Syllable-Length Path Mixture Hidden Markov Models with Trajectory Clustering for Continuous Speech Recognition

Yan Han and Lou Boves

Center for Language and Speech Technology (CLST),
Radboud University Nijmegen, The Netherlands
{Y.Han, L.Boves}@let.ru.nl

Abstract

Recent research suggests that modeling coarticulation in speech is more appropriate at the syllable level. However, due to a number of additional factors that can affect the way syllables are articulated, creating multiple acoustic models per syllable might be necessary. Our previous research on longer-length multi-path models has proved that data-driven trajectory clustering to be an attractive approach to derive multi-path models. However, the use of single distribution with unvarying covariance to model a trajectory cluster may degrade its capability of detecting pronunciation variants. In this paper, we propose a new method, namely path mixture hidden Markov model, to alleviate the adverse effects of trajectory clustering. The improvement on performance observed in continuous speech recognition experiments show path mixture model is a very effective approach.

Index Terms: continuous speech recognition, trajectory clustering, syllable-length model, path mixture HMM.

1. Introduction

Coarticulation introduces long-term spectral and temporal dependencies in speech. To model these dependencies in ASR, the use of longer-length acoustic models, based e.g. on syllables, has been proposed. However, most languages have several thousand syllables, and many of these syllables, corresponding to words which are not frequently used, will have poor coverage in the training data. As a consequence, several authors have proposed the mixed-unit model, which mixing syllable models for frequent syllables with conventional triphone models or bootstrapping longer length units from the sequence of constituent triphones [1] – [5].

However, it is unlikely that long-term coarticulation is the only, or even the most important, source of variation in triphone models. Also for syllable-length models it holds that part of the variation is due to factors such as the neighboring syllables, the presence or absence of lexical stress, the speaking rate, etc. Moreover, analyzing manual transcriptions of speech makes it obvious that syllables are frequently realized as many different phoneme sequences. Therefore, it is not a priori evident that acoustic observation densities of syllable models will model the most important sources of variation more accurately than triphones do - in particular if the syllable models are bootstrapped from a sequence of triphones, without adapting the model topology. This may explain why reports on the performance of syllable models in ASR have come to contradictory conclusions [4][6].

One way to tackle this problem is building multi-path syllable models with parallel HMMs topologies. In previous work [7], we developed a data-driven method, namely Trajectory Clustering (TC), to build multi-path parallel model topologies, and successfully applied it to the 94 most frequent syllables for continuous Dutch speech recognition. In this approach, speech observations are regarded as continuous trajectories along time in acoustic feature space, and clustered based on mixtures of regressions of these trajectories [8]. Each trajectory cluster is modeled as a prototype polynomial function with some variability around it. The variability within the clusters is described in term of a mixture of Gaussians. The EM algorithm is employed to train the cluster model in a Maximum Likelihood manner. With the results of trajectory clustering, multi-path models can be trained based on the training tokens in different clusters.

When using TC based multi-path models, some other problems arise. First, in clustering variable-length sequences data, TC is more effective than for example Mixture of Hidden Markov Models [9], in discriminating different evolutionary patterns or shapes of speech trajectories, because a continuity constraint is imposed on consecutive frames. However, TC assumes that all tokens in a cluster of trajectories are drawn from a single Gaussian, with a equal covariance for all frames. This assumption might not be so realistic, given the fact that the frames in a speech trajectory may not be equally informative. In Mixture of HHM clustering, this problem is tackled by using state-dependent covariances. The HMM state with relatively smaller covariance contribute more to the overall probability in computing the distance between speech trajectories and cluster templates. Thus, the questionable assumption made in TC may deteriorate the discriminability of pronunciation variants. Second, in large vocabulary continuous speech recognition the multi-path models for a single syllable will increase exponentially the searching load.

In this paper, we propose a novel method, namely path-mixture Hidden Markov Models, to alleviate the adverse effects and at the same time reap the benefit of the TC-based multi-path approach. The paper is further organized as follows. Section 2 describes the theoretical framework of the TC approach and of path mixture model. The experiment deigned to test the approach and the results are presented and discussed in Sections 3 and Section 4. Finally, in Section 5, we summarize the most important findings and draw conclusions about the implications for future work.

2. Method

2.1. Trajectory Clustering

In TC, speech realisations are assumed to be drawn from several components of mixture Gaussians, where the mean of each component density is a polynomial function of time. For speech realization j with a length of N_j frames, the matrix form of the regression equation for component k in D dimensional acoustic feature space

can be written as

$$\mathbf{Y}_j = \mathbf{X}_j \beta_k + \mathbf{E}_k \tag{1}$$

or:

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

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

$\mathbf{Y}_j$ 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 $\mathbf{\Sigma}_k$.

Since the speech trajectories that we will be dealing with have different durations, we normalize the trajectories to unit length by dividing the frame numbers in the second column of $\mathbf{X}_j$ by $N_j - 1$. In [7], we found that this way of handling different durations yields the most coherent clusters.

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:

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

Here, θ_k includes both the parameters of the regression model β_k and the covariance matrix of regression residual $\mathbf{E}_k$. 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:

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

in which 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 re-estimation equation for the EM algorithm can then be defined as:

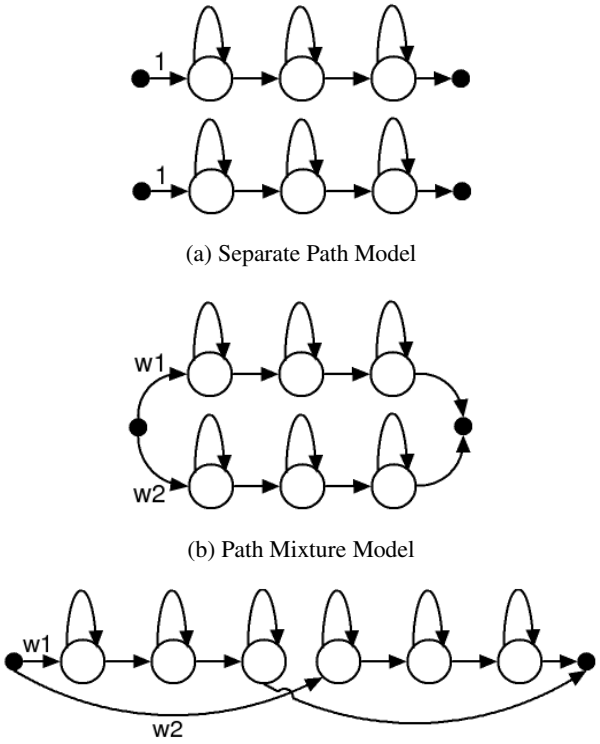
$$\hat{\beta}_k = (\mathbf{X}' \mathbf{H}_k \mathbf{X})^{-1} \mathbf{X}' \mathbf{H}_k \mathbf{Y} \tag{4}$$

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

$$\hat{w_k} = \frac{1}{M} \sum_j^M h_{jk} \tag{6}$$

where $\mathbf{Y} = [\mathbf{Y}_1' \quad \dots \quad \mathbf{Y}_M']'$ 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 segment 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 re-estimation procedure is repeated until convergence is reached.

The EM algorithm is highly sensitive to the initial values of the model parameters. We tackled this problem by using a procedure in which the number of clusters is increased incrementally until the required number of clusters is reached. Since the shapes of the trajectories are contained in the sequence of MFCC vectors, we did not include delta or delta-delta coefficients in the syllable representations that were used as input to the clustering procedure.



(a) Separate Path Model

(b) Path Mixture Model

(c) Alternative view of Path Mixture Model

Figure 1: Model topologies for Separate Path Model and Path Mixture Model.

2.2. Path Mixture Hidden Markov Model

With the results of TC, multiple HMM paths for a speech unit can be trained, based on the training tokens in different trajectory clusters. The multi-path models are integrated in the lexicon as alternative pronunciations of the words that contain this speech unit. In decoding, the paths have equal prior probabilities. We refer to this model topology as the *s*eparate path model. An example model topology with two HMM paths is illustrated in Figure.1(a).

In the path-mixture model, the TC-based separate HMM paths are combined into one entity by recruiting two non-emitting states (cf. Figure 1(b)). By doing so, the multi-path model can be regarded as a single speech unit rather than alternative pronunciations. This should decrease the searching load in decoding. For a

given observation sequence $\mathbf{Y}$, the probability that $\mathbf{Y}$ is produced by the model is given as

$$p(\mathbf{Y}|\Lambda) = \sum_{h=1}^{H} w_h p(\mathbf{Y}|\lambda_h) \quad (7)$$

where w_h are the mixture weight, subject to $\sum_{h=1}^{H} w_h = 1$, indicating how likely the hth path will be chosen. λ_h implies a HMM path, and Λ is the parameter $\Lambda = \{w_h, \lambda_h\}(h = 1 \dots H)$. By analogy with the conventional HMM, the Maximum Likelihood estimate is given as:

$$\Lambda' = \text{argmax} \sum_{h} w_h p(\mathbf{Y}|\lambda_h) \quad (8)$$

To estimate the model parameter Λ', a Baum-Welch re-estimation procedure can be directly applied, if we regard path mixture model as a single HMM chain (cf. Figure 1(c)).

The difference between path-mixture model and separate path model is not only the additional weights for parallel HMM paths, but also the way we train them. For separate path model, the HMM paths are trained by using separate sets of tokens corresponding to the trajectory clusters, whereas all the tokens are used to train path-mixture models. Thus, the training of the path-mixture model is equivalent to clustering the tokens again as in the Mixture of Hidden Markov Models approach, but now with the initialization of the parameters obtained from TC-based multi-path models. We expect that the re-clustering procedure can reap the benefit from the varying covariance property of Mixture of Hidden Markov Model. Moreover, since the segmentations of training tokens are obtained from the force-alignments of single-path models, we also expect this re-clustering based on Baum-Welch algorithm can fix the clustering error caused by segmentation faults.

In decoding, the Viterbi algorithm can also be directly used in path mixture models. It should be noted that in decoding when a search path begins with a state in a HMM path, it will end in the same HMM path, thus avoiding the trajectory folding problem [7].

3. Experiments

3.1. Speech Material

The speech material was taken from the Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN) [10]. For this study we used speech from 166 females reading books for the library for the blind. The training, development and test sets comprised non-overlapping fragments of all 166 speakers.

Feature extraction of the speech material was carried out at a frame rate of 10 ms using a 25-ms Hamming window. A pre-emphasis factor of 0.97 was employed. 12 Mel Frequency Cepstral Coefficients (MFCCs) and log-energy with corresponding first and second order time derivatives were calculated, for a total of 39 features. Channel normalization was applied using cepstral mean normalization over complete recordings, which were chunked to sentence-length entities for the purpose of further processing.

3.2. Speech Recognition

In this work, we built multi-path models for the 94 most frequent syllables. We designed experiments to compare the performance of 1) a mixed-unit system [5] with a single path for each syllable model, 2) with 2-path separate model based on TC for each syllable, and 3) with 2-path mixture model for each syllable.

The 94 context-independent syllable units of the single-path mixed-unit recognizer were initialized with the 8-Gaussian triphone models corresponding to the constituent (canonical) phonemes of the syllables. The mix of units underwent four passes of embedded re-estimation.

To build the separate path recognizer, we clustered the training tokens of each of the syllables into two clusters. The 2-path syllable models were initialized with the same 8-Gaussian single-path syllable models and re-estimated with separate sets of training tokens obtained trough TC. Since we did not find systematic relations between trajectory clusters and syllable duration, we decided to keep the number of states in the separate paths equal to the sum of the constituent triphone models.

Based on the separate-path models, the 2-path mixture of HMMs models were built. We used the parameters of the separate-path models to initialise the mixture models, starting with equal mixture weights for the two paths. The path-mixture models then underwent four passes of Baum-Welch re-estimation.

Table 1: Speech recognition results for mixed-unit recognizer, multi-path mixed unit recognizer and path mixture model recognizer.

Recognizer Type	Word Error Rate	Time
mixed-unit	$9.41\% \pm 0.5\%$	12 hours
2-path mixed-unit	$8.70\% \pm 0.5\%$	24 hours
2-path mixture model	$8.45\% \pm 0.5\%$	14 hours

Table 1 illustrates the recognition results and the time needed for decoding a test data set with 1,098 sentences consisting of 12,327 words. From the table it can be seen that the recognition performance for the 2-path mixed-unit recognizer is significantly better than the single path mixed-unit recognizer. This result confirms that although syllable models are capable to model long-term dependencies in ASR, there are other sources of variation that are more important to model [6]. By applying multi-path models based on data-driven trajectory clustering, the most important variation is accounted for in the separate models and this leads to improved performance. The performance for the path-mixture recognizer is substantially better than the separate-path recognizer. Moreover, the searching load in decoding for the path-mixture recognizer is approximately halved compared to the separate-path recognizer.

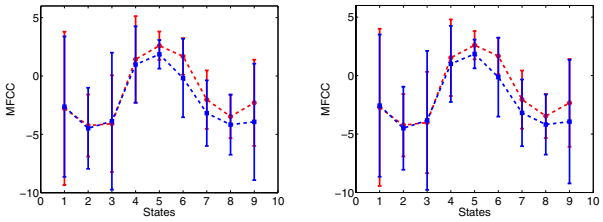
4. Discussion

In order to investigate the impact of applying path-mixture models to adapt TC-based multi-path models, we performed forced alignment of the training tokens with both the original models and the adapted models. Part of the results are illustrated in Table 2. For each syllable, we calculate the token migration rate, which indicates the percentage of training tokens aligned to a different path in adapted models than in the original TC clusters, and segmentation difference, which is the average number of frames that are different between the two alignments. These results are sorted with reference to token migration rate.

From able 2 it can be seen that the average segmentation difference for all syllables is less than 1, which indicates that the two types of two-path models are trained with essentially the same frames. The token migration rates differ considerably between the syllables. This results suggest that after adapting the TC-based

Table 2: The results of the forced-alignment of training token with both separate-path models and path-mixture models

Syllable	Token Migration	Segmentation Difference
/I_n/	15.16%	0.56
/I_s/	14.35%	0.73
/ui_t/	14.18%	0.63
/s_@/	13.87%	0.51
⋮	⋮	⋮
/d_A_n/	1.07%	0.44
/e_n/	0.90%	0.80
/s_t_@/	0.35%	0.41
/t_@_x/	0.31%	0.67
Overall	6.30%	0.63



(a) Separate Path Model (b) Path Mixture Model

Figure 2: The means and variance of HMM states for both separate-path model and path-mixture model of syllable /t_@_x/.

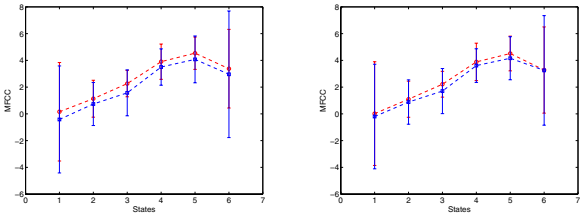
models, the parameters in some of the original models are substantially changed. In Figures 2 and 3, we compare the model parameters between the separate-path model and the path-mixture model of syllable /I_n/ (with maximum token migration rate) and /t_@_x/ (with minimum token migration rate), by plotting the means and variances of the states in the HMM paths of the first MFCC coefficient.

In Figure 2(a) we can see that the evolution of the means in two HMM paths that are substantially different. Since the HMM paths are trained using different trajectory clusters, this shows that TC is very effective in discriminating variants of speech trajectories with different forms. From Figure 2(b) it can be seen that the model parameters in the path-mixture model remain approximately unchanged. Combined with low path migration rates, this suggests that applying path-mixture models will not obscure the shape differences between speech trajectories uncovered by TC.

In Figure 3(a) we can see that two HMM paths are approximately parallel. This outcome implies that TC did not find a good split of trajectories corresponding to different patterns (explaining the high migration rate in Table 2). Because of the identical covariance matrices, the TC clustering results in two parallel cluster prototypes. However, since the variance in the boundary states is larger than in the central states, the central states are more important. Thus, after adapting the boundary states overlap almost completely, but the separation of the states in the middle of paths remained (cf. Figure3(b)).

5. Conclusion

In this paper we address the problem that single distributions with common covariance in TC-models sometimes is inconsistent with reality, and may degrade the power of uncovering the underlining pronunciation variants. This problem can be partly alleviated by adapting the original models with a re-clustering procedure by the means of Mixture of Hidden Markov Models. To this end, we apply a more flexible model topology to TC-based model. This topology allows adapting the original models by the means of Baum-Welch re-estimation, resulting in path-mixture HMMs. A recognition experiment showed that path-mixture HMMs outperform TC-based models.



(a) Separate Path Model (b) Path Mixture Model

Figure 3: The means and variance of HMM states for both separate-path model and path-mixture model of syllable /I_n/.

6. References

[1] Jones, R.J., Downey, S., and Mason J.S., "Continuous speech recognition using syllables," in *Proc. Eurospeech-97*, vol. 3, pp. 1171-1174, 1997.

[2] Ganapathiraju, A., Hamaker, J., Ordowski, M., Doddington, G., and Picone J., "Syllable-based large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9(4), pp. 358-366, 2001.

[3] Sethy, A., and Narayanan, S., "Split-lexicon based hierarchical recognition of speech using syllable and word level acoustic units", in *Proc. ICASSP-2003*, vol. 1, pp. 772-776, 2003.

[4] Messina, R., and Jouvet D., "Context-dependent long units for speech recognition," in *Proc. ICSLP-2004*, pp. 645-648, 2004.

[5] Hämäläinen, A., de Veth, J., and Boves, L., "Longer-length acoustic units for continuous speech recognition," in *Proc. EUSIPCO-2005*, Antalya, Turkey, 2005.

[6] Hämäläinen, A., Boves, L., and de Veth, J., "Syllable-length acoustic units in large-vocabulary continuous speech recognition," Submitted to *SPECOM-2005*.

[7] Han, Y., Hämäläinen, A., and Boves, L., "Trajectory Clustering of Syllable-length Acoustic Models for Continuous Speech Recognition," in *Proc. ICASSP-2006*, 2006.

[8] 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.

[9] Smyth, P., "Clustering sequences with Hidden Markov Models", *Advances in Neural Information Processing Systems*, The MIT Press, 1991.

[10] Oostdijk, N., Goedertier, W., Van Eynde, F., Boves, L., Martens, J.P., Moortgat, M., and Baayen, H., "Experiences from the Spoken Dutch Corpus Project," in *Proc. LREC-2002*, vol.1, pp. 340–347, 2002.