WHITHER LINGUISTIC INTERPRETATION OF ACOUSTIC PRONUNCIATION VARIATION

Annika Hämäläinen, Yan Han, Lou Boves & Louis ten Bosch

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

ABSTRACT

Recent research suggests that modelling pronunciation variation is more appropriate at the syllable level than at the level of context-dependent phones. Due to the large number of factors affecting syllable pronunciation, the creation of multi-path topologies is necessary. Previous research on multi-path models in connected digit recognition has proved *trajectory clustering* to be an attractive approach to deriving multi-path models. In this paper, we extend our research to large-vocabulary continuous speech recognition (LVCSR) by deriving trajectory clusters for 94 frequent syllables in a 20-hour corpus of Dutch read speech. With multi-path models based on these trajectory clusters, speech recognition performance improves significantly. We believe that recognition performance can be improved further by adapting the topologies of the parallel paths. However, the physical properties of the clusters do not provide clues to the most appropriate topology, or the best way of initialising the state observation densities. Therefore, we attempt to interpret the clusters in terms of linguistic and phonetic criteria. The results obtained so far suggest that there is no straightforward relation between physically defined trajectory clusters and linguistic and phonetic criteria.

1. INTRODUCTION

Coarticulation introduces long-span 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 in [1] – [7]. However, long-span coarticulation is not the only, or even the most important, source of pronunciation variation in speech. Part of the syllable-level variation is due to factors such as the neighbouring syllables, the position of the syllable in a multi-syllabic word, the presence or absence of lexical stress, and the speaking rate. Moreover, manual transcriptions of speech show that syllables are often realised as various phoneme sequences [15]. Therefore, the variation inherent in fluent speech cannot be modelled appropriately using acoustic observation densities that are specialised for individual syllables [7]. To account for very different pronunciations, the model topologies have to be adapted.

One way to tackle the problem of pronunciation variation is building syllable models with multi-path HMM topologies, in which each path represents a major pronunciation variant. However, because of the sheer number of factors that may play a role in syllable-level pronunciation variation, and the obscure order of their importance, it is difficult to select the variants to model in a knowledge-based manner. A bottom-up data-driven approach appears to be more promising. Yet, a data-driven approach can only be applied to a relatively small number of highly frequent syllables. To be able to cluster less frequent syllables, we do need knowledge about the linguistic and phonetic factors that have the largest and most systematic impact in acoustic modelling.

It has long been acknowledged that longer-length acoustic models might be preferable to context-dependent phone models. However, the use of longer-length units (e.g. syllables) poses an extremely difficult data sparsity problem in model training; most languages have no more than 40 phonemes, while they have several thousand syllables. One way around this problem is to bootstrap the topologies and observation densities of longer-length models using the underlying phone models. Being able to describe the variants of the longer-length units in linguistic terms is necessary also for this reason. In addition to the above-mentioned bootstrapping method, several authors have proposed mixing syllable models for frequent syllables with conventional triphone models for less frequent syllables [1] – [7].

In previous work [8][9], we developed a data-driven method, *speech trajectory clustering*, to build multi-path model topologies, and successfully applied it to longer-length acoustic models (linguistics-based Head–Body–Tail models [10]) for connected digits 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 [11]. Each trajectory cluster is modelled as a prototype polynomial function with some variability around it. The variability within the clusters is described in terms of a mixture of Gaussians. The EM algorithm is used to train the cluster model. Using the results of trajectory clustering, multi-path models can be trained based on the training tokens belonging to the different clusters.

In this paper, we investigate two aspects of multi-path syllable models for LVCSR. First, we examine whether bottom-up clusters of syllable tokens correspond to classes that can be interpreted in terms of linguistic and phonetic features. The aim of this exercise is to find clues to the best way of adapting the topologies of the parallel paths, all of which are currently equal to a sequence of triphone models corresponding to the canonical transcription of the syllables in question. Second, we investigate whether multi-path syllable models improve recognition performance as compared with a triphone recogniser and a mixed-model recogniser with single-path syllable models. This extends our previous work [6][7], which combined single-path syllable models for the 94 most frequent syllables with triphone models. To achieve the goals set for this paper, we first cluster the training tokens of the 94 most frequent syllables by means of the trajectory clustering method, and interpret the resulting clusters in terms of a number of linguistic and phonetic factors that are known to have an impact on pronunciation variation. We focus on the following factors: syllable duration, the part-of-speech (POS) tag of the word containing the syllable, lexical stress, the difference between mono- and poly-

syllabic words, and the phonetic transcription of the syllable. Using the resulting clusters, we build and test multi-path models for the 94 syllables. Since both our earlier work [8][9] and the present study have shown that trajectory clustering always detects the gender distinction as the first factor, we limit our clustering and speech recognition experiments to female speech only. We compare the results of the multi-path mixed-model recogniser with the performances of a triphone recogniser and a single-path mixed-model recogniser.

The rest of the paper is organised as follows. The trajectory clustering algorithm used to cluster the training observation sequences is described in Section 2. The data used in the experiments and their linguistic annotation are introduced in Sections 3 and 4. The results from our clustering and speech recognition experiments are presented and discussed in Sections 5 and 6. In Section 7, we summarise the most important findings and draw conclusions about the implications for future work.

2. TRAJECTORY CLUSTERING

The underlying idea of speech trajectory clustering is the Mixture of Regression Model [11]. In this model, speech realisations are considered as weighted mixtures of polynomial regression functions, each of which is defined by its regression parameters and a covariance matrix. It should be noted that this representation does not suffer from the standard Markov assumption that subsequent observation frames are independent. Also, trajectories of different tokens need not have equal length in terms of the number of frames [9]. For speech realisation 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 are of different durations, we normalise the trajectories to be of unit length by dividing the frame numbers in the second column of $\mathbf{X}_j$ by $N_j - 1$. In [9], we found that this way of handling different durations yields the most coherent clusters for the body part of connected digits. It remains to be investigated whether this also holds for clustering syllables in continuous speech. With the standard 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 alleviate this problem by using a procedure in which the number of clusters is increased incrementally until the required number of clusters is reached [8]. We start by computing the average regression function for the complete set of tokens. Then we create two sets of initial values by adding and subtracting a proportion of the standard deviation of the individual parameters. In the next step the same splitting procedure is applied to the cluster with the largest weight (which almost always happens to be the cluster with the highest number of tokens. 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. The trajectory clustering was, of course, applied to all frequent syllables individually (cf. section 3). Therefore, we also had to analyse the relation between the trajectory clusters and the knowledge-based classes for a substantial number of individual syllables.

3. SPEECH MATERIAL

The speech material was extracted from the Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN) [12], which - among other things - contains manually verified orthographic transcriptions and

Table 1. Main statistics of the CGN female speech data used for analysis.

Statistic	Training	Test	Development
Word tokens	215,810	12,327	11,822
Speakers	166	166	166
hh:mm:ss	20:15:44	01:08:54	01:06:21

POS tags. As mentioned earlier, we used speech from 166 females reading books for the Dutch library for the blind. The restriction to female speech only is due to the fact that previous experiments with trajectory clustering showed that the first two clusters delivered by the procedure invariably separate the two genders [8]. In this study, we are not interested in gender differences, but rather in linguistic and phonetic factors. We used the females because this part of the read speech corpus is somewhat larger than the male part. The training, development and test sets comprised non-overlapping fragments of all 166 speakers (cf. Table 1).

Feature extraction of the speech material was carried out at a frame rate of 10 ms using a 25-ms Hamming window and a pre-emphasis factor of 0.97. 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 normalisation was applied using cepstral mean normalisation over complete recordings, which were chunked to sentence-length entities for the purpose of further processing.

4. LINGUISTIC INFORMATION

In our earlier experiments on a smaller corpus of read speech, we used a set of 94 syllables that occurred frequently enough to allow the accurate training of single-path syllable models [6][7]. For a larger corpus, such a set of syllables would naturally be larger. However, as we needed sufficient training data for the accurate training of multi-path syllable models, we decided to use the same set of 94 syllables in this work. Each of the 94 syllables was analysed with respect to the following linguistic or phonetic information:

1) Syllable duration
2) POS tag
3) Stress
4) Monosyllabicity
5) Phonetic transcription

Syllable durations were computed by means of forced alignment. The canonical transcriptions of words were time-aligned to the speech signal using a set of triphone models trained on the 5-hour subset of the speech material used in [6][7]. The syllable durations were retrieved by mapping the triphones to the corresponding syllables. One half of the syllable realisations was defined as long and the other half short. This "definition" of long and short syllables has proved successful in our previous work on connected digits [8][9]. We also analysed normalised syllable durations by computing the average articulation rate (in terms of the number of phones per second, excluding silences) in individual sentences, and then scaling the phone durations in the sentences so that their average duration became equal to the overall average phone duration in the corpus.

The POS tagging was used to determine if the words in our data set were function or content words, and to analyse how the syllables of interest related to them. The group of function words was defined to consist of articles, adverbs, conjunctions, interjections, numerals, prepositions and pronouns. The distinction between function and content words is related, but certainly not identical, to the distinction between accented and non-accented syllables. For example, an adverb such as "veel" ('very') can occur both with and without accent. Yet, function words tend to be unaccented in continuous speech, while content words are more likely to be accented.

The feature "stress" relates to the presence of a word stress mark on the syllable in the pronunciation lexicon. Except for a small number of monosyllabic function words [13], all words in the lexicon contain one stressed syllable. Monosyllabicity marks those syllable tokens that occur as a monosyllabic word. Most of the 94 syllables occurred both as parts of polysyllabic words and as monosyllabic words on their own. Canonical transcriptions comprising syllabification and word stress information were retrieved from the CGN lexicon (in-house version of 2 May 2005) and CELEX. The CGN lexicon is built by manually verifying the pronunciation information retrieved from various existing lexical resources. A single canonical pronunciation was used per lexeme, with the CGN phone set reduced to 37 phones. The information in our lexicon was used to determine if the syllables of interest carried lexical stress or corresponded to monosyllabic words.

A 70,000-word subset of the read speech in CGN contains manually verified (broad) phonetic labels and word-level segmentations. The realised pronunciations for all the word tokens in this subset were retrieved and aligned with the syllabified canonical pronunciations. The alignment was carried out using a dynamic programming algorithm that computes the optimal alignment between two strings of phonetic symbols, taking into account the distances between the symbols in terms of articulatory features [14]. This resulted in a list of plausible pronunciation variants for each of the 94 target syllables. Using these pronunciation variants, a forced alignment of the training data was carried out to determine which pronunciation variants of the 94 syllables were most likely to have been realised in the part of the corpus that only came with orthographic transcriptions. To ensure that the complete training corpus was handled in the same manner, the forced alignment procedure was also applied to the manually transcribed part of the training corpus.

5. EXPERIMENTAL RESULTS

5.1. Trajectory Clustering

To analyse the correspondence of the trajectory clusters with the syllable duration, POS tag, stress, monosyllabicity and transcription variant, we split the acoustic observations of each syllable into two groups using trajectory clustering [8], and compared the results with the knowledge-based classification based on the above mentioned linguistic and phonetic criteria. The resulting two-way classifications were analysed visually, by examining a set of graphical representations with four-block grey scale pictures for each of the 94 syllables. In addition, the results were analysed numerically, by looking for syllables for which the large majority of the cases was concentrated in one diagonal. Fig. 1 illustrates the graphical representations of the results for four example syllable models: /t_ei_t/, /z_o/, /l_@/ and /h_a_r/. The proportion of tokens shared by a linguistic category (column) and a trajectory cluster (row) is shown as the degree of darkness of the cells, as indicated in the rightmost column. Essentially, a conspicuously dark diagonal implies a close

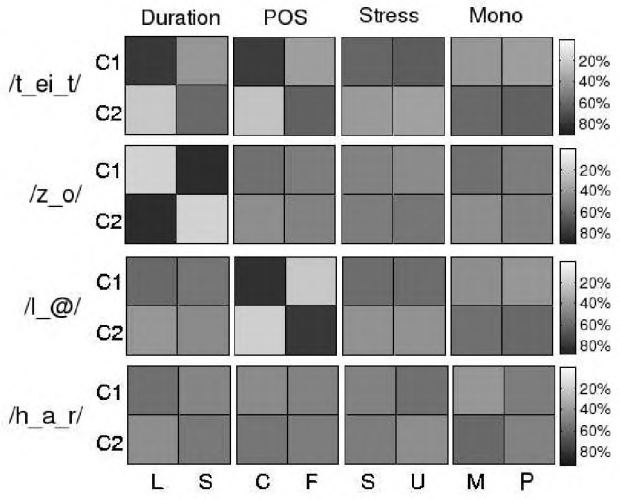

Fig. 1. The relationship of the trajectory clustering with respect to syllable duration, POS tag, stress and monosyllabicity in the case of the syllable models /t_ei_t/, /z_o/, /l_@/ and /h_a_r/. C1 = cluster 1, C2 = cluster 2. Duration: L = long, S = short. POS tag: C = content word, F = function word. Stress: S = stressed, U = unstressed. Mono: M = monosyllabic, P = polysyllabic.

correspondence between the trajectory clustering and the linguistic information under examination.

In Fig. 1, the syllable models /t_ei_t/, /z_o/, /l_@/ and /h_a_r/ demonstrate four types of correspondence between the results of the trajectory clustering and the linguistic information. For about 5% of the syllables, exemplified by /t_ei_t/, the results of the clustering corresponded with both the duration and POS. About 15% of the syllables (for example /z_o/) showed an effect of duration, and another 15% of the syllables (e.g. /l_@/) an effect of the POS. The syllable model /h_a_r/ illustrates the most typical pattern: about 65% of the syllables did not correspond to any of the five factors examined. In addition, for the factors stress and monosyllabicity, there were hardly any syllables showing a systematic connection with the results of trajectory clustering.

A more detailed analysis of the results exemplified in Fig. 1 shows that effects of the linguistic and phonetic factors may differ between syllables. We can illustrate this for the syllable /p_@/. It appears only in (numerous) multisyllabic words (e.g. "hopeloos" ('hopeless'), "repetitie" ('repetition', 'rehearsal')). Because of the vowel /@/ (schwa), /p_@/ never occurs in stressed position. The same applies to many other frequent syllables with the vowel /@/. The effect of duration in the case of /p_@/ might be related to the presence or absence of the /@/ vowel; the three-syllable word "hopeloos" can be realised as bi-syllabic, by deleting the /@/ with subsequent re-syllabification. This holds for all other /@/-syllables. The syllable /w_@/ corresponds to the monosyllabic pronoun "we" ('we', 'us') and acts as a syllable in many multisyllabic words (e.g. "huwelijk" ('marriage'), "nauwelijks" ('hardly')). In polysyllabic words /w_@/ is always unstressed, but the stress status of /w_@/ as personal pronoun is much less certain. Dutch has two forms of the personal pronoun corresponding to the English "we", the reduced form "we" and a full form "wij", pronounced /wE+/. The full form may be reduced partially, resulting in pronunciation variants that might be transcribed as either "we" or "wij". Therefore, some of the syllables transcribed as "we" may actually be somewhat reduced versions of "wij". One might expect that the forced alignment procedure always selects the correct pronunciation variant, but especially in the case of /@/ this is doubtful [15]. Similar arguments can be made for other frequent syllables.

Table 2. The proportions of pronunciation variant tokens for the syllable /O_f/ assigned to clusters produced by trajectory clustering.

Variant	Cluster 1	Cluster 2
O	57%	43%
O_v	51%	49%
O_f	52%	48%
@_v	82%	18%
@_f	83%	17%
w_O_f	82%	18%
j_O_f	100%	0%

To see if syllable tokens with different phonetic transcriptions go into different clusters, the transcription variants for each syllable were first aligned with each other and the phonetic distances between the variants were computed on the basis of articulatory features [14]. A multidimensional scaling (MDS) analysis was then carried out for syllables whose pronunciation variant distance matrices could be reduced to 1- or 2-dimensional representations. These distance representations were compared with the results of the trajectory clustering. Even though MDS produced phonetically solid distance representations, no clear correspondence could be observed between the clusters of syllable transcription variants produced by MDS, on the one hand, and the clusters produced by trajectory clustering, on the other hand. For instance, from Table 2, we can see that the majority of tokens corresponding to all variants ends up in Cluster 1 - regardless of the phonetic similarities and dissimilarities that can be observed in the distance representation of Fig. 2.

As it became obvious that none of the linguistic and phonetic criteria alone could explain the trajectory clusters on their own, we applied a classification and regression tree (CART) analysis to the data to see if combining the criteria would uncover knowledge-based structure in the data. Given a set of tokens belonging to a certain syllable, each record associated to a token had the same structure, which consisted of a number of non-category attributes corresponding to the linguistic and phonetic criteria and a category attribute representing the trajectory cluster membership. All these attributes took nominal values only: for example, "Duration: long, short", "POS: content, function". The decision trees were generated based on the ID3 algorithm [16]. We ran the CART analysis in two ways: 1) using all the syllable tokens with the syllable duration, POS tag, stress, monosyllabicity and phonetic transcription as the non-category attributes; 2) using the tokens associated with the most frequent (canonical) transcription and omitting the transcription variant attributes. With virtually no exceptions, the canonical transcription variants accounted for such a large part of the data that a systematic relation to the acoustic properties of the speech is unlikely. The decision trees were built using 10-fold cross validation. However, the high estimated error rates (larger than 40% on average) indicated that the trajectory clusters cannot be explained in a homogeneous way using the knowledge-based criteria.

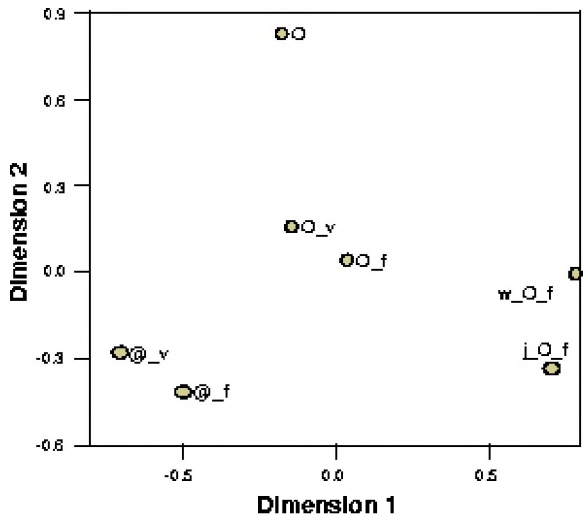

Fig. 2. The two-dimensional distance representation for the pronunciation variants of the syllable /O_f/.

The analyses described above suggest that it is difficult to devise procedures for defining the topologies of the paths associated with individual clusters, and to appropriately initialise the observation densities of the states belonging to the paths, on the basis of linguistic or phonetic knowledge. Therefore, for the time being, we decided to proceed with training multi-path models with identical topologies and identically initialised states for all parallel paths.

5.2. Speech Recognition

Based on the results of the trajectory clustering, we built multi-path models for 94 frequent syllables. We designed experiments to test whether a mixed-model recogniser with multi-path syllable models would outperform 1) a conventional triphone recogniser and 2) a mixed-model recogniser with a single path for each syllable model.

In building the triphone recogniser and the single-path mixed-model recogniser, we used the procedure described in [6]. To summarise, a standard procedure with decision tree state tying was used to train the triphone recogniser. The triphones were created based on the canonical transcriptions in the lexicon. For each HMM state, 8 Gaussian mixture components were trained. The 94 context-independent syllable models of the mixed-model recogniser were initialised with the 8-Gaussian triphone models corresponding to the constituent (canonical) phonemes of the syllables. The mixture of models underwent four passes of Baum-Welch re-estimation.

To build the multi-path mixed-model recogniser, we clustered the training tokens of each of the 94 most frequent syllables into two and three trajectory clusters. Based on the results of the trajectory clustering, we built 2-path and 3-path HMMs for each syllable. The multi-path syllable models were initialised with the 8-Gaussian single-path syllable models and re-estimated using the training tokens belonging to the clusters obtained through trajectory clustering. Since we did not find a systematic connection between trajectory clusters and the long or short duration of syllable tokens, we decided to keep the number of states in the parallel paths equal to the sum of the states in the constituent triphone models. Word entrance penalty and language model scaling factor were optimised on the independent development test set (cf. Table 1).

Table 3. Speech recognition results for the triphone recogniser, the single-path mixed-model recogniser and the multi-path mixed-model recognisers.

Recogniser Type	Word Error Rate
Triphone	$9.15\% \pm 0.5\%$
1-path mixed-model	$9.41\% \pm 0.5\%$
2-path mixed-model	$8.70\% \pm 0.5\%$
3-path mixed-model	$8.67\% \pm 0.5\%$

In order to study possible improvements due to changes in acoustic modelling only, without the risk of language modelling issues masking the effects, out-of vocabulary words were not allowed in the task. In effect, the recognition lexicon and word-level bigram network were built using all orthographic words in the training and test sets containing both female and male speech. The vocabulary consisted of about 29,700 words, and the test set perplexity, computed on a per-sentence basis using HTK, was 92. Due to the special nature of the corpus, which consists of chapters from novels, a strict separation between training and test sets would have resulted in a test set perplexity of about 350.

Table 3 illustrates the recognition results. The performance for the single-path mixed-model recogniser is slightly, but not significantly, worse than for the triphone recogniser. This replicates the results in [7] - for models trained on a substantially larger corpus. The 2-path mixed-model recogniser is significantly better than the triphone recogniser and substantially outperforms the single-path mixed-model recogniser. The confirms the hypothesis that, although syllable models are capable of modelling long-span dependencies in ASR, there are other sources of variation that are more important to model. By employing multi-path models based on data-driven trajectory clustering, the most important variation is accounted for in the parallel paths, and this leads to improved performance.

¿From Table 3, it can be seen that the recognition performance of a 3-path mixed-model recogniser is almost identical to that of the 2-path one. Analysing the results of the 3-way trajectory clustering showed that the number of training tokens for some of the paths was less than 100. Using such a limited number of training tokens does not allow the accurate training of the observation densities. At the same time, it may be that identically initialised parallel path topologies do not allow the models to reap the maximal benefits from the data-driven clustering of training tokens.

6. DISCUSSION

Generally speaking, the intrinsic variation in the speech signal can be investigated by focusing on its effect along two dimensions. The first dimension is the acoustic variation that is caused by factors such as speaker identity, vocal tract length (gender), speaking style, speaking rate, and accent. The second dimension is the symbolic variation, obtained as the result of the human perception and labelling process. Although this picture of speech variation is oversimplified, it is nevertheless useful to identify a number of research approaches.

In the case of limited symbolic variation - e.g., in the case of a set of tokens with one unique (phonetic) transcription, the symbolic variation is void whereas the acoustic variation is fully attributable to gender, speaking style etc. This means that also in the case of limited symbolic variation, the number of acoustic paths through the model must accommodate the intrinsic acoustic variation. On the other hand, even if the acoustic variation is small (such that one path is adequate), the symbolic variation may be substantial due to the noise in perception and decision making about the best transcription. This makes it evident that the relation between acoustic and symbolic variation is not straightforward. This paper presents an attempt to systematically investigate the intrinsic variation in speech by concurrently exploring the acoustic and symbolic dimensions. Considering the complexity of the issue, it is unsurprising that we were unable to find a clear linguistic interpretation for the syllable clusters.

Our speech recognition experiments suggest that syllable models with a topology equal to a sequence of triphone models do capture some, but probably not all pronunciation variation in read speech. Apparently, they do not model this variation much better than the sequences of triphone models per se. Comparisons of the observation densities in the syllable models with the densities in the corresponding states of the triphone models, which were used for bootstrapping, show that Baum-Welch re-estimation only has a small effect [7]. The fact that 2-path and 3-path syllable models do yield a small but significant improvement in performance, suggests that the gain in modelling power originates from separating different realisations of syllable tokens. The most compelling explanation for the finding that multi-path models only yield a small performance gain is the fact that all parallel paths had topologies identical to the topology of the sequence of constituent triphones.

We intend to pursue two lines of research to extend the work reported in this paper. First, we will examine the effects of implementation details in the trajectory clustering, such as the procedure for splitting clusters, and the way in which time is represented in the regression polynomials [9]. Second, we will carry out more detailed phonetic analyses of the tokens in a number of promising clusters, to identify other linguistic and phonetic criteria that might explain cluster membership and could, therefore, be used for designing appropriate techniques for adapting the topologies and initialising the observation densities.

7. CONCLUSIONS

In this paper, we addressed the issue of parallel trajectory topologies for syllable models. We showed that the results of bottom-up trajectory clustering do not clearly correspond to any, or any combination, of the linguistic or phonetic features that we tested (syllable duration, content/function word, stress, mono/polysyllabic word, phonetic transcription). This will make it very difficult, if not impossible, to design context-dependent syllable models on the basis of decision trees with linguistic and phonetic questions.

A single-path mixed-model recogniser, combining syllable and triphone models, performed slightly worse than a straightforward triphone recogniser. However, a mixed-model recogniser with multi-path syllable models outperformed the triphone recogniser, despite the fact that all parallel paths had a topology identical to the topology of the sequence of constituent triphones. This shows that it is worthwhile to try to develop techniques for designing different topologies for the paths in the multi-path models.

8. ACKNOWLEDGEMENTS

This work was carried out within the framework of the Interactive Multimodal Information eXtraction (IMIX) program, which is sponsored by the Netherlands Organisation for Scientific Research (NWO).

9. 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] Sethy, A., Ramabhadran, B., and Narayanan, S., "Improvements in ASR for the MALACH project using syllable-centric models," in *Proc. IEEE ASRU-2003*, St. Thomas, US Virgin Islands, 2003.

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

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

[7] Hämäläinen, A., Boves, L., and de Veth, J., "Syllable-length acoustic units in large-vocabulary continuous speech recognition," in *SPECOM-2005*, pp. 499-502, 2005.

[8] Han, Y., de Veth, J., and Boves, L., "Trajectory Clustering for Automatic Speech Recognition," in *Proc. EUSIPCO-2005*, Antalya, Turkey, 2005.

[9] Han, Y., de Veth, J., and Boves, L., "Speech Trajectory Clustering for Improved Speech Recognition," in *Proc. Interspeech-2005*, Lisbon, Portugal, 2005.

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

[11] 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.

[12] 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.

[13] van den Heuvel, H., van Kuijk, D., and Boves, L., "Modeling lexical stress in continuous speech recognition for Dutch," *Speech Communication*, 40(3), pp. 335-350, 2003.

[14] Elffers, B., Van Bael, C., and Strik, H. *ADAPT· Algorithm for Dynamic Alignment of Phonetic Transcriptions*, Internal report, CLST, Radboud University Nijmegen, 2005.

[15] Binnenpoorte, D. *Phonetic Transcription of Large Speech Corpora*, PhD Dissertation, Radboud University Nijmegen, 2006.

[16] Colin, A., "Building decision trees with the ID3 algorithm," *Dr. Dobbs Journal*, pp. 107-109, June 1996.