Pronunciation Variant –Based Multi-Path HMMs for Syllables

Annika Hämäläinen, Louis ten Bosch, Lou Boves

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

Abstract

Recent research suggests that it is more appropriate to model pronunciation variation with syllable-length acoustic models than with context-dependent phones. Due to the large number of factors contributing to pronunciation variation at the syllable level, the creation of multi-path model topologies appears necessary. In this paper, we propose a novel approach for constructing multi-path models for frequent syllables. The suggested approach uses phonetic knowledge for the initialisation of the parallel paths, and a data-driven solution for their re-estimation. When applied to 94 frequent syllables in a 37-hour corpus of Dutch read speech, it leads to improved recognition performance when compared with a triphone recogniser of similar complexity.
Index Terms: automatic speech recognition, pronunciation variation, multi-path syllable models

1. Introduction

Coarticulation introduces long-span spectral and temporal dependencies in speech. To model these dependencies for the purpose of ASR, the use of longer-length acoustic models, based e.g. on syllables, has been proposed [1-4]. Re-estimating the acoustic observation densities of single-path syllable models initialised with triphones underlying the canonical transcriptions of the syllables does indeed appear to capture at least some of the coarticulation-related variation; however, it seems that this is not sufficient to account for the most important effects of pronunciation variation [4]. Several authors – [5] in particular – have shown that, while syllables are seldom deleted completely, they do display considerable variation in the identity and number of phonetic symbols that best reflect their pronunciation. At the same time, it is clear that a substantial part of the variation defies modelling in the form of different sequences of symbols [6]. We believe that pronunciation variation at the syllable level is best modelled using parallel paths to capture 'major, distinct transcription variants' (hereafter MDVs), and re-estimating these parallel paths to better capture the dynamic nature of articulation.

In this paper, we propose to construct multi-path models for frequent syllables using a combination of knowledge-based and data-driven methods. The knowledge-based part of our approach uses phonetic transcriptions of the target syllables for selecting MDVs, and for initialising the observation densities of the parallel paths aimed at modelling these MDVs. The data-driven part amounts to us leaving the training entirely to the Baum-Welch algorithm, instead of predefining which training tokens to use for re-estimating the model parameters of each parallel path.

We use a mixed-model recognition scheme in which syllable models for 94 frequent syllables are combined with triphone models that cover the less frequent syllables in a Dutch read speech recognition task. We investigate whether multi-path syllable models improve recognition performance as compared with a conventional triphone recogniser and a mixed-model recogniser with single-path syllable models.

This paper is organised as follows. The speech material used in the study is described in Section 2. The selection of transcription variants for the initialisation of parallel paths is discussed in Section 3, whereas the experimental set-up is detailed in Section 4. In Section 5, results from the recognition experiments are presented and discussed. Finally, the conclusions are formulated in Section 6.

2. Speech material

The speech material used in this study was read speech extracted from the Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN) [7], which – among other things – contains manually verified orthographic transcriptions for all of the data. The data were divided into three sets comprising non-overlapping fragments of all 303 speakers: a set for training the acoustic models, a development set for optimising the language model scaling factor and word insertion penalty, and a test set for evaluating the acoustic models. Details of the data are presented in Table 1.

Table 1. *Main statistics of the speech material.*

Statistic	Train	Devel.	Test
# Word tokens	396,187	22,100	22,289
# Speakers	303	303	303
Duration (hh:mm:ss)	37:00:20	02:03:33	02:04:21

A 60,600-word subset of the data containing manually verified (broad) phonetic transcriptions and word-level segmentations was used to retrieve transcription variants for syllables. In this study, a set of 37 phone labels was used. The manually verified, non-syllabified transcriptions for all the word tokens in the subset were aligned with their syllabified canonical pronunciations 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 [8]. This procedure resulted in a list of plausible transcription variants for all the syllables in the subset. Using the transcription variants from the alignment procedure for the target syllables and canonical transcriptions for the rest of the syllables, corresponding 8-Gaussian triphones were used to perform a forced alignment of the training data in order to determine which pronunciation variants were most likely to have been realised in the part of the corpus that only came with the 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 corpus. Comparing the proportions of the different transcription variants of the target syllables in the manually verified and the automatically annotated sets of data confirmed the reliability of the automatic annotation procedure.

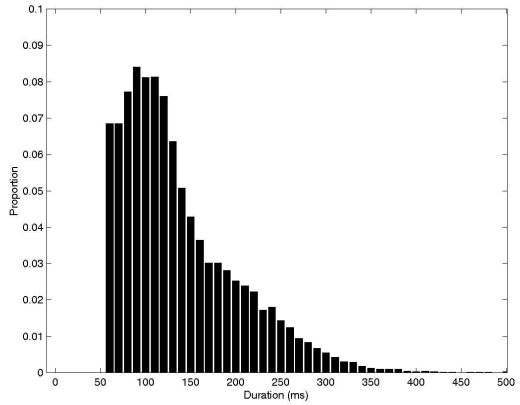

Figure 1. *Duration distribution for 40 most common CV syllables.*

3. Selection of major, distinct transcription variants

The selection of MDVs was guided by two principles. First, we wanted to keep the canonical variant as one of the MDVs, except perhaps in cases where a different variant was the most frequent variant in the training corpus. Second, we had a preference for MDVs containing fewer symbols than the canonical variant. This preference stemmed from an analysis of syllable durations obtained by using a single-path mixed-model recogniser to perform a forced alignment of the CGN data used in [4]. We observed a high proportion of syllables with the minimum duration imposed by the HMM topology (see Figure 1 for the duration distribution histogram for the 40 most common CV syllables). This finding suggests that the standard three states per underlying phone topology may have been too long. In addition, although multi-path models derived using trajectory clustering resulted in a significant improvement in recognition performance in [9], we concluded that the equal length of the parallel paths was hindering the performance gain.

The following steps were devised for selecting the optimal combination of MDVs and constructing the corresponding multi-path models:

1. Compute phonetic distance matrices – e.g. on the basis of articulatory features [8] – between all transcription variant pairs for each target syllable.
2. Identify a list of high-ranking MDV combinations on the basis of the phonetic distances between the variant pairs, weighted by their frequency of occurrence.
3. Post-process the list produced in Step 2 to take into account the preference for transcription variants shorter than the canonical: in case the canonical transcription is not mono-phonemic, pick the highest-ranking MDV combination that contains at least one transcription variant with at least one symbol less than the canonical. When none of the MDV combinations satisfies the length criterion, select the highest-ranking MDV combination.
4. Initialise HMM paths corresponding to the optimal MDV combination from Step 3 by picking the initial state parameters from the corresponding triphones [2, 4] and combine them into a multi-path model. An example of a multi-path model is shown in Figure 2.
5. Apply the Baum-Welch algorithm to re-estimate the parameters of the multi-path models in order to capture coarticulation effects.

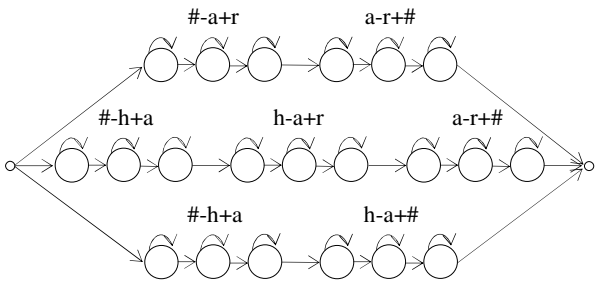

Figure 2. *Multi-path model for the syllable /har/, with the three parallel paths initialised with triphones underlying the MDVs /ar/, /har/ and /ha/, respectively.*

4. Experimental set-up

4.1. Feature extraction

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 then chunked to sentence-length entities for the purpose of further processing.

4.2. Lexicon and language model

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. The recognition lexicon consisted of a single pronunciation for each word. In the case of the triphone recogniser, the pronunciation for each word consisted of a string of canonical phones from the CGN lexicon. In the case of the mixed-model recognisers, it consisted of a) syllable units b) canonical phones, or c) a combination of a) and b). The vocabulary comprised about 29,700 words, and the test set perplexity, computed on a per-sentence basis, was 92. Due to the special nature of the corpus, which consists of excerpts from novels, a strict separation between the training and test sets would have resulted in a test set perplexity of about 350.

4.3. Acoustic modelling

Speech recognition experiments were designed to test whether a mixed-model recogniser with multi-path models for the target syllables would outperform 1) a conventional triphone recogniser and 2) a mixed-model recogniser with a single path for the target syllables. As we wanted to be able to train up to three parallel paths for each target syllable without running into data sparsity problems, we concentrated our modelling efforts on the 94 most frequent syllables in the training data. The target syllables covered 57% of all the syllable tokens in the training data, the least frequent of the target syllables occurring 850 times.

4.3.1. Triphone recogniser

A standard procedure with decision tree state tying was used to train the triphone recogniser [10]. Initial 32-Gaussian monophones were trained using linear segmentation of canonical transcriptions within automatically generated word segmentations. The monophones were used to perform a forced alignment of the training data; triphones were then bootstrapped using the resulting phone segmentations. Triphone recognisers with up to 64 Gaussian mixtures per state were trained and tested.

4.3.2. Single-path mixed-model recogniser

A procedure similar to that used in [4] was employed in building the single-path mixed-model recogniser. The context-free models for the target syllables were initialised with triphones corresponding to the canonical syllable transcriptions, and triphones were used to cover the rest of the syllables. The mix of syllable and triphone models underwent four passes of Baum-Welch re-estimation. Single-path mixed-model recognisers with up to 16 Gaussian mixtures per state were trained and tested.

4.3.3. Multi-path mixed-model recogniser

The steps described in Section 3 were followed in order to build the multi-path mixed-model recogniser. In this study, the transcription variants retrieved from the manually verified data (cf. Section 2) were aligned with each other and the phonetic distances between the variants were computed on the basis of articulatory features [8]. The parallel paths of the context-free multi-path models for the target syllables were initialised with triphones corresponding to the optimal MDV combination (cf. Section 3, Step 3), and triphones were used to cover the rest of the syllables. The mix of syllable and triphone models underwent four passes of Baum-Welch re-estimation. Multi-path mixed-model recognisers with up to 16 Gaussian mixtures per state were trained and tested.

Table 2. *Word error rates with a 95% confidence interval, and the total number of Gaussians in the recognisers.*

Recogniser	WER (%)	# Gaussians
16G triphone	10.3 ± 0.4	24,560
32G triphone	10.1 ± 0.4	49,120
16G single-path mixed-model	10.5 ± 0.4	34,304
16G multi-path mixed-model	9.9 ± 0.4	51,536

5. Results and discussion

An analysis of the MDV combinations used in building the multi-path models for the target syllables showed that the canonical transcription was always included. 85% of the bi- and tri-phonemic target syllables (81% of all the target syllables) had one or two MDVs with fewer phones than the canonical. Somewhat surprisingly, 39% of all the target syllables had one MDV with more phones than the canonical. In a third of these cases, this could be attributed to the presence of a long vowel or diphthong in the syllable. Other cases, however, seemed to be artefacts.

In Table 2, the speech recognition results and the recogniser complexities measured in the total number of Gaussians are presented for the most relevant recognisers: the 16-Gaussian triphone recogniser, the 32-Gaussian triphone recogniser (best performing triphones), and the 16-Gaussian single- and multi-path mixed-model recognisers (best performing mixed-model recognisers of each type). In terms of complexity, the 16-Gaussian single-path mixed-model recogniser lay between the 16- and 32-Gaussian triphone recognisers. Yet, it performed worse than either of them. Some of the decrease in performance will have been due to the loss of context information at some syllable boundaries, but the result still supports our finding that just retraining output pdf's is not sufficient to capture the most important effects of pronunciation variation at the syllable level [4]. Even with the loss of context information at

syllable boundaries, the 16-Gaussian multi-path mixed-model recogniser outperformed the 32-Gaussian triphone recogniser – the most comparable triphone recogniser when it comes to recogniser complexity. The reduction in WER was not significant, but the result does suggest that using multi-path models for frequent syllables is a more effective way of increasing modelling power than just increasing the number of Gaussians per state in triphones. In effect, the multi-path syllable models add prior knowledge about structure, whereas the triphone models only add detail in terms of straightforward statistics of an unstructured population.

Generally speaking, the intrinsic variation in the speech signal can be investigated in two domains. The first domain is the acoustic variation that is caused by factors such as gender, speaker identity, speaking style and accent. To a large extent, this type of variation can be captured by means of Gaussian mixture modelling. The second domain is the symbolic variation, obtained as the result of the human perception and labelling process. This type of variation, on the contrary, cannot be sufficiently accounted for by increasing the number of Gaussians per state. In the case of limited symbolic variation – for instance, in the case of a set of tokens with a unique phonetic transcription – the acoustic variation is fully attributable to gender, speaker identity etc. However, even if the acoustic variation is small, the symbolic variation might be substantial due to idiosyncrasies in the transcriptions. Therefore, the relation between acoustic and symbolic variation clearly is not straightforward.

The approach introduced in this paper utilises multi-path syllable models built using a combination of phonetic (symbolic variation) and data-driven (acoustic variation) methods; the improved recognition performance suggests that important variation is indeed accounted for in the parallel paths. To gain a better understanding of this variation, the multi-path mixed-model recogniser will be used to perform a forced alignment of the training data, and a detailed analysis of the training tokens assigned to the different parallel paths will be carried out. The results of the analysis will be used to refine our approach when it comes to the optimal number and type of MDVs used in the initialisation of the parallel paths. Ultimately, we aim to devise a method for constructing multi-path syllable models with parallel paths initialised with the triphones underlying the canonical transcriptions, and subsequently shortened (or possibly lengthened) using state merging (and splitting).

6. Conclusions

In this paper, we proposed a novel approach for constructing multi-path models for frequent syllables. The approach combines knowledge-based and data-driven techniques by using phonetic knowledge to initialise the parallel paths of the syllable models, and by subsequently leaving the further training entirely to the Baum-Welch algorithm. In essence, the approach provides a solution for initialising parallel paths of different lengths. Experiments with a mixed-model recogniser with 16 Gaussians per state suggested that multi-path syllable models capture important effects of pronunciation variation. Even though the reduction in WER was not significant, the multi-path mixed-model recogniser outperformed a 32-Gaussian triphone recogniser of comparable overall complexity. This suggests that, beyond a certain number of Gaussians per state, adapting model topologies is a more effective way of increasing modelling power than just increasing the number of Gaussians per state in triphones.

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] 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.

[2] 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.

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

[4] Hämäläinen, A., Boves, L., and de Veth, J., "Syllable-length Acoustic Units in Large-Vocabulary Continuous Speech Recognition," in *Proc. SPECOM-2005*, pp. 499-502, 2005.

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

[6] Ostendorf, M., "Moving beyond the 'beads-on-a-string' model of speech", in *Proc. IEEE ASRU-99*, Keystone, Colorado, USA. Dec 12-15, 1999.

[7] Oostdijk, N., Goedetier, 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.

[8] Elffers, B., Van Bael, C., and Strik, H., *ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcriptions*, CLST, Radboud University Nijmegen, The Netherlands, 2005.

[9] Han, Y., Hämäläinen, A., and Boves, L., "Trajectory Clustering of Syllable-Length Acoustic Models for Continuous Speech Recognition," in *Proc. ICASSP-2006*, Toulouse, France, May 14-19, 2006.

[10] Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., and Woodland, P., *The HTK Book (for HTK Version 3.2.1)*, Cambridge University, Cambridge, UK, 2002.