TRAJECTORY CLUSTERING OF SYLLABLE-LENGTH ACOUSTIC MODELS FOR CONTINUOUS SPEECH RECOGNITION

Yan Han, Annika Hämäläinen & Lou Boves

Centre for Language and Speech Technology (CLST),
Radboud University Nijmegen, The Netherlands
{Y.Han, A.Hamalainen, 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 affect the way syllables are articulated, creating multiple paths through syllable models might be necessary. Our previous research on longer-length 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 by deriving trajectory clusters for 94 very frequent syllables in a 20-hour data set of Dutch read speech. The resulting clusters are compared with a knowledge-based classification. The comparison results suggest that multi-path models for syllables are difficult to build based on phonetic and linguistic knowledge. When multi-path models based on trajectory clustering are used, speech recognition performance improves significantly. Thus, it is concluded that data-driven trajectory clustering is a very effective approach to developing multi-path models.

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 in [1] – [7]. Syllable models are assumed to be inherently capable of modeling part of the long-term dependencies in speech. However, most languages have no more than 40 phonemes, while they have several thousand syllables. Many infrequent syllables will have poor coverage in the training data. Therefore, it is unlikely that a reasonably sized training corpus would contain enough tokens to train reliable models for all syllables from scratch. As a consequence, several authors have proposed mixing syllable models for frequent syllables with conventional triphone models, or bootstrapping longer-length models from the sequence of constituent triphones [1] – [7].

However, it is unlikely that long-term coarticulation is the only, or even the most important, source of variation in speech. Also for syllable-length models it holds that part of the variation is due to factors such as the neighboring syllables, the position of the syllable in a multi-syllabic word, the presence or absence of lexical stress, and the speaking rate. Moreover, analyzing manual transcriptions of speech shows 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 those of triphones - 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 [5][7].

One way to tackle this problem is building syllable models with multi-path HMM topologies. However, because of the sheer number of different syllables in large vocabulary ASR and the large number of factors affecting their realization, it is not evident that creating multi-path models on the basis of phonetic or linguistic questions is possible. In addition, since syllables tend to appear in relatively limited phonetic and linguistic contexts, good classification criteria for some syllables may not be suitable for other syllables. Thus, a data-driven technique might be more appropriate than a knowledge-based approach.

In our previous work [8][9], we developed a data-driven method, which we named *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 modeled 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 employed to train the cluster model in a maximum likelihood manner. Using the results of trajectory clustering, multi-path models can be trained based on the training tokens in different trajectory clusters.

In this paper, we investigate two aspects of multi-path syllable models for large vocabulary ASR. First, we examine whether bottom-up clusters of syllable tokens correspond to classes that can be interpreted in terms of linguistic features. Second, we investigate whether multi-path syllable models improve recognition performance as compared to triphone models and a mixed-model system based on single-path syllable models. In doing so, we extend previous work [6][7] in which we combined 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 factors that are likely to have an impact on pronunciation variation. We focus on phonetic and linguistic factors such as syllable duration, the part-of-speech (POS) tag of the word containing the syllable, lexical stress, and the difference between mono- and polysyllabic words. 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

1-4244-0469-X/06/$20.00 ©2006 IEEE

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 large vocabulary ASR system with the performance of a triphone system and a single-path mixed-model system.

The data used in the experiments and its linguistic annotation are introduced in Sections 2 and 3. The results from our clustering and speech recognition experiments are presented and discussed in Sections 4 and 5. In Section 6, we summarize the most important findings and draw conclusions about the implications for future work.

2. SPEECH MATERIAL

The speech material was taken from the Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN) [12], which - among other things - contains manually verified orthographic transcriptions and POS tags. For this study we used speech from 166 females reading books for the Dutch library for the blind. The training, development and test sets comprised non-overlapping fragments of all 166 speakers. Details of the composition of the three sets are given in Table 1.

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

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. Feature extraction was performed using HTK.

3. LINGUISTIC INFORMATION

The set of 94 syllables from [6][7] was analyzed with respect to the following information:

- Syllable duration
- POS tag
- Stress
- Monosyllabicity

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 realizations 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].

The POS tagging was used to determine if the words in our data set were function or content words, and to analyze 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 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 which occur as a monosyllabic word. Most syllables can occur both as a part of a polysyllabic word and as a monosyllabic word of their own. Canonical pronunciations comprising syllabification and word stress information were retrieved from the CGN lexicon (in-house version of 2 May 2005) and CELEX [14]. 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.

4. EXPERIMENTAL RESULTS

In the experiment, 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 syllable duration, POS tag, stress and the monosyllabicity criteria. The resulting two-way classifications were analyzed visually, by examining a set of graphical representations with four-block grey scale pictures for each syllable of interest. In addition, the results were analyzed numerically, by checking whether the proportions of cases in the diagonal cells were comparable. Figure 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 depicted as the degree of darkness of the cells, as indicated in the rightmost column. Essentially, a conspicuously dark diagonal implies a close correspondence between the trajectory clustering and the linguistic information under examination.

4.1. Trajectory Clustering

In Figure 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_@/) showed 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 four factors examined. In addition, for the factors stress and monosyllabicity, there was hardly a syllable that would have shown a systematic connection with the

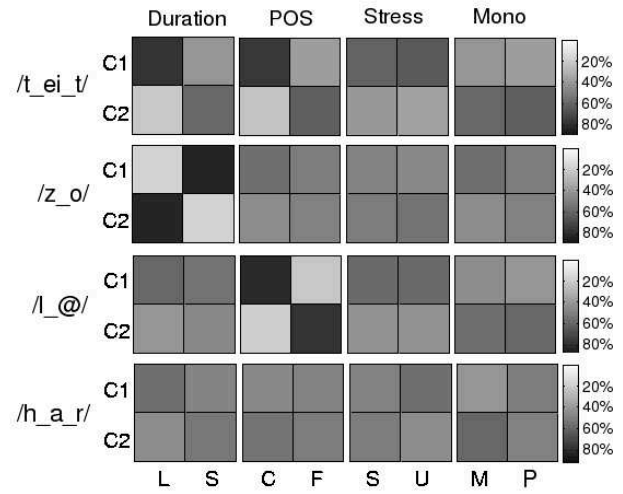

Fig. 1. The relationship of the trajectory clustering with respect to syllable duration, POS tag, stress and monosyllabicity in the case of the syllable models /t_ei_t/, /z_o/, /l_@/ and /h_a_r/. C1 = cluster 1, C2 = cluster 2. Duration: L = long, S = short. POS tag: C = content word, F = function word.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

results of trajectory clustering. These results indicate that it might be difficult to build multi-path models based on phonetic and linguistic knowledge. The data-driven trajectory clustering can automatically find the most important variant for each syllable, and effectively define an appropriate multi-path model topology.

4.2. 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 system with multi-path syllable models would outperform 1) a conventional triphone system or 2) a mixed-model system with a single path for each syllable model.

In building the triphone recognizer and the single-path mixed-model recognizer, we used the procedure described in [6]. To summarize, a standard procedure with decision tree state tying was used to train the triphone recognizer. 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 recognizer were initialized 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 reestimation.

To build the multi-path mixed-model recognizer, 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 initialized with the same 8-Gaussian single-path syllable models and reestimated with the training tokens in 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 optimized on the independent development test set (cf. Table 1).

In order to study possible improvements due to changes in acoustic modeling only, without the risk of language modeling 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 2. Speech recognition results for the triphone recognizer, the single-path mixed-model recognizer and the multi-path mixed-model recognizers.

Recognizer 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	$8.67\% \pm 0.5\%$

Table 2 illustrates the recognition results. As one can see, the recognition performance for the single-path mixed-model recognizer is slightly, but not significantly, worse than for the triphone recognizer. This replicates the results in [7], for models trained on a substantially lager corpus. The performance of the 2-path multi-path mixed-model recognizer is significantly better than the single-path mixed-model recognizer, and it substantially outperforms the triphone recognizer. The results indicate that, although syllable models are capable of modeling long-term 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 2, it can also be seen that the recognition performance of the 3-path mixed-model recognizer is almost identical to that of the 2-path recognizer. Most probably, this is caused by the undertraining of at least some of the individual HMM paths. From analyzing the results of the 3-way trajectory clustering, it appears that the number of training tokens for some HMM paths is less than 100. Using such a limited number of training tokens does not allow the accurate training of the observation densities of these paths.

5. DISCUSSION

Our experiments suggest that syllable models with a topology equal to a sequence of triphone models do not capture much more long-term coarticulation information than the sequence 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 reestimation 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 modeling power originates from separating different realizations of syllable tokens. The finding that the 3-path models did not outperform their 2-path counterparts, despite an increase in the number of model parameters, is probably due to undertraining.

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. In the experiments reported in this paper, we were not able to find appropriate techniques for defining different topologies for parallel paths. This is mainly due to the failure to find a connection between the clusters and linguistic or phonetic features that might provide clues for adapting the topologies. Contrary to our expectations, we did not find a clear connection between trajectory clusters and syllable duration. We are presently investigating two possible explanations for this lack of correspondence. One is related to the time normalization that was used in the trajectory clustering procedure [9]. The other is related to the lack of time normalization of the syllable tokens in the contingency matrix. It may well be that a much clearer connection between clusters and duration will emerge if another strategy is used for dealing with trajectories of different length in the clustering procedure, or when syllable tokens are normalized for external factors such as overall speech rate. In that case, parallel paths with different numbers of states could be trained, most likely leading to a more substantial performance gain.

Another method that can be explored to define different topologies for parallel paths is analyzing the manual phonetic transcriptions that are available in the CGN corpus for a part of the speech material. We intend to investigate whether tokens with different transcriptions end up in different clusters, and if the transcriptions could then be used as a basis for defining topologies.

So far, our trajectory clustering technique was designed so that, in each step, the cluster with the highest mixture weight is split. This can lead to clusters with a relatively small number of members, too few to allow for reliable reestimation of the observation densities. We intend to adapt the clustering procedure in order to avoid clusters which are too small for our purposes. This should help in training effective models with more than two parallel paths.

6. 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 correspond to any of the linguistic or phonetic features that we tested (duration, stress, content/function word, and mono- or polysyllabic word). 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 recognizer, combining syllable and triphone models, performed slightly worse than a straightforward triphone system. However, a mixed-model system with multi-path syllable models did outperform the triphone system, 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 and develop techniques for designing different topologies for the paths in the multi-path models. Research is under way to develop procedures for designing different topologies in the absence of clear relations to linguistic or phonetic features.

7. 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).

8. 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] Baayen, H., Piepenbrock, R., and van Rijn, H. *The CELEX Lexical Database. Release 2 (CD-ROM)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA, 1995.