

MODELLING PRONUNCIATION VARIATION USING MULTI-PATH HMMS FOR SYLLABLES

Annika Hämmäläinen, Louis ten Bosch, Lou Boves

Centre for Language and Speech Technology (CLST),
Radboud University Nijmegen, Nijmegen, The Netherlands

ABSTRACT

Recent research suggests that it is more appropriate to model pronunciation variation with syllable-length acoustic models than with triphones. 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 construct multi-path models using phonetic knowledge to initialise the parallel paths, and a data-driven solution for their re-estimation. When applied to 94 frequent syllables in a Dutch read speech recognition task, the approach leads to improved recognition performance when compared with a much more complex triphone recogniser. A detailed analysis of the pronunciation variation captured by the parallel paths pinpoints the deficiencies of the approach, and provides insights into how these may be overcome.

Index Terms— Speech recognition, hidden Markov models

1. INTRODUCTION

Coarticulation introduces long-span spectral and temporal dependencies in speech. To model these for the purpose of ASR, acoustic models based on syllables have 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 appears to capture some coarticulation-related variation, but not 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 sequences of symbols [6]. We therefore believe that syllable-level pronunciation variation 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 construct multi-path models for frequent syllables by combining knowledge-based and data-driven methods. The knowledge-based part of the approach uses phonetic transcriptions of the target syllables for selecting MDVs, and for initialising the observation densities of the parallel paths that represent these MDVs. The data-driven part amounts to us leaving the training entirely to the Baum-Welch algorithm, instead of preassigning training tokens to specific paths.

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 analyse two aspects of the multi-path syllable models. First, 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. Second, we analyse the pronunciation variation captured by the parallel paths to devise possible solutions to refining our approach when it comes to the optimal number and type of MDVs used in the initialisation of the parallel paths.

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

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 accurate orthographic transcriptions for all of the data. The read speech in CGN consists of novels read out loud for the Dutch library for the blind; this explains its relatively lively nature. A total of 41 hours of speech was divided into three non-overlapping sets comprising fragments from 303 speakers: a 37-hour set for training the acoustic models, a 2-hour development set for optimising the language model scaling factor and word insertion penalty, and a 2-hour test set for evaluating the acoustic models.

A 6.5-hour subset of the training 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. A list of plausible transcription variants for all the syllables in the subset was arrived at by aligning the manual phonetic transcriptions of word tokens with their syllabified canonical counterparts, taking into account the articulatory distance between the phones [8]. Using these transcription variants for the target syllables, and canonical transcriptions for the rest of the syllables, a forced alignment of the training data was performed with 8-Gaussian triphones to determine which pronunciation variants best represented the target syllables in the part of the corpus that only came with orthographic transcriptions. For consistency, the forced alignment procedure was also applied to the manually transcribed part of the data. Comparing the proportions of the different transcription variants of the target syllables in the manually verified and the automatically transcribed data confirmed the reliability of the automatic transcription procedure.

3. EXPERIMENTAL SET-UP

3.1. Feature extraction

Feature extraction 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 first and second order derivatives were calculated, for a total of 39 features. Channel normalisation was applied using cepstral mean normalisation over complete recordings.

3.2. Lexicon and language model

The recognition lexicon comprised a single pronunciation for each of the 29,700 words in the recognition task. In the case of the triphone recogniser, the pronunciations 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). A word-level bigram network was built using the relevant part of the CGN corpus. The test set perplexity, computed on a per-sentence basis using HTK [9], was 92.

3.3. Acoustic modelling

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 test the approach without running into data sparsity problems, we concentrated our modelling efforts on a set of frequent syllables in the training data. In our earlier experiments on a smaller corpus of read speech, we used a set of 94 target syllables that occurred frequently enough to allow the accurate training of single-path syllable models [4]. In order to have sufficient training data for the robust training of multi-path syllable models, we decided to use the same set of syllables in this work. The 94 target syllables covered 57% of all the syllable tokens in the training data, the least frequent of them occurring 850 times and therefore warranting reliable estimation of a maximum of three parallel paths. The ‘major, distinct transcription variants’ used for the initialisation of these parallel paths were selected using the procedure described in Section 3.3.3.

3.3.1. Triphone recogniser

A standard procedure with decision tree state tying was used to train the triphone recogniser [9]. 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.

3.3.2. Single-path mixed-model recogniser

Two single-path mixed-model recognisers were experimented with: an 8-Gaussian and a 16-Gaussian recogniser. A procedure similar to that used in [4] was employed in building the recognisers. The context-free models for the target syllables were initialised with the 8/16-Gaussian triphones corresponding to the canonical syllable transcriptions. Triphones from the 8/16-Gaussian triphone

recogniser were used to cover the rest of the syllables. The resulting mix of syllable and triphone models underwent four passes of Baum-Welch re-estimation.

3.3.3. Multi-path mixed-model recogniser

Two multi-path mixed-model recognisers were experimented with: an 8-Gaussian and a 16-Gaussian recogniser. The MDVs used for the initialisation of the parallel paths of the syllable models were selected using the procedure elaborated in [10]. In short, we chose a combination of transcription variants that were maximally dissimilar to each other, with the provision that the canonical transcription should be kept (except if another variant was more frequent in the training corpus), and that variants with fewer phones than in the canonical should be preferred. An example of a multi-path syllable model is shown in Fig. 1. The parallel paths of the context-free multi-path models for the target syllables were initialised with the 8/16-Gaussian triphones corresponding to the optimal MDV combination. Triphones from the 8/16-Gaussian triphone recogniser were used to cover the rest of the syllables. The resulting mix of syllable and triphone models underwent four passes of Baum-Welch re-estimation.

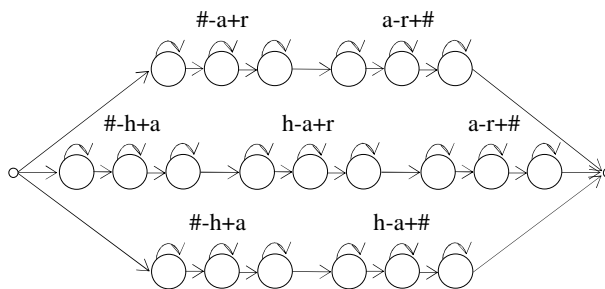


Fig. 1. Multi-path model for the syllable /har/, with the three parallel paths initialised with triphones underlying the MDVs /ar/, /har/ and /ha/, respectively.

4. EXPERIMENTAL RESULTS

4.1. Speech recognition

In Table 2, the speech recognition results and the recogniser complexities in terms of the total number of states 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). The complexity of the syllable models was estimated with the same tying ratio as that used in building the triphone models. The performance of the 16-Gaussian single-path mixed-model recogniser was slightly, but not significantly worse than that of the 16-Gaussian triphone recogniser, the most comparable triphone recogniser from the point of view of the total number of Gaussians in the system. Considering the loss of context information at some syllable boundaries, the small decrease in performance is not surprising. Yet, the result 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]. This is further supported by the fact that, even with the

loss of context information, the 16-Gaussian multi-path mixed-model recogniser outperformed both the 16-Gaussian triphone recogniser and the 32-Gaussian triphone recogniser, which is much more complex in terms of the total number of Gaussians in the system. In the latter case, the reduction in WER was not significant, but the result does suggest that using multi-path models for frequent syllables might be a more effective way of increasing modelling power than just increasing the number of Gaussians per state in triphones. Thanks to their richer topology, multi-path syllable models appear to be better at modelling observed variation in speech than models that merely have more parameters organised along a single path.

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

Recogniser type	WER (%)	# States
16-G triphone	10.3 ± 0.4	1,535
32-G triphone	10.1 ± 0.4	1,535
16-G single-path mixed-model	10.5 ± 0.4	1,603
16-G multi-path mixed-model	9.9 ± 0.4	1,764

Table 3. Percentages of the variant tokens of the syllable /en/ assigned to parallel paths initialised with the triphones underlying the MDVs /en/, /eN/ and /e/. The articulatory distances between the variants and the MDVs are shown in parentheses.

MDV	Variant /en/	Variant /em/	Variant /eN/	Variant /e/
/en/	76.2% (0)	82.5% (1)	10.9% (2)	10.0% (3)
/eN/	21.2% (2)	5.3% (3)	89.1% (0)	5.0% (3)
/e/	2.7% (3)	12.3% (3)	0% (3)	85.0% (0)

4.2. Linguistic analysis of the parallel paths

Even though the use of multi-path syllable models led to improved recognition performance relative to the other recognisers experimented with, our approach clearly is oversimplified. We chose to train up to three parallel paths per target syllable; that is, the optimal MDV combination [10] was used when constructing the parallel paths for most of the syllables, whereas all transcription variants were used for those with no more than three variants (10% of the syllables). This may not have been appropriate, as more paths may have been warranted for some syllables, while fewer may have sufficed for others, such as those with two or three transcription variants only.

To gain a better understanding of the pronunciation variation captured by the parallel paths, and to obtain insights into how the approach may be refined when it comes to the optimal number and type of MDVs used in the initialisation of the parallel paths, the multi-path mixed-model recogniser was used to perform a forced alignment of the training data, and a meticulous analysis of the training tokens assigned to each of the paths was carried out. The token-to-path assignment appeared to be clearly related to the articulatory similarity, or dissimilarity, between the syllable variant tokens and the MDVs used to initialise the parallel paths. Table 3 illustrates the situation for the syllable /en/; the majority of the /en/, /em/, /eN/ and /e/ tokens were assigned to the path that had been initialised with an MDV that was articulatorily the closest (in terms of a Levenshtein distance based on articulatory features [8]). Whenever this is the case, the acoustic information coincides with

the articulatory information. The articulatory distance does not, however, fully account for the observed token-path assignment due to the many confounding factors that are known to influence acoustic path modelling – for instance, gender, accent, context effects, and speaking rate. The effects of these factors are accounted for by the data-driven re-estimation pass of our approach.

In general, the total percentage of all syllable tokens assigned to a path indicated the necessity of the path. A path was considered redundant if less than 5% of all syllable tokens were assigned to it. Straightforward phonetic and linguistic explanations for the necessity, or equivalently the redundancy, of the paths could often be found. These explanations provided us with insights into the relationship between symbolic and acoustic variation in speech. An initial analysis of the MDV combinations used in building the multi-path models for the target syllables showed that the canonical transcriptions were always included, and that the corresponding paths were indeed necessary. Somewhat unexpectedly, 39% of all the target syllables had one MDV with more phones than the canonical, with the corresponding paths being redundant in only 5% of the cases. 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. For 30% of these syllables, at least one of the corresponding paths was redundant.

A closer investigation unveiled three recurring phenomena that were modelled by the non-redundant long paths: syllable-initial glide insertion (26% of the cases), syllable-final /n/ insertion (26% of the cases), and syllable-initial /@/ insertion (17% of the cases). Syllable-initial /j/ insertions were caused by strong articulations of the diphthong /E+/ in the preceding syllable. Examples of such cases are the monosyllabic personal pronouns "hij" ('he') and "wij" ('we'). Syllable-initial /w/ insertions were modelling connecting sounds between the final vowel of the preceding syllable and the initial vowel of the syllable in question. Such a /w/ insertion might, for instance, occur between the first two words of the word string "hoe is 't weer" ('how is the weather'), which might be pronounced as /huwIs@twer/, instead of the canonical /huIs@twer/. With hindsight, the glides would perhaps have been more appropriate to align with the preceding syllable. The syllable-final /n/ insertions were due to the fact that verbs and plural forms of nouns that orthographically end in "-en" did not contain the /n/ in their canonical phonetic transcription. Even though the /n/ deletion does occur very frequently in Dutch, it could arguably be claimed that the canonical transcription should contain the syllable-final /n/ – also because omitting it in the canonical transcription results in syllables that may or may not end with /n/ being bundled together with syllables that cannot end with it. Examples of these two types of syllables are the second syllable of the word "hebben" ('to have') that can be pronounced as either /b@n/ or /b@/, and the first syllable of the word "begeleider" ('supervisor') that cannot be pronounced as /b@n/. Modelling epenthesis, the syllable-initial /@/ insertion was perhaps the most interesting of the three recurring phenomena. For instance, the word "werken" ('to work') was sometimes pronounced as /wEr@k@/ instead of /wErk@/. In fact, epenthesis could be considered to result in the resyllabification of the word: the epenthesised pronunciation is tri-syllabic (/wE-r@-k@/) instead of bi-syllabic (/wEr-k@/). As the procedure for extracting the transcription variants for syllables did not account for resyllabification, the inserted /@/ was again aligned with the latter

syllable of the canonical transcription. The phenomena described above manifest themselves on the threshold between sub-symbolic and symbolic representations of speech, and suggest that variation in speech can be described adequately only by utilising both representations.

Syllable structure seemed to play a prominent role when syllables that exhibited the aforementioned three phenomena were excluded from the analysis. Two thirds of the CV and V syllables had redundant paths, whereas this was the case for only a fourth of the CVC and VC syllables. This finding can probably be explained by the pronunciation variation patterns that syllables typically present; syllable onsets generally maintain their canonical form, whereas the coda elements often get deleted (or assimilated with the onset of the following syllable), and the nuclei remain but might change quality [5]. Because of the presence of the consonantal coda, CVC and VC syllables have more room for variation. An example of a CV syllable that very clearly had a redundant path is the syllable /he/. 84.5% of the syllable tokens were assigned to the path initialised with the MDV /he/, 14.4% to the path initialised with /hE/ and the remaining 1.1% to the redundant path initialised with /@/. The path initialised with /hE/ models a change in the quality of the nucleus, whereas the path initialised with /@/ is not needed due to the consonantal onset hardly ever being deleted. On the contrary, a good example of a CVC syllable with all the three paths necessary is the syllable /vor/. The MDVs used for the initialisation of the parallel paths were /vor/, /v@r/ and /vo/. In the case of the syllable /vor/, the necessity of the three paths is even more evident when the words it appears in are examined. First, it appears as the highly frequent monosyllable "voor" ('for'), which is also likely to be realised in reduced forms, thereby warranting the MDVs /vo/ and /v@r/. Second, it appears as a part of several multisyllabic words, in which it might or might not have word stress. An example of a multisyllabic word in which both options occur is the word "voornaam", which translates as 'first name' if the stress is on the first syllable, and as 'distinguished' if the stress is on the second syllable. Even though the syllable-final /r/ might not always be complete, the quality change in the nucleus merits modelling.

To summarise, the analysis of the pronunciation variation modelled by the parallel paths showed that the necessity of parallel paths for a given syllable could often be explained by factors such as its syllabic structure, phonetic context, lexical stress, and position in a multisyllabic word, as well as the part(s)-of-speech of the word(s) it appears in. Therefore, we intend to experiment with a more sophisticated, multi-pass version of our approach in future research. To ensure that a sufficient number of paths is constructed for each syllable, we will start with a combination of more than three MDVs and incrementally remove potentially redundant paths based on token-to-path assignment statistics. Recognition performance on development test data will be used as the final criterion for the optimal MDV combination. During this process, part-of-speech and stress information will be utilised to determine whether multiple syllable models are warranted for syllables with different linguistic functions or highly varying stress patterns.

5. CONCLUSIONS

In this paper, we constructed multi-path models for frequent syllables. The employed 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. Promising recognition results were achieved with a multi-path mixed-model recogniser containing 16 Gaussians per state: it significantly outperformed both a triphone recogniser and a single-path mixed-model recogniser of comparable complexity. Even though the reduction in WER was not significant, the multi-path mixed-model recogniser also outperformed a much more complex triphone recogniser. This suggests that adapting model topologies might be a more effective way of increasing modelling power than just increasing the number of Gaussians per state in triphones. Finally, a detailed analysis of the pronunciation variation captured by the parallel paths suggests that refining both the knowledge-based and the data-driven aspects of the approach may lead to further performance gains.

6. ACKNOWLEDGEMENTS

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

7. REFERENCES

- [1] A. Ganapathiraju, J. Hamaker, M. Ordowski, G. Doddington, and J. Picone, "Syllable-based large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, Vol. 9(4), pp. 358-366, 2001.
- [2] A. Sethy, and S. Narayanan, "Split-lexicon based hierarchical recognition of speech using syllable and word level acoustic units," in *Proceedings of ICASSP-2003*, Vol. 1, pp. 772-776, 2003.
- [3] R. Messina, and D. Jouviet, "Context-dependent long units for speech recognition," in *Proceedings of ICSLP-2004*, pp. 645-648, 2004.
- [4] A. Hämäläinen, L. Boves, and J. de Veth, "Syllable-length Acoustic Units in Large-Vocabulary Continuous Speech Recognition," in *Proceedings of SPECOM-2005*, pp. 499-502, 2005.
- [5] S. Greenberg, "Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation", *Speech Communication*, 29:159-176, 1999.
- [6] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech", in *Proceedings of IEEE ASRU-99*, Keystone, CO, USA. Dec 12-15, 1999.
- [7] N. Oostdijk, W. Goedetier, F. Van Eynde, L. Boves, J.P. Martens, M. Moortgat, and H. Baayen, "Experiences from the Spoken Dutch Corpus Project," in *Proceedings of LREC-2002*, Vol. 1, pp. 340-347, 2002.
- [8] B. Elffers, C. Van Bael, and H. Strik, *ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcriptions*, Centre for Language and Speech Technology, Radboud University Nijmegen, The Netherlands, 2005.
- [9] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.2.1)*, Cambridge University, Cambridge, UK, 2002.
- [10] A. Hämäläinen, L. ten Bosch, and L. Boves, "Pronunciation Variant -Based Multi-Path HMMs for Syllables," in *Proceedings of Interspeech-2006*, Pittsburgh, PA, USA. Sep 17-21, 2006.