector machines for various choices of σ^2.

4.2 RBF kernels

Another important choice in the adoption of the support vector framework for problems such as these involves the choice of kernel, K. In the experiments below, we considered Radial Basis Function kernels of the sort

$$K(\mathbf{x}, \mathbf{y}) = \exp -\frac{||\mathbf{x} - \mathbf{y}||^2}{\sigma^2}$$

Experiments were conducted with values of $\sigma^2 = 100, 250, 500$. Shown in fig. 6 are the ROC curves generated for each such choice of σ^2. For the values chosen, we do not see much variation in performance. However, the performance of the non-linear SVMs using such RBF kernels is significantly better than that of the linear SVMs that we have experimented with. Fig. 7 shows the histograms on the training set of the quantity $d(\mathbf{x}) = \sum_i (y_i \alpha_i K(\mathbf{x}, \mathbf{x_i})) + b$ where K is chosen to be an RBF kernel with $\sigma^2 = 500$. Recall that $d(\mathbf{x})$ is proportional to the distance from the optimal hyperplane in the higher dimensional space into which the data is non-linearly mapped. It is therefore a measure of non-linear separability just as it was a measure of linear separability in fig. 2. Again, positive training examples (stops) with values of $d < 0$ and negative training examples

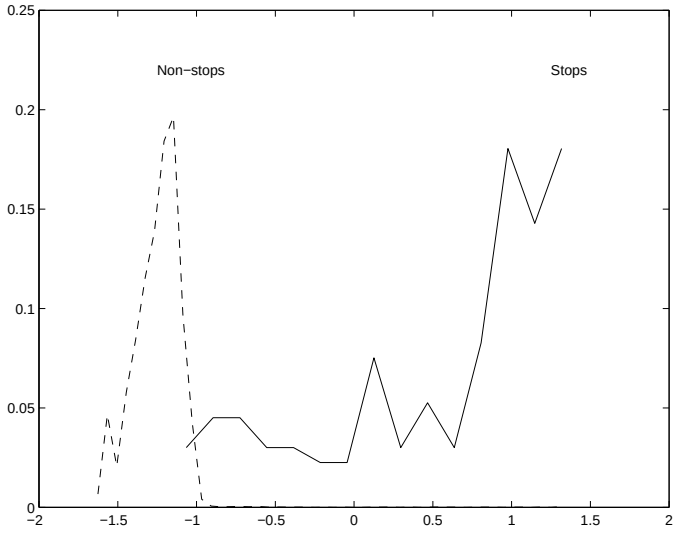

Figure 7: Histogram of $d(x)$ for stops and non-stops on the training set. $d(x)$ is computed using a non-linear SVM for an RBF kernel with $\sigma^2 = 250$. There are 10760 non-stops and only 133 stops on which the support vectors were trained.

(non-stops) with values of $d > 0$ are misclassified in the training set. The separability has improved over the linear case earlier. While this may not be obvious from the plot, the number of misclassifications for such a non-linear classifier is 51 compared to 62 for the linear classifier outputs plotted in fig. 2.

4.3 Converting SVM outputs to probabilities

From the preceding discussion, we see that the SVM output, $d(\mathbf{x}) = \mathbf{w}.\mathbf{x} + b$, is a real number that is proportional to the distance of the point $\mathbf{x}$ to the optimal maximum-margin hyperplane. The weight vector $\mathbf{w}$ is seen to have an expansion in terms of the support vectors and most generally, $d = \sum_i y_i \alpha_i K(\mathbf{x}, \mathbf{x}_i) + b$. If $K(\mathbf{x}, \mathbf{x}_i) = \mathbf{x}.\mathbf{x}_i$ then we have the linear separators in the original space X, otherwise we have non-linear separators.

As a result of this formulation, d is unbounded and it might be desirable to normalize it in some appropriate fashion for integration into larger systems which are composed of several support vectors for different features. Following [2], one possibility is to construct probability density estimates

on the one-dimensional space into which the outputs of the SVMs can be mapped. Thus, we have the following approximation:

$$P(\mathbf{x}|\text{stop}) \approx P(d(\mathbf{x})|\text{stop})$$

In other words, rather than doing density estimation in the high-dimensional space X where the data originally lives, we do density estimation in the lower dimensional space obtained by performing non-linear discrimination using SVMs. Due to drastic dimensionality reduction, the second density estimation problem is considerably easier than the first. At the same time, the power of non-linear discrimination has been used in a pattern classification framework to obtain this lower dimensional space on which density estimation is then performed. To obtain *a-posteriori* density estimates, we have

$$P(\text{stop}|\mathbf{x}) \approx \frac{P(\text{stop})P(d(\mathbf{x})|stop)}{P(\text{stop})P(d(\mathbf{x})|stop) + P(\text{non-stop})P(d(\mathbf{x})|\text{non-stop})}$$

Alternatively, one can obtain a likelihood ratio – $\frac{P(\mathbf{x}|stop)}{P(\mathbf{x}|non-stop)}$ – which by the Neyman Pearson lemma is sufficient for optimal detection. Shown in fig. 8 are the speech signal, the raw outputs of an RBF based SVM, the renormalized outputs that can be interpreted as the *a-posteriori* probability of a stop at each point in time (panels 3 and 4), and the log-likelihood-ratio $log(\frac{P(\mathbf{x}|stop)}{P(\mathbf{x}|non-stop)})$ at each point in time. The line densities for the stops were computed as a gamma-distribution and for the non-stops were computed as a gaussian using the data displayed in histogram form in fig. 7. The prior probability of a stop $P(stop)$ was chosen to be $\frac{133}{10893}$ (computed from number of closure-burst frames versus number of non-stop frames) and $\frac{1}{6}$ (prior at the phonemic level). The difference between panels 2 and 3 reflect this difference in the prior and show how it can significantly affect the scores. Consequently, the prior has be to carefully chosen with regard to the role the particular detector (in this case, stop detector) has to play in the overall system. In general, the MAP estimates seem to renormalize the outputs to highlight all peaks including spurious ones, the log-likelihood-ratio renormalizes to suppress all peaks except the strongest ones. Thus the one favors insertions and the other deletions.

4.4 A Bottomline

We have discussed how Support Vector Machines can be applied to the problem of stop detection and investigated some of the issues that arise when

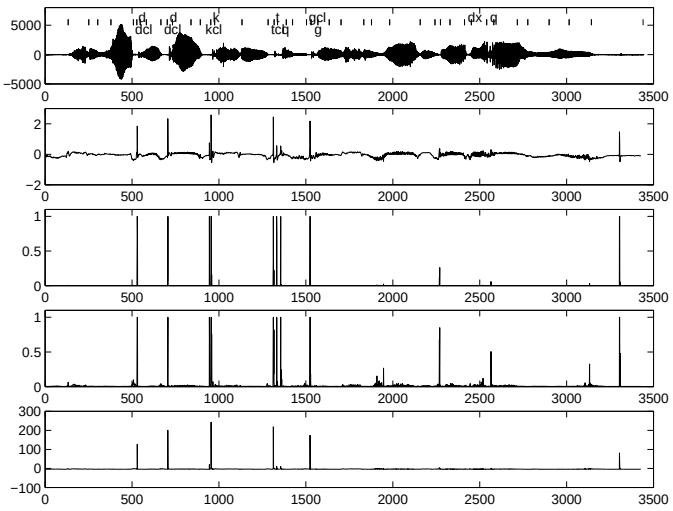

Figure 8: An example of raw SVM output (panel 2),renormalized as an *a-posteriori* probability (panel 3) aligned with the original speech signal (panel 1).

linear and non-linear variants of them are deployed for the task. In conclusion, it is worthwhile to summarize by noting three curves and a point in fig. 9 that represent the bottomline of this investigation. The three curves are ROC curves obtained by applying a non-optimized differential operator, an optimized linear operator (linear SVMs) and an optimized non-linear operator (non-linear SVM) on the same task of stop detection. The point represents the performance of an HMM on the same task. The gradual improvement from the differential operator to a non-linear SVM characterizes the improvement obtained as more sophisticated learning paradigms are employed on this task.

5 Kernels and the Perceptual Magnet Effect

From the discussion in the preceding sections, it is clear that the non-linear detectors that are created by introducing kernels in the support vector formalism provide significantly better performance for stop detection problems. The importance of non-linearities in speech perception has long been emphasized and has received some renewed attention in the context of the so called "perceptual magnet effect". In this section, we explore some of

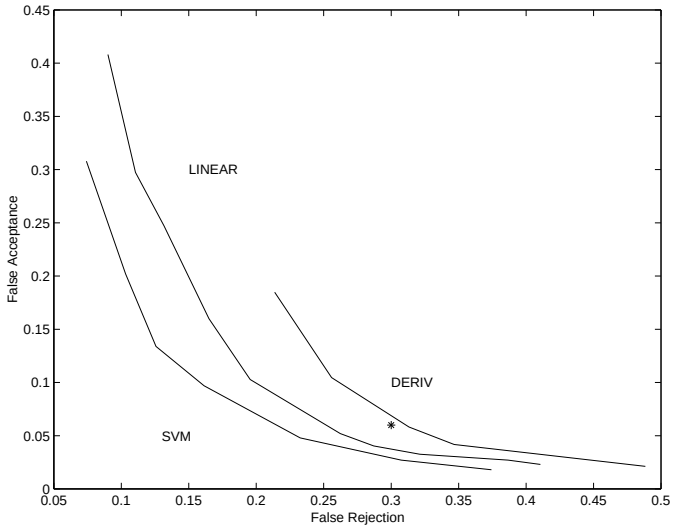

Figure 9: Performance of linear and non-linear support vector machines against derivative operators and HMMs.

the implications of the support vector formalism for speech perception and phonetics. In particular, we show how the kernel based transformations effectively define new metrics on the space of speech sounds and we see how these metrics are related to each other. We see that the final stop detector that is created utilizes a metric that is consistent with the magnet effect. The discussion here contributes towards the attempts (Ghitza and Sondhi, 1997; Nocerino et al, 1985; Harnad, 1987) to explicate the nature of the relationship between perceptual metrics, categorical perception, and speech recognition.

5.1 Support Vectors and their Phonetic Status

Let us begin by briefly examining the support vectors picked out by the algorithm and provide a phonetic interpretation for them. For the case where kernels based on radial basis functions are used and $\sigma^2 = 250$, there are 223 different support vectors that are found by the algorithm. 81 of these correspond to positive exemplars (closure burst transitions of stop consonants) and the rest correspond to negative exemplars. Since negative exemplars come from arbitrary 11 ms chunks in the utterance, their phonetic status is harder to interpret and we do not attempt a systematic analysis

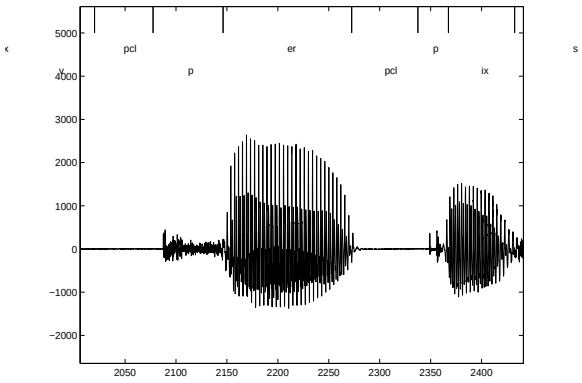

Figure 10: Two examples of /p/ in the same sentence. The one on the right is chosen as a support vector, the one on the left is not.

here.

On the other hand, positive exemplars all correspond to closure-burst transitions that have been hand segmented for the training database. Hence, they are easier to compare against each other and it is interesting to see which positive exemplars are chosen as support vectors by the algorithm.

The exemplars picked out by support vector machines essentially correspond to poorly articulated stop consonants. Shown in fig. 10 are two instances of /p/ following one another in the same utterance in the training set. The one on the left has a well articulated phonetic realization with a very strong closure burst transition. This does not correspond to a support vector. The one on the right has a much poorer realization and does correspond to a support vector. In general, since closure burst transitions are more pronounced for unvoiced stop consonants, we find, perhaps unsurprisingly, that 57% of all unvoiced stop consonants and 70% of all voiced stop consonants are chosen as support vectors.

It is worthwhile to make one additional remark. Support vectors are data points close to the decision boundary. This is clearly seen for stop consonants where perfect exemplars of stop consonants are far from the decision boundary and not chosen as support vectors. It is interesting to note that we have far more negative exemplars (10760) than positive (133). The support vectors are not distributed in the same proportion however. Only 1.32% of negative exemplars were chosen as support vectors while 60.1% of positive exemplars were chosen. This is presumably because vast parts of the speech signal are quite unlike closure burst transitions and

therefore far from the decision boundary. On the other hand, many more stop consonants (closure burst transitions) are confusible with transitions from a non-fricative sound to a fricative one and therefore chosen as support vectors.

5.2 Perceptual Magnet and Categorical Perception

Let us now consider two related but distinct aspects of human perceptual behavior in the context of the interpretation of speech like stimuli.

The first of these is *categorical perception*. This relates to the formation of linguistic categories and the interpretation (perception) of speech sounds as belonging to one of these categories. Various psychophysical experiments have been conducted from the 1950's onwards (see Harnad, 1987, for an extensive discussion) that elaborate on this behavior. Categorical perception is elicited, for example, when a varying acoustic stimulus is presented to a human subject and the latter is asked to identify the nature of the phoneme. For example, consider the distinction between the voiced labial stop /b/ and the unvoiced labial stop /p/. Imagine a one-dimensional acoustic continuum[5] where prototypical exemplars of /b/ lie at one end and those of /p/ at the other. As the sound is varied *gradually* from one end to the other along this continuum, the interpretation changes *abruptly* at a point somewhere in the middle that marks the transition boundary between one class and the other.

The second, and of greater interest to us here, is the *perceptual magnet effect* discovered and articulated by Kuhl, 1991 and elaborated on by others. This refers to the relationship between the perceptual distance and the acoustic distance between speech samples. Conceptually, one might imagine the space of speech sounds X with two different metrics imposed upon it: (i) the *acoustic* metric $d_A(x, y)$ which is the distance between the sounds measured in the acoustic or physical domain, and (ii) the *perceptual* metric $d_P(x, y)$ in the psychological domain which is a measure of how far the human perceptual system thinks these two sounds are. Consider a point $x \in X$ and another point $x' \in X$ such that $d_A(x, x') = \epsilon$. In other words, the two speech sounds are ϵ-apart in the acoustic domain. One elicits from the human subject[6] an estimate of $d_P(x, x')$. One repeats this experiment

[5] The most studied one-dimensional continuum for the voiced-unvoiced distinction between stop consonants is the Voice Onset Time (VOT).

[6] How one does so experimentally is a complicated story well beyond the scope of the current paper.

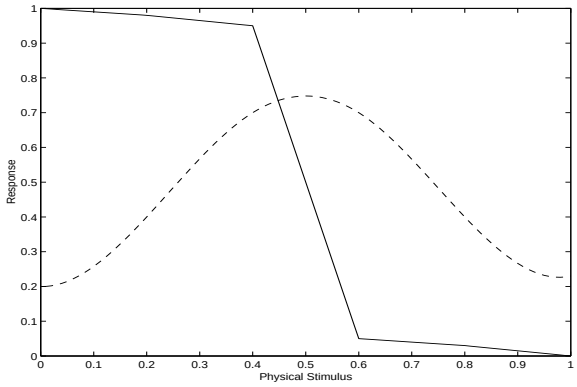

Figure 11: A schematic of categorical perception and the closely related magnet effect. The x-axis denotes the hypothetical value of the physical stimulus. The solid line plots the hypothetical response of over subjects as they are asked to categorize the stimulus into one of two linguistic categories. The y- value may be interpreted as the probability of making a judgement as belonging to category 1. This changes very dramatically in categorical perception. The dotted line denotes hypothetical discriminability between two equally spaced sounds as the location of the two neighboring physical stimuli is varied. The y-value may be interpreted as the "perceptual distance" between the two sounds. Discriminability ("perceptual distance") is high at the boundary and low away from it.

for different pairs x and x' in the space X. Thus all pairs x and x' are chosen so that their acoustic distance is held fixed at $d_A(x, x') = \epsilon$. Interestingly, it is found that the perceptual distance varies with location (x). In some regions of the space the perceptual distance is quite high and in others much lower. There is thus a warping of the space X. It is also found that the regions of high perceptual distance correspond to category boundaries and low perceptual distance correspond to *prototypical* exemplars of phonetic categories. For example, if one were to move along the one-dimensional continuum between /b/ and /p/ discussed earlier, one would get a picture as in fig. 11. In other words, discrimination is acute at category boundaries whereas far away from the boundaries, it is less so.

It is worthwhile to clarify the distinction between the more recent notion of the perceptual magnet effect and classical categorical perception which has been known for a long time. Phenomena surrounding categorical perception invoke the notion of categories and typically ask human subjects

to make category judgements – typically categorize (classify) the presented speech token into a phonemic class. In contrast, phenomena surrounding the perceptual magnet effect do not require the subject to ever make a category judgement[7]. Thus from a mathematical point of view, the perceptual magnet effect is simply the relationship between the topologies implied by the d_A and d_P metric on the space of sounds X. On the other hand, categorical perception (into two categories, for example) implies that some region of the space $X_1 \subset X$ is categorized as belonging to class 1 while the rest is categorized as belonging to class 2. In other words, it is enough to be able to define a discrimination function $f : X \to R$ such that $f(x) > 0$ for $x \in X_1$ and $f(x) < 0$ otherwise. The decision boundary is then given by the set $\{x|f(x) = 0\}$.

In principle, it may therefore be possible to define a discriminant function and thus realize categorical perception without any sort of perceptual magnet effect. In the next section, we consider a simple example where several natural discriminant functions exist but they might not give rise to a perceptual magnet effect. The human perceptual system, of course, displays both categorical perception and the magnet effect.

5.3 Prototypes and Category Boundaries

The perceptual magnet effect found that perceptual discriminability was acute near the boundary and less near prototypical exemplars of phonemes. Imagine a situation where the space X is one-dimensional, i.e., $X = R$. Let the acoustic distance between two sounds x and y by simply $|x-y|$. Further, let the sounds of class 1 be distributed according to a Gaussian distribution with a mean μ (variance $\sigma^2 = 1$) and the sounds of class 2 be distributed with a mean $-\mu$ (variance $\sigma^2 = 1$) as shown in the figure 12

[7] It should be pointed out that while classical categorical perception usually involved identification tasks, closely related experiments studying perceptual discriminability for sounds that were equally spaced have been conducted well before the formulation of the perceptual magnet effect. Thus while "classical" categorical perception would predict that two instances of the same category would be indistinguishable, later studies have clarified the more graded discriminability that exists between sounds of the same class with discriminability increasing towards the category boundary. (Again, see Harnad, 1987 for extensive discussion on such issues.) Usually such studies were conducted along a one dimensional physical continuum. In some sense, the perceptual magnet effect may be viewed as a generalization of this idea to higher dimensional spaces.

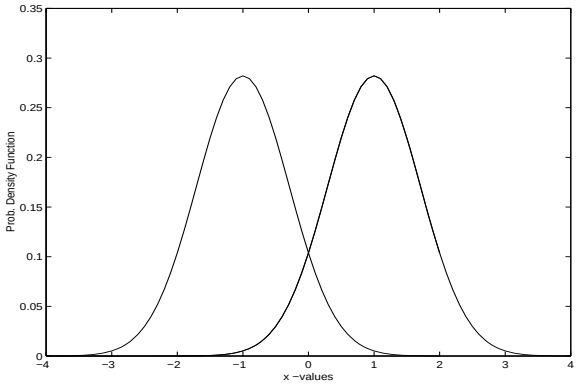

Figure 12: An example for distribution of sounds. There are two classes denoted by the two normal distributions. The probability distribution on the right is $P(x|class\ 1)$ and denotes the distribution of exemplars of class 1. Similarly, the distribution on the left is the distribution of exemplars of class 2.

Therefore,

$$P(x|class\ 1) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}$$

and

$$P(x|class\ 2) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x+\mu)^2}{2}}$$

In other words, the prototypes for the two classes are $x = \mu$ and $x = -\mu$. Assuming both classes are equally likely, the optimal decision boundary is simply given by $x = 0$. Thus any function $f(x)$ where $f(x) > 0 \Leftrightarrow x > 0$ is a candidate discriminant function that will realize categorical perception.

In general, it is natural to think of the discriminant function $f(x)$ as a measure of the "category goodness" of different exemplars. Thus, when x is near the boundary, the value of $f(x)$ is close to zero – far away from the boundary, it is large and either positive or negative as the case may be. We can therefore define the *iso-perception* sets associated with such a discriminant function to be

$$X_{f,\theta} = \{x \in X | f(x) = \theta\}$$

All elements of $X_(f, \theta)$ have the same degree of category goodness. Given the notion of the iso-perception sets, it is now reasonable to think of the

perceptual distance between two sounds x and y to be simply

$$d_P(x,y) = |f(x) - f(y)|$$

Let us consider the most natural discriminant function for our toy example with two classes – the *log likelihood ratio* function, i.e.,

$$f(x) = \log(\frac{e^{-\frac{(x-\mu)^2}{2}}}{e^{-\frac{(x+\mu)^2}{2}}}) = 2\mu x$$

It is easy to see that for this discriminant function, d_P is given by

$$d_P(x,y) = 2\mu|x-y| = 2\mu d_A(x,y)$$

Thus the perceptual distance defined by this discriminant function is a simple linear scaling of the acoustic distance between the tokens.

Another natural discriminant function is provided by the *a posteriori density* function as follows (recall, priors are equal, i.e., $P(y=1) = \frac{1}{2}$):

$$f(x) = P(y=1|x) - \frac{1}{2} = \frac{P(x|y=1) - P(x|y=0)}{P(x|y=1) + P(x|y=0)} = \frac{1}{1+e^{-2xu}} - \frac{1}{2}$$

Now consider the situation where $y = x + \epsilon$ so that the acoustic distance between x and y is always held fixed at ϵ. We can vary x and compute $d_P(x, x+\epsilon)$ for each x. This is given by

$$d_P(x, x+\epsilon) = |f(x+\epsilon) - f(x)| \approx \epsilon|f'(x)|$$

Using the expression for $f(x)$ above, it is easy to show that

$$d_P(x, x+\epsilon) \approx \epsilon(2\mu)(\frac{1}{2} + f(x))(\frac{1}{2} - f(x))$$

Since $f(x) \in (-1/2, 1/2)$, the perceptual distance is maximum at the decision boundary ($f(x) = 0$) and decreases symmetrically away from this boundary on both sides. This time the discriminant function realizes the perceptual magnet effect.

The point of this example is to merely show that it might be possible to define discriminant functions in many different ways yet not all will give rise to perceptual metrics[8] that are consistent with the magnet effect. In the next section, we consider the geometry of the non-linear support vector machines and the metrics implied by them. We also show that they display something like a perceptual magnet effect.

[8]Since the articulation of the perceptual magnet effect in Kuhl (1991), there have been a few attempts to explain and characterize this in computational terms (see Lacerda, 1995 and Guenther and Gjaja, 1996).

5.4 The Geometry of Support Vector Machines

Consider the situation with support vector machines. There is an original space $X = R^k$ in which the acoustic data lie with its natural topology implied by the Euclidean distance in this acoustic space. Points in X are now mapped into the infinite dimensional Hilbert space l_2 (the space of square summable sequences). Thus, each $x \in X$ is mapped into a sequence given by

$$x \rightarrow_\psi \begin{pmatrix} \sqrt{\lambda_1}\phi_1(x) \\ \sqrt{\lambda_2}\phi_2(x) \\ \cdots \\ \sqrt{\lambda_i}\phi_i(x) \\ \cdots \end{pmatrix}$$

where the λ_i's are the eigenvalues and the ϕ_i's are the eigenfunctions of the kernel $K(x, y) : R^k \times R^k \rightarrow R$. The support vector machine then constructs a hyperplane in this infinite dimensional space (the vector orthogonal to the hyperplane is given by $w \in l_2$) and therefore defines a discriminant function as

$$f(x) = (w, x') + b = \sum_i \alpha_i K(x, x_i) + b$$

where x' is the image of x under the mapping specified above.

The acoustic data space X is mapped nonlinearly (by the mapping ψ) into an infinite dimensional space. This embedding is mediated via the functions ϕ_i's. For Radial Basis Function kernels of the sort we use in this paper, the eigenfunctions are smooth and differentiable. As a result, the image of X (denoted by $\psi(X)$) is a manifold of dimension at most k and so does not span the entire higher dimensional space. In other words, the original acoustic space is twisted and embedded as a low dimensional manifold in an infinite dimensional space. As one moves in the original acoustic space $X = R^k$, the image move around in the manifold $\psi(X)$.

An optimal hyperplane separating positive and negative exemplars is constructed in the Hilbert space l_2 in accordance with the principles of structural risk minimization outlined in earlier sections. Points on the manifold are projected in the direction perpendicular to this optimal hyperplane. The discriminant function $f(x)$ is measure of distance from the optimal hyperplane along this direction and ultimately a measure of category goodness. Therefore, it is possible, as before, to define iso-perception regions in the original acoustic space and associated perceptual distances. Thus, consider two points x and $x + \Delta x$ where both x and Δx are k-dimensional vectors in

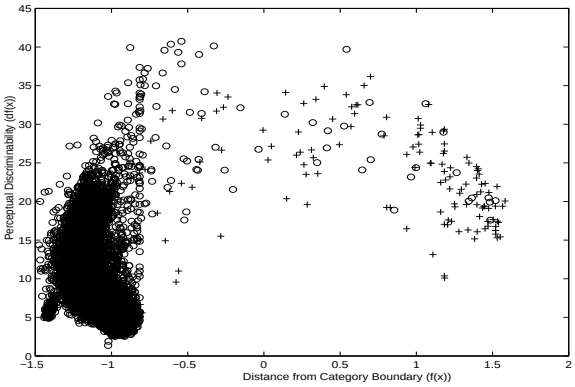

Figure 13: The perceptual magnet effect and SVMs. Each data point corresponds to an acoustic example denoted by "+" for stop consonants and "o" for the rest. The x-value for each data point denotes the distance to the optimal hyperplane found by the SVM technique. The$y$-value for the same data point denotes the magnitude of the perceptual change for a unit change in acoustics in the direction of greatest perceptual change along the perceptual manifold.

the acoustic space. The perceptual distance may be defined as

$$d_P(x, x + \Delta x) = |f(x) - f(x + \Delta x)| \approx |(\Delta x).(\nabla f)|$$

At the point x, the greatest perceptual change clearly occurs if one moves in a direction parallel to ∇f. Along this direction, we have (where $|\nabla f|_x$ is simply the gradient of f ($|\nabla f|$) evaluated at x)

$$d_P(x, x + \Delta x) \approx ||\Delta x|||\nabla f||_x = d_A(x, x + \Delta x)||\nabla f||_x \tag{11}$$

Equation 11 relates perceptual distance to acoustic distance for each point x in the acoustic space. Shown in fig. 13 is a plot of this relationship as x is varied. Each point in the scatterplot of fig. 13 corresponds to an acoustic data point. The data points corresponding to stop consonants are denoted by a "+" while the rest are denoted by "o". For each acoustic token x, the quantity $|\nabla f|_x$ is plotted against $|f(x)|$. Since $|\nabla f|_x$ is proportional to perceptual distance (keeping acoustic distance constant) and $|f(x)|$ is proportional to distance from the category boundary, fig. 13 essentially shows how perceptual discriminability changes with distance to the category boundary for different points in the original acoustic space X. Perceptual discriminability is high near the category boundaries and low away from it.

5.5 The Riemannian Structure of the Manifold

We see that the image $\psi(X)$ is a low-dimensional manifold in l_2. It turns out that this manifold admits a natural Riemannian structure. We explore this structure in this section and show how the embedding gives rise to a number of different distance metrics which we can then relate to each other.

Recall that the mapping $\psi : X \to Z$ (where Z is the space l_2) is performed via a positive definite kernel $K(\mathbf{x}, \mathbf{y}) : X \times X \to R$ such that $(,)_Z$ denotes the inner product in the space Z

$$(z_1, z_2)_Z = K(\mathbf{x}, \mathbf{y}) = \sum_i \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{y})$$

An infinitesmal change in $\mathbf{x}$ in the original acoustic space therefore gives rise to an infinitesmal change in the image space Z given by

$$||\psi(\mathbf{x}) - \psi(\mathbf{x} + d\mathbf{x})||^2 = (\psi(\mathbf{x}) - \psi(\mathbf{x} + d\mathbf{x}), \psi(\mathbf{x}) - \psi(\mathbf{x} + d\mathbf{x}))$$

$$= (\psi(\mathbf{x}), \psi(\mathbf{x})) - (\psi(\mathbf{x}), \psi(\mathbf{x}+d\mathbf{x})) - (\psi(\mathbf{x}+d\mathbf{x}), \psi(\mathbf{x})) + (\psi(\mathbf{x}+d\mathbf{x}), \psi(\mathbf{x}+d\mathbf{x}))$$

Since the inner product is given by the symmetric kernel K, this is equal to

$$K(\mathbf{x}, \mathbf{x}) - 2K(\mathbf{x}, \mathbf{x} + d\mathbf{x}) + K(\mathbf{x} + d\mathbf{x}, \mathbf{x} + d\mathbf{x})$$

Taking a Taylor expansion around $\mathbf{x}$ of $K(\mathbf{x}, \mathbf{x})$ we have

$$||\psi(\mathbf{x}) - \psi(\mathbf{x} + d\mathbf{x})||^2 = (d\mathbf{x})^T G(d\mathbf{x})$$

where G is a $k \times k$ matrix such that

$$G_{ij} = (\frac{1}{2} \partial_{x_i} \partial_{x_j} K(\mathbf{x}, \mathbf{x}) - \{\partial_{y_i} \partial_{y_j} K(\mathbf{x}, \mathbf{y})\}_{\mathbf{y}=\mathbf{x}}$$

Here, x_i and y_i refer to the ith component of $\mathbf{x}$ and $\mathbf{y}$ respectively. The partial derivative with respect to x_i is denoted as ∂_{x_i}.

Clearly, G defines the metric tensor and one can now compute geodesic distances on this manifold as $\mathbf{x}$ varies in the original space $X = R^k$.

Thus, given (i) the original space X, (ii) the space l_2 and (iii) the embedding in l_2 of the manifold given by $\psi(X)$, we can now define four different metrics between speech tokens as

1.
$$d_1(\mathbf{x},\mathbf{y}) = ||\mathbf{x}-\mathbf{y}||_X = (\sum_i (x_i - y_i)^2)^{\frac{1}{2}}$$

This is the Euclidean distance in the original space $X = R^k$ and is the natural measure of acoustic distance between the speech tokens $\mathbf{x}$ and $\mathbf{y}$.

2. The acoustic tokens $\mathbf{x}$ and $\mathbf{y}$ are mapped into l_2 and one can define the Euclidean distance between $\psi(\mathbf{x})$ and $\psi(\mathbf{y})$ as

$$d_2(\mathbf{x},\mathbf{y}) = ||\psi(\mathbf{x})-\psi(\mathbf{y})|| = (\sum_i \lambda_i(\phi_i(\mathbf{x})-\phi_i(\mathbf{y}))^2)^{\frac{1}{2}} = (K(\mathbf{x},\mathbf{x})+K(\mathbf{y},\mathbf{y})-2K(\mathbf{x},\mathbf{y}))^{\frac{1}{2}}$$

In this paper, we have taken the kernel K to be the radial basis function kernel given by
$$K(\mathbf{x},\mathbf{y}) = e^{-\frac{||\mathbf{x}-\mathbf{y}||^2}{\sigma^2}}$$

Therefore, $K(\mathbf{x},\mathbf{x}) = K(\mathbf{y},\mathbf{y}) = 1$ and we see that

$$d_2^2(\mathbf{x},\mathbf{y}) = (2(1 - e^{\frac{d_1^2(\mathbf{x},\mathbf{y})}{\sigma^2}}))$$

3. The geodesic (Riemannian) distance between $\psi(\mathbf{x})$ and $\psi(\mathbf{y})$ computed on the manifold $\psi(X)$ with the metric tensor G as derived above. For the radial basis function kernels of the sort we use in this paper, we can compute G and it is easily seen that $G_{ij} = 0$ if $i \neq j$ while $G_{ii} = \frac{1}{\sigma^2}$. In other words, $G = \frac{1}{\sigma^2}I$ where I is the identity matrix or

$$(d\mathbf{x})^T G(d\mathbf{x}) = \frac{1}{\sigma^2}(d\mathbf{x})^T(d\mathbf{x})$$

Performing the path integral along the shortest path, we see that the geodesic distance reduces simply to

$$d_3(\mathbf{x},\mathbf{y}) = \frac{1}{\sigma}||\mathbf{x}-\mathbf{y}|| = \frac{1}{\sigma}d_1(\mathbf{x},\mathbf{y})$$

Thus the mapping is such that the geodesic distance along the manifold is proportional to the Euclidean distance in the original space X between the points $\mathbf{x}$ and $\mathbf{y}$.

4. The distance measure that is relevant for the classification of stops and non stops projects the points on the manifold along the direction perpendicular to the optimal hyperplane chosen according to the construction of Support Vector Machines. Therefore, the fourth natural distance measure between $\mathbf{x}$ and $\mathbf{y}$ is given by

$$d_4(\mathbf{x}, \mathbf{y}) = |f(\mathbf{x}) - f(\mathbf{y})|$$

We have already studied the behavior of this distance measure as a function of $\mathbf{x}$ where $\mathbf{y} = \mathbf{x} + \Delta\mathbf{x}$ and $\Delta\mathbf{x}$ is small change in the direction of greatest change.

We can now provide a fairly complete picture of the operations that are performed on the acoustic data by using kernel based detectors such as Support Vector Machines. The acoustic data lie in the original acoustic space X with a distance d_1 naturally defined between the elements of X. The technique of Support Vector Machines proceeds by mapping the data non-linearly into a manifold that admits a Riemannian structure. This gives rise to three other natural distance measures. Of these d_2 and d_3 are seen to be a function of the acoustic distance d_1 between the speech tokens. Thus, if the acoustic distance were held fixed, then d_2 and d_3 would not vary at different points in the space. These distance measures therefore do not display the perceptual magnet effect. In contrast, the distance d_4 that arises out of projections in the direction perpendicular to the optimal hyperplane displays the magnet effect. This is the distance measure that is used in the ultimate classification or detection of stop consonants from continuous speech.

6 Conclusions

Stop detection belongs to a class of feature detection problems that naturally arise in a feature based approach to speech recognition. It is particularly challenging because stops present a transient signal with a period of rapid change that is often poorly characterized by standard cepstral features of the sort that are used in traditional speech recognition systems.

We have discussed how the problem can be cast as trying to discriminate between positive and negative examples of patterns of energy and Wiener entropy corresponding to stops and non-stops respectively. Consequently, we now need to search over a class of operators ($\mathcal{H}$) that allow us to perform

such a discrimination. An important issue that then has to be resolved is to choose the class $\mathcal{H}$ of the appropriate complexity, i.e., rich enough to capture potentially complex decision surfaces and yet small enough so that its parameters can be estimated well from the data at hand. We have introduced the principle of structural risk minimization that provides a framework that allows us to do this. We have also discussed the application of support vector machines within such a framework to the stop detection problem.

Two major issues have to be kept in mind for the successful deployment of this technology for detection problems. One is the choice of an appropriate U — this allows us to trade-off the fit to data with the model complexity; the other is the choice of an appropriate kernel K. We have investigated a few such choices and discussed their effect on our detection task. We have also discussed how to convert the output of the SVM detectors into *a-posteriori* probabilities so that they can be integrated into a larger system composed of many such detectors.

Our performance results are summarized in fig. 9 that depicts the bottomline for our recognition experiments. There is a steady improvement from the linear to non-linear SVMs. Furthermore, both mark a significant improvement over derivative functions and one kind of HMM used for the same task.

Not surprisingly, the non-linear SVMs show the best performance. In another significant part of this paper, we have discussed in considerable detail the nature of these non-linearities. We have elaborated on how these non-linearities are generated by the kernel and have provided an interpretation of this in geometric terms. We have shown a single kernel function gives rise to three additional distance metrics on the class of speech sounds in addition to the original acoustic distance between these sounds. Much of this discussion has particular relevance in the light of the perceptual magnet effect. We have shown how only one of these new distance measures is consistent with the notions of the magnet effect.

References

[1] Burges, C.J., "A Tutorial on Support Vector Machines for Pattern Recognition," Data Mining and Knowledge Discovery, Vol. 2, No. 2, pp. 1-47, 1998.

[2] Burges, C.J., Chari, S., Niyogi, P., C. Nohl, " Discriminative Gaussian Mixture Models for Speaker Identification," *preprint.*, submitted to IEEE Transactions on Neural Networks.

[3] Cortes, C. and Vapnik, V. N., "Support Vector Networks," Machine Learning Journal, Vol. 20, pp. 1-25.

[4] Chou, W., Lee, C.H., Juang, B.H., Soong, F.K., "A Minimum Error Rate Pattern Recognition Approach to Speech Recognition," International Journal of Pattern Recognition and Artificial Intelligence, Vol. 8, NO. 1, pp. 5-31, 1994.

[5] Courant, R. and Hilbert, D., *Methods of Mathematical Physics,* J. Wiley, New York, 1953.

[6] Ghitza, O. and Sondhi, M.M., "On The Perceptual Distance Between Speech Segments," *Journal of the Acoustical Society of America,* 101(1):522-529, Jan. 1997.

[7] Guenther, F. and Gjaja, M. N., "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.

[8] S. Harnad (ed.), *Categorical Perception: The Groundwork of Cognition,* Cambridge University Press, 1987.

[9] Jakobson, R., Fant, G., Halle, M., *Preliminaries to Speech Analysis: The Distinctive Features and their Acoustic Correlates,* MIT Press Monograph, 1952.

[10] F. Jelinek, *Statistical Methods in Speech Recognition*, MIT Press, 1997.

[11] P. Kuhl, "Human adults and human infants show a 'perceptual magnet effect' for the prototypes of speech categories, monkeys do not." *Perception and Psychophysics,* 50, pp. 93-107, 1991.

[12] Lacerda, F., "The perceptual magnet effect: An emergent consequence of exemplar based phonetic memory," in Elenius, K. and Branderud, P. (ed), *Proceedings of the XIIIth International Congress of PHonetic Sciences,* vol. 2., pp. 140-147. Stockholm:KTH and Stockholm University.

[13] Niyogi, P. and Girosi, F., "On the Relationship between Generalization Error, Hypothesis Complexity and Sample Complexity for Regularization Networks," Neural Computation, Vol. 8.4, 1996.

[14] Niyogi, P. and Ramesh, P., "Incorporating Voice Onset Time to Improve Letter Recognition Accuracies," Proceedings of the ICASSP, Seattle, May 1998.

[15] Niyogi, P., Mitra, P., and Sondhi, M., "A Detection Framework for Locating Phonetic Events," Proceedings of ICSLP-98, Sydney, Australia.

[16] Niyogi, P. and Sondhi, M.M., "Detecting Stop Consonants in Continuous Speech," Bell Labs Tech. Report.

[17] Niyogi, P. and Sondhi, M.M., " Detecting Stop Consonants in Continuous Speech," submitted, *Journal of Acoustical Society of America.*

[18] N. Nocerino, F.K. Soong, L.R. Rabiner, and D.H.Klatt, "Comparative Study of Several Distortion Measures for Speech Recognition," *Speech Communication,* 4:317-331, 1985.

[19] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition,* Prentice-Hall, 1993.

[20] Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C., and Anthony, M., "A Framework for Structural Risk Minimization," Proceedings, 9th Annual Conference on Computational Learning Theory, pp. 68-76, 1996.

[21] Shawe-Taylor, J., Bartlett, P. L., Williamson, R. C., and Anthony, M., "Structural Risk Minimization over Data-Dependent Hierarchies," NeuroCOLT Technical Report NC-TR-96-053, 1996.

[22] Stevens, K.N. *Acoustic Phonetics*, MIT Press, 1998.

[23] Vapnik, V. N. *The Nature of Statistical Learning Theory,* Springer-Verlag, 1995.

[24] Vapnik, V. N. *The Estimation of Dependencies based on Empirical Data,* Springer-Verlag, 1982.

[25] Vapnik, V.N., *Statistical Learning Theory,* John Wiley, 1998.