DIPHONE MULTI-TRAJECTORY SUBSPACE MODELS

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 the extension of capturing speech transitions embedded in diphones using trajectory models. The slowly varying dynamics of spectral trajectories carry much discriminant information that is very crudely modelled by traditional approaches such as HMMs. We improved our methodology of explicitly capturing the trajectory of short time spectral parameter vectors introducing multi-trajectory concepts in a probabilistic framework. Optimal subspace selection is presented which finds the most discriminant plane for classification. Using the E-set from the TIMIT database results suggest that discriminant information is preserved in the subspace.

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 or the use of context dependent models (triphones) has become common practice. However, both specialised 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. 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 [3].

Clearly, a phone model for the vowel [i:] derived from all contexts would be noisy, due to the different spectral trajectories into the vowel [i:], for example in the CV transition /bee/ and /gee/. Hence we start from a slightly different premise that attempts to focus on the transition between phones. Diphone units capture these transitions being defined 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. The work described in this paper is an extension of our earlier approach to model speech transitions in a subspace where the temporal ordering is preserved which may be found in [4].

This paper introduces multi-trajectory concepts and a probabilistic trajectory scoring to define an objective trajectory mapping method. Different methods of trajectory clustering and subspace selection are shown to optimise our subspace models. Findings are that much of the discriminatory information is retained even in a continuous speech environment. We illustrate this on a simple problem involving the discrimination of the E-set, on the TIMIT database. The confusion matrix is used to show the potential complementary information by modeling phone transitions explicitly.

2. SUBSPACE PROJECTION MODEL

Projecting a sequence of short term spectral parameters onto a subspace with $l \leq 3$, where the temporal sequence of these vectors are preserved, makes it possible to visualise and model trajectories of important speech dynamics. The parameter requirements for such a subspace model is reduced to $l \times (n + p)$, where p is the dimensionality of the spectral representation and n is the average number of spectral frames for a speech unit.

An adaptation of the well known technique for dimensionality reduction, principal component analysis or *Karhunen-Loéve-Transform* [1], is used to generate projections onto a l-dimensional subspace where the temporal ordering of the data sequence is preserved. This method is called time-constrained PCA (TC-PCA) [4].

Considering 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. TC-PCA can be described by solving the covariance matrix of the set of temporal extended vectors. The subspace projection for each diphone a results in a transformation matrix $\mathcal{P}^a_\tau$. Finding the best subspace for a particular diphone model which is

represented by a particular τ is achieved by finding the best plane among all models, described below in section *Optimal Subspace*.

2.1. Data Importance Adjustment

The extraction of the diphones from a phonetically labeled database, like TIMIT, was performed using the start and end sample information for the different phones involved. This information made it possible to calculate the average frame length of each phone involved. Considering Schwartz et al. [6], who proposed the importance of the transitions within diphones, each training token was adjusted to preserve the important regions. Here the inner region of a diphone was treated as an inelastic area so that for expansion or shrinkage the area around the phone boundary was preserved (see Figure 1). The average length for each phone involved was used to adapt the training data for one diphone so that each token had the same length. The resulting data set $\mathcal{T}$ was then used to calculate the needed transformation matrices in the TC-PCA process.

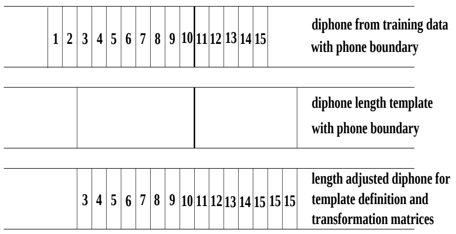


Figure 1: Extension and shrinkage rule for diphones having different length in the process of template generation and transformation matrix calculation using Schwartz elastic/inelastic region theory.

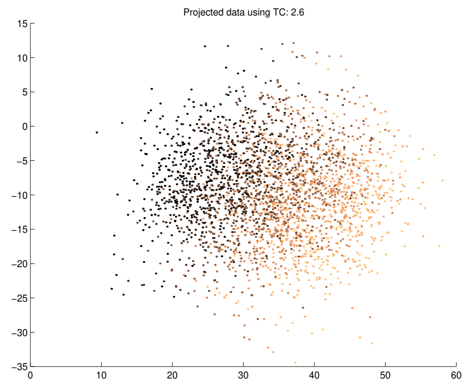
The normalisation of the individual trajectories for all diphones was performed to reveal the characteristic spectral shapes of the projected short term spectral parameters. Speaker characteristics distort the underlying process by noise and scaling problems to which the TC-PCA process is sensitive. Figure 2 shows the projection results for raw and normalised diphone scatter plots.

2.2. Gaussian Trajectory Model

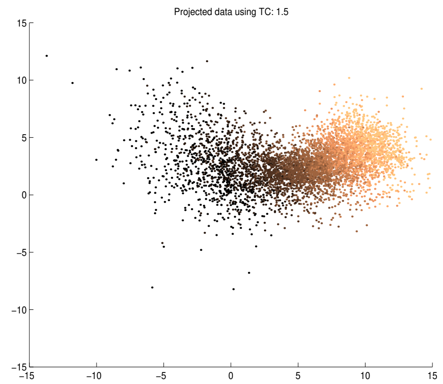
Assuming an N length sequence of observation vectors $t_1^N = [t_1, \cdots, t_N]$ generated by diphone a, where t_j is a p-dimensional observation vector at time j, this sequence defines a segment corresponding to diphone a. The subspace trajectory is than formed by transforming the p-dimensional observation segment into a l-dimensional one $\hat{t}_1^N$, using $\hat{t}_1^N = t_1^N \times \mathcal{P}_\tau^a$. The segment is represented by the subspace trajectory model:

$$\hat{\mathbf{t}}_j = \hat{\mu}_j^a + \hat{\epsilon}_j \qquad 1 \leq j \leq N \tag{1}$$

where $\hat{\mu}_j^a$ and $\hat{\epsilon}_j$ are the l-dimensional mean vector and residual error vector at time j. Considering the error vectors $\hat{\epsilon}_j$ i.i.d., representing a Gaussian with zero mean and an



(a) Projected raw data



(b) Projected normalised data

Figure 2: Scatterplot of the data set for the diphone /d-ih/, changing gray scales from dark to bright indicates temporal evolution. (a) shows the raw data projected on a time constrained subspace. (b) shows a projection of the same data normalised. Normalisation reveals the underlying spectral shape of the temporal evolution of the projected short term spectral parameters.

invariant covariance matrix Σ_a, the likelihood of a sequence of vectors can be expressed as:

$$P(\hat{\mathbf{t}}_1, \cdot, \hat{\mathbf{t}}_N | a) = \prod_{j=1}^{N} f(\hat{\mathbf{t}}_j)$$

$$= \prod_{j=1}^{N} \frac{1}{(2\pi)^{\frac{l}{2}} |\Sigma_a|^{\frac{l}{2}}} e^{-\frac{1}{2}(\hat{\mathbf{t}}_j - \hat{\mu}_j^a)^T \Sigma_a^{-1} (\hat{\mathbf{t}}_j - \hat{\mu}_j^a)}. \tag{2}$$

Extending this scenario to a multi trajectory case, where M trajectories represent a diphone segment as a sequence of anchor points represented as Gaussian the mixture density trajectory model can be described as:

$$f(\hat{\mathbf{t}}_j) = \sum_{k=1}^{M} w_k f_k(\hat{\mathbf{t}}_j) \tag{3}$$

The mixture weights w_k sum up to one, and the model parameters $\hat{\mu}_1^{a_k}, \cdots, \hat{\mu}_N^{a_k}$ and Σ_a can be estimated using a maximum likelihood criterion described in the following section.

2.3. Trajectory Clustering

Characteristic trajectories captured in diphones are modelled in the high-dimensional spectral parameterisation. Optimising the trade-off between parameter requirements and accuracy performances, the most suitable subspace can be chosen using any best subspace projection onto a l-dimensional subspace $l < p$ by adaptation of the projection matrix.

2.3.1. K-means Trajectories

One simple way of clustering trajectories which still maintains the natural trajectories within a cluster is using the K-means algorithm. The algorithm is applied to data belonging to the initial frame of a diphone a, which is located in a spectral stationary region. The algorithm partitions the data set $\mathcal{T}_a$ into M disjoint subsets $\mathcal{S}_{a_k}$ containing $\mathbf{T}_k$ data points. Subsequently clustering the initial frame into M centers, the consecutive points are also pooled according to the subsets found, hence:

$$\begin{aligned} \mu^{a_k} &= \frac{1}{\mathbf{T}_k} \sum_{n \in \mathcal{S}_k} \mathbf{t}_n \\ (\mu_1^{a_k}, \cdots, \mu_N^{a_k}) &= \left(\frac{1}{\mathbf{T_{k_1}}} \sum_{n \in \mathcal{S}_k} \mathbf{t}_{n_1}, \cdots, \frac{1}{\mathbf{T}_{k_N}} \sum_{n \in \mathcal{S}_k} \mathbf{t}_{n_N}\right) \end{aligned}$$

Depending on the number of clusters found by the K-means trajectory method, the algorithm results in M different trajectory templates which were used for trajectory classification described below.

2.3.2. EM Trajectories

Trajectory templates for model parameters can also be obtained using a maximum likelihood criterion, that maximises $P(\mathbf{t}_1^N|a)$. The re-estimation formulas based on the EM algorithm were derived by Fukada et al. [2], who maximised the following auxiliary function Q:

$$\begin{aligned} Q(\bar{\Phi}|\Phi) &= \mathcal{E}[logP(\mathbf{t}_1^N, k|\bar{\Phi})|\mathbf{t}_1^N, \Phi] \\ &= \sum_{k=1}^{M} \frac{P(\mathbf{t}_1^N, k|\Phi)}{P(\mathbf{t}_1^N|\Phi)} logP(\mathbf{t}_1^N, k|\bar{\Phi}) \end{aligned} \tag{4}$$

where Φ and $\bar{\Phi}$ are the sets of the current model parameter and the re-estimated model parameter. Maximising Eq. 4 will lead to the different model parameters for k different mixture components. The results of the k-means trajectory clustering are used as initial model parameters for the EM approach.

3. TRAJECTORY CLASSIFICATION

When a test trajectory is received, it is time warped to match the length of the template. This enables the scoring of test trajectories of different length. The model score is defined by the log-likelihood of the template which results in the maximum score $\mathcal{V}_a$. The trajectory is than classified by finding the diphone model with the best score, discriminating between all competing diphones within the set of models. The scoring is performed in a arbitrary subspace and the projection matrix is applied to the templates and the test trajectory before the scoring process is performed. The likelihood score $\mathcal{V}_a$ for a individual diphone a can be expressed as:

$$\mathcal{V}_a = \max_k \{ \frac{1}{N_a} \sum_{j=1}^{N_a} logP(\hat{\mathbf{t}}_1^N | \hat{\mu}_{1_k}^{N_k}) \} \tag{5}$$

3.1. Optimal Subspace

The aim is to choose a plane which scores maximal in terms of resulting likelihoods for training tokens and is also maximal discriminant when competing with other models. During the training process of choosing an optimal projection plane, two different approaches are introduced to find the optimal plane. The first method which is computational much cheaper ($\mathcal{O}(N)$) is the ML transformation approach. Whereas the maximum discriminant method with its $\mathcal{O}(N^2)$ computational requirements is much more expensive. Both algorithms are described below.

3.1.1. ML Transformation

A simple way of finding an optimal plane for data transformation is to consider the training data for a specific diphone only, finding a maximum likelihood (ML) solution. Determination of the optimal plane is performed by maximising the sum over all likelihood scores within the training set which results in a particular plane index. No data from different diphones are considered for this selection.

3.1.2. MD Transformation

Transformation plane selection can also be performed considering the data for all competing models to obtain a plane which is most discriminant (MD). The scores for all training data and for all diphones is used for a particular model to find the projection plane which results in the most discriminant transformation.

4. EXPERIMENTAL WORK

The experimental illustrations in this section, focus on the inter-model discriminative accuracy, showing how well one can distinguish between diphone trajectory models. Despite a significant information loss during the dimensionality reduction, the important dynamic information is still preserved. TIMIT is a convenient database to demonstrate this approach of using diphones as speech segments because TIMIT is phonetically labeled which makes it possible to extract all occurring diphones. In the test scenario the diphones of the complete E-set are used (e.g. /b-ih/, /d-ih/, /jh-ih/, /p-ih/, /s-ih/, /t-ih/ and /v-ih/). There are in total 2948 training tokens and 1076 testing tokens available. Utilising the finding from our earlier work that higher order MFCCs have an oscillating nature, five Mel-frequency cepstral coefficients (MFCC) are used as speech feature vectors. With an average trajectory sequence length of approximately 12 frames within the E-set the parameter requirements for a mixture model of k mixtures using an l-dimensional subspace can be calculated by $\mathcal{M} = [(k \times 12) + (5+1)] \times l$, which results in an average model size of about 100 parameters. In Figure 3 the confusion matrices is shown for the classification task in the original parameter space and the projected space.

	TIMIT E-set Accuracy			
	ML Plane		MD Plane	
Subspace	Kmeans	EM	Kmeans	EM
2D(normed)	49.8%	50.9%	41.7%	43.0%
Org(normed)	57.4%	57.1%	57.4%	57.1%
2D(unnormed)	45.6%	31.6%	47.3%	34.0%
Org(unnormed)	53.9%	41.8%	53.9%	41.8%

Table 1: Accuracy measures for the complete E-set using the TIMIT database. Results are shown for a 2-dimensional subspace and the original space using the different optimal subspaces and the different trajectory modeling methods. Separate measures were given for normed and unnormed trajectories.

5. DISCUSSION

In this paper we presented an extension of our work of modeling temporal trajectories on a low-dimensional subspace, which results in models of very low complexity and reduced memory requirements in comparison with models involving context-dependent speech units. The results suggest that discriminant information is preserved in the subspace focusing on the temporal ordering. Future work will concentrate on the usefulness of the trajectory information whether modeling transitions provides one with complementary knowledge in comparison with the information obtained by standard HMM systems. Using a N-best rescoring scheme [5] the proposed algorithm can be extended to deal with continuous speech, which is proposed to incorporate the subspace mode into a phone-based system. The rescoring mechanism can be used to emphasise paths in the lattice of hypotheses using 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 lead to better performances.

6. REFERENCES

[1] C.M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, 1995.

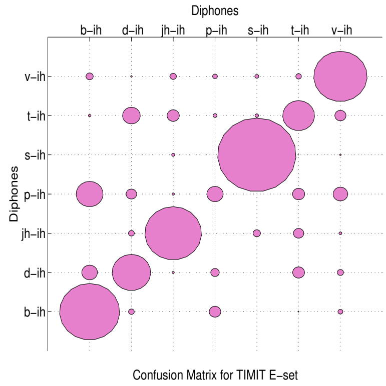
[2] T. Fukada, Y. Sagisaka, and K.K. Paliwal. Model parameter estimation for mixture density polynomial segment models. *Int. Conf. in Acoustics, Speech and Signal Processing*, 1997.

[3] 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.

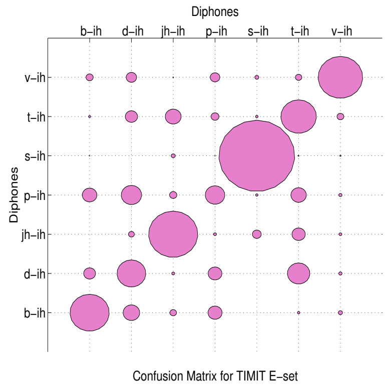
[4] K. Reinhard and M. Niranjan. Parametric subspace modeling of speech transitions. *Int. Conf. in Acoustics, Speech and Signal Processing*, 1998.

[5] 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.

[6] 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.



(a) Original space



(b) 2D subspace

Figure 3: Confusion matrix for diphone classification of the E-set within TIMIT. (a) shows the best results using the unprojected parameter space in which the classification is performed. (b) shows the best results of projected test trajectory onto a 2-dimensional space.