PARAMETRIC SUBSPACE MODELING OF SPEECH TRANSITIONS

K. Reinhard and M. Niranjan

CUED/F-INFENG/TR.308

February 1998

Cambridge University Engineering Department
Trumpington Street
Cambridge CB2 1PZ
England

E-mail: kr10000@eng.cam.ac.uk, niranjan@eng.cam.ac.uk

Abstract

This report describes an attempt at capturing segmental transition information for speech recognition tasks. The slowly varying dynamics of spectral trajectories carries much discriminant information that is very crudely modelled by traditional approaches such as HMMs. In approaches such as recurrent neural networks there is the hope, but not the convincing demonstration, that such transitional information could be captured. The method presented here starts from the very different position of explicitly capturing the trajectory of short time spectral parameter vectors on a subspace in which the temporal sequence information is preserved. We approach this by introducing a temporal constraint into the well known technique of Principal Component Analysis. On this subspace, we attempt a parametric modelling of the trajectory, and compute a distance metric to perform classification of diphones. We use the principal curves method of Hastie and Stuetzle and the Generative Topographic map (GTM) technique of Bishop, Svenson and Williams to describe the temporal evolution in terms of latent variables. On the difficult problem of /bee/, /dee/, /gee/ we are able to retain discriminatory information with a small number of parameters. Experimental illustrations present results on ISOLET and TIMIT database.

Zusammenfassung

Dieser Bericht beschreibt den Versuch Informationen über dynamische Transitionen in phonetischen Sprachsegmenten zu erfassen, um sie für die Spracherkennung nutzbar zu machen. Gerade die dynamischen Prozesse der spektralen Trajektoren repräsentieren charakteristische Unterscheidungsmerkmale, welche durch die traditionellen statistischen Mustererkenner, wie z.B. `Hidden Markov Model`, ungenügend berücksichtigt werden. Man hoffte, durch die Anwendung von rekursiven neuronalen Netzen (RNNs) diese dynamischen Informationen besser in Systeme integrieren zu können, welches aber nicht überzeugend belegt werden konnte. In diesem Bericht wird von einem unterschiedlichen Blickwinkel aus gezeigt, wie Trajektoren, die aus spektralen Parametervektoren gebildet werden, explizit modelliert werden können. Diese Modellierung erfolgt in einem Unterraum, der zeitlich-sequenzielle Informationen erhält. Dies wird durch die Integrierung einer zeitbezogenen Nebenbedingung in die Standardmethode `Principal Component Analysis` realisiert. In diesem Unterraum erfolgt eine parametrische Modellierung der Trajektoren. Mittels einer Abstandsmetrik wird eine Klassifizierung von Diphonen vorgenommen. Mit den Methoden `Principal Curves` von Hastie/Stuetzle und der `Generative Topographic Map` (GTM) von Bishop, Svenson und Williams wird die zeitliche Entwicklung der Vektoren mit Hilfe von latenten Variablen beschrieben. An der Problematik zur Unterscheidung der Diphone /bee/, /dee/ und /gee/ mit Hilfe von charakteristischen Trajektoren zeigen wir, daß eine hohe Klassifizierungsrate erreichbar ist, wobei eine sehr geringe Anzahl von Parametern benötigt wird. Unsere Ergebnisse werden mit Hilfe der Datenbanken ISOLET und TIMIT experimentell illustriert, die in den Bericht integriert sind.

Contents

1 Introduction

One important feature of the complex temporal structure of speech signals is the systematic variation in the realisation of phones in different acoustic contexts. Smooth, and continuous, movement of the articulators towards and away from some notional target positions produces the acoustic signal that is rich in structure. It encodes not just the underlying linguistic content to be conveyed, but also much more information relating to the context in which it is spoken. In fluent speech the target positions towards which the articulators move may often not be realised, because of the movement towards the subsequent position may already have begun in parts of the system. In the corresponding acoustic signal, it would then be hard to isolate steady state regions that could be uniquely identified with phones. Discriminatory information enabling the decoding process is not localized in the steady states, but is likely to be smoothly distributed in the sequence of transitions of the signal.

The popular hidden Markov model (HMM) of speech signals, in its simplest form, approximates the signal as a sequence of statistically stationary regions. Once a segmentation is assigned, either at some stage of the iterative training process or at the Viterbi alignment in the test stage, the probabilistic score of the model is insensitive to the temporal ordering of the acoustic vectors that get assigned to a particular state. This clearly is a poor approximation to the dynamics of the vocal tract. The hard segmentation imposed by a finite number of states is also a poor model of the complex generation process.

Such weaknesses of the HMM approach have long been recognised. Techniques to deal with these include building models for context sensitive phones, and expanding the feature vector to include derivatives of spectral parameters. These refinements have resulted in highly successful speech recognition systems [45, 44] that can produce impressive accuracies on very large tasks. However, both specialized models for dealing with context and dimensionality expansion to capture ordering results in an explosion in the number of parameters. Robust estimation of a very large number of parameters then becomes the challenging task, requiring techniques such as tied mixtures.

Approaches such as Neural Networks [35, 7] attempt to optimise a nonlinear discriminant function that assigns a phone class membership probability to each spectral frame. Here too the slowly varying temporal dynamics is essentially ignored. Techniques to capture some dynamics include the use of a moving window (TDNN, Waibel et al. [43]) and the Recurrent Neural Network with state feedback [36].

In the speech research community it is well known that an adequate description of the temporal evolution of speech parameters are essential for robust and efficient synthesis and recognition. Work in the speech research community that emphasizes the importance of spectral transitions include that of Ahlbom et al. [3]. They showed using resynthesis experiments that segmental transitions may be used in reconstructing speech with minimal coarticulatory effect. Marteau et al. [30] show that dynamic information is of great importance for the recognition of high speed or very coarticulated transitions where it is difficult to detect any targets. They already suggested diphone-like segments with trajectory concepts.

Recently there has been much interest in the use of segmental models [32]. These attempts to model the time variation of a particular feature within a segment. Most approaches use phones as a segment. Stochastic trajectory models were used for modeling phone-based speech units as clusters of trajectories in parameter space. The trajectories are modeled by mixture of state sequences of multivariate Gaussian density functions to explain inter-frame dependencies within a segment. Similar results and methods for phone segments were reported being successfully used. Afify et al. [1, 2] and Gong et al. [22, 23] focused on trajectories which are sampled into n points within a segment and are represented by a mean and covariance vector for each point. Fukada et al. [17] represented the mean and covariance matrix by a polynomial fit within a segment. All of them found the mean and covariance matrix by employing a k-means algorithm to the representation space. In contrast Gish et al. [20], Goldenthal [21] and Holmes et al. [25] modelled each feature dimension directly using additional delta coefficients. Gish et al. modelled the mean vectors within a segment as a quadratic function but having only a limited covariance matrix variation per segment available. Holmes et al. modelled the trajectories using slope and mean to form a linear model within a segment using only a Gaussian mixture specific covariance matrix to represent the segmental variance. Goldenthal being aware of the statistical coefficient dependencies used the error component to enhance recognition results. Deng et al. [10] showed that the stationary-state assumption appears to be reasonable when a state is intended to represent a short segment of sonorant or fricative speech sound but in continuously spoken sentences, even vowels contain virtually no stationary portions [46]. They showed the importance of transitional acoustic trajectories for word segments reporting superior results over traditional HMMs on a limited task recognising 36 CVC words. A dynamical system segment model was proposed by Digalakis et al. [11, 12, 13] which resulted in significant improvement over the independent frame model for phone recognition.

Although all approaches try to circumvent the frame independence assumption within a segment and report improved results in comparison to frame independent models, the inter-segment correlation between segments is still modelled using the statistical independent assumption. This in particular doesn't hold for phones as segments where the acoustic transitions are located at the segment boundaries rather than in the segment centers. The spectral trajectory of say the vowel [i:] is quite different in the CV syllable /bee/ from that in the syllable /gee/. Clearly, a model for the phoneme [i:] derived from occurrences [i:] in all contexts would be noisy due to co-articulation. In this work we focus on diphones as units of speech carrying transitional information between acoustic targets. The motivation is partly due to the work of Ghitza and Sondhi [19], who also used diphones to represent non-stationary acoustic information. They used diphone units as states in an hidden Markov model framework to circumvent the independent and identical distribution assumption for successive observations within a state. Further diphones as units of concatenation has been very effective in producing synthetic speech [37].

In a parametric space (i.e. cepstral space) a speech signal can be represented as a point which moves as articulatory configuration changes. The sequence of moving points is called a trajectory of speech. We address the problem of acoustic modeling of speech at a diphone level. The model is motivated by the following ideas:

1. Context affects the trajectory of speech signals. Models for speech recognition should rely on the trajectory of speech vectors rather than on the geometrical position of observations in the parameter space, since a given point can belong to different trajectories.
2. The realisation of trajectories of a diphone form characteristic transitions that relate to acoustic context.
3. If diphones are modelled as a sequence of states, then, due to contextual variability, the distribution variance at the boundaries of a speech model is smaller than that of the center

part of the model. Joining models together will make the inter-model independency assumption less important. A weighting giving more importance to the extremities of the model in the recognition decision would thus improve the accuracy.

4. Diphones as speech model implies a certain inherent syntactic constraint on possible state sequences, quite apart from any additional grammatical constraints that might be imposed.

In this report we describe an attempt to capture segmental transition information of diphones in a speech recognition context. We look for the trajectory of the spectral parameter vector, projected on a subspace. On this subspace we impose a parametric model of the trajectory. Transitions corresponding to different diphones result in different representations in the subspace. We quantify the discriminant information retained on the subspace by demonstrations on small scale speech recognition tasks on the ISOLET and TIMIT database.

We present a method of modeling transitions in diphones in a subspace framework which is easy to model and requires a small amount of training data. We illustrate our hypothesis on a typical problem in speech recognition, the discrimination of /b/, /d/ and /g/ in the context of /ee/, an ambiguous problem in phone classification. This method shares the same idea of modeling dynamic transitions of speech with many other methods developed in the recent years mentioned above [21, 41, 11, 28]. However, because we are deriving our model in a low dimensional space, this approach does not increase model complexity for modeling the dynamics in speech. The incorporation of this information in existing recognition systems could be made with an N-best rescoring scheme, proposed by Schmid and Barnard [38, 39] and Rayner [33], to improve recognition results. This report is organised as follows. Section 2 starts with the basic details of the front end parameterisation. In 2.2 we discuss the subspace projection technique used. We introduce a simple idea to enforce temporal ordering information into principal component projection of the data. In section 3 we describe three approaches: a simple representation in terms of an average, the principal curves idea of Hastie & Stuetzle and the Generative Topographic map of Bishop as mechanisms employed to model trajectories. Section 4 describes the distance computations required in classification tasks, and in section 5 we describe the experimental illustration of these ideas.

2 Subspace Model

In this section we show that our acoustic transition can be represented in a low-dimensional space. A clue to how we can expect trajectory availability in low-dimensional space is given to us by the spectrogram, a two-dimensional representation of the short time Fourier transform, with frequency on the vertical, time on the horizontal axis and amplitude represented by a gray or colour scale. Within vocalised sounds and in particular at the boundaries between them, the spectrogram is characterised by smooth trajectories. In earlier work [34] we presented some preliminary results illustrating the visualisation of dynamic information within speech transitions in a lower dimensionality. Our aim was to find a suitable subspace representation where the discriminant information can be modelled. This approach would lead to a very simple model. Each speech unit consists of a transformation matrix which projects the p-dimensional data onto a L-dimensional plane which needs $L * p$ parameter (L $\ll$ p). The dynamics on this plane can be modelled by a stored template consisting of N frames needing $L * N$ parameters. Parameter requirements per model would be very small, since a maximum of $L * (p + N)$ parameters would be needed.

2.1 Feature Extraction

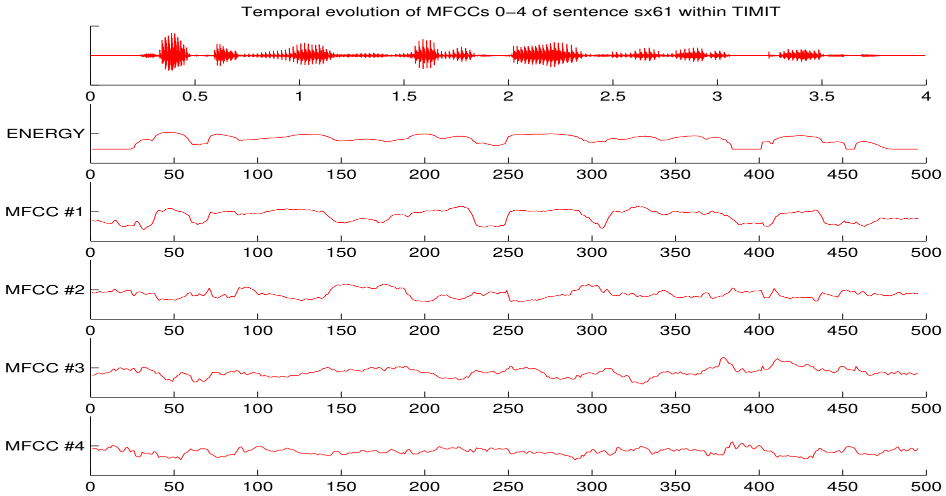
For short time spectral analysis, we use mel frequency cepstral coefficients (MFCC). A window size of 25ms which is usually referred to as a frame. The analysis is repeated at short steps. We use two different step sizes to show the importance and significance of our transition resolution. Step sizes of 5ms and 10ms were employed to capture the short-scale events, which results in 4 different feature representations. These are shown in Table 1. The MFCCs are calculated from the log filterbank amplitudes m_j using Discrete Cosine Transform where N is the number of filterbank channels.

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} m_j \cos\left(\frac{\pi i}{N}(j - 0.5)\right) \tag{1}$$

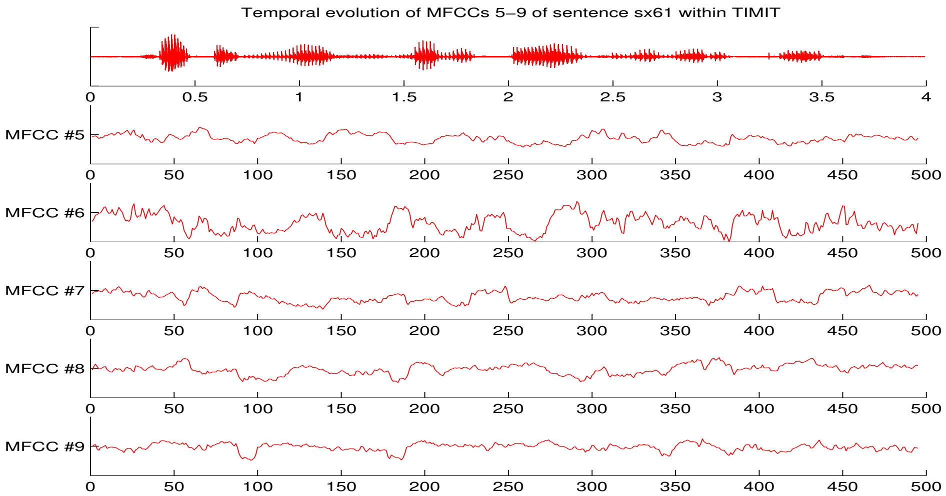
The number of coefficients used varys in our experimental illustration because recent findings showed that higher order MFCCs have oscillating characteristics and may not be suitable for trajectory modeling, which was shown by Hu and Barnard [26]. Lower order MFCCs represent speech transitions in a smoother way which will be shown in the results obtained for different speech segment representations (see Figure 1). Hence we use either 5 MFCCs or 10 MFCC (including energy) for our speech frame representation to distinguish the most suitable data space which results in optimal performance (see Section 5).

Name	**Speech frame representation**
REP1	5MFCC, 25ms window, 5ms step size
REP2	5MFCC, 25ms window, 10ms step size
REP3	10MFCC, 25ms window, 5ms step size
REP4	10MFCC, 25ms window, 10ms step size

Table 1: Speech frame representation used in the experimental evaluation



(a) MFCC 0-4



(b) MFCC 5-9

Figure 1: Scatterplot of the MFCC representation over time for the TIMIT (see Section 5.1) sentence **sx61** "Chocolate and roses never fail as a romantic gift" uttered from the male speaker **ABCO** from the training set using 25ms analysis window which is moved 5ms for each frame calculating the first 9 MFCCs plus energy (MFCC #0).

2.2 Subspace Definition

Since linear effects are directly captured by the covariance structure of the variable pairs the emphasis here is on the discovery of non-linear effects such as temporal transitions or other general non-linear associations among the variables. Friedman [16] and Huber [27] argued that, although arbitrary non-linear effects are impossible to parameterise in full generality, they are easily recognised when presented in a low-dimensional visual representation of the data density.

2.2.1 Time Constrained PCA (TC-PCA)

A very popular unsupervised technique for dimensionality reduction is the principal component analysis (PCA) or *Karhunen-Loéve-Transform* [14], the principal directions being given by the eigenvectors of the data covariance matrix. The aim is to find a set of M orthogonal vectors in data space that account for as much as possible of the data's variance. Projecting the data from their original p-dimensional space onto the M-dimensional subspace spanned by these vectors performs a dimensionality reduction. We add an additional constraint to the data to force the PCA algorithm to focus on temporal dependencies.

Given a data set $\mathcal{T}$ which consists of D sequences of N p-dimensional points $\mathcal{T} = \mathbf{T_1}, \cdots, \mathbf{T_D}$ with $\mathbf{T_k} = \mathbf{t_{k1}}, \cdots, \mathbf{t_{kN}}$, it is the temporal evolution of these vectors in each sequence that is of interest. In order to preserve the temporal sequence information, we expand the dimensionality of the data by one, using $\mathbf{t_{k*}} = \tau * (\mathbf{1}, \cdots, \mathbf{N})$. Hence $\mathbf{T_*} = \mathbf{t_{k*}}, \mathbf{t_{k1}}, \cdots, \mathbf{t_{kN}}$, the extra dimension representing a scalable frame ordering as time constraint. The scale factor τ is introduced to control the weighting imposed by this arbitrary choice of incorporating the order information. Tuning this parameter is achieved by an exhaustive search to determine the most discriminant subspace among all models during performance tests (see Section 5). Our subspace definition using TC-PCA can be described by solving the covariance matrix of the set of temporal extended vectors and is given by:

$$\Sigma^{\tau} = \sum_{\mathbf{k=1}}^{\mathbf{D}} \left[\sum_{\mathbf{i=1}}^{\mathbf{N+1}} (\mathbf{t_{ki}} - \bar{\mathbf{t}})(\mathbf{t_{ki}} - \bar{\mathbf{t}})^{\mathbf{T}} \right] \tag{2}$$

the solution of the minimisation problem with respect to the choice of basis vectors $\mathbf{u_i^{\tau}}$ leads to the equation

$$\Sigma^{\tau} \mathbf{u_i^{\tau}} = \lambda_{\mathbf{i}}^{\tau} \mathbf{u_i^{\tau}} \tag{3}$$

which is satisfied by $\mathbf{u_i^{\tau}}$ being the eigenvectors of the covariants matrix. Optimal dimensionality reduction can be performed in terms of projecting the data onto the eigenvectors corresponding to the largest eigenvalues λ_i^{τ}. Defining our 2-dimensional subspace which is dependent on the time constraint τ introduced above our transformation matrix is characterised by the following equation assuming that $\lambda_1^{\tau} \geq \lambda_2^{\tau} \geq \cdots \geq \lambda_{N+1}^{\tau}$.

$$\mathcal{P}_{\tau} = \begin{bmatrix} \mathbf{u}_1^{1^{\tau}} & \mathbf{u}_2^{1^{\tau}} \\ \vdots & \vdots \\ \mathbf{u}_1^{p^{\tau}} & \mathbf{u}_2^{p^{\tau}} \end{bmatrix} * \begin{bmatrix} \sqrt{\lambda_{\mathbf{1}}}^{\tau} & 0 \\ 0 & \sqrt{\lambda_{\mathbf{2}}}^{\tau} \end{bmatrix} \tag{4}$$

Such projection onto 2D enables interesting visualisation, which we think is an important first step. This is not, however, a necessary restriction. The subspace projection analysis can be carried out in any arbitrary dimension smaller than that of the original parameterisation.

3 Trajectory Models

The constrained projection outlined in the previous section, leads to a sequence of points in the subspace. The next stage is to characterise the evolution of these points in a manner that enables us to extract a distance metric with which we can classify these sounds. Three attempts at implementing such characterisation are described in this section. Later in this report we show experimental comparisons.

3.1 Meanframe Model

Our first representation is a simple average trajectory that minimises the frame-wise squared error. Denoting the training set $\mathcal{T}$ consists of D sequences of N p-dimensional points for a specific diphone model, $\mathcal{T} = \{\mathbf{T_1} \cdots \mathbf{T_D}\}$ with $\mathbf{T_k} = \{\mathbf{t_{k1}} \cdots \mathbf{t_{kN}}\}$ and $\mathbf{t_{kl}} \in \mathcal{R}^{\mathbf{p}}$. This leads to a sequence of average points where $\mathcal{E}_{\mathcal{T}}[\cdot]$ denotes the expectation, or ensemble average.

$$\begin{aligned} \mathcal{M}_{AV} &= \{\mathcal{E}_{\mathcal{T}}[\{\mathbf{t_{11}} \cdots \mathbf{t_{D1}}\}] \cdots \mathcal{E}_{\mathcal{T}}[\{\mathbf{t_{1N}} \cdots \mathbf{t_{DN}}\}]\} \\ &= \{\hat{\mathbf{t}}_1^{av}, \cdots, \hat{\mathbf{t}}_N^{av}\}. \end{aligned} \tag{5}$$

3.2 Principal Curves Model

The algorithm for principal curves by Hastie/Stuetzle [24] describes a method of extracting a smooth one-dimensional curve that passes through the "middle" of a p-dimensional data set providing a non-linear summary of the data. The algorithm for constructing principal curves starts with an initial guess, originally the first principal line, to define the initial ordering of the data and hence the neighbourhood relation. This line is then smoothly bent according to the actual data representation, by locally averaging p-dimensional points and iteratively minimising the orthogonal distances to the new curve. Below, we give a very brief description of the principal curves idea. For details, the reader is referred to Hastie/Stuetzle [24].

Given $\mathbf{T} = \{\mathbf{t_1} \cdots \mathbf{t_n}\}$ random vectors in $\mathcal{R}^{\mathbf{p}}$. Then principal curves is defined by $\mathbf{f}$, where $\mathbf{f}$ denote a smooth curve in $\mathcal{R}^{\mathbf{p}}$ parameterised over $\Phi \subseteq \mathcal{R}^{\mathbf{1}}$, a closed interval, that does not intersect itself ($\phi_1 \neq \phi_2 \Longrightarrow \mathbf{f}(\phi_1) \neq \mathbf{f}(\phi_2)$). The projection index $\phi_{\mathbf{f}} : \mathcal{R}^{\mathbf{p}} \longrightarrow \mathcal{R}^{\mathbf{1}}$ is defined as:

$$\phi_{\mathbf{f}}(\mathbf{t_i}) = \sup_{\phi}\{\phi : ||\mathbf{t_i} - \mathbf{f}(\phi)|| = \inf_{\mu} ||\mathbf{t_i} - \mathbf{f}(\mu)||\}. \tag{6}$$

Equation (6) is read as follows: A search for each data point $\mathbf{t_i}$ is performed over all points on the principal curve $\mathbf{f}$ parameterised over μ. For the closest point on the curve its parameter μ is assigned to ϕ as the best fit. The projection index $\phi_{\mathbf{f}}(\mathbf{t_i})$ is the value of ϕ for which $\mathbf{f}(\phi)$ is closest to $\mathbf{t_i}$. If there are several values, the largest is picked. The definition can be interpreted as starting with an initial guess to provide an ordering and collecting for any particular parameter value ϕ all observations that have $\mathbf{f}(\phi)$ as their closest point on the curve. If $\mathbf{f}(\phi)$ is the average of those observations, and if this holds for all ϕ, then $\mathbf{f}$ is called a principal curve. Because there is in general only one observation $\mathbf{t}_i$ related to a certain ϕ_i the observations projecting into a neighbourhood are locally averaged.

The principal curve algorithm iterates through *Projection-Expectation* steps until the relative change in the distance from the data points to its projections is below a certain threshold (see Figure 2(b)). The initial guess is very important because it should already represent temporal neighbourhood. Therefore our principal curve starts as a line segment which is generated by connecting each frame average by a straight line. The initial projection step of the data is hence calculated by projecting the data onto the line segments which defines a neighbourhood relationship using the parameterisation ϕ of the projected data points along the one-dimensional curve $\mathbf{f}(\phi)$. This carefully chosen initial line delivers a good first guess for the temporal ordering of the data (see Figure 2(a)).

$$\frac{|D^2(\mathbf{T},\mathbf{f}^{i-1}) - D^2(\mathbf{T},\mathbf{f}^{i})|}{D^2(\mathbf{T},\mathbf{f}^{i-1})} < \text{threshold}. \tag{7}$$

with

$$D^2(\mathbf{T},\mathbf{f}^i) = \sum_{i=1}^{n} ||\mathbf{t}^i - \mathbf{f}^i(\phi_i)||^2. \tag{8}$$

3.2.1 Expectation

For the expectation step, a locally weighted running line smoother is employed to estimate a new $\mathbf{f}(\phi)$ which will be used to calculate the new projected points and hence the new distance from the data points to its projections. In principle one considers a number of data points which project onto the neighbourhood of each projected point which is defined by a kernel function over ϕ to perform a linear regression. The linear regression delivers a new projected point which results in a new principal curve. A new projecting step is then performed to define a new temporal ordering. We used the algorithm for robust locally weighted regression suggested by Cleveland [8]. We calculated the parameters using linear regression, minimising the following expression, such that $\hat{\beta}_j(\phi_i)$ are the values of β_j:

$$\sum_{k=1}^{n} w_k(\phi_i)(\mathbf{f}(\phi_k) - \beta_0 - \beta_1\phi_k)^2 \quad \text{with} \quad w_k(\phi_i) = W(h_i^{-1}(\phi_k - \phi_i)). \tag{9}$$

and

$$\begin{aligned} W(x) &= (1-|x|^3)^3 \quad \text{for} \quad |x| < 1 \\ &= 0 \quad\quad\quad\quad\;\; \text{for} \quad |x| \geq 1. \end{aligned} \tag{10}$$

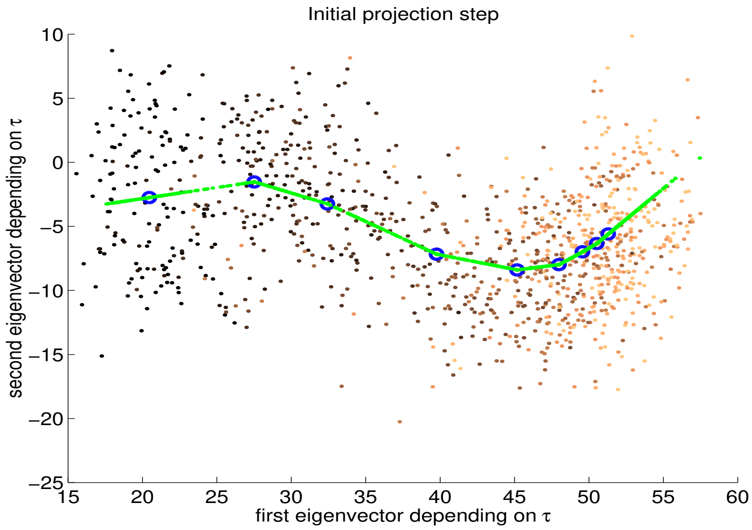
For each ϕ_i, weights, $w_k(\phi_i)$, are defined for all ϕ_k, $k = 1,\ldots,n$ using the weight function W. This is done by centering W at ϕ_i and scaling it so that the point at which W first becomes zero is at the r^{th} nearest neighbour of ϕ_i. For each i let h_i be the distance from ϕ_i to the r^{th} nearest neighbour of ϕ_i. That is, h_i is the r^{th} smallest number among $|\phi_i - \phi_j|$, for $j = 1, \ldots, n$. This procedure for computing the initial fitted values is referred to as locally weighted linear regression, where $\hat{\mathbf{f}}(\phi_i) = \hat{\beta}_0(\phi_i) + \hat{\beta}_1(\phi_i)\phi_i$.

3.2.2 Projection

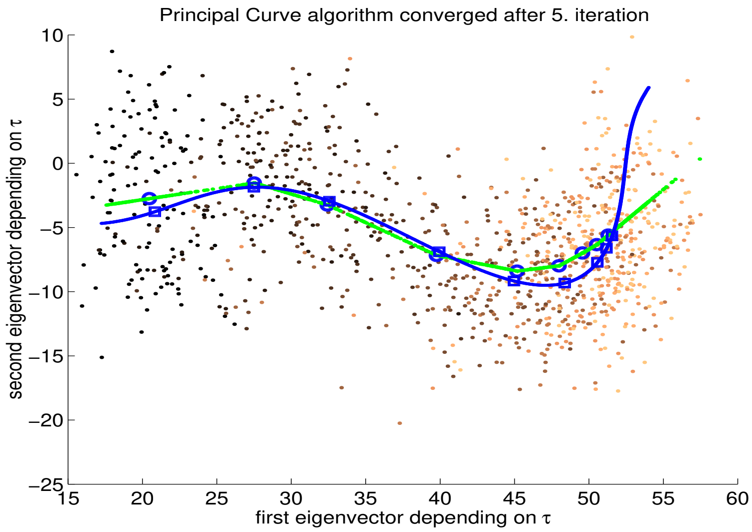
The aim of the projection step is to reorder the data according to its distance relationship to the new generated curve. Starting from the curve represented by n points $\mathbf{f}^{(\mathbf{j})}(\cdot)$ we compute for each $\mathbf{t_i}$ in the sample the value $\phi_i = \phi_{\mathbf{f}^{(\mathbf{j})}}(\mathbf{t_i})$ which represents the temporal ordering of our data. First we find the closest point on the line segment joining each pair $(\mathbf{f}(\phi_k), \mathbf{f}(\phi_{k+1}))$. The closest point to the curve is then either the projection onto a line segment or one of the $\mathbf{f}(\phi_k)$. Using these values to represent the curve, we replace ϕ_i by the arc length from $\mathbf{f}_\mathbf{1}^{(\mathbf{j})}$ to $\mathbf{f}_\mathbf{i}^{(\mathbf{j})}$. Potential problems with this projection step can be found where the expected values of the observations project onto $\mathbf{f}(\phi_{min})$ or $\mathbf{f}(\phi_{max})$. To circumvent this problem the corresponding data points generate a new $\mathbf{f}(\phi_{min})$ or $\mathbf{f}(\phi_{max})$ by projecting the data point onto the first or last line segment extending the original principal curve.

This algorithm doesn't use frame specific information but adjust the curve according to the density distribution of the training data. Hence it is very important that the initial guess supports the time correlation of the data points. To find a suitable trajectory model $\mathcal{M}_{PC}$, one has to find the closest point on the principal curve for each average frame point. This data driven approach adjusts the frame specific average points to the underlying data distribution.

$$\begin{aligned} \mathcal{M}_{PC} &= \{\mathbf{f}(\arg\min_{\mu} \|\hat{\mathbf{t}}_1^{av} - \mathbf{f}(\mu)\|) \cdots \mathbf{f}(\arg\min_{\mu} \|\hat{\mathbf{t}}_N^{av} - \mathbf{f}(\mu)\|)\} \\ &= \{\hat{\mathbf{t}}_1^{pc}, \cdots, \hat{\mathbf{t}}_N^{pc}\}. \end{aligned} \tag{11}$$



(a) Algorithm after initial temporal ordering



(b) Converged Principal Curve algorithm

Figure 2: Principal Curves generated for the diphone */dee/* for male speaker extracted from ISOLET database (see Section 5.1). The time-constrained transformation to emphasize temporal ordering is using $\tau = 2.8$. (a) shows the initial guess which represents the first temporal ordering. (b) plots the converged principal curve and the found trajectory model as squares on the curve. Time evolution is shown in the scatterplot of the data in changing colours from dark to bright.

3.3 GTM Model

A probabilistic generative model to explain high dimensional data in terms of a lower dimensional, or hidden variable representation is the Generative Topographic Mapping (GTM) algorithm of Bishop [6, 5]. The hidden space is called latent space, this term is new in speech literature, and the generative mapping is non-linear. The goal of the latent variables $\mathbf{x} = (\mathbf{x_1}, \cdots, \mathbf{x_L})$ is to define the data distribution $p(\mathbf{t})$ of data in a p-dimensional space $\mathbf{t} = (\mathbf{t_1}, \cdots, \mathbf{t_p})$, with $L <$p. Since in reality the data will only approximately live on a lower-dimensional manifold, it is appropriate to include a noise model for the $\mathbf{t}$ vector. The non-linear mapping function $\mathbf{y}(\mathbf{x}, \mathbf{W})$ transforms the latent variables into data space where they represent centers of radially-symmetric Gaussians having variance β^{-1}.

$$p(\mathbf{t} \mid \mathbf{x}, \mathbf{W}, \beta) = \Big(\frac{\beta}{\mathbf{2}\pi}\Big)^{\mathbf{p/2}} \mathbf{exp}\Big\{-\frac{\beta}{\mathbf{2}}\|\mathbf{y}(\mathbf{x},\mathbf{W}) - \mathbf{t}\|^{\mathbf{2}}\Big\}. \tag{12}$$

If we specify a prior distribution $p(\mathbf{x})$ on the latent-variable space, this will induce a corresponding distribution $p(\mathbf{y}|\mathbf{W})$ in the data space. The distribution in $\mathbf{t}$-space, for a given value of $\mathbf{W}$, is then obtained by integration over the $\mathbf{x}$-distribution.

$$p(\mathbf{t}|\mathbf{W}, \beta) = \int \mathbf{p}(\mathbf{t}|\mathbf{x}, \mathbf{W}, \beta)\mathbf{p}(\mathbf{x})\mathbf{dx}. \tag{13}$$

For a given data set $\mathcal{T} = (\mathbf{t^1}, \cdots, \mathbf{t^N})$ one can determine the transformation matrix $\mathbf{W}$ and the inverse variance β by finding a maximum log-likelihood solution using the EM-algorithm. The likelihood is given by

$$\mathcal{L}(\mathbf{W}, \beta) = \mathbf{ln}\prod_{\mathbf{n=1}}^{\mathbf{N}} \mathbf{p}(\mathbf{t_n}|\mathbf{W}, \beta). \tag{14}$$

Choosing a sum of delta functions centered at the latent points $p(\mathbf{x}) = \frac{\mathbf{1}}{\mathbf{K}} \sum_{\mathbf{i=1}}^{\mathbf{K}} \delta(\mathbf{x} - \mathbf{x_i})$ the integral in equation 13 can be performed analytically, giving

$$p(\mathbf{t}|\mathbf{W}, \beta) = \frac{\mathbf{1}}{\mathbf{K}} \sum_{\mathbf{i=1}}^{\mathbf{K}} \mathbf{p}(\mathbf{t}|\mathbf{x_i}, \mathbf{W}, \beta). \tag{15}$$

The log likelihood then becomes

$$\mathcal{L}(\mathbf{W}, \beta) = \sum_{\mathbf{n=1}}^{\mathbf{N}} \mathbf{ln}\Big\{\frac{\mathbf{1}}{\mathbf{K}} \sum_{\mathbf{i=1}}^{\mathbf{K}} \mathbf{p}(\mathbf{t_n}|\mathbf{x_i}, \mathbf{W}, \beta)\Big\}. \tag{16}$$

Using a particular parameterised form of $\mathbf{y}(\mathbf{x}, \mathbf{W})$ by a generalised linear regression model of the form $\mathbf{y}(\mathbf{x}, \mathbf{W}) = \mathbf{W}\Phi(\mathbf{x})$ where the elements of $\Phi(\mathbf{x})$ consists of M fixed basis functions $\Phi_j(\mathbf{x})$, one is able to maximise $\mathcal{L}(\mathbf{W}, \beta)$ finding the optimal weight matrix $\mathbf{W}^*$ and a inverse variance β^*.

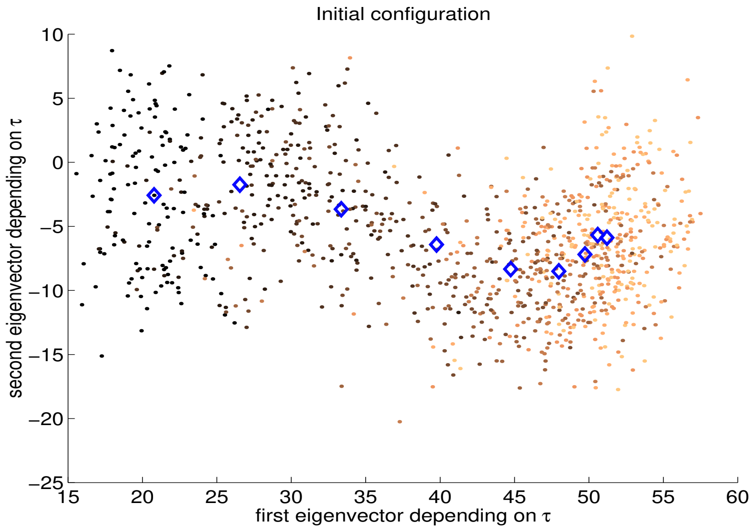
For our trajectory model $\mathcal{M}_{GTM}$ we are defining a one dimensional latent space (L=1) with as many latent variables as numbers of frames per speech unit. Using as an initial weight matrix $\mathbf{W}$, the transformation from the latent space to the average points in each frame in data space (see Figure 3 (a)), we expect the positions of the Gaussian centers to be optimised according to the data density distribution (see Figure 3 (b)).

$$\begin{aligned}\mathcal{M}_{GTM} &= \left\{\mathbf{y}(\mathbf{x_1}, \mathbf{W}^*), \cdots, \mathbf{y}(\mathbf{x_N}, \mathbf{W}^*)\right\} \\ &= \{\hat{\mathbf{t}}_1^{GTM}, \cdots, \hat{\mathbf{t}}_N^{GTM}\}. \end{aligned} \tag{17}$$

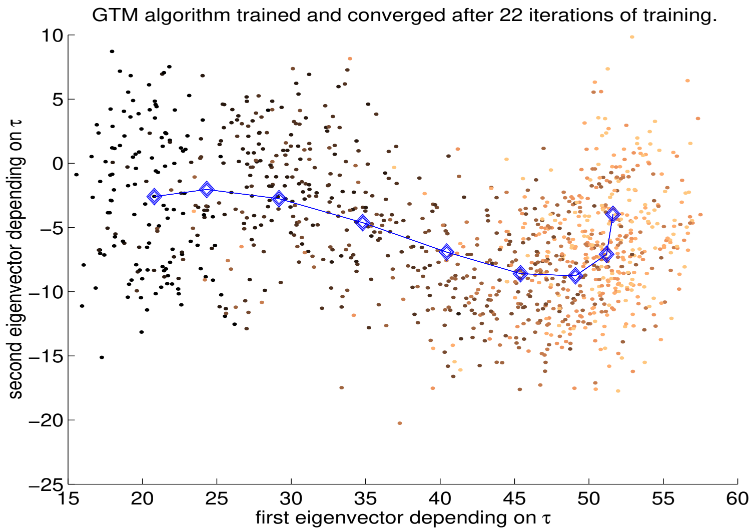
The Principal Curve algorithm discussed in section 3.2 can be interpreted as a latent space model for density distribution estimation shown by Tibshirani [42], similar to the Generative Topographic Mapping model. The similarity between both approaches was shown by Bishop [6]. Rather than defining as many latent points as number of frames present in the trajectory model, Principal Curves is using as many latent points as data points are available. In both cases one likes to find a sequence of points in latent space which correspond to center of Gaussians in data space representing the local density distribution. Principal Curves is starting from our initial location of projecting the data points onto the line segments between the average values for each frame found for our meanframe model. We adjust the location of the latent points according to the real data distribution which is represented by data points which project onto the neighbourhood of each latent point. Hence, moving along the principal curve we adjust each latent point which represents a center of a density model iteratively by considering the local weighted average of data points projecting onto the neighbourhood of the latent point. The projected neighbourhood defines a kernel function to define influencing data. After each iteration the expectation step adjusts the position of each latent point in p-dimensional space and the projection step defines the new neighbourhood relationship. Finally our Principal Curve is the sequence of latent points.

In the case of GTM we start with an initial guess for the location of the Gaussian centers using our average frame points, these positions in the p-dimensional space are then adjusted according to the responsibilities of the local data points. The GTM algorithm is using spherical Gaussians with a few latent points. It constrains the solution to equidistant points within the high dimensional data which is not necessarily a good reflection of the real underlying generation process.

An attempt to model time varying signals using the GTM framework is reported in Bishop et al [4]. It is worth noting the difference between their approach and ours. Bishop et al. use a hidden Markov model in which the hidden states are given by a two dimensional latent variable of the GTM model. Our model, on the other hand, characterises the trajectory through time by a sequence of latent variable in a one dimensional space. Here the temporal sequence of data points is represented by the ordering of the latent space. In our temporal constrained subspace the latent space induces a sequence of trajectory knots which represents the temporal ordering through time which can be used for similarity measures.



(a) Algorithm after weight matrix initiation



(b) Converged GTM algorithm

Figure 3: Projected latent space generated by GTM for the diphone /*dee*/ for male speaker extracted from ISOLET database (see Section 5.1). The time-constrained transformation to emphasize temporal ordering is using $\tau = 2.8$. (a) shows the initial guess which represents the first temporal ordering. (b) plots the converged GTM algorithm and the found trajectory model as centers of Gaussians. Time evolution is shown in the scatterplot of the data in changing colours from dark to bright.

4 Trajectory Mapping

To use the ideas discussed above in a classification setting we impose a smoothing of the test data by fitting a constrained natural spline through it before projecting onto the subspace. This section describes the smoothing spline algorithm and the computation of distances in the projected space. This trajectory comparison method is motivated from the observation that the speech signal tends to follow certain paths corresponding to the underlying phonemic units. This was utilised by Sun [41] who modelled the target positions for phonetic units. He found that it is less important to model accurately the path of the intermediate positions that are results of the coarticulation process.

4.1 Smoothing Splines

This section introduces the idea of fitting a smoothing spline to find an optimal trade-off between accuracy and smoothness. Smoothing out a speech transition follows our motivation that diphones consist of larger contextual variability at the center of the speech unit whereas the extremities represent acoustically steady states.
Suppose we have N observations $t_1, \cdots, t_N$ of N distinct knots $a = x_1 < x_2 < \cdots < x_N = b$ and we assume that the data can be represented by the following model

$$t_i = S(x_i) + \epsilon_i; \qquad\qquad i = 1, \cdots, N \tag{18}$$

where $S(\cdot)$ is a deterministic function characterising the relationship between x's and t's and the ϵ_i's are independent random variables assumed to have Gaussian distributions.
The objective is to estimate the function $S(\cdot)$ which is as close to the data path and as smooth as possible. This idea can be formulated through the following objective function:

$$L(S) = \int_a^b \left[S^{''}(x)\right]^2 dx + \sum_{i=1}^{N} \omega_i \left(t_i - S(x_i)\right)^2 \tag{19}$$

where the ω_i's are the weights associated with t_i's representing the relative contributions of the i_{th} observation to the model estimation.
It is well known that the problem of minimising the objective function in (19) has a unique explicit solution [29], which is the natural cubic spline function. The estimation procedure is briefly described as follows. Let

$$\begin{aligned}
(\mathbf{T})_{\mathbf{N\times 1}} &= (t_1, \cdots, t_N)^T \\
(\mathbf{F})_{\mathbf{N\times 1}} &= (f_1, \cdots, f_N)^T = (S(x_1), \cdots, S(x_N))^T \\
(\mathbf{A})_{\mathbf{(N-2)\times 1}} &= (A_2, \cdots, A_{N-1})^T = (S^{''}(x_2), \cdots, S^{''}(x_{N-1}))^T
\end{aligned}$$

denote the sequence of data points, function values of $S(\cdot)$ and the second derivatives. Let $h_i = x_i - x_{i-1}$ denote the spacing of the x variable. The estimate of $S(\cdot)$ is obtained through the estimation of f_i's and A_i's. Hence fitting a smoothing spline to the given data points, one compensates for the high variance and makes a comparison to a given generalised trajectory model (see Section 3) more reliable [29].

$$
B = \begin{pmatrix}
\frac{h_2+h_3}{3} & \frac{h_3}{6} & 0 & \cdots & 0 \\
\frac{h_3}{6} & \frac{h_3+h_4}{3} & \frac{h_4}{6} & \cdots & \vdots \\
0 & & \ddots & & 0 \\
\vdots & & \cdots & & \frac{h_{N-1}}{6} \\
0 & \cdots & 0 & \frac{h_{N-1}}{6} & \frac{h_{N-1}+h_N}{3}
\end{pmatrix}.
$$

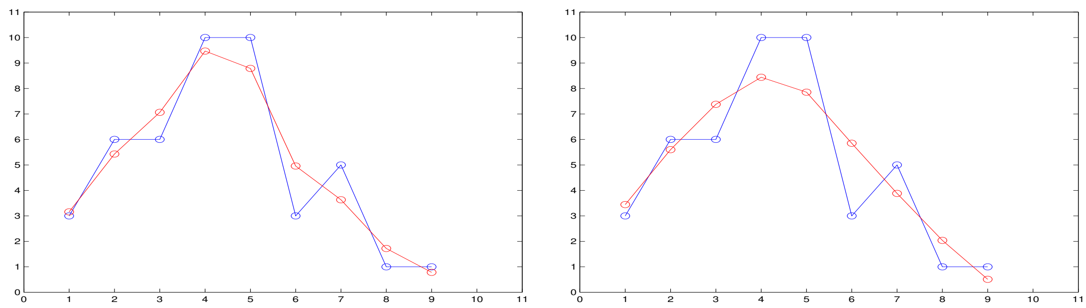
$$
D = \begin{pmatrix}
\frac{1}{h_2} & -\frac{1}{h_2}-\frac{1}{h_3} & 0 & \cdots & & 0 \\
0 & \frac{1}{h_3} & -\frac{1}{h_3}-\frac{1}{h_4} & \frac{1}{h_4} & \cdots & \vdots \\
\vdots & & \ddots & & \frac{1}{h_{N-1}} & 0 \\
0 & \cdots & 0 & \frac{1}{h_{N-1}} & -\frac{1}{h_{N-1}}-\frac{1}{h_N} & \frac{1}{h_N}
\end{pmatrix}.
$$

$$
\Omega = \mathbf{diag}(\omega_{\mathbf{1}}, \cdots, \omega_{\mathbf{N}})
$$

Here B is an $(N-2)\times(N-2)$ matrix and D is an $(N-2)\times N$ matrix. The solution to the minimisation problem is:

$$
\begin{aligned}
\mathbf{A} &= (\mathbf{D}\Omega^{-\mathbf{1}}\mathbf{D}^{\mathbf{T}} + \mathbf{B})^{-\mathbf{1}}\mathbf{DT} &\quad (20)\\
\mathbf{F} &= \mathbf{T} - \Omega^{-\mathbf{1}}\mathbf{DA} &\quad (21)
\end{aligned}
$$

By defining a regularisation factor Ω one controls the smoothness of the fitted spline. We observe at this point that if Ω is chosen large, then roughly speaking Ω^{-1} is close to zero and the solution is close to interpolating with a natural cubic spline using all given points as knots.



(a) Smoothing spline with $\Omega = 10$

(b) Smoothing spline with $\Omega = 1$

Figure 4: Influence of smoothness control parameter Ω during generation of smoothing spline.

For a subspace solution one can now define the smoothing spline for all coordinates in each dimension. In the case of our trajectory $\mathbf{T} = (\mathbf{t_1}, \cdots, \mathbf{t_N})$ with $\mathbf{t_i} \in \mathbf{R^P}$ we can compute for each dimension k our smoothed coordinates $\mathbf{f_i} \in \mathbf{R^P}$ of the N trajectory points assuming that the temporal ordering forms the abscissa:

$$\begin{aligned} \mathcal{S}_\mathbf{T} &= (\mathbf{f_1}, \cdots, \mathbf{f_N})^\mathbf{T} \\ &= (\tilde{\mathbf{t}}_1, \cdots, \tilde{\mathbf{t}}_N)^T. \end{aligned} \tag{22}$$

4.2 Distance Measure for Classification

Our next step is to define a distance measure for our parametric approach of measuring the similarity of test trajectories to our found trajectory models. In our subspace representation we have to find a measurement which takes time evolution as well as geometrical position of the sequence of observations into account. We therefore define a distance measure which compares individual frames geometrically, normalising its overall distance by its arc length [31].

The initial operation on a subspace is a projection of a vector. This is performed by $\tilde{\mathbf{t}}_P = (\tilde{\mathbf{t}})^T \mathbf{P}_\tau$, where the projection matrix $\mathbf{P}_\tau$ is the matrix defined by the training data and a certain time constrained factor τ (see Section 2). Using our individual trajectory models $\mathcal{M}$ we can compute a normed squared orthogonal distance $d_{sub}(\hat{\mathbf{t}}, \tilde{\mathbf{t}})$ from our trajectory model $\hat{\mathbf{t}}_P = (\hat{\mathbf{t}})^T \mathbf{P}_\tau$:

$$d_{sub}(\hat{\mathbf{t}}_P, \tilde{\mathbf{t}}_P) = \sum_{i=1}^{N} d_{sub}(\hat{\mathbf{t}}_P^i, \tilde{\mathbf{t}}_P^i) = \frac{\sum_{i=1}^{N} ||\hat{\mathbf{t}}_P^i - \tilde{\mathbf{t}}_P^i||^2}{arclength(\tilde{\mathbf{t}}_P)}. \tag{23}$$

The distance score can then be used to decide the model preferences. The distance measure classification picks the model with a minimum distance of the test trajectory to the templates. Normalisation of the distance score is made taking into account the possibly different lengths of the test trajectory. That implies that longer trajectory contribute more to a distance score without necessarily being the incorrect model hypothesis, hence by normalising using the arc-length of the trajectory one computes a distance score which is relative to the size of the trajectory. Further improvements suggest a probabilistic approach leading to a more objective similarity measure for trajectories.

5 Experimental Illustrations

5.1 Databases

ISOLET Database

We use a subset of the ISOLET [9] database to illustrate the idea, using the isolated spoken characters /B/, /D/ and /G/ to obtain the diphone /bee/, /dee/ and /gee/. The complete database is an isolated speech, alphabet database and consists of two tokens of each letter produced by 150 American English speaker 75 female and 75 male. Hence there were in total 240 training tokens and 60 test tokens for each diphone, which can be split into 120 training tokens and 30 test tokens per gender and diphone. Because ISOLET is not phonetically transcribed and time aligned, we hand-labelled the start of the spoken character and extracted a fixed number of frames for each diphone to capture the acoustic transition. We used this sub-optimal approach to obtain initial results. Using the TIMIT database (see later) we were able to compensate the fixed frame assumption using time alignment and dynamic time warping. The available data was split into a training and test set. We used 80% of the data for training (ISOLET1-4) and 20% for tests (ISOLET5), as recommended by the originators of the dataset.

TIMIT Database

The DARPA (Defence Advanced Research Project Agency) TIMIT database [15, 18] is an acoustic-phonetic database consisting of data, that is phonetically transcribed and time aligned. TIMIT contains a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States of America. Each speaker utters 2 calibration sentences (prefix `sa`) , 5 phonetically compact sentences (prefix `sx`) and 3 phonetically diverse sentences (prefix `si`). The data is split into train and test subsets. The test set contains 1.344 utterances from 168 speakers. The data from the remaining 462 speakers forms the training set. No sentence or speaker appears in both the training and test set.

TIMIT is a convenient database to demonstrate our approach of using diphones as speech segments because of TIMITs phonetic labelling which makes it possible to extract all occurring diphones. In our test scenario we extracted the diphones /bee/, /dee/ and /gee/, for which the number of occurrences and its corresponding phones are listed in Table 2.
TIMIT being a phonetically balanced database which doesn't imply that it is balanced with respect

Diphone	**TIMIT phones**	**Training**		**Testing**		**Total**
		Male	**Female**	**Male**	**Female**	
/bee/	b-iy, b-ih	367	172	149	87	775
/dee/	d-iy, d-ih	498	227	162	103	990
/gee/	jh-iy, jh-ih	203	107	41	23	374

Table 2: Occurrences of example diphones within TIMIT and its corresponding phones.

to our speech segments diphones, as well. To get as many training/test examples of diphones within TIMIT, we merged the 61 TIMIT symbols onto a set of 39 symbols, which was also done in experiments by other researchers [36]. The TIMIT symbols and its reduction to the set of 39 symbols is given in Table 7 and Table 8 in the appendix. Furthermore for our experimental tests we merged the phones /`iy`/ and /`ih`/ to get sufficient training and testing data for reliable results. This was done under the assumption that the acoustic transitions of the diphones /b-iy/ and /b-ih/ are very similar, as well as for the diphones having as their initial phone /d/ and /g/.

Using TIMIT we calculated our diphone from the phonetic labelling, using the start sample and end sample information for the different phones. Schwartz et al. [40] proposed that the transition within the diphone is very important. Hence one has to treat the inner region of a diphone as an inelastic region so that for expansion or shrinkage this area has to be preserved. Beside the standard time warping scheme to adjust different length of test trajectories to the length of the template trajectory, we also warped the test trajectory into a trajectory template using Schwartz elastic/inelastic idea. For trajectory model generation we treat the training data identical and warped each training example using Schwartz et al. [40] into a transition representation of equal length. This was necessary to perform the calculation of the transition matrix which makes use of both male and female training data. For test purposes we used two different approaches. Firstly we warped the test trajectory into the corresponding diphone specific trajectory model size for comparison reasons. Secondly Schwartz et al. [40] idea was applied preserving the inelastic regions of the diphone (see Figure 5) to show performances for non-time-aligned speech segments.

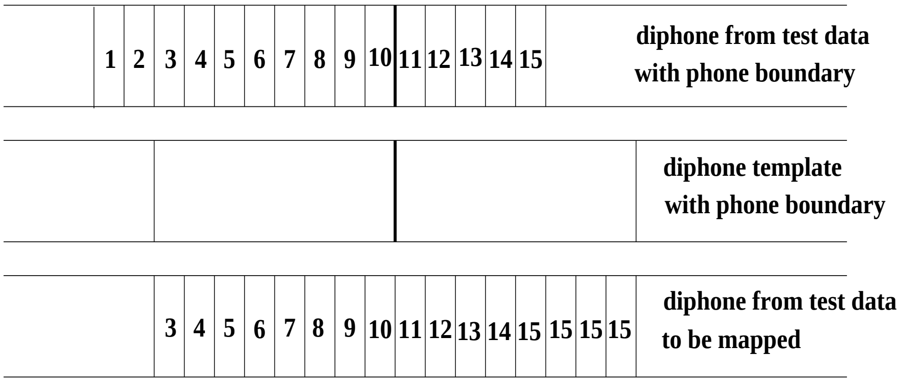


Figure 5: Extension and shrinkage algorithm for diphones having different length in the process of template generation using Schwartz elastic/inelastic region theory.

5.2 Transformation Optimisation

The training process can be described by finding an appropriate transformation matrix for each diphone which enables the classification process to be optimal. We have to search through all possible combinations of planes per diphone which leads to an exhaustive search. If k is the number of available planes which is equal to the time constrained eigenvectors of our TC-PCA approach using k different time constraints and l the number of diphones we have to evaluate $\mathcal{O}(k^l)$ combinations.

In each combination the gender specific test data for all different diphones must be evaluated on how good a specific model discriminates among all given models. Each plane is characterised by the diphone used and the specific time constraints τ. In Figure 6 we illustrate the behaviour of different temporal constraints towards accuracy results.

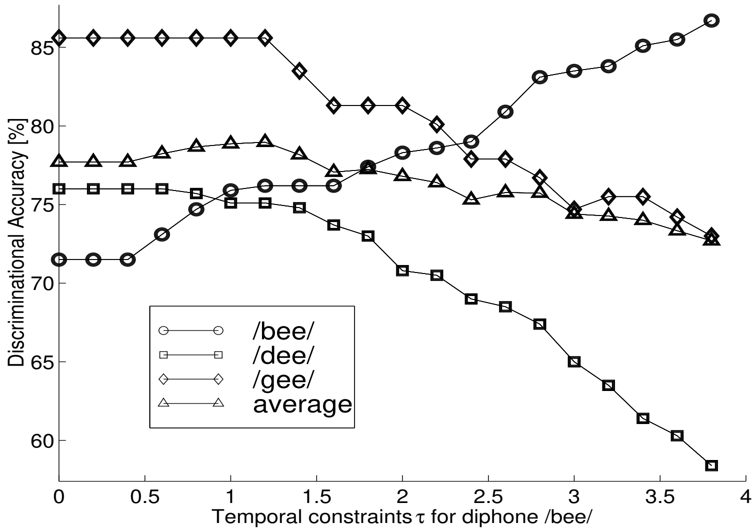


Figure 6: Discrimination monitor for the feature representation REP1 within TIMIT database. We used the $Model_{AV1}$ (see section 5.4.1) to illustrate the trend of accuracy figures by changing the time constraints τ for the diphone /bee/ plane only.

5.3 Parameter Requirements

The subspace approach we describe requires a very small number of parameters. These parameters are determined by the size of the transformation matrix P_τ and the number of points for the trajectory. If one uses a speech representation of p MFCCs, which leads to an p-dimensional speech vector per frame, one needs $2 * (p+1)$ parameters to describe the transformation matrix. Including the extra time constrained parameter, this transformation matrix performs a dimensionality reduction from $p+1$ to 2 dimensional space.

Furthermore we need a sequence of N points in 2-dimensional space to describe our model trajectory, requiring additional $2 * N$ parameters. The number N is found in the training process. For ISOLET we choose a fixed number of frames because of the missing phone labels (see Table 3) whereas for the TIMIT database the length of the individual trajectories is determined by the average length of the training examples (see Table 4). Obviously, gender independent models used require a smaller number of parameters because the gender dependent ones have the transformation matrix in common.

ISOLET	/bee/			/dee/			/gee/		
	male	female	total	male	female	total	male	female	total
REP1	48	48	84	48	48	84	48	48	84
REP2	30	30	48	30	30	48	30	30	48
REP3	58	58	94	58	58	94	58	58	94
REP4	40	40	58	40	40	58	40	40	58
HMM	-	-	300	-	-	300	-	-	300

Table 3: Parameter requirements for the diphone models depending on the used representation and the number of needed data points per trajectory. For ISOLET we choose a fixed number of frames per diphone representing 90ms of speech transitions. For comparison reasons we added the parameter requirements for the baseline HMM mentioned in section 5.4.1.

TIMIT	/bee/			/dee/			/gee/		
	male	female	total	male	female	total	male	female	total
REP1	34	34	56	32	30	50	38	38	64
REP2	24	24	36	22	22	32	26	28	42
REP3	44	44	66	42	40	60	48	48	74
REP4	34	34	46	32	32	44	36	38	52

Table 4: Parameter requirements for the diphone models depending on the used representation and the number of needed data points per trajectory. In case of TIMIT the number of data points per trajectory is determined by the training data and its given phone labels.

5.4 Performance Evaluation

We present the experimental results in this section There are two key issues which need to be experimentally proved. One is the inter-model discriminational accuracy showing how well one can distinguish between diphone trajectory models. We want to show that, despite a significant information loss during the dimensionality reduction, we preserve important dynamic information. This information is useful to distinguish categories of acoustic transitions, in our case diphones. Furthermore will we show that diphone transitions are very characteristic and can be located in an unknown sentence without time alignment using matched filters.

5.4.1 Discriminational Accuracy

For the discriminational accuracy evaluation task we run an exhaustive search with different time constrained factors τ to find a combination of time constrained planes which is most discriminant for individual trajectory models. A further parameter in the optimisation task was the smoothness of the test trajectory defined by Ω. We used different Ω to show the influence of the smoothing operation to the accuracy in the following tables. The plane index PI indicates the time constraints used for the planes /bee/-/dee/-/gee/ which was found most discriminant and can be transformed to τ using $\tau = [PI - 1] * 0.2$, hence in the case of $PI = 1$ ordinary PCA is used to determine the dimensionality reduction transformation.

ISOLET

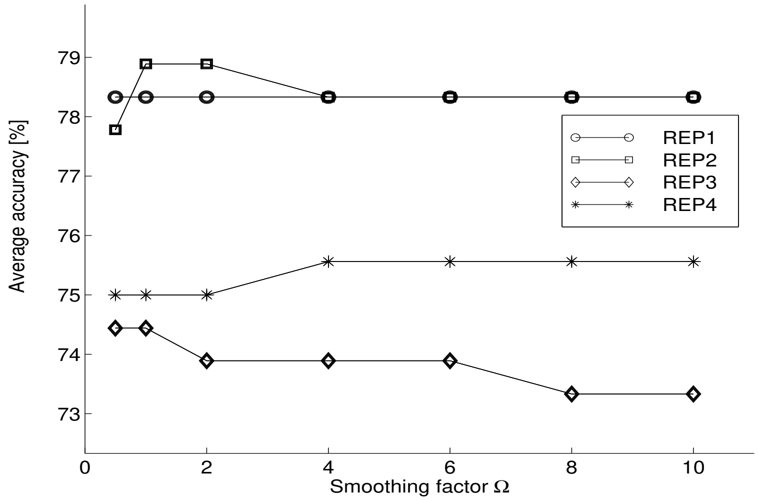
The tables in appendix B (see Tables 9-11) show the recognition accuracy results for the ISOLET database using the proposed models from section 3. ISOLET is not phone labelled hence we marked the start of the acoustic event and took a sequence of frame vectors corresponding to 90 ms of speech to capture the characteristic transition. Depending on the used representation the speech trajectories consist of 9 (REP2, REP4) or 18 (REP1, REP3) frames. In the following table we give the best results for each model and representation whereas a more accurate performance figures can be found in the tables mentioned above. Our smoothing scheme to compensate noise influence to test trajectories

ISOLET	$Model_{AV}$	$Model_{PC}$	$Model_{GTM}$
REP1	78.33%	80.56%	77.78%
REP2	78.89%	81.11%	78.89%
REP3	74.44%	78.89%	74.44%
REP4	75.56%	76.11%	69.44%

Table 5: Optimal performances for each model and representation using the ISOLET database.

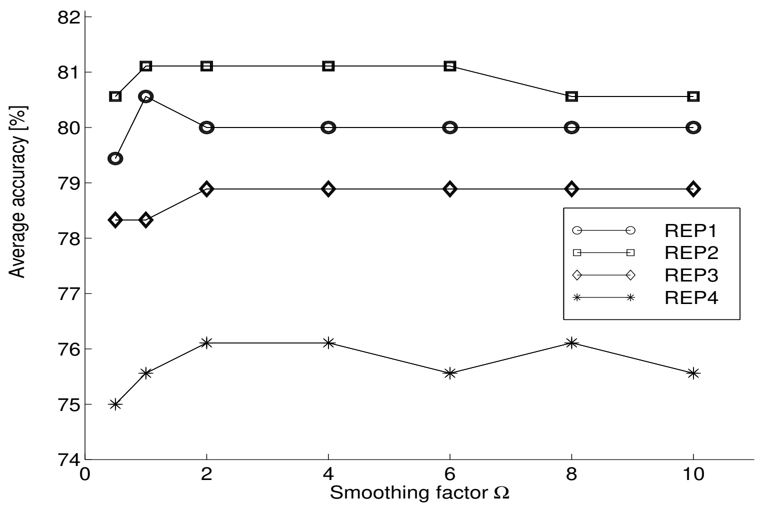
was monitored using different smoothing parameter Ω. We show the dependencies by plotting the average accuracy over the used Ω to get trends in choosing an optimal Ω. In Figures 7,8 we show for all trajectory models the evolution of the average accuracy using different Ω for our different representations. The results showed no significant difference in the resulting error rate. These findings follow the results obtained for different representation. Focusing on the different temporal resolutions, there was also no significant accuracy difference between representations using a 5ms shift or a 10ms shift to get a more accurate transitional representation. We furthermore monitored performance differences using different speech representations. Lower order MFCCs (REP1 + REP2) translated into substantial better results than representations including higher order MFCCs (REP3 + REP4).

With the best results reaching an average accuracy of 81.11%, using our principal curve model $\mathcal{M}_{PC}$ with an appropriate smoothing factor Ω (see Table 10) , we obtain similar results to a baseline HMM system using one mixture and a diagonal covariance matrix on a BTL E-SET giving 84.5% accuracy for this task. The difference translates into 6 test examples less correctly classified which is not substantial. However, what is important is to note that the representation here is very simplistic, namely, a projection onto two dimensions. In comparison to 300 parameters per model of the HMM system, the subspace trajectory approach uses only $2 * (6 + 9 + 9) = 48$ parameters for a gender independent diphone model because the trajectory is built by 9 anchor points.

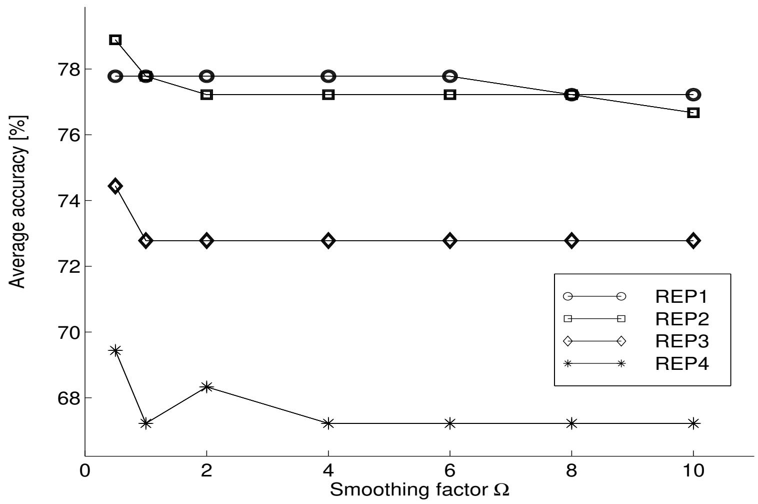


(a) Ω-monitor for $\mathcal{M}_{AV}$

Figure 7: Evaluation of the influence of smoothing parameter to the obtained average accuracy for our average trajectory models defined in section 3. The accuracy trend for different smoothing factors is displayed using the ISOLET database. The dependencies are plotted for all used representations and for a number of different Ωs.



(a) Ω-monitor for $\mathcal{M}_{PC}$



(b) Ω-monitor for $\mathcal{M}_{GTM}$

Figure 8: Evaluation of the influence of smoothing parameter to the obtained average accuracy for our trajectory models defined in section 3. The accuracy trend for different smoothing factors is displayed using the ISOLET database. The dependencies are plotted for all used representations and for a number of different Ωs.

TIMIT

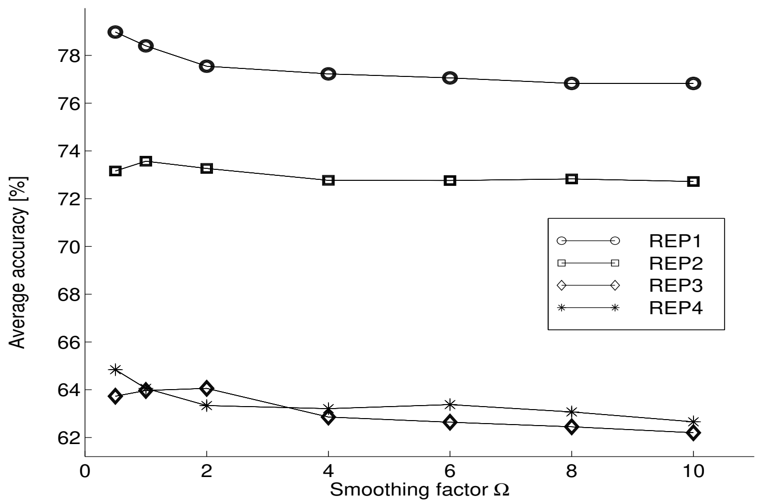
The tables in the appendix C (Tables 12-15) show the obtained accuracy results for the TIMIT database using the proposed models from section 3. For TIMIT database we consider a special case of model, where the average points are calculated in two distinct ways. The problem originates from the different diphone template lengths according to the labelled phone data. Here one has to adjust every test trajectory to the length of the stored template trajectory during performance tests. This can be done in two different ways.
Firstly our meanframe model is computed by warping the training examples into the average length of the diphone. This model will be addressed as M_{AV1} in the following tables and discussion. The second method considers the Schwartz et al. [40] inelastic regions as important and cut the average length from the natural occurrence of our diphone from our training sentences. A calculation of the averages per frame is performed afterwards in the same way for both approaches. This model will be addressed as M_{AV2} in the following tables and discussion.

During the experiments the first algorithm uses test data which is warped into a template before comparing it to our average frame model. The second approach is using Schwartz's idea to preserve in particular the transition and adjusts the test trajectory according to the inelastic region theory. The latter method doesn't warp the data into a given template by minimising its difference. It is considering only the length of the two involved phones to construct the test trajectory of equal length from the given sequence of data preserving the natural occurrence in the test sentence. Hence time alignment is not performed during tests which makes the knowledge of diphone boundaries unnecessary. A frame-wise comparison using our distance score is performed afterwards. Classification is done by finding the minimum score within each model. In Table 6 we give the best results for each model and representation whereas a more accurate performance figures can be found in the tables mentioned above. We furthermore considered also for TIMIT a smoothing scheme which we

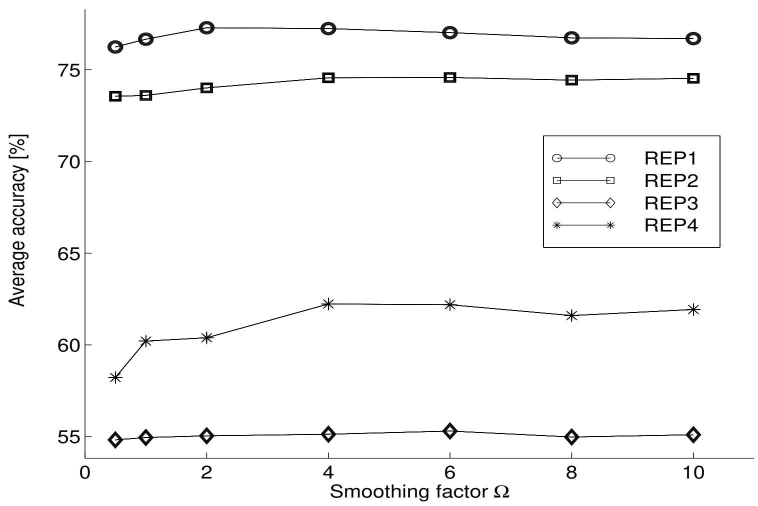
TIMIT	$Model_{AV1}$	$Model_{AV2}$	$Model_{PC}$	$Model_{GTM}$
REP1	78.89%	77.29%	76.84%	77.10%
REP2	73.57%	74.58%	73.67%	72.81%
REP3	64.06%	55.30%	63.27%	62.43%
REP4	64.84%	62.23%	65.03%	64.55%

Table 6: Optimal performances for each model and representation using the TIMIT database.

monitored for all models. Its influence of the test results is shown in Figures 9,10. Although an influence is present there is no substantial improvement using smoothing splines to compensate noise influence on the acoustic trajectory.
For the choice of speech representation used, the results using the TIMIT database showed even stronger significants than for ISOLET. Choosing a speech representation based on higher order MFCCs translates to upto 30% performance degradation having at least an absolute average performance loss of 10%. This can be observed for all models. Additionally are the performance values increased for lower order MFCC representations when a higher temporal resolution is used as well. Contradictionary a better time resolution for higher order MFCC representations translates into worse results.

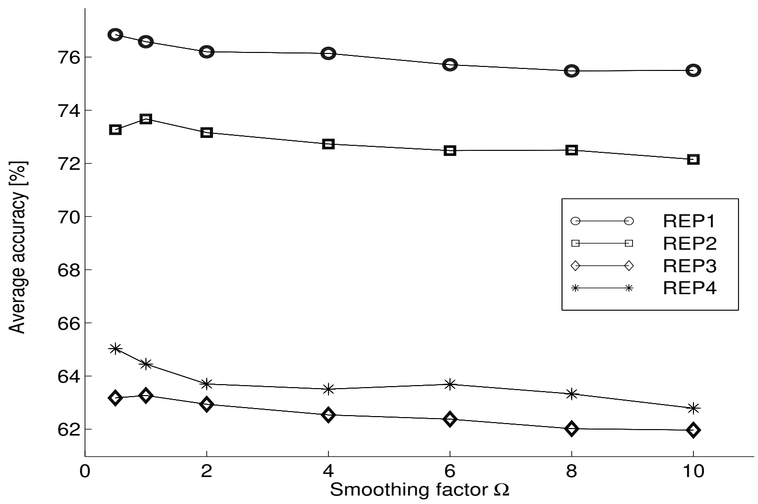


(a) Performance monitor in respect to Ω for $\mathcal{M}_{AV1}$

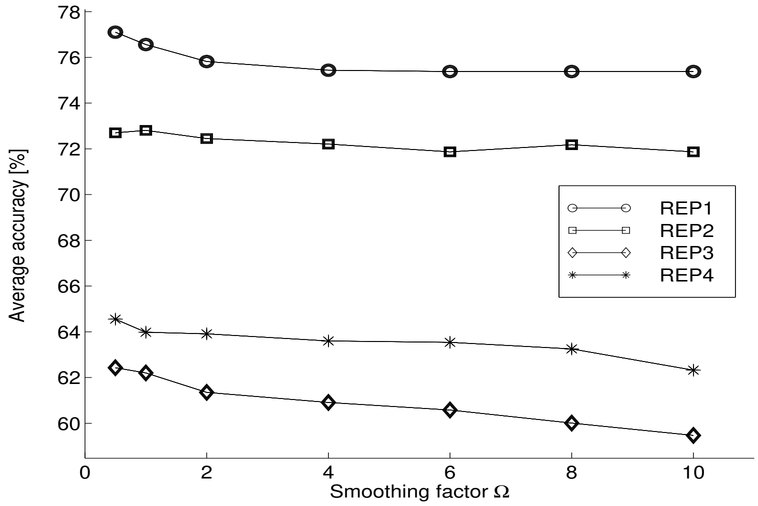


(b) Performance monitor in respect to Ω for $\mathcal{M}_{AV2}$

Figure 9: Evaluation of the influence of smoothing parameter to the obtained average accuracy for our trajectory models M_{AV1} and M_{AV2} defined in section 3. The accuracy trend for different smoothing factors is displayed using the TIMIT database. The dependencies are plotted for all used representations and for a number of different Ωs.



(a) Performance monitor in respect to Ω for $\mathcal{M}_{PC}$



(b) Performance monitor in respect to Ω for $\mathcal{M}_{GTM}$

Figure 10: Evaluation of the influence of smoothing parameter to the obtained average accuracy for our trajectory models M_{PC} and M_{GTM} defined in section 3. The accuracy trend for different smoothing factors is displayed using the TIMIT database. The dependencies are plotted for all used representations and for a number of different Ωs.

5.4.2 Matched Filters

We now look at the problem of spotting a particular diphone in continuous speech, using the projective representations as matched filters. Our diphone models are used as "matched filters" sweeping over an unknown utterance, where each model should peak at the location of its appearance (see Figure 11). This approach in particular gives some insight how important time alignment is for diphone recognition. This is obtained by a sweep over each sentence which is done without warping phone aligned segments into the length of the diphone model but is done just by shifting over the sentence frame by frame. Hence time alignment during recognition is not necessary if one can detect the diphones adequately within this framework.

A measurement of how reliable the models match the correct diphone location, we use ROC (Receiver Operating Characteristic) curves for each diphone model over all test sentences. ROC curves plot the true-positive rate over false-positive rate giving an estimate how discriminant the diphone model is over all other possible acoustic transitions (see Figure 12). The curve is parameterised by a threshold which in our case is the used distance score. Each point, which represents a used threshold, is shown in Figure 12. It is increased for each step along the curve. For each unknown utterance all test sentences which include an occurence of the questioned transition were considered.

This information source can be used to determine how reliable one can locate typical transitions for incorporation into existing systems using techniques like N-best rescoring mechanisms. Hence diphone model information could be used as compensational models to complement existing systems which are not modeling the inter-phone dependencies.

Beside a location matching scheme used for the matching filter experiments one can also obtain a discriminational accuracy measure using inelastic region theory. Here we are interested in the discriminant abilities between our diphone models rather than the location traceability of single models. In the accuracy approach we do not warp our test trajectories into the length of our individual diphone model lengths but are using the real sequence of frames within an unknown sentence which is mapped onto the length of the model. This is equivalent to sweeping over the unknown utterances using our matching filter approach running all diphone models in parallel to obtain comparable results. Discriminational accuracy rates for this approach is given in Table 13 which indicates that detection and recognition of transitional regions modeled as diphones can be done without warping time-aligned segments into our templates.

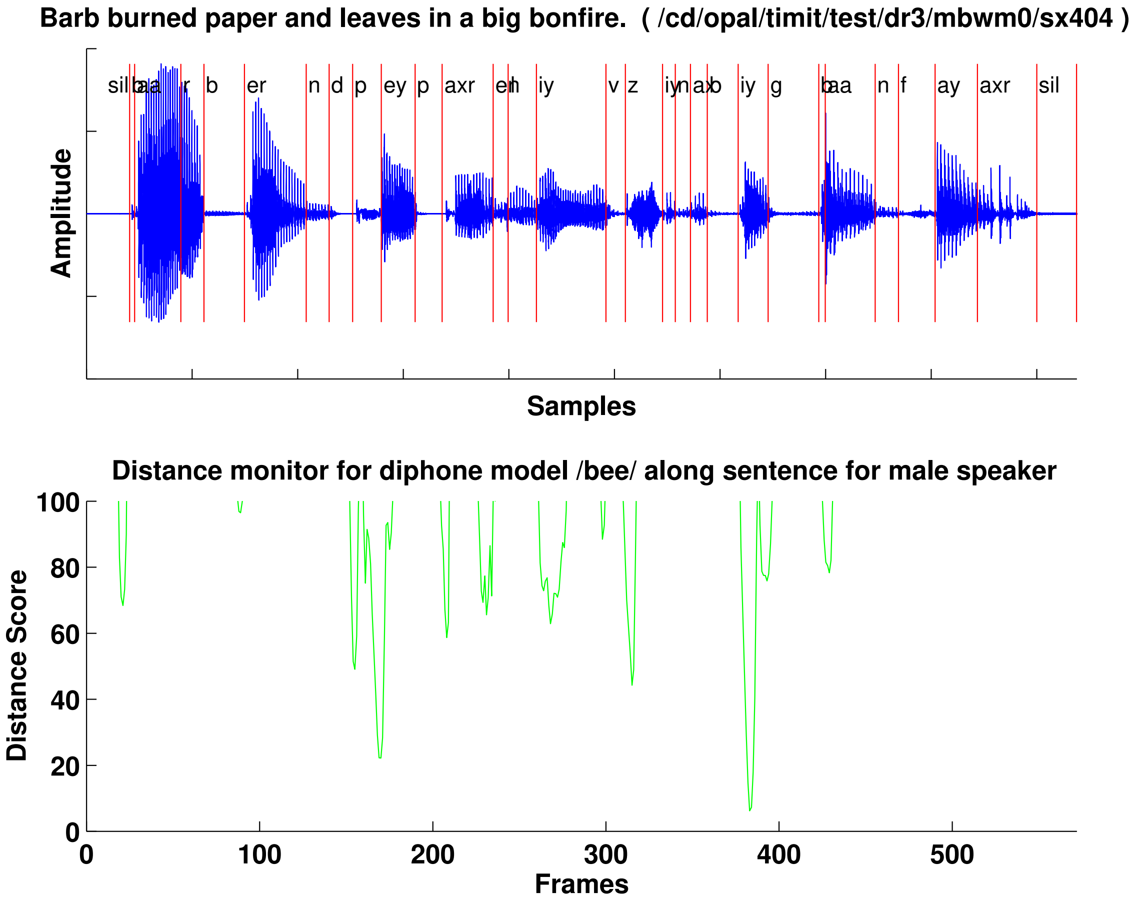
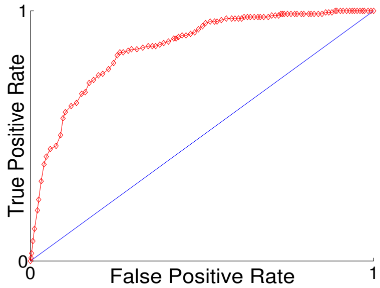
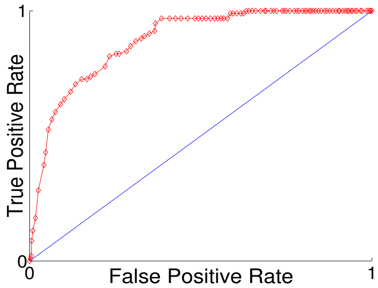


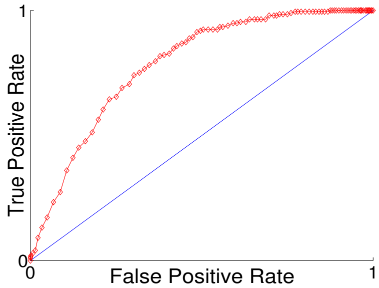
Figure 11: Distance monitor for male diphone model /bee/ along a sentence from the TIMIT database. Small distances result in good matches between model and test trajectory. As model parameter we used the transforamtion found for the best accuracy percentage, setting $\Omega = 0.5$ and using representation **REP1**. The model is moved over the sentence shifting one frame each time step.



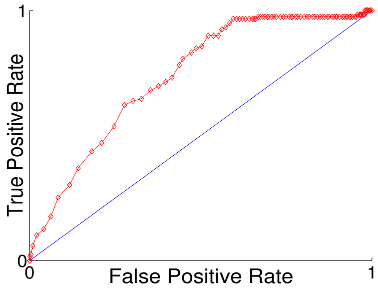
(a) ROC curve for male /bee/



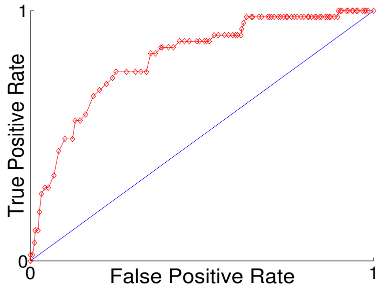
(b) ROC curve for female /bee/



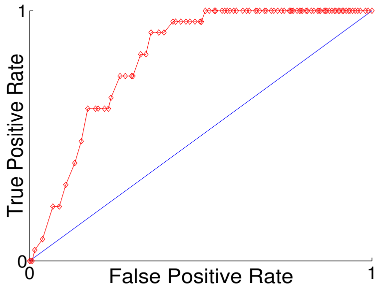
(c) ROC curve for male /dee/



(d) ROC curve for female /dee/



(e) ROC curve for male /gee/



(f) ROC curve for female /gee/

Figure 12: ROC curves obtained for female and male diphone models. Upper distance threshold was 100 which leads to a more pessimistic curve because all other false-positive examples with a distance score higher than 100 will not appear in this curve. We used the settings for Ω, plane index and the representation according to the found optimal accuracy measurement.

6 Discussion

ISOLET

These experiments demonstrate that the very simple representation adopted retains a reasonable amount of discrimination. The most accurate model is the principal curve model with its most flexible interpretation. It represents the underlying data distribution most adequately because of its number of latent points which allows principal curve to adjust accurately and results in superior error rates in comparison to our other methods. The slowly moving kernel along our principal curve allows precise adaptation to the local data distribution maintaining good generalisation using a sufficient wide kernel. The required temporal distance of consecutive points is maintained by the related ordering of our initial guess. Temporal ordering in our subspace can translate to non-equidistant ordering of anchor points of the trajectory model depending how strong the temporal influence is used for our subspace projection.

Our second latent point model GTM, in contrast, considers only a fixed number of latent points which is much smaller than the number of data points existing in the training data. It uses spherical Gaussians with a fixed variance to model the probability distribution function (pdf) of the data. This puts a considerable number of constraints on the resulting Gaussian distributions. A further limit is the equidistant assumption of Gaussian centers along the curve which does not reflect the temporal occurrence within the plane, which leads clearly to worse performance. Here possible adjustments are needed using different variances for the translated latent points resulting in Gaussian centers in data space. The standard GTM algorithm has to be tailored to count more for the temporal dependencies than just using the temporal constrained subspace.

The average frame model represents an intermediate model which takes the temporal ordering into account but assumes that all test data belongs to fixed frames which is not true. In particular for ISOLET where we dealt with a fixed number of frames per speech unit, where there are diphones with different temporal lengths, we are summarising most likely frames which doesn't match. Hence the average frame model cannot represent the real underlying frame sequence of our transition. Here most likely data points are averaged which do not belong to the same trajectory anchor point because of the lack of phonetic labelling.

Although arguing that noise disrupted speech transitions would not match accurately the trajectory model without smoothing out the path, the results indicate that there is no difference in the resulting error rates. Hence smoothing out our test trajectories in case of the normalised frame-wise distance measure seems not to affect the results for ISOLET data. We showed in our experiments that the use of Ω doesn't influence the obtained results significantly. This might be due to the simple approach using Ω as a scalar rather than as a vector. Here we can focus our smoothing efforts to the region we believe is important to smooth out which might result in better performances. Further tests are necessary to rule out a necessity of a smoothing factor in our trajectory framework. In comparison to a baseline HMM, obtaining 84.5% accuracy, our best model with 81.1% classifies only 6 examples less correctly. By using just a 1/6 of the parameters which are used for each model in a baseline HMM approach, using one mixture and a diagonal covariance matrix, we obtained most of the discriminant information within our subspace. Clearly a more complex task is needed to evaluate the use of our transitional models. We applied our method for the TIMIT database which will be discussed in the next paragraph.

TIMIT

The results obtained in diphone accuracy were far better from what we had expected because of the influences of realistic continuous speech to the acoustic transitions in comparison to isolated spoken letters. Nevertheless could we show that the accuracy trends using TIMIT follow similar patterns. With an optimal accuracy of approximately 79% for our meanframe trajectory model $\mathcal{M}_{AV1}$ using warped test data we obtained no substantial worse results in comparison to the ISOLET results performing this particular task.

It is interesting to note the fact that using our meanframe trajectory model with the test data generated using Schwartz inelastic region theory, we obtained best results of 77.3% accuracy. This reflects the fact that transitions are very characteristic within continuous speech because our test data was not warped into the trajectory template. Here the sequence of frames in the unknown sentence was preserved, suggesting that the time alignment in a diphone framework doesn't perhaps contribute so much.

The difference in accuracy between the different models used was quite small. This is perhaps due to the modest size of the experiments. They suggest that more complex experiments are needed to determine the most suitable model, knowing their individual drawbacks discussed above. Here principal curves or GTM might be more suitable to introduce a probabilistic trajectory similarity measurement to compensate the subjective distance measure. The TIMIT results show that trajectories in continuous speech are more disturbed. A more accurate trajectory model counting for the real data distribution does not necessarily translate into better results which is due to the high variance in the test trajectories. As for ISOLET the results show that the use of a smoothing scheme doesn't translate into superior results although the influence can be measured giving a fluctuation of $\pm 2\%$ in the accuracy figures where more smoothing provides better results. As one can see in Figures 9,10 the monitor for all trajectory models over Ω doesn't peak sharply, suggesting an optimal choice of Ω. Further investigations have to be performed to exclude the possibility of an useful Ω which leads to better results. The smoothing spline scheme allows Ω to be vectorial which enables one to emphasize smoothing on individual trajectory points or regions. That might be more appropriate because we are dealing with highly dynamic transitions with less variance at the boundaries.

The TIMIT results support the hypothesis that it is very important which speech representation one chooses when using trajectory models. The figures clearly reflect the fact that higher order MFCCs leads to worse recognition results. With its oscillating character higher order MFCCs seem not to be suitable for trajectory modeling. The expected difference between representations using higher order MFCCs and those representations using more suitable MFCCs for transitional modeling were confirmed. Our results showed a significant error rate increase when using the representation including higher order MFCCs.

7 Conclusions

In this study, we have proposed a new method of modeling speech transitions with a subspace model. We show that temporal transitions in speech can be visualised and modeled in a low dimensional space. This approach has the advantages of reduced memory requirements in comparison with models involving context-dependent speech units. In addition, the subspace models require relatively little data compared to HMMs.

The results show that discriminant information is preserved in our subspace focusing on the temporal ordering. In particular for TIMIT we could show the characteristics of transitions which is available through our subspace modeling without any forced alignment. The results are encouraging to further investigate the use of subspace models for speech transitions, which could be used as compensational models in respect to the inter-segment independent assumption used in state-of-the-art recognition systems.

Our future work concentrates on the usefulness of the trajectoral information whether modeling transitions provides one with orthogonal knowledge in comparison with the information obtained by standard HMM systems. It includes also research into the distance score or similarity measurement used to compare trajectories. One needs to define a more objective function. Because of the projection of different trajectory onto different time constrained planes for each model, one needs to have a trajectory size independent measure of how likely the sequence of test points is generated by the underlying sequence of model points. Here a probabilistic approach would help to find a suitable objective function which is independent on the size and could incorporate also the particular variances within the optimal planes which would increase the accuracy of finding the correct model. Further interest will focus on the exploration of the dimensionality used for our subspace projection and if it is possible to establish a link between the used subspace dimensionality and accuracy results obtained for different speech representations.

We therefore have to extend our experimental work using a larger set of models and compare results with standard systems using the same speech units. Alternatively we propose a N-best rescoring scheme [38, 39, 33] to incorporate the subspace mode into a phone-based system. The rescoring mechanism can be used to emphasize paths in the lattice of hypothesis using our transitional models which might avoid pruning out the correct sequence of phones. The modelled inter-phone characteristics, which are captured by diphones, should complement baseline systems and should lead to better performances.

Acknowledgement

We thank the Neural Computing Research Group at Aston University for making the Matlab code of the GTM algorithm available in the public domain. KR acknowledges financial support from Girton College, Cambridge European Trust and the EPSRC.

References

[1] M. Afify, Y. Gong, and J.-P. Haton. Nonlinear time alignment in stochastic trajectory models for speech recognition. *Int. Conf. in Spoken Language Processing*, 1:291–293, 1994.

[2] M. Afify, Y. Gong, and J.-P. Haton. Stochastic trajectory models for speech recognition: An extension to modelling time correlation. *European Conference on Speech Communication and Technology (Eurospeech)*, pages 515–518, 1995.

[3] G. Ahlbom, F. Bimbot, and G. Chollet. Modeling spectral speech transitions using temporal decomposition techniques. *Int. Conf. in Acoustics, Speech and Signal Processing*, 1:13–16, 1987.

[4] C.M. Bishop, G.E. Hinton, and I.G.D. Strachan. GTM through time. *IEE International Conference on Artificial Neural Networks*, 1997.

[5] C.M. Bishop, M. Svensen, and C.K.I. Williams. GTM: A principled alternative to the self-organizing map. *To appear: Neural Computation*, 1996.

[6] C.M. Bishop, M. Svensen, and C.K.I. Williams. GTM: The Generative Topographic Mapping. NCRG/96/015, Aston University, Birmingham, 1996.

[7] H.A. Boulard and N. Morgan. *Connectionist Speech Recognition: A Hibrid Approach*. Kluwer Academic Publishers, 1994.

[8] W.S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368):829–836, December 1979.

[9] R. Cole, Y. Muthusamy, and M. Fanty. The ISOLET spoken letter database. Technical Report CSE 90-004, Oregon Graduate Institute, 1994.

[10] L. Deng, M. Aksmanovic, X. Sun, and C.F.J. Wu. Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states. *IEEE Transactions on Speech and Audio Processing*, 2(4):507–520, 1994.

[11] V.V. Digalakis. *Segment-based stochastic models of spectral dynamics for continuous speech recognition*. PhD thesis, Boston University, 1992.

[12] V.V. Digalakis, R. Rohlicek, and M. Ostendorf. A dynamical system approach to continuous speech recognition. *Int. Conf. in Acoustics, Speech and Signal Processing*, 1:289–292, 1991.

[13] V.V. Digalakis, R. Rohlicek, and M. Ostendorf. ML estimation of a stochastic linear system with EM algorithm and its application to speech recognition. *IEEE Transactions of Speech and Audio Processing*, 1:431–442, 1993.

[14] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. New York, Wiley, 1973.

[15] W. Fisher, G. Doddington, and K. Goudie-Marshall. The DARPA speech recognition research database : Specification and status. *In Proceedings Speech Recognition Workshop*, 1986.

[16] J.H. Friedman. Exploratory projection pursuit. *Journal of the American Statistical Association*, 82:249–266, 1987.

[17] T. Fukada, Y. Sagisaka, and K.K. Paliwal. Model parameter estimation for mixture density polynomial segment models. *Int. Conf. in Acoustics, Speech and Signal Processing*, 1997.

[18] J.S. Garofolo. Getting started with the DARPA TIMIT CD-ROM: an acoustic phonetic continuous speech database. Technical report, National Institute of Standards and Technology (NIST), 1988.

[19] O. Ghitza and M.M. Sondhi. Hidden Markov models with templates as non-stationary states: an application to speech recognition. *Computer Speech and Language*, 2:101–119, 1993.

[20] H. Gish and K. Ng. Parametric trajectory models for speech recognition. *Int. Conf. in Spoken Language Processing*, 1:466–469, 1996.

[21] W. Goldenthal. *Statistical trajectory models for phonetic recognition.* PhD thesis, Department of Aeronautics and Astronautics, MIT, 1994.

[22] Y. Gong and J.-P. Haton. Stochastic trajectory modeling for speech recognition. *Int. Conf. in Acoustics, Speech and Signal Processing*, 1:57–60, 1994.

[23] Y. Gong, I. Illina, and J.-P. Haton. Modeling long term variability information in mixture stochastic trajectory framework. *Int. Conf. in Spoken Language Processing*, 1996.

[24] T. Hastie and W. Stuetzle. Principal curves. *Journal of the American Statistical Association*, 84(406):502–516, 1989.

[25] W.J. Holmes and M.J. Russell. Linear dynamic segmental HMMs: Variability representation and training procedure. *Int. Conf. in Acoustics, Speech and Signal Processing*, 1997.

[26] Z. Hu and E. Barnard. Smoothness analysis for trajectory features. *Int. Conf. in Acoustics, Speech and Signal Processing*, 1997.

[27] P.J. Huber. Projection pursuit. In *The Annals of Statistics*, volume 13(2), pages 435–475, 1985.

[28] A. Kannan and M. Ostendorf. Adaptation of polynomial trajectory segment models for large vocabulary speech recognition. *Int. Conf. in Acoustics, Speech and Signal Processing*, 1997.

[29] P. Lancaster and K. Salkauskas. *Curve and Surface Fitting: An introduction.* Academic Press, 1986.

[30] P.F. Marteau, G. Bailly, and M.T. Janot-Giorgetti. Stochastic model of diphone-like segments based on trajectory concepts. *Int. Conf. in Acoustics, Speech and Signal Processing*, 1988.

[31] E. Oja. *Subspace Methods of Pattern Recognition.* Research Studies Press, Letchworth, U.K., 1983.

[32] 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 Transaction on Speech and Audio Processing*, 4(5):360–378, 1996.

[33] M. Rayner, D. Carter, V.V. Digalakis, and P. Price. Combining knowledge sources to reorder N-best speech hypothesis lists. *Proceedings of the 1994 ARPA Workshop on Human Language Technology*, 1994.

[34] K. Reinhard and M. Niranjan. Non linear speech transition visualization. *IEE Int. Conf. in Artificial Neural Networks*, 1997.

[35] A. Robinson, H. Boulard, M. Hochberg, D. Kershaw, D. Morgan, and S. Renals. A Neural Network based Speaker Independent, Large Vocabulary, Continuous Speech Recognition System: The WERNICKE Project. *European Conference on Speech Communication and Technology (Eurospeech)*, pages 1941–1944, 1996.

[36] T. Robinson and F. Fallside. A recurrent error propagation network speech recognition system. *Computer Speech and Language*, 5:259–274, 1991.

[37] P.L. Salza, E. Foti, L. Nebbia, and M. Oreglia. MOS and pair comparison combined methods for quality evaluation of text-to-speech systems. *Acustica*, 82:650–656, 1996.

[38] P. Schmid. *Explicit N-best formant features for segment based speech recognition.* PhD thesis, Oregon Graduate Institute, 1996.

[39] P. Schmid and E. Barnard. Explicit N-best formant features for vowel classification. *Int. Conf. in Acoustics, Speech and Signal Processing*, pages 991–994, 1997.

[40] R. Schwartz, J. Klovstad, J. Makhoul, D. Klatt, and V. Zue. Diphone synthesis for phonetic vocoding. *Int. Conf. in Acoustics, Speech and Signal Processing*, pages 891–894, 1979.

[41] D.X. Sun. Statistical modeling of co-articulation in continuous speech based on data driven interpolation. *Int. Conf. in Acoustics, Speech and Signal Processing*, 1997.

[42] R. Tibshirani. Principal curves revisited. *Statistics and Computing*, 2:183–190, 1992.

[43] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K.J. Long. Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics,Speech and Signal Processing*, 37(3), March 1989.

[44] S.J. Young, J.Jansen, J.J. Odell, D. Ollason, and P.C. Woodland. *The HTK Book.* Version 2.0. Entropic Cambridge Research Laboratory and University of Cambridge, 1995.

[45] S.J. Young, J.J. Odell, and P.C. Woodland. Tree-based state tying for high accuracy acoustic modeling. *In ARPA Workshop on Human Language Technology*, pages 307–312, March 1994.

[46] V.W. Zue. Speech spectrogram reading: An acoustic study of the English language. *Lecture Notes, MIT*, August 1991.

A TIMIT phone description

Vowels		Semi-vowels		Nasals	
eh	b**e**t	l	**l**at	m	**m**om
ih	b**i**t	el	bott**le**	em	bott**om**
ao	b**ough**t	r	**r**ay	n	**n**oon
ae	b**a**t	w	**w**ay	en	butt**on**
aa	b**o**tt	y	**y**acht	ng	si**ng**
ah	b**u**t	hh	**h**ay	eng	wash**ing**ton
uw	b**oo**t	hv	a**h**ead	nx	wi**nn**er
uh	b**oo**k				
er	b**ir**d	**Fricatives**		**Stops**	
ux	t**oo**t	s	**s**ea	p	**p**ea
ay	b**i**te	sh	**sh**e	b	**b**ee
oy	b**oy**	z	**z**one	t	**t**ea
ey	b**ai**t	zh	a**z**ure	d	**d**ay
iy	b**ee**t	th	**th**in	k	**k**ey
aw	b**ou**t	dh	**th**en	g	**g**ay
ow	b**oa**t	f	**f**in	dx	mu**dd**y
ax	about	v	**v**an	q	
axr	butter				
ix	debit	**Affricates**		**Closures**	
ax-h	suspect	ch	**ch**oke	pcl	**p**ea
		jh	**j**oke	bcl	**b**ee
Silence				tcl	**t**ea
h#				dcl	**d**ay
pau				kcl	**k**ey
epi				gcl	**g**ay

Table 7: 61 Phones of the TIMIT Database

1.	p	14.	t	27.	k
2.	pcl tcl kcl bcl dcl	15.	dx	28.	b
	gcl q epi h# pau				
3.	d	16.	g	29.	dh
4.	m em	17.	n en nx	30.	ng eng
5.	s	18.	z	31.	ch
6.	th	19.	f	32.	sh zh
7.	jh	20.	v	33.	l el
8.	r	21.	y	34.	hh hv
9.	w	22.	eh	35.	ow
10.	ao aa	23.	uw ux	36.	er axr
11.	ay	24.	ey	37.	aw
12.	ax ah ax-h	25.	ix ih	38.	ae
13.	uh	26.	oy	39.	iy

Table 8: 39 Phones of the merged TIMIT Database

B ISOLET recognition results

$Model_{AV}$		/bee/	/dee/	/gee/	Average	PI
REP1	$\Omega = 0.5$	71.67%	75.00%	88.33%	78.33%	6-10-8
	$\Omega = 1$	71.67%	75.00%	88.33%	78.33%	6-10-8
	$\Omega = 2$	71.67%	70.00%	93.33%	78.33%	1-1-5
	$\Omega = 4$	71.67%	68.33%	95.00%	78.33%	1-1-8
	$\Omega = 6$	71.67%	68.33%	95.00%	78.33%	1-1-8
	$\Omega = 8$	71.67%	68.33%	95.00%	78.33%	1-1-8
	$\Omega = 10$	71.67%	68.33%	95.00%	78.33%	1-1-8
REP2	$\Omega = 0.5$	68.33%	81.67%	83.33%	77.78%	20-15-8
	$\Omega = 1$	68.33%	76.67%	91.67%	78.89%	13-2-10
	$\Omega = 2$	70.00%	81.67%	85.00%	78.89%	11-16-7
	$\Omega = 4$	70.00%	80.00%	85.00%	78.33%	7-11-1
	$\Omega = 6$	70.00%	75.00%	90.00%	78.33%	13-11-14
	$\Omega = 8$	70.00%	75.00%	90.00%	78.33%	16-13-16
	$\Omega = 10$	70.00%	75.00%	90.00%	78.33%	16-13-16
REP3	$\Omega = 0.5$	61.67%	70.00%	91.67%	74.44%	1-7-10
	$\Omega = 1$	63.33%	68.33%	91.67%	74.44%	1-5-9
	$\Omega = 2$	63.33%	71.67%	86.67%	73.89%	2-7-8
	$\Omega = 4$	63.33%	71.67%	86.67%	73.89%	4-8-9
	$\Omega = 6$	63.33%	66.67%	91.67%	73.89%	1-7-12
	$\Omega = 8$	63.33%	65.00%	91.67%	73.33%	1-6-11
	$\Omega = 10$	63.33%	70.00%	86.67%	73.33%	1-7-8
REP4	$\Omega = 0.5$	60.00%	75.00%	90.00%	75.00%	11-2-15
	$\Omega = 1$	61.67%	75.00%	88.33%	75.00%	3-2-11
	$\Omega = 2$	61.67%	75.00%	88.33%	75.00%	1-3-9
	$\Omega = 4$	61.67%	73.33%	91.67%	75.56%	1-4-18
	$\Omega = 6$	60.00%	78.33%	88.33%	75.56%	1-15-19
	$\Omega = 8$	61.67%	73.33%	91.67%	75.56%	1-3-17
	$\Omega = 10$	63.33%	71.67%	91.67%	75.56%	1-1-16

Table 9: Resulting discriminant accuracy using $\mathcal{M}_{AV}$ and different smoothing parameters Ω. The mapping scheme for the ISOLET database is based on the assumption that all diphones consists of the same number of frames because of the missing phone labels. The used plane index for the optimal accuracy is given for each representation which translates to a certain time constraints τ.

$Model_{PC}$		/bee/	/dee/	/gee/	Average	PI
REP1	$\Omega = 0.5$	71.67%	78.33%	88.33%	79.44%	1-8-1
	$\Omega = 1$	71.67%	76.67%	93.33%	80.56%	6-1-11
	$\Omega = 2$	71.67%	75.00%	93.33%	80.00%	1-8-4
	$\Omega = 4$	71.67%	75.00%	93.33%	80.00%	1-8-3
	$\Omega = 6$	71.67%	75.00%	93.33%	80.00%	1-8-3
	$\Omega = 8$	71.67%	75.00%	93.33%	80.00%	1-8-3
	$\Omega = 10$	71.67%	75.00%	93.33%	80.00%	1-8-3
REP2	$\Omega = 0.5$	68.33%	81.67%	91.67%	80.56%	3-8-10
	$\Omega = 1$	71.67%	80.00%	91.67%	81.11%	3-7-8
	$\Omega = 2$	73.33%	80.00%	90.00%	81.11%	3-7-8
	$\Omega = 4$	73.33%	80.00%	90.00%	81.11%	3-7-8
	$\Omega = 6$	73.33%	81.67%	88.33%	81.11%	1-9-5
	$\Omega = 8$	73.33%	81.67%	86.67%	80.56%	3-7-1
	$\Omega = 10$	73.33%	81.67%	86.67%	80.56%	3-7-1
REP3	$\Omega = 0.5$	73.33%	78.33%	83.33%	78.33%	1-4-6
	$\Omega = 1$	73.33%	81.67%	80.00%	78.33%	1-4-3
	$\Omega = 2$	73.33%	81.67%	81.67%	78.89%	1-4-3
	$\Omega = 4$	73.33%	81.67%	81.67%	78.89%	1-4-3
	$\Omega = 6$	73.33%	81.67%	81.67%	78.89%	1-4-3
	$\Omega = 8$	73.33%	81.67%	81.67%	78.89%	1-4-3
	$\Omega = 10$	73.33%	81.67%	81.67%	78.89%	1-4-3
REP4	$\Omega = 0.5$	61.67%	76.67%	86.67%	75.00%	1-4-15
	$\Omega = 1$	60.00%	76.67%	90.00%	75.56%	1-9-17
	$\Omega = 2$	58.33%	76.67%	93.33%	76.11%	8-19-15
	$\Omega = 4$	60.00%	76.67%	91.67%	76.11%	3-15-16
	$\Omega = 6$	60.00%	76.67%	90.00%	75.56%	1-10-19
	$\Omega = 8$	60.00%	76.67%	91.67%	76.11%	5-17-16
	$\Omega = 10$	61.67%	76.67%	88.33%	75.56%	1-10-19

Table 10: Resulting discriminant accuracy using the Principal Curve model $\mathcal{M}_{PC}$ and different smoothing parameters Ω. The mapping scheme for the ISOLET database is based on the assumption that all diphones consists of the same number of frames because of the missing phone labels. The used plane index for the optimal accuracy is given for each representation which translates to a certain time constraints τ.

$Model_{GTM}$		/bee/	/dee/	/gee/	Average	PI
REP1	$\Omega = 0.5$	71.67%	65.00%	96.67%	77.78%	6-1-11
	$\Omega = 1$	71.67%	65.00%	96.67%	77.78%	10-5-12
	$\Omega = 2$	71.67%	65.00%	96.67%	77.78%	10-5-12
	$\Omega = 4$	71.67%	65.00%	96.67%	77.78%	10-5-12
	$\Omega = 6$	71.67%	65.00%	96.67%	77.78%	10-5-12
	$\Omega = 8$	70.00%	65.00%	96.67%	77.22%	7-1-11
	$\Omega = 10$	70.00%	65.00%	96.67%	77.22%	7-1-11
REP2	$\Omega = 0.5$	70.00%	73.33%	93.33%	78.89%	1-1-20
	$\Omega = 1$	70.00%	83.33%	80.00%	77.78%	1-19-12
	$\Omega = 2$	70.00%	83.33%	78.33%	77.22%	1-20-10
	$\Omega = 4$	70.00%	83.33%	78.33%	77.22%	1-20-9
	$\Omega = 6$	70.00%	83.33%	78.33%	77.22%	1-20-10
	$\Omega = 8$	70.00%	83.33%	78.33%	77.22%	1-20-10
	$\Omega = 10$	70.00%	85.00%	75.00%	76.67%	1-20-6
REP3	$\Omega = 0.5$	75.00%	60.00%	88.33%	74.44%	13-2-12
	$\Omega = 1$	71.67%	66.67%	80.00%	72.78%	1-1-7
	$\Omega = 2$	75.00%	56.67%	86.67%	72.78%	13-1-12
	$\Omega = 4$	75.00%	60.00%	83.33%	72.78%	14-4-14
	$\Omega = 6$	75.00%	60.00%	83.33%	72.78%	14-3-14
	$\Omega = 8$	75.00%	60.00%	83.33%	72.78%	15-3-14
	$\Omega = 10$	75.00%	60.00%	83.33%	72.78%	15-3-15
REP4	$\Omega = 0.5$	60.00%	61.67%	86.67%	69.44%	17-1-19
	$\Omega = 1$	55.00%	60.00%	86.67%	67.22%	1-1-19
	$\Omega = 2$	58.33%	63.33%	83.33%	68.33%	20-10-18
	$\Omega = 4$	58.33%	60.00%	83.33%	67.22%	7-1-14
	$\Omega = 6$	58.33%	60.00%	83.33%	67.22%	7-1-16
	$\Omega = 8$	58.33%	60.00%	83.33%	67.22%	8-1-17
	$\Omega = 10$	60.00%	55.00%	86.67%	67.22%	15-1-19

Table 11: Resulting discriminant accuracy using the Generative Topographic Mapping model $\mathcal{M}_{GTM}$ and different smoothing parameters Ω. The mapping scheme for the ISOLET database is based on the assumption that all diphones consists of the same number of frames because of the missing phone labels. The used plane index for the optimal accuracy is given for each representation which translates to a certain time constraints τ.

C TIMIT recognition results

$Model_{AV1}$		/bee/	/dee/	/gee/	Average	PI
	$\Omega = 0.5$	76.23%	75.01%	85.63%	78.98%	7-3-5
	$\Omega = 1$	74.98%	74.59%	85.63%	78.40%	6-3-6
	$\Omega = 2$	74.32%	72.70%	85.63%	77.55%	7-3-9
REP1	$\Omega = 4$	74.89%	77.94%	78.84%	77.23%	12-10-6
	$\Omega = 6$	75.56%	75.56%	80.06%	77.06%	9-2-5
	$\Omega = 8$	73.98%	74.28%	82.23%	76.83%	11-9-10
	$\Omega = 10$	73.98%	74.28%	82.23%	76.83%	11-9-10
	$\Omega = 0.5$	73.50%	63.74%	82.24%	73.16%	5-2-7
	$\Omega = 1$	77.71%	60.74%	82.24%	73.57%	16-1-9
	$\Omega = 2$	78.05%	60.74%	81.02%	73.27%	20-8-15
REP2	$\Omega = 4$	78.39%	60.12%	79.80%	72.77%	20-1-15
	$\Omega = 6$	78.96%	60.74%	78.58%	72.76%	20-2-14
	$\Omega = 8$	78.05%	61.85%	78.58%	72.83%	18-6-15
	$\Omega = 10$	78.96%	60.61%	78.58%	72.72%	20-1-14
	$\Omega = 0.5$	41.78%	60.39%	89.02%	63.73%	6-3-15
	$\Omega = 1$	39.63%	63.25%	89.02%	63.97%	1-2-15
	$\Omega = 2$	39.87%	63.30%	89.02%	64.06%	1-2-16
REP3	$\Omega = 4$	39.63%	64.53%	84.41%	62.86%	2-1-8
	$\Omega = 6$	39.29%	64.22%	84.41%	62.64%	2-1-8
	$\Omega = 8$	37.71%	61.84%	87.80%	62.45%	1-2-16
	$\Omega = 10$	37.47%	65.94%	83.19%	62.20%	1-3-9
	$\Omega = 0.5$	52.47%	67.14%	74.92%	64.84%	4-15-20
	$\Omega = 1$	53.38%	65.11%	73.70%	64.06%	5-12-20
	$\Omega = 2$	53.71%	62.60%	73.70%	63.34%	3-3-17
REP4	$\Omega = 4$	53.38%	61.14%	71.26%	61.93%	3-3-17
	$\Omega = 6$	54.86%	60.34%	74.92%	63.38%	10-1-20
	$\Omega = 8$	56.01%	58.27%	74.92%	63.07%	19-8-20
	$\Omega = 10$	56.01%	58.27%	73.70%	62.66%	20-7-20

Table 12: Resulting discriminant accuracy using $\mathcal{M}_{AV1}$ and different used smoothing parameters Ω and mapping schemes with the TIMIT database. The used plane index for the optimal accuracy is given for each representation which translates to a certain time constraints τ. In this experiment we warped the test data into the template before comparison.

$Model_{AV2}$		/bee/	/dee/	/gee/	Average	PI
REP1	$\Omega = 0.5$	80.78%	66.92%	81.02%	76.24%	7-1-8
	$\Omega = 1$	80.54%	70.58%	78.84%	76.66%	7-2-6
	$\Omega = 2$	80.88%	69.96%	81.02%	77.29%	8-2-7
	$\Omega = 4$	80.54%	71.37%	79.80%	77.24%	7-1-6
	$\Omega = 6$	80.21%	71.07%	79.80%	77.02%	7-1-6
	$\Omega = 8$	79.54%	70.89%	79.80%	76.74%	6-1-7
	$\Omega = 10$	79.54%	72.00%	78.56%	76.70%	6-1-6
REP2	$\Omega = 0.5$	74.22%	62.02%	84.41%	73.55%	3-1-10
	$\Omega = 1$	74.79%	62.82%	83.19%	73.60%	1-1-9
	$\Omega = 2$	74.79%	62.82%	84.41%	74.01%	5-14-15
	$\Omega = 4$	80.02%	60.48%	83.19%	74.56%	20-5-13
	$\Omega = 6$	77.62%	61.71%	84.41%	74.58%	15-1-12
	$\Omega = 8$	79.78%	61.54%	81.97%	74.43%	19-3-6
	$\Omega = 10$	79.78%	61.85%	81.97%	74.53%	19-1-4
REP3	$\Omega = 0.5$	47.10%	61.31%	56.04%	54.82%	3-7-11
	$\Omega = 1$	49.64%	53.59%	61.61%	54.95%	9-3-17
	$\Omega = 2$	50.89%	52.63%	61.61%	55.04%	10-2-17
	$\Omega = 4$	49.88%	53.90%	61.61%	55.13%	10-3-17
	$\Omega = 6$	50.22%	54.08%	61.61%	55.30%	9-1-6
	$\Omega = 8$	46.53%	59.90%	56.04%	54.98%	9-1-16
	$\Omega = 10$	51.89%	55.18%	58.22%	55.10%	9-1-12
REP4	$\Omega = 0.5$	56.68%	58.54%	59.44%	58.22%	1-7-11
	$\Omega = 1$	58.84%	56.77%	65.00%	60.21%	3-1-12
	$\Omega = 2$	58.60%	57.57%	65.00%	60.39%	2-3-6
	$\Omega = 4$	58.27%	56.90%	71.53%	62.23%	2-1-11
	$\Omega = 6$	58.94%	56.11%	71.53%	62.19%	5-1-9
	$\Omega = 8$	58.70%	55.80%	70.31%	61.60%	4-1-5
	$\Omega = 10$	58.70%	56.77%	70.31%	61.93%	2-1-5

Table13: Resulting discriminant accuracy using $\mathcal{M}_{AV2}$ and different used smoothing parameters Ω and mapping schemes with the TIMIT database. The used plane index for the optimal accuracy is given for each representation which translates to a certain time constraints τ. In this experiment we adjusted the test trajectory to the template using Schwartz inelastic region idea.

$Model_{PC}$		/bee/	/dee/	/gee/	Average	PI
REP1	$\Omega = 0.5$	72.49%	72.39%	85.63%	76.84%	9-3-9
	$\Omega = 1$	74.99%	71.28%	83.46%	76.58%	12-1-9
	$\Omega = 2$	72.49%	76.04%	80.06%	76.20%	12-7-7
	$\Omega = 4$	71.82%	76.53%	80.06%	76.14%	10-1-6
	$\Omega = 6$	74.41%	76.05%	76.67%	75.71%	13-3-4
	$\Omega = 8$	73.41%	70.80%	82.24%	75.48%	12-2-10
	$\Omega = 10$	70.82%	76.84%	78.84%	75.50%	10-2-6
REP2	$\Omega = 0.5$	72.93%	63.43%	83.46%	73.27%	5-2-7
	$\Omega = 1$	76.81%	60.74%	83.46%	73.67%	16-1-9
	$\Omega = 2$	77.72%	60.74%	81.02%	73.16%	20-8-15
	$\Omega = 4$	76.23%	62.15%	79.80%	72.73%	13-2-15
	$\Omega = 6$	78.63%	59.02%	79.80%	72.48%	20-2-17
	$\Omega = 8$	78.63%	60.30%	78.58%	72.50%	20-2-16
	$\Omega = 10$	77.38%	61.71%	77.36%	72.15%	17-1-16
REP3	$\Omega = 0.5$	38.72%	65.20%	85.63%	63.18%	1-3-6
	$\Omega = 1$	39.29%	63.65%	86.85%	63.27%	1-2-13
	$\Omega = 2$	38.72%	65.68%	84.41%	62.94%	1-1-7
	$\Omega = 4$	36.90%	66.30%	84.41%	62.54%	1-4-8
	$\Omega = 6$	36.56%	66.17%	84.41%	62.38%	1-4-9
	$\Omega = 8$	36.90%	64.75%	84.41%	62.02%	1-1-10
	$\Omega = 10$	36.56%	66.17%	83.19%	61.97%	1-4-9
REP4	$\Omega = 0.5$	53.04%	67.14%	74.92%	65.03%	4-15-20
	$\Omega = 1$	55.77%	63.87%	73.70%	64.45%	10-5-14
	$\Omega = 2$	54.62%	62.77%	73.70%	63.70%	7-6-17
	$\Omega = 4$	56.01%	60.83%	73.70%	63.51%	10-1-20
	$\Omega = 6$	56.59%	59.55%	74.92%	63.69%	13-1-20
	$\Omega = 8$	56.92%	58.14%	74.92%	63.33%	17-2-20
	$\Omega = 10$	53.71%	60.96%	73.70%	62.79%	2-8-20

Table 14: Resulting discriminant accuracy using the Principal Curve model $\mathcal{M}_{PC}$ and different used smoothing parameters Ω and mapping schemes with the TIMIT database. The used plane index for the optimal accuracy is given for each representation which translates to a certain time constraints τ. In this experiment we adjusted the test trajectory to the size of the template using time warping.

$Model_{GTM}$		/bee/	/dee/	/gee/	Average	PI
REP1	$\Omega = 0.5$	75.90%	74.10%	81.28%	77.10%	9-8-7
	$\Omega = 1$	73.41%	71.59%	84.68%	76.56%	3-1-7
	$\Omega = 2$	72.40%	71.59%	83.46%	75.82%	1-1-8
	$\Omega = 4$	73.31%	75.39%	77.26%	75.44%	11-11-8
	$\Omega = 6$	74.65%	71.42%	80.06%	75.38%	8-3-6
	$\Omega = 8$	74.65%	71.42%	80.06%	75.38%	8-2-6
	$\Omega = 10$	74.65%	71.42%	80.06%	75.38%	8-2-6
REP2	$\Omega = 0.5$	75.42%	61.40%	81.28%	72.70%	15-2-6
	$\Omega = 1$	76.57%	60.57%	81.28%	72.81%	16-1-7
	$\Omega = 2$	78.05%	61.67%	77.62%	72.45%	20-10-13
	$\Omega = 4$	78.39%	60.61%	77.62%	72.21%	20-1-15
	$\Omega = 6$	78.29%	60.92%	76.41%	71.87%	20-1-13
	$\Omega = 8$	78.29%	61.85%	76.41%	72.18%	20-6-16
	$\Omega = 10$	77.96%	63.43%	74.23%	71.87%	18-3-7
REP3	$\Omega = 0.5$	53.95%	46.49%	86.85%	62.43%	13-1-11
	$\Omega = 1$	54.10%	42.26%	90.24%	62.20%	15-2-18
	$\Omega = 2$	53.19%	40.62%	90.24%	61.35%	14-1-19
	$\Omega = 4$	42.60%	55.71%	84.41%	60.91%	1-2-5
	$\Omega = 6$	42.60%	54.74%	84.41%	60.58%	1-1-5
	$\Omega = 8$	49.54%	41.46%	89.02%	60.01%	14-1-19
	$\Omega = 10$	44.75%	50.46%	83.19%	59.47%	6-1-9
REP4	$\Omega = 0.5$	53.86%	64.89%	74.92%	64.55%	3-15-20
	$\Omega = 1$	55.68%	62.55%	73.70%	63.98%	1-2-12
	$\Omega = 2$	55.68%	61.14%	74.92%	63.91%	10-10-20
	$\Omega = 4$	57.74%	59.37%	73.70%	63.60%	17-12-20
	$\Omega = 6$	58.89%	56.81%	74.92%	63.54%	14-1-19
	$\Omega = 8$	54.53%	60.30%	74.92%	63.24%	1-10-20
	$\Omega = 10$	55.10%	59.68%	72.48%	62.42%	2-12-20

Table 15: Resulting discriminant accuracy using the Generative Topographic Mapping model $\mathcal{M}_{GTM}$ and different used smoothing parameters Ω and mapping schemes with the TIMIT database. The used plane index for the optimal accuracy is given for each representation which translates to a certain time constraints τ. In this experiment we adjusted the test trajectory to the size of the template using time warping.