Abstract

In this paper, we construct multi-path syllable models using phonetic knowledge for initialisation of 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 5-hour subset of the training data contained manually verified broad phonetic 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.

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 or 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
4.2. Speech recognition

Table 1. \( \text{WERs with a } 95\% \text{ confidence interval.} \)

<table>
<thead>
<tr>
<th>Recogniser type</th>
<th>WER (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>64-G triphone</td>
<td>7.6 ± 0.4</td>
</tr>
<tr>
<td>64-G mixed-model – untrained</td>
<td>10.6 ± 0.4</td>
</tr>
<tr>
<td>64-G mixed-model – re-trained</td>
<td>8.7 ± 0.4</td>
</tr>
</tbody>
</table>
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


