TRAJECTORY CLUSTERING FOR AUTOMATIC SPEECH RECOGNITION

Yan Han, Johan de Veth, and Louis Boves

Center for Language and Speech Technology,
Department of Language and Speech,
Radboud University, Nijmegen, The Netherlands
{Y.Han, J.deVeth, L.Boves}@let.ru.nl

ABSTRACT

In this paper, we present an approach for automatic clustering of multi-dimensional dynamic trajectories corresponding to speech data that is based on Trajectory Clustering (TC). TC uses the Expectation Maximization algorithm (EM) for clustering with the mixtures of Multiple Linear Regression model. Since the initial values of the model parameters are critical to the clustering performance, a successive splitting algorithm was developed to incrementally increase the number of clusters. We define multipath HMM topologies using the trajectory clusters found. Based on the hypothesis that pronunciation variation in speech is more systematic at a unit level that is longer than a phone, we used modelling units defined in terms of Head-Body-Tail (HBT) models for connected digit recognition for the Dutch language. It appears that multipath HMM topologies based on TC clusters outperform multi-path HMM topologies based on prior knowledge about speaker gender and speaking rate.

1. INTRODUCTION

Over the past two decades, hidden Markov models (HMMs) have been the dominant methodology for modelling speech acoustics in automatic speech recognition. It is known, however, that the high degree of articulatory variation that HMMs must account for gives rise to what is known as the trajectory folding phenomenon[1]. Under the first order Markov assumption that adjacent acoustic observation vectors are independent, there is little means to impose continuity constraints on acoustic observation vector sequences during recognition, beyond inclusion of delta and delta-delta coefficients in the observation vectors. As a result, an actual observation sequence can be recognized with high probability as a sequence of densities that is physically implausible, which, in turn, increases the probability of recognition errors.

To overcome the adverse effect of trajectory folding, previous research investigated multipath HMMs [2, 3, 4]. A multipath HMM approach can alleviate the effect of trajectory folding if it can be assumed that variations in acoustic observation sequences are systematic, and can be modelled separately. In a multipath HMM, observation sequences are explicitly disallowed to switch between parallel HMM paths. In [2] it was assumed that speaker gender and speed of articulation cause systematic variation. These features were used successfully to train independent paths in HMMs. However, such a heuristically inspired classification approach cannot discover systematic variation that is not related to meta-data (such as gender) or to characteristics of the speech that are easily established in the training database (e.g., speaking rate). For this reason, an automatic method for discovering systematic variation due to pronunciation differences directly from acoustic observation sequences is a promising approach to multipath HMM topology design. [3] demonstrated an automatic detection method for systematic variation in the context of continuous digit recognition, in which the training tokens were clustered based on a Dynamic Programming approach. In [4], training tokens were clustered based on forced alignment scores instead of directly using the acoustic feature vectors. Both of these automatic detection methods were shown to be successful for automatic digit recognition.

In our work, we are developing another automatic detection approach (called Trajectory Clustering based HMMs, TCHMMs) for developing multipath HMM topologies. In our approach, training tokens are clustered into distinct groups in terms of polynomial trajectories in acoustic feature space. In [5] the idea of polynomial feature trajectory has been applied to vowel classification, with limited success. The best result amounted to 66% correct classification of vowels in the TIMIT database. In [5], a mixture model for trajectories was trained that is similar to a conventional vector-based Gaussian mixture model. The vector-based Gaussian mixture model has been widely used for probabilistic clustering, and the mixture model for trajectories can also be extended in this manner, which leads to the Trajectory Clustering Model proposed in [6]. In this approach, a trajectory cluster is modelled as a prototype polynomial function, and the variability around the prototype is described as a mixture of Gaussians.

Most conventional HMM based systems for automatic speech recognition are based on a description of speech as a sequence of phones. This choice for short duration units imposes an unrealistic constraint, because only correlation of speech as short as 30-40 ms duration can be modeled. However, recent research shows that articulation variation in speech is more appropriately described at the level of the syllable than at the level of the phone [7]. Systematic variation with a typical time span as long as a syllable is difficult to describe in terms of a sequence of (context dependent) phones. For this reason, we investigate the use of Trajectory Clustering HMMs (TCHMMs) that are based on longer length modelling units. For this paper, we studied TCHMMs based on Head-Body-Tail (HBT) model units (cf. [8]) for a connected digit recognition task.

Trajectory clustering in a high dimensional feature space is far from trivial. Our first experiments indicated that clustering can be impeded by the large degree of overlap in some of the features, even if the classes are clearly distinct for other features. In this paper we describe the method we used to tackle this problem. Furthermore, we studied how well automatically derived cluster assignment correspond to a priori classification for male and female and long and short utterances. Finally, we compared the recognition results for TCHMMs to a priori knowledge based Multipath models.

This paper is further organized as follows. Section 2 describes the theoretical framework of the Trajectory Clustering Model. In Section 3, the experimental set-up is described that we used to arrive at different Multipath HMM topologies for connected Dutch digit recognition. Section 4 discusses our results. Finally, in Section 5, we summarise our main conclusions.

2. TRAJECTORY CLUSTERING MODEL

2.1 Mixture of Regression Model

The underlying idea of speech trajectory clustering is the Mixture of Regression Model [6]. In this model, the speech realizations are assumed to be drawn from several components of mixture Gaussians, where the mean of each component density is a ploynomial function of time. For speech realization j with the length of N_j, the matrix form of the regression equation for component k in D dimensional acoustic feature space can be written as

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

or:

$$\begin{bmatrix} y_j^{(d)}(1) \\ y_j^{(d)}(2) \\ \vdots \\ y_j^{(d)}(N_j) \end{bmatrix} = \begin{bmatrix} 1 & x_j(1) & \dots & x_j(1) \\ 1 & x_j(2) & \dots & x_j(2)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_j(N_j) & \dots & x_j(N_j)^p \end{bmatrix} \begin{bmatrix} \beta_{k,0}^{(d)} \\ \beta_{k,1}^{(d)} \\ \vdots \\ \beta_{k,p}^{(d)} \end{bmatrix} + \begin{bmatrix} e_k^{(d)} \\ e_k^{(d)} \\ \vdots \\ e_k^{(d)} \end{bmatrix}$$

$$\text{for} \quad d = 1, \dots, D$$

$\mathbf{Y}_j$ is the feature vector matrix, which is $N_j \times D$; $\mathbf{X}_j$ is an $N_j \times (p+1)$ matrix whose second column contains the frame numbers corresponding to the feature vector in $\mathbf{Y}_j$, and p is the highest order of the regression model, in our case $p = 3$; β_k is a matrix of regression coefficients; $\mathbf{E}_k$ is $N_j \times D$ residual error matrix which is assumed to be zero-mean multivariate Gaussian with covariance matrix $\mathbf{\Sigma}_k$.

With the standard regression assumption that the error is conditionally independent at different x points along the trajectory, the probability that a complete trajectory is generated by component k is:

$$P(\mathbf{y}_j | x_j, \theta_k) = \prod_i^{N_j} f_k(\mathbf{y}_j(i) | x_j(i), \theta_k) \quad (2)$$

Here, θ_k includes both the parameters of the regression model $g_k(x)$ and the error covariance matrix $\mathbf{e}_k$.

Once $P(\mathbf{y}_j | x_j, \theta_k)$ is computed for all K components, the membership probability h_{jk}, which corresponds to the posterior probability that trajectory $\mathbf{y}_j(i)$ is generated by component k, can be expressed as:

$$h_{jk} = \frac{w_k \prod_i^{N_j} f_k(\mathbf{y}_j(i) | x_j(i), \theta_k)}{\sum_k^K w_k \prod_i^{N_j} f_k(\mathbf{y}_j(i) | x_j(i), \theta_k)} \quad (3)$$

where w_k is the weight of the mixture densities. The trajectory will be assigned to the component yielding the highest membership probability.

With this notation, the reestimation equation for the EM algorithm can then be defined as:

$$\hat{\beta}_k = (\mathbf{X}' \mathbf{H}_k \mathbf{X})^{-1} \mathbf{X}' \mathbf{H}_k \mathbf{Y} \quad (4)$$

$$\hat{\mathbf{\Sigma}}_k = \frac{(\mathbf{Y} - \mathbf{X}\hat{\beta}_\mathbf{k})' \mathbf{H}_k (\mathbf{Y} - \mathbf{X}\hat{\beta}_\mathbf{k})}{\sum_j^M \mathbf{h}_{jk}^*} \quad (5)$$

$$\hat{w_k} = \frac{1}{M} \sum_j^M h_{jk}, \quad (6)$$

in which $\mathbf{Y} = [\mathbf{Y}_1' \quad \dots \quad \mathbf{Y}_M']'$, and $\mathbf{X} = [\mathbf{X}_1' \quad \dots \quad \mathbf{X}_M']'$, so that $\mathbf{Y}$ contains all the feature vectors of the data set, one realization after another, corresponding to the frame numbers in $\mathbf{X}$. $\mathbf{H}_k = \text{diag}([\mathbf{h}_{1k}^* \quad \dots \quad \mathbf{h}_{Mk}^*])$, where $\mathbf{h}_{jk}^*$ is a row vector consisting of N_j copies of the membership probability h_{jk}. The estimated parameters are then used to compute new values of h_{jk} for the next step in the iteration. This iterative reestimation procedure is repeated until convergence is reached.

2.2 EM Implementation: Successive Splitting

One of the issues for EM algorithm is how to compute the initial values for the model parameters. In a series of initial experiments with Trajectory Clustering, we initialized model parameters by randomly assigning observed speech trajectories to one of K clusters. The clusters we obtained indicated that the parameter estimation procedure is heavily dependent on the initial cluster assignments: Different initial assignments of speech trajectories lead to different clusters after EM estimation. Looking at the observed trajectories together with the clusters that were trained, we found that there is a high degree of overlap for some of the dimensions of the acoustic feature vectors, whereas for other dimensions distinct trajectory clusters were clearly visible. Apparently, the high degree of discrimination in some of the coordinates of the acoustic feature vectors can be masked by the high degree of overlap in other feature coordinates, to the extent that the trajectory clusters found after EM training can be dominated by the non-discriminant feature coordinates.

One way to address this issue could be to emphasize the highly discriminant feature coordinates (and de-emphasize the less discriminant feature coordinates) during clustering. However, such an approach would have to rely on prior knowledge about what feature coordinates are most discriminant. We sidestepped the problem of identifying the most discriminant feature coordinate by means of a procedure in which the number of clusters is increased incrementally. Therefore, 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. The newly obtained polynomial functions are then used to compute the initial values of the parameters of the model with two clusters. The splitting is iterated, each time splitting the cluster with the largest w_k, until K clusters are obtained.

3. METHOD AND MATERIAL

In order to test the performance of TCHMM, a number of comparison experiments were carried out on a connected Dutch digit recognition task. In what follows, the type of speech material, the method for feature extraction, and the design of the model topologies are briefly described.

3.1 Speech Material

The speech material for our experiments was taken from the Dutch POLYPHONE corpus, the Dutch SESP corpus and the Dutch CASIMIR corpus [9]. For each of the three corpora, speech was recorded over the public switched telephone network in the Netherlands. Speech signals were recorded from a primary rate ISDN telephone connection. Among other things, the speakers were asked to read several connected digit strings. The number of digits in a string varied between 1 to 14. For training we used a set of 9,753 strings containing 61,592 digits. Care was taken to balance the training material with respect to:

(1) an equal number of male and female speakers,
(2) an equal number of speakers from each of the 12 provinces in the Netherlands, and
(3) an equal number of tokens per digit.

All models were evaluated with an independent set of 10,000 test utterances comprising 80,016 digits. The independent test set was balanced according to the same criteria as the training material. None of the original utterances used for training or testing had a high background noise level.

3.2 Acoustic Features Extraction

We computed 16 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, 16 filter band energy values were calculated, with the filter bands triangularly shaped and uniformly distributed on a Mel-frequency scale. In addition, we also computed the total log-energy values for each frame. These signal processing steps were performed using HTK3.1. Next, 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 and added these to the 12 channel normalized Mel-frequency

cepstral coefficients. Together with log-energy and first and second order delta log-energy we obtained 39 dimensional transformed feature vectors.

3.3 Acoustic Model Topologies

In this work, we made use of Head-Body-Tail (HBT) models as the baseline system. In this approach each digit is split up into three parts. The middle part of the word – the body – is assumed to be context-independent. The first part – the head – and the last part – the tail – are dependent of the previous and subsequent digit (or silence), respectively. Thus, for each digit one context-independent body model and 11 context-dependent head and tail models were trained. The head and tail models consisted of three states, whereas the number of states in body models is based on the mean duration of the digit as observed in the train corpus. In addition to digit models, one silence and one noise model, both consisting of 3 states, were built. All the HMM paths have the standard left-to-right no-skip topology.

In this paper we investigate the feasibility of TCHMM for finding structure in dynamic speech data. Therefore, we constrained ourself to clustering the speech segments corresponding to the body parts of the digits in the training corpus. These segments were obtained by decoding the training material with the baseline version of the HBT recognizer. Then, for each cluster a separate Body model was trained. Merging the models for the clusters into one Body model with parallel paths yields what we call Multipath models. Other Multipath models were trained on the basis of information about speaker gender and articulation rate. In all experiments we followed the same procedure:

(1) bootstrap the transcription by forced alignment with the baseline HBT models;
(2) cluster the training tokens of the body units into groups;
(3) relabel the modelling units in the transcription according to the resulting clusters;
(4) define an independent HMM path for each cluster;
(5) set the numbers of states in parallel HMMs as the minimum duration, i.e. the number of frames, as observed in the resulting clusters.

To build Multipath models based on a priori knowledge, we examined the classification criteria with respect to the gender of the speakers, duration of the realizations and the combination of both. Many previous experiments with gender dependent models have shown significant improvement in recognition performance. This shows that gender is one of the major sources of variation in speech acoustics. Thus, we built gender dependent Multipath models with two equal length HMM paths. Articulation rate is known to be another important source of acoustic variation, because fast and slowly uttered words are likely to have different co-articulation patterns. In this work, the median of the duration distribution of the body unit was taken as the threshold value to divide the training tokens into short and long realizations. Duration dependent Multipath models with two different length HMM paths were trained based on this classification. Finally, we trained 4-path gender and duration dependent models, according to the combination of gender and duration classifications. The training tokens were first split into two groups with respect to the gender criterion, and each group was further split with respect to the duration criterion.

Rather than using the a priori knowledge of training material, TCHMM allows us to build Multipath models on the basis of a classification supported by acoustic data. Considering that the dependence between frames is explicitly modelled in trajectory clustering, we only used 12 MFCCs as the acoustic feature vector in TCHMM clustering. Since the Trajectory Clustering method used in this study requires that the number of clusters is set a priori, we formed up to 16 clusters for the 10 sets of Body segments. Thus, we obtained 16 sets of TCHMM Body models. The number of states in the separate paths for all kind of Multipath models range from 6 to 23.

Table 1: *Association between knowledge based classification and 2-path TCHMM for the Body unit of the Dutch digit /nul/.*

	Cluster 1	Cluster 2
female-short	1146	130
female-long	1732	57
male-short	91	1418
male-long	156	1556
total	3125	3161

Table 2: *Association between knowledge based classification and 4-path TCHMM for the Body unit of the digit /nul/.*

	Cluster 1	Cluster 2	Cluster 3	Cluster 4
female-short	417	737	82	40
female-long	602	1120	23	44
male-short	53	32	972	452
male-long	95	51	572	994
total	1167	1940	1649	1530

4. EXPERIMENTAL RESULTS

The first experiment is to evaluate the TCHMMs with reference to the a priori knowledge based models. Tables 1, 2 and 3 show the association between the a priori classification and the trajectory clustering for the body part of the digit /nul/ (*zero*). The column categories in the tables represent the a priori classification with respect to the combination of Gender and Duration criteria, and the row categories represent the clusters derived from Trajectory Clustering. Each cell shows the number of tokens coexisting in the corresponding row and column categories. ¿From Table 1 we see, that, when the trajectories of realizations are automatically divided into 2 clusters, the numbers of tokens in the two clusters are approximately equal, and the observed variability reflects the gender difference. From Table 2, the discriminability for gender is still visible in the 4-path TCHMM. The physical interpretation of the two-way division of the gender-based clusters is not yet completely evident. Thus, TCHMM has discovered structure in the training data that cannot be covered by a priori knowledge based classification. In Table 3, differences between Cluster 1 and Cluster 2 seem to suggest that there is a low degree of correlation with long and short realizations of female utterances. Moreover, the tokens in Cluster 7 and Cluster 8 show that the variability present in some female-long and male-long realizations are considered as important by Trajectory Clustering. The similar associations between priori classification and trajectory clustering are found in other digits beside /nul/.

The comparison of the recognition performance of TCHMM and the a priori knowledge based models is shown in Fig. 1. In this figure, the Word Error Rates are presented as a function of the total number of Gaussian mixtures per system. From Fig. 1, we see that for the digit recognition task all Multipath models perform substantially better than the baseline model. The 2-path TCHMM performs as good as 2-path gender dependent models, which is predictable from the high association of these subsets of training tokens. Substantial improvement of WER is found for both 4-path TCHMM and 4-path gender and duration dependent model compared to 2-path gender dependent model. The similarity of the recognition performance between these 4-path models gives us evidence that the variability uncovered by trajectory clustering is as important as duration variability in this speech data set. The lowest WER found in 4-path TCHMMs has a relative reduction of 16.8% over the baseline model.

Fig. 2 shows the recognition performance of TCHMMs with

Table 3: *Association between knowledge based classification and 8-path TCHMM for the Body unit of the digit /nul/.*

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8
female-short	293	492	321	34	18	40	8	70
female-long	463	311	304	7	12	6	40	646
male-short	30	27	50	542	275	524	61	0
male-long	66	21	88	282	461	244	541	9
total	852	851	759	865	766	814	650	725

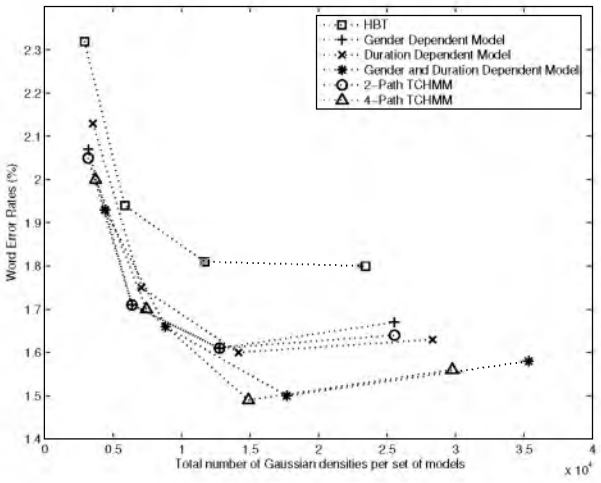

Figure 1: *Digit recognition results comparing TCHMM derived and a priori knowledge based Multipath models.*

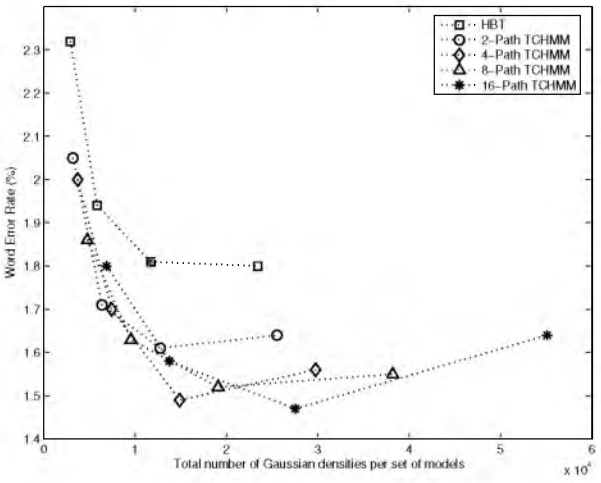

Figure 2: *Digit recognition results for TCHMMs with different number of separate paths.*

different number of separate paths. From Fig. 2, we see that no substantial improvement on WER could be observed if the number of Gaussian mixtures exceed 15,000 for baseline HBT model and for TCHMMs with up to 8 paths. Most probably, lack of training data for the separate HMM paths causes this under-training. The substantial improvement of performance is found among the models with 4 Gaussian mixtures per state as the number of separate paths increases. This shows that Trajectory Clustering is effective in uncovering relevant variability in speech data and results in the sharper models. However, when the number of mixtures increased, under-training occurs due to the small size of the trajectory clusters. 16-path TCHMM with 16 Gaussian mixtures gives the lowest WER of 1.47% . But, the difference on WER compared to the 4-path TCHMM with 16 Gaussians is not significant.

5. CONCLUSIONS

In this paper, a new approach was investigated for automatic clustering of the training tokens in a connected digit database to define Multipath HMM topologies. We introduced a probabilistic mixture regression model for speech observation sequences, and showed how an incremental cluster splitting strategy sidestepped the initialization sensitivity problem in EM framework. A number of experiments were carried out on a connected digit recognition task for Dutch. The best recognition result presented in this work was obtained in 16-path TCHMM with 16 Gaussian mixtures per state. This indicates that TCHMM is a very effective method to identify pronunciation variation.

Future research will be aimed at improving TCHMM performance by using the derivatives of MFCCs during clustering. Furthermore, we will apply TCHMM in the Head and Tail parts of HBT model, where the contextual variability is dominant.

REFERENCES

[1] I. Irina and Y. Gong, "Elimination of Trajectory Folding Phenomenon: HMM, Trajectory Mixture HMM and Mixture Stochastic Trajectory Modle", *Proc. ICASSP-97*, pp. 1395–1398, 1997

[2] O. Scharenborg, A.G.G. Bouwman and L. Boves, "Connected Digit Recognition with Class Specific Word Models", *Workshop on Voice Operated Telecom Services*, Ghent, Belgium, pp. 71-74, 2000.

[3] J.Picone, "Duration in Context Clustering for Speech Recognition", *Speech Communication*, Vol.9, pp. 119-128, 1990.

[4] F.Korkmazskiy, "Generalized Mixture of HMMs for Continuous Speech Recognition", *Proc. ICASSP-97*, pp. 1443–1446, 1997.

[5] H. Gish and Kenny Ng, "Parametric trajectory models for speech recognition", *Proc. ICSLP-96*, pp. 466–469, 1996.

[6] S. Gaffney and P. Smyth, "Trajectory clustering with mixtures of regression models", *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 63-72, 1999.

[7] S. Greenberg, "Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation", *Speech Communication*, Volume 29, pp. 159-176, 1999.

[8] Chou, W., Lee,C.-H., and Juang, B.-H., "Minimum error rate training of inter-word context-dependent acoustic model units in speech recognition," *Proc. ICSLP-94*, pp. 439-442, 1994.

[9] J. Sturm and E. Sanders, "Modelling Phonetic Context using Head-Body-Tail Acoustic Models for Connected Digit Recognition", *Proc. ICSLP-2000*, pp. I-429-432, 2000