PARAMETRIC SUBSPACE MODELLING OF SPEECH TRANSITIONS

Klaus Reinhard and Mahesan Niranjan

Cambridge University Engineering Department
Cambridge CB2 1PZ, England, U.K.
kr10000@eng.cam.ac.uk niranjan@eng.cam.ac.uk

ABSTRACT

In this paper we report on attempting to capture 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 attempts such as recurrent neural networks there is the hope, but not convincing demonstration, that such transitional information could be captured. We start 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 (Time Constrained Principal Component Analysis). On this subspace, we attempt a parametric modelling of the trajectory, compute a distance metric to perform classification of diphones. Much of the discriminant information is still retained in this subspace. This is illustrated on the isolated transitions /bee/, /dee/ and /gee/.

1. INTRODUCTION

The temporal evolution of the short time spectrum is an important characteristic of speech signals. This time variation is caused by the movement of the vocal tract and is a rich source of information not only of the phonetic content of what is spoken, but also other information, such as the speaker. State of the art statistical models make crude approximations to the temporal variation, essentially by a piecewise constant approximation that is inherent in the hidden Markov model. Small extensions to this approximation, such as the inclusion of *delta* and *delta-delta* parameters has become common practice, but one is immediately faced with the problem of reliability in parameter estimation caused by the expansion in dimensionality. The use of Recurrent Neural Networks is seen as one plausible mechanism to capture such transitional information. An alternative approach is the use of segmental models that model the time evolution of feature vectors within a segment. Typically, these approaches use the phone as the unit of segmentation [5].

We start from a slightly different premise that attempts to focus on the transition between phones. The spectral trajectory into the vowel [i:], for example is quite different in the CV transition /bee/ than in that of /gee/. Clearly, a phone model for the vowel [i:] derived from all contexts would be noisy. Hence we focus on diphone units, defining the diphone as half of one phone followed by half of the next phone. While the number of segments to model increases rapidly, the hope is that one has a greater chance of capturing the transitional information explicitly. We adopt a modified Principal Component Analysis (PCA) approach, to compute projections onto a two dimensional subspace of the diphone transitions. The two dimensional projections were a starting point for this work, enabling the visualisation of the trajectory in the projected space, but the method itself is not restricted to two dimensions. The work described in this paper shows that much of the discriminatory information is retained in the projection we propose. We illustrate this on a simple problem involving the discrimination of /b/, /d/ /g/, on the ISOLET database [1]. Receiver Operating Characteristic (ROC) curve is used to present the compromise between detection of a transition and false alarms.

2. SUBSPACE MODEL

We find a two dimensional projection of the diphone data. The number of parameters required to perform this projection is $2*(n+p)$, where p is the dimensionality of the parametric spectral representation and n is the number of spectral frames in the diphone. Information of the temporal ordering of the data frames is captured by introducing a constrained PCA, described below.

2.1. Time-constraint PCA (TC-PCA)

A very popular unsupervised technique for dimensionality reduction is the principal component analysis or *Karhunen-Loéve-Transform* [2], the principal directions being given by the eigenvectors of the data covariance matrix. 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 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 extra time dimensionality. Tuning this parameter is achieved by an exhaustive search to determine the most discriminant subspace among all models during performance tests. 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:

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

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

$$\Sigma^{\tau} \mathbf{u_i^{\tau}} = \lambda_{\mathbf{i}}^{\tau} \mathbf{u_i^{\tau}} \tag{2}$$

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 eigenvlues 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 $\lambda_1^{\tau} \geq \lambda_2^{\tau} \geq \cdots \geq \lambda_{N+1}^{\tau}$.

$$\mathcal{P}_{\tau} = \begin{bmatrix} \mathbf{u}_1^{1^{\tau}} & \mathbf{u}_2^{1^{\tau}} \\ \vdots & \vdots \\ \mathbf{u}_1^{p^{\tau}} & \mathbf{u}_2^{p^{\tau}} \end{bmatrix} * \begin{bmatrix} \sqrt{\lambda_1^{\tau}} & 0 \\ 0 & \sqrt{\lambda_2^{\tau}} \end{bmatrix} \tag{3}$$

2.2. Trajectory Model

The aim is to extract quantitative non-linear dynamics from the training data to form a trajectory model. In the time constraint planes described above we found strong indicator for typical trajectories for the different CV syllables. This is illustrated in Fig. 1.

The method of extracting the underlying sequence of points in representation space to form a trajectory model is inspired by the Bias-Variance discussion. Assuming that 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 $\mathbf{t_{kl}} \in \mathbf{R^p}$, one likes to find the best representationm resulting in a minimum error solution. Because we are looking for an N point trajectory model to compare similiar test trajectories frame-wise with our model we consider each frame separately and calculate an average p-dimensional point for each frame which leads to a sequence of average points where $\mathcal{E}_{\mathcal{T}}[\cdot]$ denotes the expectation, or ensemble average.

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

3. TRAJECTORY MAPPING

Trajectory mapping is motivated from the observation that although the speech signal is highly dynamic, its movement tends to follow certain paths, corresponding to the underlying involved phonemic units [7]. Hence we should find typical *shapes* of trajectories for specific diphones which could be corrupted by noise or speaker characteristics. Here an appropriate distance measure has to be found which counts for the *shape* of the speech transition.

3.1. Smoothing Spline

The idea of smoothing splines is 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

$$t_i = S(x_i) + \epsilon_i; \qquad i = 1, \cdots, N. \tag{4}$$

where $S(\cdot)$ is a deterministic function characterising the major 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 $S(\cdot)$ such that it is close to the data path as possible and on the other hand as smooth as possible. This idea can be formulated through the following objective function:

$$L(S) = \int_a^b \left[S^{''}(x) \right]^2 dx + \sum_{i=1}^{N} \omega_i \left(t_i - S(x_i) \right)^2. \tag{5}$$

where the ω_i's are the weight associated with t_i's representing the relative contributions of the ith observation to the model estimation.

It is well known that the problem of minimising the objective function in (5) has a unique explicit solution [3], which is indeed 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, funtion 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 the high variance and makes a comparison to a given generalised trajectory model more reliable [3].

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

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

$$\Omega = \mathbf{diag}(\omega_1, \cdots, \omega_N)$$

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

$$\begin{aligned} \mathbf{A} &= (\mathbf{D}\Omega^{-1}\mathbf{D}^{\mathbf{T}} + \mathbf{B})^{-1}\mathbf{DT} & (6) \\ \mathbf{F} &= \mathbf{T} - \Omega^{-1}\mathbf{DA} & (7) \end{aligned}$$

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

$$\mathcal{S}_{\mathbf{T}} = (\mathbf{f_1}, \cdots, \mathbf{f_N})^{\mathbf{T}} = (\tilde{\mathbf{t}}_1, \cdots, \tilde{\mathbf{t}}_N)^{\mathbf{T}}.$$

3.2. Distance Measure Classification

In our subpace representation we have to find a similiarity 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 individual distances by its arc length [4].

The basic operation for our subspace method is a projection of the high-dimensional vector $\tilde{\mathbf{t}}$, and is given by $\tilde{\mathbf{t}}_P = (\tilde{\mathbf{t}})^T \mathbf{P}_\tau$ where the projection matrix $\mathbf{P}_\tau$ is the found subspace matrix defined by the training data and a certain time constraint factor τ. Using our trajectory model $\mathcal{M}$ we can compute a normed squared orthogonal distance $d_{sub}(\hat{\mathbf{t}}, \tilde{\mathbf{t}})$ from our trajectory model $\hat{\mathbf{t}}_P = (\hat{t})^T \mathbf{P}_\tau$:

$$d_{sub}(\hat{\mathbf{t}}_P, \tilde{\mathbf{t}}_P) = \sum_{i=1}^{N} d_{sub}(\hat{\mathbf{t}}_P^i, \tilde{\mathbf{t}}_P^i) = \frac{\sum_{i=1}^{N} \|\hat{\mathbf{t}}_P^i - \tilde{\mathbf{t}}_P^i\|^2}{arclength(\tilde{\mathbf{t}}_P)}. \quad (8)$$

Performing the distance measure for all models we obtain our classification result by finding the diphone template with the minimum distance.

4. EXPERIMENTAL ILLUSTRATION

We use a subset of the ISOLET [1] 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. The available data was split into a training and test set. We used 80% of the data for training (ISOLET1-4), 20% for tests (ISOLET5), as recommended by the originators of the dataset.

With an average recognition accuracy of 78.3% the results produced with our subspace models are worse in comparison to the results for a baseline HMM using one mixture and a diagonal covariance matrix on a BTL E-SET giving 84.5 % accuracy. 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 of the HMM system, the subspace trajectory approach uses only $2 * (6 + 18) = 48$ parameters when the trajectory is built by 18 anchor points.

Representation	**Accuracy**			
	BEE	DEE	GEE	Average
4 MFCC + energy	71.7%	75%	88.3%	78.3%

Table 1: Obtained results for our trajectory models from diphones /bee/,/dee/ and /gee/.

5. DISCUSSION

In this study, we have proposed a new method of modeling speech transitions with a subspace model. We showed that temporal transitions in speech can be visualised and modeled in a low dimensional space. This approach has the advantages that the memory requirements for our subspace model is much less demanding in comparison with models involving context-dependent speech units, which furthermore leads to a model which is easier trainable with the presents of a limited amount of training data which is true for diphones in speech. 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 obtained information whether modelling transitions provides one with orthogonal information in comparison with the information obtained by standard HMM systems. We will extend our experimental results using TIMIT to retrieve more reliable results while optimising our approach according to employ more suitable speech representations, trajectory mapping methods and subspace projections [6].

6. REFERENCES

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

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

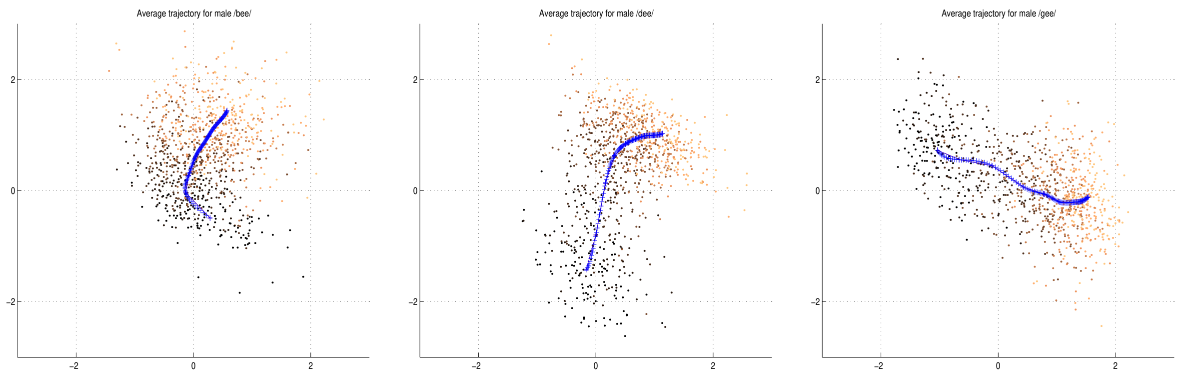
[3] P. Lancaster and K. Salkauskas. *Curve and Surface Fitting: An introduction.* Academic Press, 1986.

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

[5] 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.

[6] K. Reinhard and M. Niranjan. Subspace Modelling of Speech Transitions. Technical report, Cambridge University Engineering Department, 1997, http://svr-www.eng.cam.ac.uk/kr10000.

[7] 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.



(a) scatterplot of the diphone /bee/ for male speaker and their average trajectory plotted onto most discriminant plane

(b) Scatterplot of the diphone /dee/ for male speaker and their average trajectory plotted onto most discriminant plane

(c) Scatterplot of the diphone /gee/ for male speaker and their average trajectory plotted onto most discriminant plane

Figure 1: Optimised planes for /bee/, /dee/ and /gee/: The trajectory model for male speakers and a scatter plot of the data for each diphone is shown. The planes are spanned by the largest eigenvectors found in the described TC-PCA framework. Evolution in time is characterised in the scatterplot as a change in colours from dark to bright.

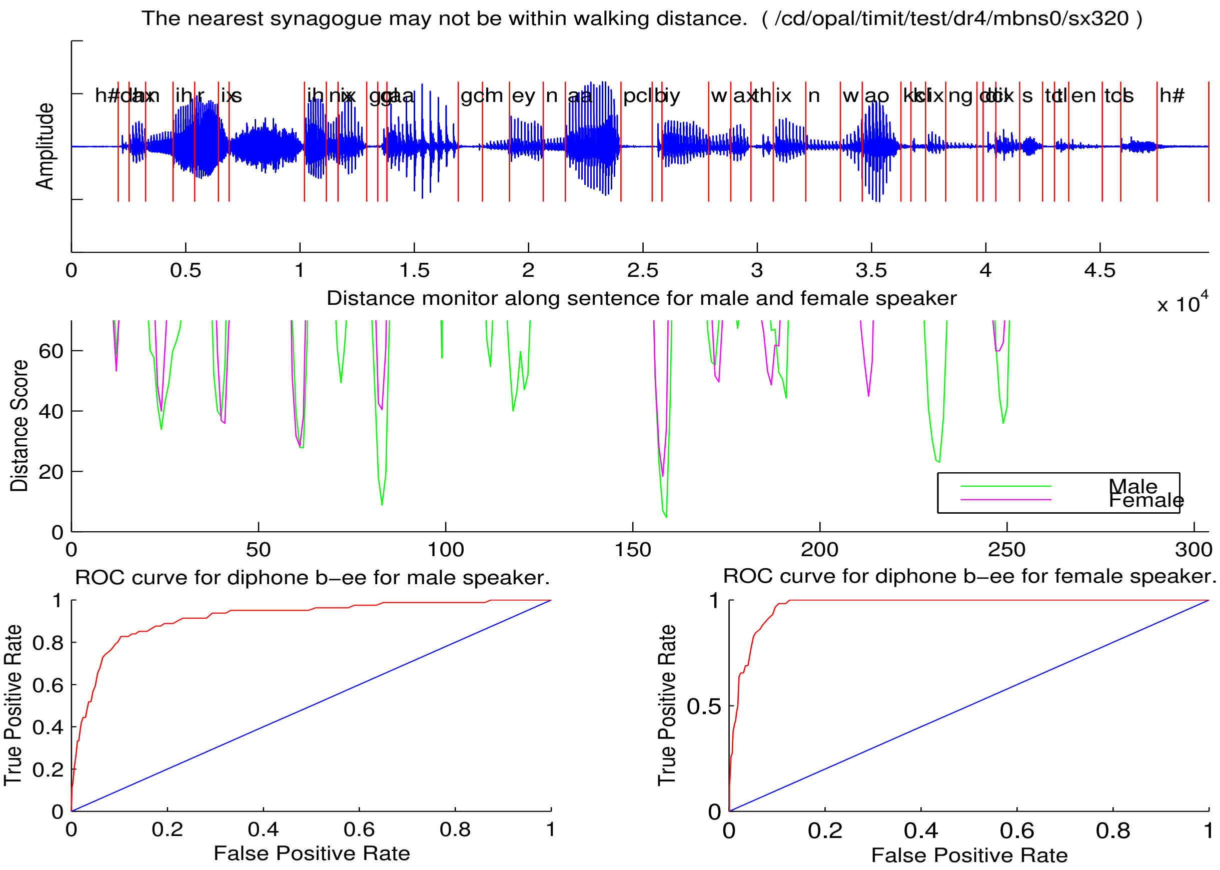


Figure 2: Distance monitor for male and female diphone model /bee/ along a sentence from the TIMIT database. Small distances result in good matches between model and test trajectory. The corresponding ROC curve for male and female /bee/ over the whole TIMIT database is plotted underneath.