

Construction and Analysis of Multiple Paths in Syllable Models

Annika Hämäläinen, Louis ten Bosch, Lou Boves

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

{A.Hamalainen, L.tenBosch, L.Boves}@let.ru.nl

Abstract

In this paper, we construct multi-path syllable models using phonetic knowledge for initialising the parallel paths, and a data-driven solution for their re-estimation. We hypothesise that the richer topology of multi-path syllable models would be better at accounting for pronunciation variation than context-dependent phone models that can only account for the effects of left and right neighbours. We show that parallel paths that are initialised with phonetic knowledge and then re-estimated do indeed result in different trajectories in feature space. Yet, this does not result in better recognition performance. We suggest explanations for this finding, and provide the reader with important insights into the issues playing a role in pronunciation variation modelling with multi-path syllable models.

Index Terms: speech recognition, hidden Markov models, multi-path syllable models, Kullback-Leibler distance, pronunciation variation

1. Introduction

Coarticulation introduces long-span spectral and temporal dependencies in speech that syllable models – unlike context-dependent phone models – may be able to model [1-6]. 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]. Greenberg [7] – amongst others authors – has 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 [8]. Therefore, it would seem profitable to combine segmental and non-segmental approaches, using ‘major, distinct transcription variants’ (MDVs) for initialising the parallel paths of multi-path syllable models and Baum-Welch re-estimation for capturing coarticulation effects.

The segmental part of our approach utilises manual phonetic transcriptions of frequent syllables when selecting MDVs. The observation densities of the parallel paths are then initialised using the triphones underlying these MDVs. The non-segmental part leaves further training to the Baum-Welch algorithm. Multi-path models for 94 frequent ‘target syllables’ are incorporated into a mixed-model recogniser in which triphone models cover the less frequent syllables in a Dutch read speech recognition task.

The goal of this paper is to illustrate the challenges of using MDV-based multi-path syllable models to model pronunciation variation. To this end, we analyse the shift from a sequence of initialisation triphones to re-trained parallel paths from two points of view. First, we analyse the degree to which the HMM states of the re-trained paths differ from

those of the untrained paths. Second, we compare the speech recognition performance of the untrained and the re-trained multi-path syllable models with each other, and relate the changes in the speech recognition performance to the changes in the HMM states. Finally, we compare the performance of the multi-path syllable models with that of triphones.

2. Speech material

We used read speech extracted from the Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN) [9], consisting of novels read out loud for a library for the blind. 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 contained manually verified broad phonetic transcriptions. 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 [10]. Using these transcription variants for the 94 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 complete corpus (including the part that came with manual transcriptions). 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 recogniser, 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 [11], was 92.

3.3. Acoustic modelling

To analyse the effect of the re-estimation on recognition performance, the performance of the mixed-model recogniser was tested both before and after Baum-Welch re-estimation. In addition, the performance of the mixed-model recogniser was compared with that of a triphone recogniser. 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 3.3.2.

3.3.1. Triphone recogniser

A standard procedure with decision tree state tying was used to train the word-internal triphone recogniser [11]. Initial 32-Gaussian monophones were trained for 37 ‘native’ Dutch phones 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 128 Gaussian mixtures per state were trained and tested.

3.3.2. Mixed-model recogniser

Mixed-model recognisers with up to 64 Gaussian mixtures per state were trained and tested. The MDVs used for the initialisation of the parallel paths of the context-free syllable models were selected using the procedure elaborated in [5]. In short, we chose a combination of transcription variants that were maximally dissimilar to each other, with the provisions that the canonical transcription should be kept (unless 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 Figure 1. The parallel paths of the multi-path models for the target syllables were initialised with the triphones corresponding to the optimal MDV combination. Triphones from the 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.

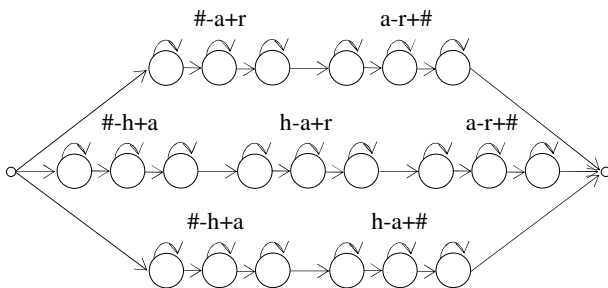


Figure 1: Multi-path model for the syllable /har/, with the three parallel paths initialised with triphones underlying the MDVs /ar/, /har/ and /ha/, respectively.

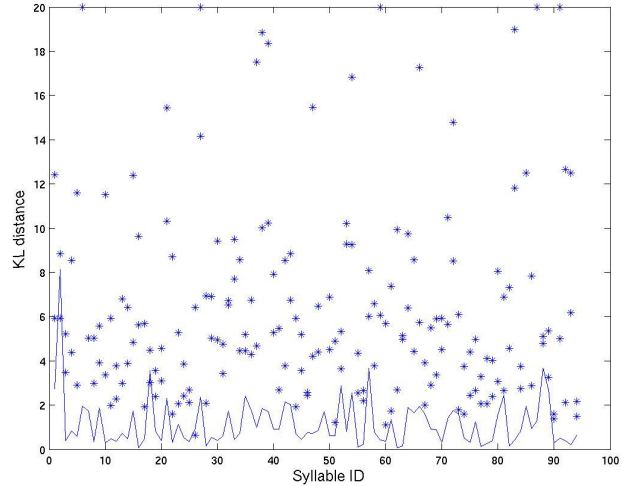


Figure 2: KLD between the initialised and the re-trained parallel paths for the 94 target syllables. The solid line represents the paths initialised with the canonical transcriptions, and the asterisks represent the paths initialised with the non-canonical transcriptions.

4. Experimental results

4.1. Effect of further training

The first aim of this paper was to investigate what happens when the parallel paths of the multi-path syllable models are trained further from the sequences of triphones used for their initialisation. To this end, we calculated the distances between the probability density functions (pdfs) of the HMM states of the re-trained paths and the pdfs of the corresponding states of the initialised paths. We used the Kullback-Leibler distance (KLD, [12]) as our distance measure. Figure 2 illustrates the distances for each of the 94 target syllables. The solid line represents the paths initialised with the canonical transcriptions, whereas the asterisks represent the paths initialised with the non-canonical transcriptions. Apparently, the paths initialised with the canonical transcriptions almost always change the least. In the majority of cases, the paths initialised with the non-canonical transcriptions change at least twice as much. Since 89% of the syllables had more than three transcription variants, this suggests that the added paths account for most of the pronunciation variation that is not captured by the canonical path.

In some respects, our approach of constructing multi-path syllable models is not very sophisticated. We chose to train up to three parallel paths per target syllable; that is, the optimal MDV combination [5] 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. In [6], we used the mixed-model recogniser to perform a forced alignment of the training data and carried out an analysis of the training tokens assigned to each of the parallel paths. It may be assumed that the total percentage of all syllable tokens assigned to a path is a measure of its importance. In our analysis, we considered a path redundant if less than 5% of all

syllable tokens were assigned to it. With this definition, 32% of the syllables had at least one redundant path.

Figure 2 shows that a number of paths change drastically in the further training. Interestingly, 79% of the paths that have a KLD of 10 or higher correspond with paths that were deemed redundant in [6]. Only for 20% of the syllables with redundant paths is the KLD for the redundant path well under 10. Even in these cases, the redundant path has changed considerably more than the canonical path. As discussed in [6], we could find rather straightforward phonetic and linguistic explanations for the redundancy. Hence, we can conclude that dramatic changes during re-estimation indicate that the path is not relevant from the point of view of WER reduction, or that the path is relevant but initialised incorrectly and in need of serious correction.

Figure 2 does not tell which direction the non-canonical paths change during re-estimation – whether they move closer to or further away from the canonical. We investigated this issue for the subset of eleven syllables whose parallel paths were of equal length with each other. 75% of the paths moved away from the canonical, while 25% of them got closer to it. In most cases, the MDVs of the converging paths were very close to the MDVs of the canonical paths in terms of articulatory distance. For instance, the syllable /a/ had a converging path corresponding to the MDV /A/, which is articulatorily very close to the MDV /a/. Therefore, it appears that the MDV selection procedure should enforce a minimum articulatory distance between the MDVs, to avoid heavily overlapping paths and the attendant reduction of the ratio between the number of states to be trained and the number of training tokens available.

To summarise, the KLD analysis illustrated the acoustic stability of the canonical paths. In other words, there is a close relationship between the symbolic and the acoustic variation in speech in the case of the canonical transcriptions. In the case of the non-canonical transcriptions, the relationship is not always straightforward. Large changes during re-estimation could be attributed to larger acoustic variation within the non-canonical transcriptions, as well as suboptimal introduction or initialisation of parallel paths. It is also worthwhile mentioning that the KLD is useful in analysing the extent of acoustic differences between the non-canonical and the canonical paths. This measure might be interesting, for example, when studying the acoustic reduction of speech units.

4.2. Speech recognition

The second aim of this paper was to investigate what happens in terms of WER when going from the untrained to the re-trained mixed-model recogniser. In addition, we wanted to compare the WER of the mixed-model recogniser to that of a triphone recogniser. In Table 1, the most relevant speech recognition results are presented. The 64-Gaussian triphone recogniser and the 64-Gaussian mixed-model recogniser were the best performing instantiations of their respective types. The performance of the 64-Gaussian mixed-model recogniser is shown both before and after Baum-Welch re-estimation.

Table 1. WERs with a 95% confidence interval.

Recogniser type	WER (%)
64-G triphone	7.6 ± 0.4
64-G mixed-model – untrained	10.6 ± 0.4
64-G mixed-model – re-trained	8.7 ± 0.4

The 64-Gaussian triphone recogniser significantly outperformed the re-trained 64-Gaussian mixed-model recogniser. To understand why this is the case, we must consider the 64-Gaussian mixed-model recogniser *before* further training. It consists of a mixture of biphones, triphones, and context-free multi-path syllable models the parallel paths of which consist of sequences of biphones and triphones. The biphones and triphones originate from the 64-Gaussian triphone system. The essential difference between the triphone recogniser and the untrained mixed-model recogniser is twofold.

First, compared with the word-internal triphone recogniser, some or all context information is lost at the syllable boundaries in the case of syllables that do not correspond with monosyllabic words. 50% of all the target syllable tokens in the training data corresponded with monosyllabic words and did not therefore lose any context information. 17% of the tokens occurred as the first syllable and 24% as the last syllable of a multisyllabic word. Hence, right context information was lost for the last phones of the word-initial syllables and left context information for the first phones of the word-final syllables. 9% of the tokens appeared word-internally and lost both left and right context information. Second, adding parallel paths to the syllable models essentially translates into adding pronunciation variants into the recogniser. It is well known that modelling pronunciation variation by adding transcription variants in the lexicon is not straightforward because of the resulting increase in lexical confusability [13]. Similarly, the parallel paths of the multi-path syllable models are increasing the lexical confusability. The extent of the problem becomes clear when one considers the fact that half of the target syllable tokens corresponded with monosyllabic words and that 91% of these tokens corresponded with function words. Function words are typically of lower information valence than content words and, therefore, pronounced in a highly reduced fashion [7]. Consequently, our approach produced short, easily confusable model paths for monosyllabic function words. For instance, the transcription variant /d/ was one of the MDVs for both of the Dutch definite articles ‘de’ and ‘het’. In cases where a definite article is directly followed by a noun, the bigram language model should be able to help. However, if there is an adjective between the article and the noun, the bigram language model is left powerless. In other words, all the confusability that the parallel paths caused in such cases translated into confusability at the word level, and – when the language model could not assist in solving the problem – had a direct impact on the WER. In the case of multisyllabic words, the syllables that are modelled with triphones may save the word from being misrecognised. However, this is more likely if the syllable that is modelled with a multi-path syllable model is not a word-initial or word-final one.

We can see the effect of the lost context information and the increased lexical confusion as the dramatic 3-percentage-point increase in WER between the triphone recogniser and the untrained mixed-model recogniser. Interestingly, Baum-Welch re-estimation is able to recover from the problems to a large extent, decreasing the WER by 1.9 percentage points. In effect, the re-estimation can incorporate coarticulation-related variation into the syllable models, repairing the effect of suboptimal initialisation of the parallel paths, and re-introduce at least some context information (some of the target syllables appear in very homogeneous contexts). The KLDs that we see between the initialised and the re-trained parallel paths in Figure 2 are directly related to these changes in the paths. However, Baum-Welch re-estimation will never

be able to alleviate the problem of lexical confusability. Therefore, the performance of the re-trained mixed-model recogniser remains significantly lower than that of the triphone recogniser.

One might argue that we could improve the performance of the mixed-model recogniser by refining our MDV selection approach. We could certainly devise ways of eliminating suboptimal transcription variants from being used as MDVs and avoiding MDVs that would result in overlapping pronunciations with existing words in the lexicon. However, it is difficult to see how pronunciation variants could be added without increasing the confusability of the lexicon. Accounting for pronunciation variation by means of (context-independent) syllable models seems to introduce an unexpected problem. All variants are invariably applied to all words in which a given syllable occurs, even if some of the variants may only occur in other contexts. In this sense, adding variants to a strictly phonemic lexicon offers a much higher degree of control.

To conclude, we started from the hypothesis that the richer topology of multi-path syllable models would be better at accounting for pronunciation variation than triphone models that merely have more model parameters organised along a single path. We assumed that re-estimating multi-path syllable models initialised with MDVs would ‘specialise’ the model paths to such an extent that lexical confusability would not be a problem. However, this turned out not to be the case. The re-estimation essentially takes us from the symbolic level to a subsymbolic level but this is not enough to avoid the problem of lexical confusability. To a large extent, the problem could be attributed to syllables that corresponded with monosyllabic function words and had short, easily confusable paths. Yet, these are the words that have the highest amount of pronunciation variation and have a sufficient amount of training data available for constructing syllable-length models.

5. Conclusions

In this paper, we constructed multi-path models for frequent syllables. From a set of manual phonetic transcriptions, we automatically selected up to three ‘major, distinct transcription variants’ (MDVs) for each syllable. We then used triphones underlying these MDVs for initialising the topologies and observation densities of the parallel paths of the multi-path syllable models. The model parameter re-estimation was left to the Baum-Welch algorithm. We analysed the shift from a sequence of initialisation triphones to re-trained parallel paths from two points of view. First, we investigated how the probability density functions of the HMM states of the parallel paths change during re-estimation. Second, we compared the speech recognition performance of the untrained and the re-trained multi-path syllable models with each other. In addition, we compared the performance of the multi-path syllable models with that of triphones. The analysis of the evolution of the syllable models paths illustrated the changes taking place during model parameter estimation. These changes corresponded with changes in the recognition performance when going from the initialised to the re-trained multi-path syllable models. The re-estimation did incorporate coarticulation-related variation into the syllable models, repairing the effect of suboptimal initialisation of parallel paths, and introduced at least some context information to the initially context-free syllable models. However, the addition of parallel paths into the syllable models introduced unexpected lexical confusability in

the recogniser. Therefore, compared with the triphone recogniser, the lexical confusability increased, resulting in a significant decrease in the recognition performance. The main contribution of this paper, then, is to provide insights into the issues playing a role in pronunciation variation modelling with multi-path syllable-models. These issues illustrate the inherent difficulty of pronunciation variation modelling, whatever the approach.

6. Acknowledgements

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

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] L. Deng, D. Yu, and A. Acero. “Structured speech modeling,” *IEEE Transactions on Audio, Speech and Language Processing (Special Issue on Rich Transcription)*, Vol. 14(5), pp. 1492-1504.
- [4] A. Härmäläinen, 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] A. Härmäläinen, 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.
- [6] A. Härmäläinen, L. ten Bosch, and L. Boves, “Modelling pronunciation variation using multi-path HMMs for syllables”, in *Proceedings of ICASSP-2007*, Honolulu, HI, USA. Apr 15-20, 2007.
- [7] S. Greenberg, “Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation”, *Speech Communication*, 29:159-176, 1999.
- [8] 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.
- [9] 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.
- [10] B. Elffers, C. Van Bael, and H. Strik, *ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcriptions*, technical report, Radboud University Nijmegen, The Netherlands, 2005.
- [11] 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, UK, 2002.
- [12] S. Kullback, and R. Leibler, “On information and sufficiency”, *Annals of Mathematical Statistics*, 22:79-86, 1951.
- [13] J. Kessens, C. Cucchiari, and H. Strik, “A data-driven method for modeling pronunciation variation,” *Speech Communication*, 40:517-534, 2003.