

# Trajectory Clustering for Solving the Trajectory Folding Problem in Automatic Speech Recognition

Yan Han\*, Johan de Veth, and Lou Boves

Center for Language and Speech Technology (CLST),

Radboud University Nijmegen, The Netherlands.

Post address: P.O. Box 9103, 6500 HD Nijmegen, The Netherlands.

Telephone: +31 (0)24 3616069

Fax: +31 (0)24 3612907

Email: {Y.Han, J.deVeth, L.Boves}@let.ru.nl

## **Abstract**

In this paper we introduce a novel method for clustering speech gestures, represented as continuous trajectories in acoustic parameter space. Trajectory Clustering allows us to avoid the conditional independence assumption that makes it difficult to account for the fact that successive measurements of an articulatory gesture are correlated. We apply the Trajectory Clustering method for developing multiple parallel HMMs for a continuous digits recognition task. We compare the performance obtained with data-driven clustering to the recognition performance obtained with conventional Head-Body-Tail models, which use knowledge-based criteria for building multiple-HMMs in order to obviate the trajectory folding problem.

The results show that Trajectory Clustering is able to discover structure in the the training database that is different from the structure assumed by the knowledge-based approach. In addition, the data-derived structure gives rise to significantly better recognition performance, and results in a 10% word error rate reduction.

**EDICS Category: SPE-RECO; SPE-GASR**

# Trajectory Clustering for Solving the Trajectory Folding Problem in Automatic Speech Recognition

## I. INTRODUCTION

Over the past decades, hidden Markov models (HMMs) have been the dominant methodology for modeling speech acoustics in automatic speech recognition. HMMs provide a flexible and powerful structure, in which the time-varying nature of speech is accounted for by an underlying Markov process, and the short time spectral variability is modelled by the statistical processes associated to the model states [1]. HMMs assume that all acoustic observation vectors in a given state depend only on the state, and are not dependent on neighboring vectors. However, this assumption is at odds with the fact that speech is produced by continuous movements of the articulators.

The discrepancy between the statistical independence assumption and the physics of speech production gives rise to the so called trajectory folding phenomenon [2], illustrated in Fig.1. For a sound  $s$ , a two state (A, B) left-to-right HMM with two Gaussian pdfs per state has been trained with observations from male (M) and female (F) speakers. It is reasonable to assume that, in each state, one Gaussian models the male and the other the female voice. For the male training data the model yields high probability for the path  $A(M) \rightarrow B(M)$ , and for female utterances for the path  $A(F) \rightarrow B(F)$ . However, when presented with previously unseen input the model may yield a high probability for the path  $A(M) \rightarrow B(F)$ , which has never been observed in the training set of  $s$  and that might correspond to another sound  $s'$ . As consequence,  $s'$  could be misrecognized as  $s$ , resulting in performance degradation.

To overcome the adverse effect of trajectory folding, several approaches to multiple-paths modeling have been proposed, such as multiple-HMMs modeling [3] [4] [5], and Mixture of Stochastic Trajectory Modeling [6]. The idea underlying multiple-paths acoustic modeling is to use model topologies with multiple parallel paths to account for the structure inherent in the acoustic variability. By constructing parallel paths for different pronunciation variants, trajectory folding is explicitly disallowed. All multi-path HMM approaches appear to result in some improvement of recognition performance [2]. Although the approaches differ in the methods used to build parallel paths, they share the problem of how to discover the structure of the acoustic variability, or more technically, how to cluster the training tokens

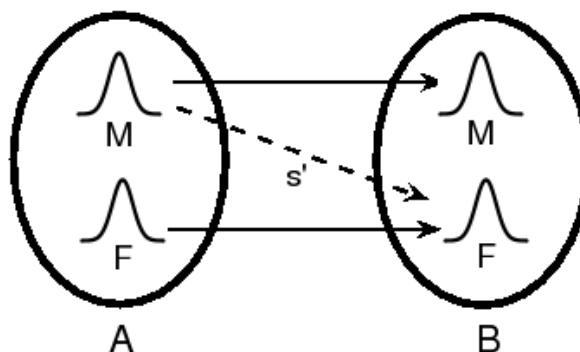


Fig. 1. An example to illustrate the trajectory folding phenomenon

belonging to the same acoustic unit (e.g. a phoneme, a syllable, or a word) into subgroups that are as homogenous as possible.

The usual way to account for the structure in the variation in speech is to build gender dependent models and context dependent triphones. These models use a priori knowledge (e.g. gender, linguistic context) as the criterion to cluster the training data, after which an HMM is built for each cluster. However, this top-down method is not necessarily suitable for all sources of variation. First of all, it is very hard to decide what is the most important source of variation in a certain speech database. Inter-speaker variation, for example, may well be more important than linguistic context variation for a small vocabulary recognition task. Secondly, even within one speech database, the most important variation for different acoustic units may be due to different factors, such as speaking style, speed or accent of the speakers. Unit dependent ranking of factors that cause pronunciation variation can be handled by means of decision trees, but only to the extent that the factors are accurately labeled in the training database. However, some important sources of variation may not be amenable to top-down modeling because they cannot be reliably annotated in the training data. Seemingly subtle local variations in speaking style and speaking rate, for instance, are important for many speech recognition tasks, but it is very hard to label parts of utterances in a database for these features. These limitations of the knowledge-based methodology limit the power of conventional multiple-HMM acoustic modeling.

To overcome the limitations of the knowledge-based approach, a data-driven approach that does not rely on linguistic knowledge would be attractive. Contrary to a knowledge-based approach, a data-driven approach automatically derives the most salient pronunciation variation classes from the training tokens of individual acoustic units. In this way, the most important variants can be uncovered directly from the

acoustic data. However, given the fact that speech tokens are time series data with different length, it is not possible to directly apply the traditional distance measure for fixed dimensional vectors to speech tokens clustering. One way to overcome this problem is to map each speech token to fixed length vector sequences. In [5], a single HMM for a certain acoustic unit is first trained using all the training tokens. Each training token is then converted into a vector sequence with the fixed length equal to the number of states in the resulting HMM by averaging the corresponding distances over all the frames assigned to the same HMM state. In [6], the different length speech tokens are mapped to fixed-length vector sequences by using linear time scaling. With the fixed length vector sequences, traditional clustering procedures, such as K-means and the LBG algorithm, can be easily applied. Nevertheless, in these methods the length of the vector sequences must be priori determined, and normally the most appropriate length is hard to decide. Furthermore, the mapping of speech tokens may sacrifice relevant acoustic detail, especially with respect to the temporal evolution of speech pattern. An alternative distance measure is based on dynamic time wrapping (DTW) [3]. However, all these methods effectively assume statistical independence between subsequent frames of a speech token.

A more appropriate way for clustering speech tokens is to develop a distance measure that is defined on continuous trajectories, instead of what is essentially a bag of acoustic observation vectors. By doing so, the prototypes of clusters can be modeled as continuous and smooth curves rather than sequences comprising different numbers of frame vectors. This would allow measuring the distances between tokens with different length and the cluster prototypes without any length conversion. One suitable representation of such a continuous curve is a polynomial. The use of polynomial regression for the purpose of speech recognition has been proposed in [7] [8] [9] [10] [11]. In [7] [8], parameter trajectory models were built, in which mixture of polynomial regressions were used as the templates of acoustic units. In [9] [10] [11], trended hidden Markov models have been proposed, where the state-dependent polynomial regression functions were embedded in conventional HMMs. These approaches have been successfully applied to phone classification and word spotting, and offer an improved performance over conventional HMMs. In this work, we further extended the application of mixture of polynomial regressions in clustering speech trajectories.

The major contribution of this paper is to introduce a novel data-driven method to cluster training tokens, namely *Trajectory Clustering* (TC). In this approach, the training tokens belonging to the same acoustic unit are represented in terms of continuous trajectories in acoustic parameter space along time. The speech trajectories are then clustered into a number of classes using the Mixture of Polynomial Regression framework [12]. Based on the results obtained from TC, multiple HMMs in a parallel topology

can be trained for the acoustic unit. In order to evaluate the performance of the proposed approach, we compared the performance of the proposed TC based models (TCHMMs) with knowledge-based models in connected digit recognition.

This paper is organized as follows: Section II introduces the mathematics needed to define distances on continuous trajectories of arbitrary length, together with the overall clustering strategy that we used in our experiments. Section III describes the design and the results of the experiments in which we compared TCHMMs for connected digit recognition with the knowledge based Head-Body-Tail (HBT) model approach [13]. Section IV discusses our results and the possibilities of further improvements. Finally, in Section V, our main conclusions are drawn.

## II. METHODOLOGY

### A. Speech as Trajectories

After front-end feature extraction, a training token for a particular acoustic unit is represented by a sequence of  $N$  acoustic vectors,  $\mathbf{Y} = \mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_{N-1}$ . Each of these vectors reflects the short-time speech spectrum during a time interval of typically 10 ms. Thus, a training token  $\mathbf{Y}$  can be considered as a trajectory, which is a function of an independent variable  $x$  corresponding to time, with the response variable  $\mathbf{y}$  corresponding to points in acoustic feature space over time. An obvious way to represent time is to use the sequence number of the acoustic vectors, such that  $(x = 0, 1, \dots, N - 1) \rightarrow (\mathbf{y} = \mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_{N-1})$ . Fig. 2(a) shows an example of speech trajectories in this representation for five different tokens belonging to the same acoustic unit. The  $x$ -axis represents time and the  $y$ -axis is the first MFCC coefficient. The time intervals between successive acoustic vectors are equal for all speech trajectories.

The use of sequence numbers of acoustic vectors is not the only possibility to represent time for speech trajectories. One may consider a speech unit as a continuous articulatory gesture from an initial to a final position. Realizations of the same speech unit should have similar initial and final positions, independent of the speed with which they are produced. Therefore, it is fair to consider the first and last acoustic vectors of  $\mathbf{Y}$  as corresponding to the initial and final articulator positions, respectively. The vectors in between can be regarded as the sampled measurements of the movement of the articulators with a uniform sampling interval. From this point of view, a slow token has a higher sampling rate than a fast token, due to the fixed time interval between the samples in the front-end feature extraction. Based on this idea, an alternative representation of speech trajectories is illustrated in Fig. 2(b) where the same tokens are shown as in Fig. 2(a). The  $x$ -axis represents sequence numbers mapped onto a fixed interval  $[0, 1]$ . Thus, the

relation between  $x$  and  $y$  can be presented as  $(x = \frac{0}{N-1}, \frac{1}{N-1}, \dots, \frac{N-1}{N-1}) \rightarrow (y = \mathbf{y}_0, \mathbf{y}_1, \dots, \mathbf{y}_{N-1})$ . In this representation, the time intervals between acoustic vectors within one speech trajectory are uniform, but they may differ between different speech trajectories.

In both forms of trajectory representations, the spectral measurements of a token  $\mathbf{Y}$  remain unchanged. The only difference is the method of mapping time to acoustic vectors. The trajectory clustering algorithm that we introduce in this paper can be applied successfully to either representation, but they might give rise to different clustering results. In Fig. 2(a), the separation of long and short speech trajectories is clearly visible. However, in Fig. 2(b) they seem to be differences between the slopes of the curves. Thus, it may well be that the difference in duration (number of frames) that dominates the first representation may mask the different dynamics underlying the continuous movement of the articulators. Because we expect that dynamics is more important than duration, we decided to use the method for representing time exemplified in Fig. 2(b) in this paper.

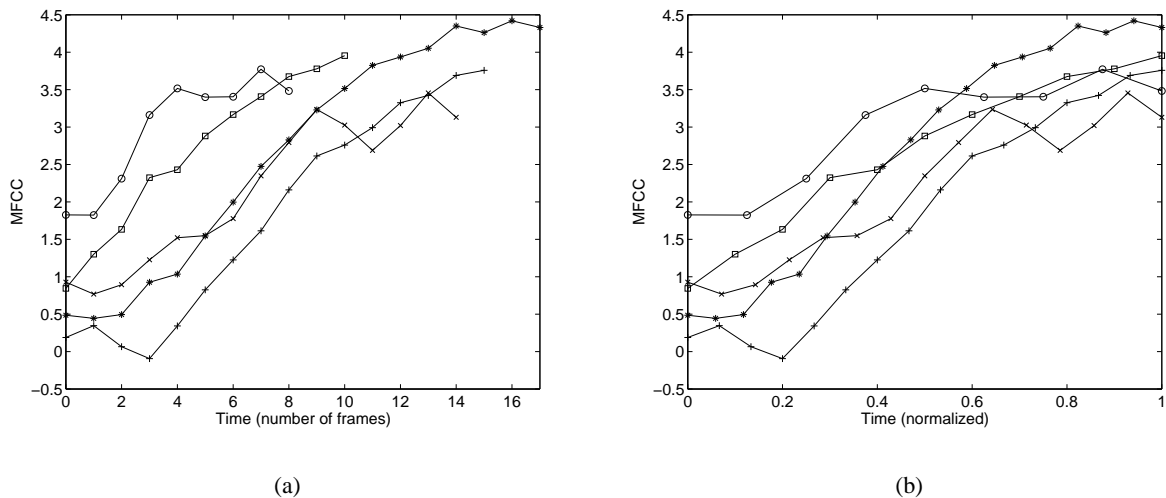


Fig. 2. Two representations of speech trajectories.

### B. Polynomial Regression of Speech Trajectories

Any time series can be approximated by a polynomial function  $g(x) = \beta_0 + \beta_1x + \beta_2x^2 + \dots + \beta_px^p$ . Thus, the standard regression relationship between data  $y$  and time  $x$  can be described as  $y = g(x) + e$ , in which  $e$  represents the residue between the actual data  $y$  and the regression approximation  $g(x)$ .

Analogously, for a speech trajectory  $j$  with a length of  $N_j$  acoustic vectors representing  $D$ -dimensional

features, the regression equation can be written in matrix form as

$$\mathbf{Y}_j = \mathbf{X}_j \beta + \mathbf{E}_j \quad (1)$$

shown in expanded form in equation (2).

$$\begin{bmatrix} y_j^1(0) & \dots & y_j^D(0) \\ y_j^1(1) & \dots & y_j^D(1) \\ \vdots & \ddots & \vdots \\ y_j^1(N_j-1) & \dots & y_j^D(N_j-1) \end{bmatrix} = \begin{bmatrix} 1 & \dots & (\frac{0}{N_j-1})^p \\ 1 & \dots & (\frac{1}{N_j-1})^p \\ \vdots & \ddots & \vdots \\ 1 & \dots & (\frac{N_j-1}{N_j-1})^p \end{bmatrix} \begin{bmatrix} \beta_0^1 & \dots & \beta_0^D \\ \beta_1^1 & \dots & \beta_1^D \\ \vdots & \ddots & \vdots \\ \beta_p^1 & \dots & \beta_p^D \end{bmatrix} + \begin{bmatrix} e_j^1(0) & \dots & e_j^D(0) \\ e_j^1(1) & \dots & e_j^D(1) \\ \vdots & \ddots & \vdots \\ e_j^1(N_j-1) & \dots & e_j^D(N_j-1) \end{bmatrix} \quad (2)$$

Here,  $\mathbf{Y}_j$  is the acoustic feature matrix, which is  $N_j \times D$ .  $\mathbf{X}_j$  is an  $N_j \times (p+1)$  matrix whose second column contains the normalized sequence numbers corresponding to the acoustic vectors in  $\mathbf{Y}_j$ ;  $\beta$  is a matrix of regression coefficients.  $\mathbf{E}_j$  is the  $N_j \times D$  residual error matrix, and  $p$  is the order of the polynomial regression model. In general, the residual error will become smaller as the order  $p$  of the polynomial regression increases.

Similar to the polynomial regression for a single speech trajectory, it is also possible to fit one polynomial to a set of trajectories. Suppose we have a set  $\mathfrak{S}$  of  $M$  speech trajectories; the regression equation can then be written as

$$\mathbf{Y} = \mathbf{X} \beta + \mathbf{E} \quad (3)$$

where  $\mathbf{Y} = [\mathbf{Y}'_1 \dots \mathbf{Y}'_M]'$  and  $\mathbf{X} = [\mathbf{X}'_1 \dots \mathbf{X}'_M]'$ , so that  $\mathbf{Y}$  contains all the acoustic vectors of all the speech trajectories, one after another, corresponding to the time contained in  $\mathbf{X}$ .  $\mathbf{E}$  is the overall residual error matrix. Since the articulator movements can be regarded as a physical process, it is reasonable to model the residual errors in terms of a Gaussian density. Thus, we assume that  $\mathbf{E}$  is a zero-mean Gaussian with a time-invariant covariance matrix  $\Sigma$ .

Now we are ready to define a probabilistic model for a data set  $\mathfrak{S}$  comprising  $M$  speech trajectories. Suppose we have a trajectory  $j$  ( $1 \leq j \leq M$ ) with  $N_j$  acoustic vectors. The probability density function of observing the  $i$ th ( $0 \leq i \leq N_j-1$ ) acoustic vector  $\mathbf{y}_j(i)$ , given  $x_j(i)$ , is defined as  $f(\mathbf{y}_j(i)|x_j(i), \theta)$ , where  $\theta$  represents the model parameters, including both the regression coefficients  $\beta$  and the covariance matrix  $\Sigma$ , or symbolically  $\theta = \{\beta, \Sigma\}$ . With the standard regression assumption that the error is conditionally independent at different  $x$  along the trajectory, the density of the complete trajectory can then be defined as

$$P(\mathbf{Y}_j|\mathbf{X}_j, \theta) = \prod_{i=0}^{N_j-1} f(\mathbf{y}_j(i)|x_j(i), \theta) \quad (4)$$

Assuming conditional independence between individual speech trajectories, which is reasonable because one speech realization has no direct impact on any other, the probability density function of all speech trajectories is the full joint density of the individuals:

$$P(\mathbf{Y}|\mathbf{X}, \theta) = \prod_{j=1}^M \prod_{i=0}^{N_j-1} f(\mathbf{y}_j(i)|x_j(i), \theta) \quad (5)$$

The log-likelihood of the parameters  $\theta$  given the speech trajectory set  $\mathfrak{S}$  can be defined directly from Eq. (5)

$$L(\theta|\mathfrak{S}) = \sum_{j=1}^M \log \prod_{i=0}^{N_j-1} f(\mathbf{y}_j(i)|x_j(i), \theta) \quad (6)$$

The model parameters  $\theta = \{\beta, \Sigma\}$  can be estimated straightforwardly by maximizing the log-likelihood (Eq. (6)). In fact, the solution for  $\beta$  and  $\Sigma$  is obtained from the well known Least Squares Fit:

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (7)$$

$$\hat{\Sigma} = \frac{(\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})}{\sum_{j=1}^M N_j} \quad (8)$$

### C. Mixture of Polynomial Regression

Fig. 3 illustrates the regression result with a cubic polynomial for a set of five speech trajectories. The bold solid line represents the polynomial with the ML estimate of the regression parameters. Conventionally, a cluster of data is defined as a mean or centroid in combination with a distance metric. For polynomial regression, the best fitting polynomial is defined as the centroid, and the distance metric is the probability that a speech trajectory is generated by the regression model. By virtue of this distance measure, a probabilistic clustering algorithm for speech trajectories can be developed based on the representation of trajectories as a mixture of polynomial regressions.

In the standard Mixture of Gaussians Model it is assumed that data is generated by a mixture of  $K$  Gaussian components. The density of the data  $y$  generated by the mixture model is a linear combination of the component densities  $P(y|\theta_k) = \sum_{k=1}^K \omega_k f_k(y|\theta_k)$ , where  $\omega_k$  is the weight of the  $k^{th}$  component with  $\sum_{k=1}^K \omega_k = 1$ , and  $f_k(y|\theta_k)$  is the density given that an  $y$  belongs to component  $k$ . Following the

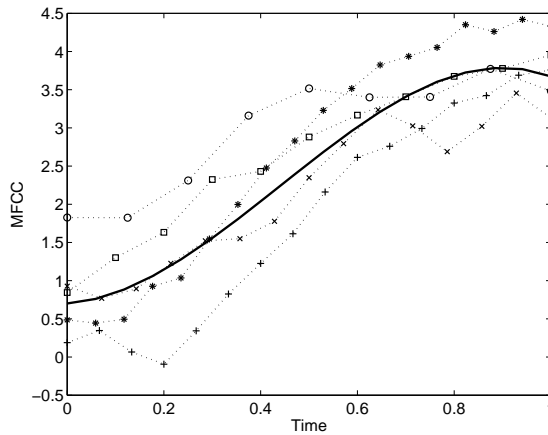


Fig. 3. Polynomial regression for a set of speech trajectories.

same idea, we can construct a mixture model for polynomial regressions. Assume that a speech trajectory can be modeled by a Gaussian mixture with  $K$  components, each of which is a regression model with polynomial mean and Gaussian residue. Then the density of a speech trajectory  $\mathbf{Y}_j$  generated by the mixture model is a linear combination of the component regression models (Eq. (4)), which can be written as

$$P(\mathbf{Y}_j|\mathbf{X}_j, \theta) = \sum_{k=1}^K \omega_k \prod_{i=0}^{N_j-1} f_k(\mathbf{y}_j(i)|x_j(i), \theta_k) \quad (9)$$

where the  $\omega_k$ 's are the weights of the components,  $f_k(\mathbf{y}_j(i)|x_j(i), \theta_k)$  is the density generated by trajectory  $\mathbf{Y}_j$  belongs to component  $k$ , and  $\theta_k = \{\beta_k, \Sigma_k\}$  are the model parameters for the  $k^{\text{th}}$  regression component. Similar to Eq.(6), the log-likelihood of the parameter  $\theta$  given the speech trajectory set  $\mathfrak{S}$  can be defined as

$$L(\theta|\mathfrak{S}) = \sum_{j=1}^M \log \sum_{k=1}^K \omega_k \prod_{i=0}^{N_j-1} f_k(\mathbf{y}_j(i)|x_j(i), \theta_k) \quad (10)$$

To find the maximum likelihood estimates of the parameters of a mixture model, Expectation Maximization (EM) [14] is the most general algorithm. The EM algorithm consists of two steps: the E-step calculates the expectation of Eq. (10) with respect to the current parameter set  $\theta^{t-1}$ , and the M-step maximizes the expectation to yield new parameters  $\theta^t$ . The expectation of  $L(\theta|\mathfrak{S})$  is

$$\begin{aligned}
E[L(\theta|\mathfrak{S})] &= \sum_{j=1}^M \sum_{k=1}^K h_{jk} \log \omega_k \\
&+ \sum_{j=1}^M \sum_{k=1}^K \sum_{i=0}^{N_j-1} h_{jk} \log f_k(\mathbf{y}_j(i)|x_j(i), \theta_k)
\end{aligned} \tag{11}$$

where

$$h_{jk} = \frac{\omega_k \prod_{i=0}^{N_j-1} f_k(\mathbf{y}_j(i)|x_j(i), \theta_k)}{\sum_{k=1}^K \omega_k \prod_{i=0}^{N_j-1} f_k(\mathbf{y}_j(i)|x_j(i), \theta_k)} \tag{12}$$

Here,  $h_{jk}$  can be thought of as the membership probability which is the posterior probability that trajectory  $\mathbf{Y}_j$  is generated by component  $k$ .

The solution for the regression parameter  $\hat{\beta}_k$ , the covariance  $\hat{\Sigma}_k$  and the mixture weights  $\hat{\omega}_k$  that maximize Eq.(11) can be obtained straightforwardly from the weighted least squares regression [15]:

$$\hat{\beta}_k = (\mathbf{X}'\mathbf{H}_k\mathbf{X})^{-1}\mathbf{X}'\mathbf{H}_k\mathbf{Y} \tag{13}$$

$$\hat{\Sigma}_k = \frac{(\mathbf{Y} - \mathbf{X}\hat{\beta}_k)'\mathbf{H}_k(\mathbf{Y} - \mathbf{X}\hat{\beta}_k)}{\sum_{j=1}^M N_j h_{jk}} \tag{14}$$

$$\hat{\omega}_k = \frac{1}{M} \sum_{j=1}^M h_{jk} \tag{15}$$

In the equations above,  $\mathbf{H}_k$  is a diagonal matrix with  $[\mathbf{h}_{1k}^* \quad \mathbf{h}_{2k}^* \quad \dots \quad \mathbf{h}_{Mk}^*]$  as the diagonal, where  $\mathbf{h}_{jk}^*$  is a row vector containing  $N_j$  copies of the membership probability  $h_{jk}$ . By default, the EM algorithm starts with randomly initialized model parameters. Then the E-step and M-step are iteratively performed until convergence on Eq.(10) is reached. Finally, each speech trajectory is assigned to the cluster with the highest membership probability  $h_{jk}$ . Fig. 4 illustrates the trace of the EM algorithm as applied to splitting the five different speech trajectories in Fig. 2(a) into two clusters at various iterations.

#### D. Speech Trajectory Clustering with Successive Split

One of the issues with the EM algorithm is how to guess the initial values for the model parameters. In a series of initial experiments with Trajectory Clustering, we initialized model parameters by randomly assigning speech trajectories in our training database to one of  $K$  clusters. The clusters we obtained

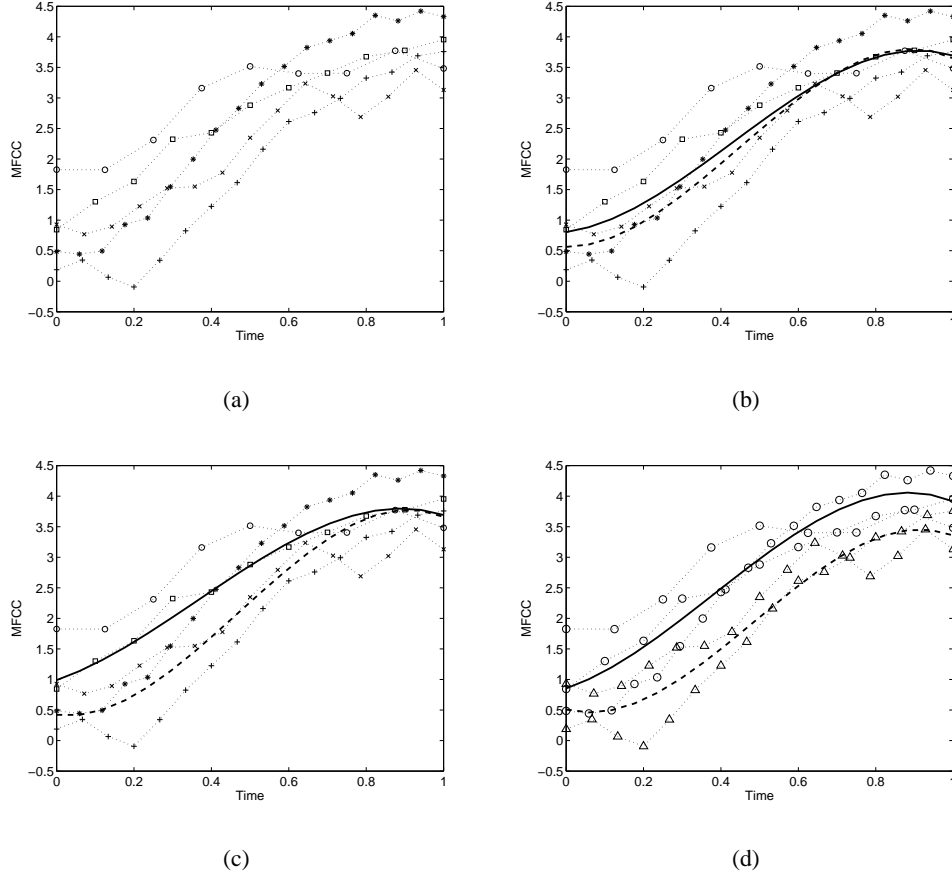


Fig. 4. The trace of the EM algorithm as applied to splitting five speech trajectories into two clusters at various iterations. (a) The original speech trajectories; (b) The initial location of the polynomials; (c) The location of the polynomials after the first iteration; (d) The location of the polynomials after the convergence of EM and the resulting cluster assignment.

indicated that the parameter estimation procedure is highly sensitive to the initial cluster assignments: Different initial assignments of speech trajectories led to different clusters after EM estimation. We solved the problem by means of a procedure in which the number of clusters is increased incrementally. To that end, we start by computing the best fitting polynomial function for the complete data set. Then, the polynomial function is split by adding and subtracting a fraction of the estimated standard deviation from all the mean values. The resulting polynomial functions are then used as the initial values of the parameters of the model with two clusters. The splitting is iteratively applied in the cluster with the largest  $w_k$  until  $K$  clusters are obtained. In all our experiments with TC that we have conducted so far, we have found that the component with the largest  $w_k$  is always related to the cluster with the largest number of trajectories. Thus, it appears that taking cluster size as the criterion for selecting the cluster

to be split would have yielded identical results.

### III. EXPERIMENTS

#### A. *Speech Material*

The performance of the proposed TC based models was evaluated by applying the approach to a connected Dutch digit recognition task. The speech material for our experiments was taken from the Dutch POLYPHONE [16], SESP [17] and CASIMIR corpora [18]. For each of the corpora, speech was recorded over the public switched telephone network in the Netherlands. Among other things, the speakers were asked to read several connected digit strings. The number of digits in a string varied from 1 to 14. For training we used a set of 9,753 strings containing 61,592 digits. All models were evaluated with an independent set of 10,000 test utterances comprising 80,016 digits. None of the original utterances used for training or testing had a high background noise level.

We computed 12 Mel-frequency log-energy coefficients using a 25 ms Hamming window shifted with 10 ms steps and a pre-emphasis factor of 0.98. Based on a Fast Fourier Transform, 12 filter band energy values were calculated, with the filter bands triangularly shaped and uniformly distributed on a Mel-frequency scale. Mel-frequency cepstra were computed from the raw Mel-frequency log-energy coefficients using the DCT. Channel normalization was done by means of cepstrum mean subtraction over the entire utterance. Finally, we computed the first and second order time derivatives. Together with log-energy and first and second order delta log-energy we obtained 39 dimensional feature vectors.

#### B. *Experimental Design*

In our experiments we used Head-Body-Tail (HBT) [13] models as the baseline system. HBT models account for pronunciation variation in a knowledge based manner. Because context induced pronunciation variation at the boundaries of a digit is much larger than in the middle, each digit is split up into three parts. The middle part of a digit (the Body) is assumed to be context-independent. The first part (the Head) and the last part (the Tail) are dependent on the previous and subsequent digit (or silence), respectively. Thus, each digit is modeled as one context-independent body HMM and 11 context-dependent head and tail HMMs that can be conflated in models with 11 parallel paths. In all our experiments the head and tail HMMs consisted of three states, whereas the number of states in body models was based on the mean duration of the digit as observed in the train corpus [18]. In addition to digit models, one silence and one noise model, both consisting of three states, were built. All the HMM paths have the standard left-to-right no-skip topology.

To build TC-based multiple-HMMs models, we made use of the Head, Body and Tail parts of the ten digits as basic acoustic units, and applied Trajectory Clustering to the parts of the training tokens associated to each of these units. The segmentation of the training tokens was obtained from the baseline HBT models by means of forced alignment. In clustering, we adopted polynomial regression with the order  $p = 3$ . Experiments with regression models with order  $p > 3$  did not improve the performance, because the regression parameters  $\beta$  of the terms with higher order were not significantly different from zero. In addition, the average length of the training tokens for all Head and Tail units was approximately equal; thus, it is reasonable to use the same order of polynomial regression for all units. The results obtained from Trajectory Clustering were used to train the separate paths of TC-based multiple-HMMs models.

We conducted two experiments to compare TC-based multiple-HMMs with knowledge-based multiple-HMMs. In the first experiment we compared knowledge-based and TC-derived parallel paths to the Body parts of the digits, i.e., the parts where we do not expect a large degree of context-induced variation. In the second experiment we compared knowledge-based and TC-derived parallel paths to the Head and Tail parts, where we do expect a substantial amount of context-induced variation.

In the first experiment, we compared the performance of knowledge-based models and the TC-based models by replacing the single HMM Body model by four path multiple-HMMs. To that end we classified the tokens belonging to each Body unit into four classes, based on the combination of gender and duration. The median duration of the tokens was taken as the threshold value to create long and short duration clusters. The resulting token subsets were then used to train gender and duration dependent multiple-HMMs with four parallel paths. The number of states in the paths was determined by the shortest training token in each cluster. The tokens of the Body units were also split into four groups with Trajectory Clustering. Considering that the dependence between frames is explicitly modeled in TC, we only used the 12 MFCCs as the acoustic feature vector in clustering. Based on the resulting clusters, TC-based models with four parallel paths for the Body units were trained, with the full 39 parameter acoustic vectors. Again, the number of states in each path was determined by the duration of the shortest token in each cluster. Both the knowledge-based and the TC-based multiple-HMMs for the Bodies were combined with the baseline Head and Tail models.

In the second experiment, we evaluated the performance of the TC-based multiple-HMMs in modeling the pronunciation variation in the Head and Tail parts of the digits. Compared to the baseline HBT models, the TC-based models had exactly same model topologies with 11 parallel HMM paths in Head and Tail parts and single HMM path in Body part. The only difference is that in training the parallel

paths, the baseline models used 11 subsets of training tokens based on linguistic context, whereas the TC-based models used 11 subsets obtained with Trajectory Clustering.

All the models in these experiments were trained and evaluated with HTK [19]. In order to study the improvements due to changes in acoustic modeling only, without the risk that the language model could mask the effects, we used a language model that only specifies that all digits have equal prior probability, and that each digit can follow each of the ten digits and silence with equal prior probability. We also gave the parallel paths in the multiple-HMMs equal prior probability.

### C. Experimental Results

Fig. 5 illustrates the recognition performance of gender and duration dependent models and four-paths TC-based models in the Body units of HBT. In this figure, the WER is presented as a function of the total number of Gaussian mixtures in the models. In each curve, the points correspond to 1, 2, 4, 8, 16, 32 Gaussians in one HMM state. The vertical bars represent the 95% confidence interval of the measurements. From Fig. 5, it can be seen that the TC-based models result in a slightly better recognition accuracy, with a smaller number of model parameters than in the knowledge-based models. This is because of the lower number of states in the TC-based models, due to the fact that we set the number of states equal to the number of frames in the shortest token in a cluster. In most of the cases the TC-based clusters comprised tokens that were shorter than the median length of the corresponding Body part. These results confirm that without using any priori knowledge, the proposed TC based model is capable to discover the intrinsic variation in the training speech and to account for the underlying structure of speech dynamics.

The comparison of the recognition performance of TC-based models for the Head and Tail parts with the baseline context-dependent HBT models is shown in Fig. 6. From this figure, we see that by applying TC, we obtain a minimum word error rate of 1.60%, compared to a minimum word error rate of 1.77% for the HBT model. This improvement is significant at the 95% significance level. Moreover, it can also be seen from Fig. 6 that with only one Gaussian in each HMM state, the TC-based models outperform the HBT models by a large margin. Apparently, the TC-based clustering succeeds in disclosing variation in the training database that is more important for recognition performance than the classes formed on the basis of linguistic criteria.

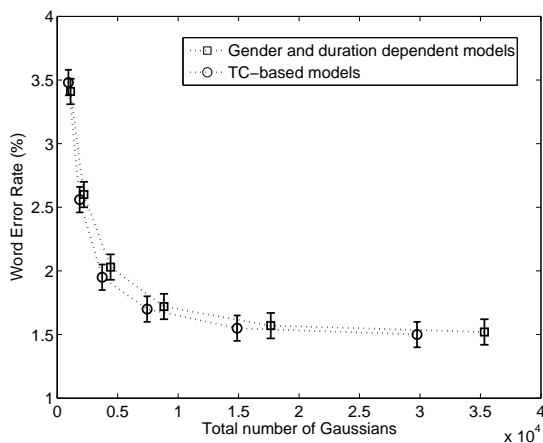


Fig. 5. Digit recognition results comparing the gender and duration dependent models and the TC-based models for the Body parts.

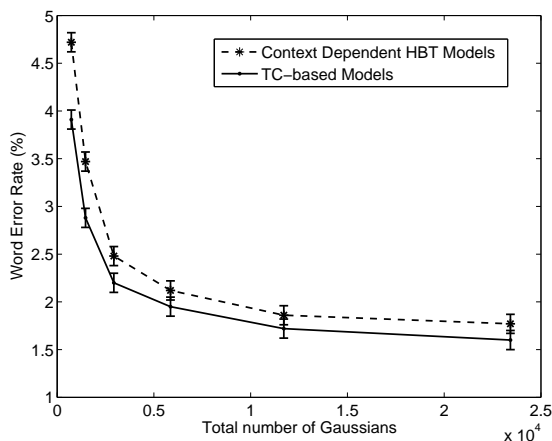


Fig. 6. Digit recognition results comparing the HBT and the TC-based models for the Head and Tail parts.

#### IV. DISCUSSION AND FUTURE WORK

The experiments showed that the performance of TC-based HBT models was at least equal to the knowledge-based models. In order to interpret the cause of the improvement, we had a closer look at the clustering results of TC by comparing the tokens collected in each cluster with the corresponding knowledge-based classification. Table I illustrates the correspondence between results obtained from the knowledge-based classification and TC for the Body unit of digit /nul/ (*zero*). The row categories in the table represent the knowledge-based classification with respect to the combination of gender and

TABLE I  
CORRESPONDENCE BETWEEN KNOWLEDGE-BASED CLASSIFICATION AND TRAJECTORY CLUSTERS FOR THE BODY UNIT OF  
THE DIGIT /NUL/.

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
female-short	<b>417</b>	<b>737</b>	82	40
female-long	<b>602</b>	<b>1120</b>	23	44
male-short	53	32	<b>972</b>	<b>452</b>
male-long	95	51	<b>572</b>	<b>994</b>

duration, and the column categories represent the trajectory clusters. The figures show the number of tokens coexisting in the corresponding row and column categories.

From Table I it can be seen that the results of TC clearly reflect the gender split of the training tokens. However, TC does not reproduce the classification with respect to duration very well. It seems that other as yet unknown factors are also important and intervene with duration. So far, we have not been able to identify linguistic or phonetic features that correspond to the TC-clusters. Moreover, the fact that duration seems to behave differently in the male and female parts of the training database suggests that the role of a given knowledge-based feature may well be dependent on other concomitant features. This makes it harder for knowledge-based classification to produce the actual separation of tokens supported by the data.

Fig. 7 demonstrates the correspondence between results obtained from the classification based on linguistic context and from TC for the Tail unit of the digit /nul/. In this figure, the left and right panels correspond to the female and male part of the training database, respectively. We show the two panels because it was evident that gender was the single most important feature in TC-clustering. This outcome shows that the assumption made in HBT that the linguistic context is the most important source of variation is probably not true. The 11 rows correspond to the 11 different linguistic contexts, and the 11 columns correspond to the 11 TC clusters. Note that the TC clusters with identical labels in the two panels are indeed identical. The proportion of tokens shared by one knowledge-based cluster and one TC cluster is reflected by the degree of blackness of the cells, as indicated in the leftmost column in both panels. It can be seen that the TC clusters reflect some variation related to linguistic context. For example, the Tails of the digit /nul/ followed by the digit /een/ (*one*) are likely to be clustered into C1 (for both males and females). A large number of training tokens followed by the digits /vier/ (*four*) and /vijf/ (*five*) are clustered into C3 for female speakers and C10 for male speakers. The tokens followed by

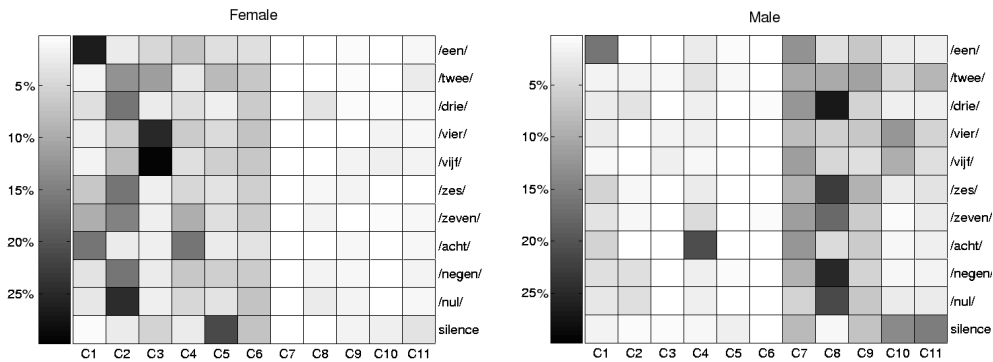


Fig. 7. Correspondence between knowledge-based classification and trajectory clusters for the Tail unit of digit /nul/.

a digit starting with a vowel tend to be clustered into other groups than the tokens followed by a digit starting with a consonant.

To investigate the difference between multi-path HMMs for the tail part of digit /nul/ trained by the token clusters obtained from linguistic context and TC, a  $22 \times 22$  distance matrix was computed, based on the Kullback-Leibler (KL) distance between each pair of HMM paths among 11 context dependent paths and 11 TC-based paths. To present these distances, a multidimensional scaling (MDS) analysis was carried out to reduce the  $22 \times 22$  matrix to a 2-dimensional representation, which is shown in Fig. 8. A small distance between two points in this figure is equivalent to a small KL distance between two HMM paths, which implies that the acoustic observations modeled by both paths have similar location and dispersion in acoustic space. The figure shows a high degree of similarity between the tail part of /nul/ followed by the /v/ of /vier/ and /vijf/. The same holds for the /z/ of /zes/ and /zeven/, and for /n/ of /negen/ and /nul/. It can be assumed that in the case of connected digits coarticulation across word boundaries does not extend far beyond the first phoneme of the following word. But both for /vier/ and /vijf/, and for /zes/ and /zeven/ also the vowels are somewhat similar, which make coarticulation even more similar. Thus, it is obvious that the identity of the following digit is not the best criterion to define clusters, precisely because of this similarity. The large distances between TC-based paths suggest that data driven clustering yields models that differ more between each other than when knowledge-based clustering was applied in the HBT.

The comparison of knowledge-based and TC-based clusters strongly suggests that the most important factors that determine the difference between internally homogeneous clusters of speech data cannot always be reproduced by linguistic criteria. The fact that the improvement in recognition performance

for the Body parts is smaller than for the Head and Tail parts suggests that the advantage of TC-based clustering is even larger in speech segments that are characterized by substantial dynamics. The fact that we do find this result also suggests that the capability to treat dynamically changing trajectories as units is extremely important. Trajectory Clustering provides this capability.

In this paper we applied the novel Trajectory Clustering approach for developing multi-path HMMs to a relatively simple task: connected digit recognition. Although TC-based clusters consistently outperformed knowledge-based clustering, it can be argued that we did not use all prior knowledge to the maximum possible extent in the experiments with Head and Tail models, where we might have integrated context information in the language model by requiring that Tails for the context /nul/ can only be followed by the digit /nul/. Because we have not been able so far to identify similar information that might also be integrated in the language model with TC-based clusters, it might seem that in the end knowledge-based clusters are to be preferred. However, we have also applied the proposed TC approach to a large vocabulary continuous speech recognition task [20]. Here, we built parallel models for the 94 most frequent syllables in a read speech corpus, which together cover over 50% of the syllable tokens in that corpus, and triphone models to cover the remaining syllables. Despite this large coverage of the high frequency syllables we have not been able to define useful linguistic criteria for clustering syllable tokens. On the other hand, using models with three parallel paths obtained with TC clustering yielded a significant performance improvement compared to a triphone-only system with the same number of parameters. This result suggests that the approach can be generalized to large vocabulary continuous speech recognition and that the absence of symbolic labels to identify the individual paths for integration in higher level models does not annihilate the gain in acoustic modeling accuracy.

In our future work, we will consider two possibilities to improve TC-based multiple-HMM acoustic modeling. Firstly, we will reconsider the mapping from time to acoustic vectors. In Section II-A we discussed two types of representations. In this work we assumed that the start and end points of the trajectories coincide in acoustic space, but may be traversed with different speed. However, these assumptions may not be true, because of potential systematic differences related to articulatory gestures between (groups of) speakers that may be reflected in the segmentations of training tokens. Thus, we will assume that all the tokens have different start and end points in time and articulatory space, and automatically derive a shifting and a scaling coefficient from the speech data. Secondly, there is no strong evidence that the choice of  $w_k$  as the criterion to select the component to be split in building the clusters is optimal. In our experiments, the cluster with largest  $w_k$  always had the largest number of speech tokens. Always splitting the largest cluster had the additional advantage that it tends to result in clusters

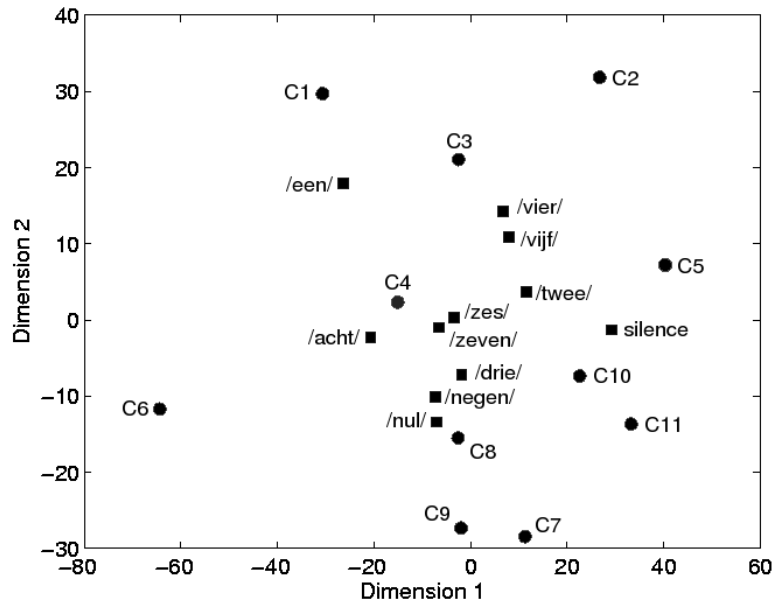


Fig. 8. The two-dimensional distance representation for the context-dependent and data-driven based HMM paths for the tail part of digit /nul/.

with approximately equal size. However, the largest cluster is not necessarily always the one with highest level of heterogeneities. Yet, intuitively it seems preferable to split heterogeneous clusters rather than big ones, if the latter are fairly homogeneous. Thus, we plan to investigate other criteria for selecting the cluster to be split, such as the determinant of the covariance matrix [21] and the distance between the estimated distribution and the empirical distribution of the data [22].

## V. CONCLUSIONS

This paper introduced a novel data-driven approach, namely Speech Trajectory Clustering, for building multiple-HMM acoustic models. In this method, we model each training token as a continuous trajectory along time in acoustic parameter space. The speech trajectories belonging to the same acoustic unit are clustered by means of a mixture of polynomial regression, where polynomial functions represent the means of clusters. The parameters of the polynomial functions and the residual covariance are estimated by using the EM algorithm. The resulting trajectory clusters are then used to train multiple-HMMs in a parallel topology for each acoustic unit.

To evaluate the performance of Speech Trajectory Clustering based multiple-HMM acoustic models, two experiment were carried out to compare their performance with knowledge based multiple-HMMs

in a connected digits recognition task. The results show that the performance of the TC-based models is at least equal to knowledge-based models and outperforms knowledge-based models in regions where the articulatory dynamics is highest. In addition, we found that Speech Trajectory Clustering models variation that cannot be captured by knowledge-based clusters. These findings suggest that the capability to treat dynamically changing trajectories as modeling units is very promising.

#### ACKNOWLEDGEMENTS

The research is part of the Interactive Multimodal Information eXtraction (IMIX) program, which is funded by the Netherlands Organization for Scientific Research (NWO).

#### REFERENCES

- [1] L. Rabiner and B. Juang, "An introduction to hidden markov models," *IEEE Acoust., Speech, Signal Processing Mag.*, pp. 4–16, Jan. 1986.
- [2] I. Illina and Y. Gong, "Elimination of trajectory folding phenomenon: HMM, Trajectory Mixture HMM and Mixture Stochastic Trajectory model," *In Proceedings of ICASSP-97*, vol. 2, pp. 1395–1398, 1997.
- [3] J. Picone, "Duration in context clustering for speech recognition," *Speech Communication*, vol. 9, pp. 119 – 128, 1990.
- [4] J. Su, H. Li, J. P. Haton, and K. T. Ng, "Speaker time-drifting adaptation using trajectory mixture hidden Markov models," *In Proceedings of ICASSP-1996*, vol. 2, pp. 709 – 712, 1996.
- [5] F. Korkmazskiy, "Generalized mixture of HMMs for continuous speech recognition," *In Proceedings of ICASSP97*, pp. 1443–1446, 1997.
- [6] Y. Gong, "Stochastic trajectory modelling and sentence searching for continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 5, no. 1, pp. 33 – 44, January 1997.
- [7] H. Gish and K. Ng, "Parametric trajectory models for speech recognition," *In Proceedings of ICASSP-93*, pp. 466–469, 1996.
- [8] —, "Segmental speech model with applications to word spotting," *In Proceedings of ICSLP-96*, vol. 2, pp. 447–450, 1993.
- [9] L. Deng, M. Aksmanovic, D. Sun, and C. F. J. Wu, "Speaker-independent phonetic classification using hidden Markov models with state-conditioned mixtures of trend functions," *IEEE Transactions on Speech and Audio Processing*, vol. 5, pp. 319–324, 1997.
- [10] L. Deng and M. Aksmanovic, "Speech recognition using hidden Markov models with polynomial regression functions as nonstationary states," *IEEE Transactions on Speech and Audio Processing*, vol. 2, pp. 507–520, 1994.
- [11] M. Aksmanovic and L. Deng, "Speech trajectory discrimination using the minimum classification error learning," *IEEE Transactions on Speech and Audio Processing*, vol. 6, pp. 505–515, 1998.
- [12] S. Gaffney and P. Smyth, "Trajectory clustering with mixtures of regression models," *In Proceedings of the fifth ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 63–72, 1999.
- [13] W. Chou, C. Lee, and B. Juang, "Minimum error rate training of inter-word context-dependent acoustic model units in speech recognition," *In Proceedings of ICSLP-94*, pp. 439–442, 1994.

- [14] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society*, vol. 39, pp. 1–38, 1977.
- [15] W. Cleveland and S. Devlin, "Locally weighted regression: An approach to regression analysis by local fitting," *Journal of the American Statistical Association*, vol. 83, pp. 596–610, 1998.
- [16] E. den Os, T. Boogaart, L. Boves, and E. Klabbers, "The Dutch Polyphone corpus," *In Proceedings of EuroSpeech-95*, pp. 825–828, 1995.
- [17] F. Bimbot, "An overview of the CAVE project research activities in speaker verification," *Speech Communication*, vol. 31, pp. 155–180, 2000.
- [18] J. Sturm and E. Sanders, "Modelling phonetic context using Head-Body-Tail acoustic models for connected digit recognition," *In Proceedings of ICSLP-2000*, vol. 1, pp. 429–432, 2000.
- [19] S. Young, G. Evermann, and T. Hain, *The HTK Book (for HTK version 3.2.1)*. Cambridge University Engineering Department, 1997.
- [20] Y. Han, A. Hämmäläinen, and L. Boves, "Trajectory clustering of syllable-length acoustic models for continuous speech recognition," *In Proceedings of ICASSP-2006*, vol. I, pp. 1169–1170, 2006.
- [21] A. Sankar, "Experiments with a Gaussian merging-splitting algorithm for HMM training for speech recognition," *In Proceedings of DARPA Speech Recognition Workshop 1998*, 1998.
- [22] N. Ueda and R. Nakano, "EM algorithm with split and merge operations for mixture models," *Systems and Computers in Japan*, vol. 32, pp. 1–11, 2000.