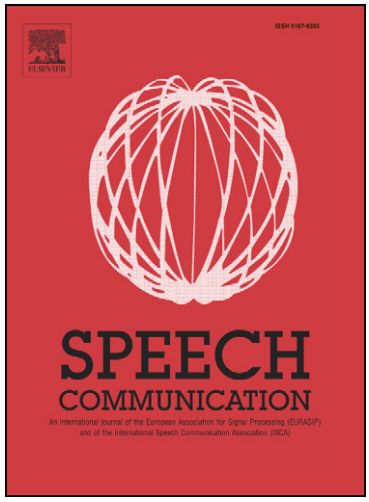
Modelling pronunciation variation with single-path and multi-path syllable models: Issues to consider

Annika Hämäläinen, Louis Ten Bosch, Lou Boves



Modelling pronunciation variation with single-path and multi-path syllable models: Issues to consider

Annika Hämäläinen, Louis ten Bosch, Lou Boves

PII: S0167-6393(08)00109-X
DOI: 10.1016/j.specom.2008.07.001
Reference: SPECOM 1738

To appear in: *Speech Communication*

Received Date: 15 August 2007
Revised Date: 7 July 2008
Accepted Date: 8 July 2008

Please cite this article as: Hämäläinen, A., Bosch, L.t., Boves, L., Modelling pronunciation variation with single-path and multi-path syllable models: Issues to consider, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.07.001

MODELLING PRONUNCIATION VARIATION WITH SINGLE-PATH AND MULTI-PATH SYLLABLE MODELS: ISSUES TO CONSIDER

Annika Hämäläinen, Louis ten Bosch and Lou Boves

Centre for Language and Speech Technology (CLST)

Faculty of Arts

Radboud University Nijmegen

P.O. Box 9103

6500 HD Nijmegen

The Netherlands

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

Corresponding author: Annika Hämäläinen

Tel. +31 24 361 60 45

Fax: +31 24 361 29 07

Abstract

In this paper, we construct context-independent single-path and multi-path syllable models aimed at improved pronunciation variation modelling. We use phonetic transcriptions to define the topologies of the syllable models and to initialise the model parameters, and the Baum-Welch algorithm for the re-estimation of the model parameters. 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 the left and right neighbours, or single-path syllable models whose power of modelling segmental variation would seem to be limited. However, both context-dependent phone models and single-path syllable models outperform multi-path syllable models on a large vocabulary continuous speech recognition task. Careful analyses of the errors made by the recognisers with single-path and multi-path syllable models show that the most important factors affecting the speech recognition performance are syllable context and lexical confusability. In addition, the speech recognition results suggest that the benefits of the greater acoustic modelling accuracy of the multi-path syllable models can only be reaped if the information about the syllable-level pronunciation variation can be linked with the word-level information in the language model.

Keywords: ASR, HMM, topology, syllable, pronunciation variation

1. Introduction

One of the most fundamental characteristics of speech is its variability. In fact, the way a word is pronounced is different each time that it is uttered – whether by different speakers or by the same speaker (Strik and Cucchiarini, 1999). The inter-speaker variation results from differences in the speakers' vocal tract length, age, gender, accent etc. The intra-speaker variation, on the other hand, can be caused by, for instance, coarticulation, prosodic factors, articulation rate, and changes in the emotional and physical state of the speaker (Wester, 2002).

Because of pronunciation variation and the complex acoustic patterns following from it, and because of the practical limitations that until recently have prevented the use of exemplar-based models of speech, speech has conventionally been decomposed into shorter segments for the purpose of automatic speech recognition (ASR). Consequently, the same way as phonological analysis, most large-vocabulary continuous speech recognisers rely on the assumption that speech can adequately be represented as a sequence of discrete phones ('beads on a string') (Ostendorf, 1999). The most obvious problem with this assumption, i.e. the fact that the articulatory and acoustic properties of those 'beads' strongly depend on their neighbours in the 'string', is dealt with by introducing context-dependent phone models, such as triphones. With reasonable amounts of training data and state tying to deal with unseen triphones, triphones allow for robust training. Detailed analysis of natural speech (Greenberg, 1999; Johnson, 2004; Saraclar and Khudanpur, 2000) has, however, shown that a single string of triphones is often not enough for dealing with pronunciation variation. Therefore, 'explicit' pronunciation variation modelling involves listing multiple alternative phonetic representations of words in phonetic lexicons (Wells, 2000), as well as in the lexicons used in large vocabulary automatic speech recognisers. In ASR, explicit pronunciation variation

modelling has, however, met with limited success because of the increased lexical confusability (Kessens et al., 2003). Furthermore, while triphones are able to capture short-span contextual effects such as phoneme substitution and reduction (Jurafsky et al., 2001b), there are complexities in speech that triphones fail to capture. Coarticulation effects, for instance, often stretch beyond the left and right neighbouring phones. The corresponding long-span spectral and temporal dependencies are not easy to capture with models that have as limited a window size as triphones (Ganapathiraju et al., 2001). Moreover, the pronunciation variants in the lexicon do not cover all variation in actual speech production (McAllaster and Gillick, 1999; Saraclar and Khudanpur, 2000; Saraclar et al., 2000).

To alleviate the problems of the 'beads on a string' representation of speech, several authors propose modelling the spectral and temporal variation in speech 'implicitly' by using longer-length linguistic units as the basic building blocks of speech (Ganapathiraju et al., 2001; Hämäläinen et al., 2007a; Jones et al., 1997; Jouvet and Messina, 2004; Plannerer and Ruske, 1992; Sethy and Narayanan, 2003; Sethy et al., 2003). For various reasons, most of these authors (Ganapathiraju et al., 2001; Hämäläinen et al., 2007a; Jones et al., 1997; Jouvet and Messina, 2004; Sethy and Narayanan, 2003; Sethy et al., 2003) suggest using syllable-length models. First, using syllables allows for a relatively compact representation of speech, while maintaining a manageable level of recogniser complexity. Second, support for syllables (or their articulatory and perceptual reality) comes from studies of human speech production and perception. Interestingly, Sethy and Narayanan's (2003) experimental findings also suggest that most of the long-span acoustic correlations are limited to the duration of syllables. Third, syllables are relatively stable as linguistically relevant units, as illustrated by Greenberg's (1999) finding that the syllable deletion rate of spontaneous speech is as low as 1%, as compared with the 12% deletion rate of phones. Johnson (2004) reported a syllable mismatch

rate of 7.6% for content words and 5% for function words in a corpus of spontaneous interviews. A 'mismatch' is a word that has a different number of syllables in its actual realisation than in its canonical lexical representation. The large majority of the mismatches in Johnson's corpus were deletions. Although this may cast some doubt on the stability of the syllable as a linguistic unit, Johnson also advocates a 'nonsegmental modelling' (i.e. implicit) approach to pronunciation variation modelling. More specifically, he suggests that modelling pronunciation variation with phoneme-based segmental models in the lexicon – whether it is with one or more pronunciation variants – is not sufficient to capture the highly detailed nature of acoustic variability. Instead, he speaks for nonsegmental multiple-entry models of speech that are able to capture this kind of detailed acoustic variability.

The most important challenge of using syllable models in large-vocabulary continuous speech recognition is the inevitable sparseness of data in the model training. Many languages – including Dutch – have several thousands of syllables, some of which will have very low occurrence counts in a medium-sized training corpus (such as the 37-hour corpus used in this research) and will therefore not have enough acoustic data for reliable model parameter estimation. The data sparseness problem is more severe for syllables than for triphones: on average, syllables cover a much longer stretch of speech than triphones and their modelling, therefore, requires a much larger number of states. Furthermore, as the syllables comprise more phones, increasingly complex types of articulatory variation must be accounted for. Because of the large number of syllables and the large number of syllable contexts they may appear in, it is very difficult to create context-dependent syllable models. Thus, more accurate modelling of the acoustic patterns within the syllable boundaries may go at the cost of modelling the effects of the contexts in which the syllables appear. This raises the question

whether the advantage of more accurate modelling of within-syllable variation may be annihilated by the lack of context modelling.

The solutions suggested for the data sparseness problem are two-fold. First, syllable models with a sufficient amount of training data are used in combination with triphones (Ganapathiraju et al., 2001; Hämäläinen et al., 2007a; Jouvet and Messina, 2004; Sethy and Narayanan, 2003; Sethy et al., 2003). In other words, triphones are backed off to when a given syllable does not occur frequently enough for reliable model parameter estimation. Second, to ensure that a relatively small amount of training data is sufficient, the syllable models are cleverly initialised (Hämäläinen et al., 2007a; Jouvet and Messina, 2004; Sethy and Narayanan, 2003; Sethy et al., 2003). Sethy and Narayanan (2003), for instance, suggest initialising the single-path syllable models with the parameters of the biphones and triphones underlying the canonical transcription of the syllables (see Figure 1). Subsequent Baum-Welch re-estimation is expected to incorporate the coarticulation- and reduction-related spectral and temporal dependencies in speech into the initialised models by adjusting the means and variances of the Gaussian components of the mixtures associated with the HMM (Hidden Markov Model) states of the syllable models.

FIGURE 1 HERE

Figure 1: Single-path model for the syllable /har/, with the single path through the model initialised with the biphones and triphones underlying the canonical syllable transcription (Hämäläinen et al., 2007a; Sethy and Narayanan, 2003). The phones before the minus sign and after the plus sign in the notation denote the left and right context in which the context-dependent phones have been trained. The hashes in the biphones denote the boundaries of the context-independent syllable model.

Because of the data sparseness problem mentioned above, most previous studies of implicit pronunciation variation modelling with syllable models (Ganapathiraju et al., 2001;

Hämäläinen et al., 2007a; Sethy and Narayanan, 2003; Sethy et al., 2003) have used context-independent single-path syllable models. To the best of our knowledge, only Jouvet and Messina (2004) have attempted to build context-dependent single-path syllable models. However, the improvements in recognition performance that they achieved on tasks with a limited vocabulary size were, overall, comparable with those achieved in studies with context-independent single-path syllable models. This may be an indication that the amount of training data they had available was not enough to capture all the relevant context effects. However, it may also be the case that model topologies with a single path are not able to capture the relevant variation, irrespective of the amount of training data available. This is because syllable-length speech segments display considerable variation in the identity and number of phonetic symbols that best reflect their pronunciation (Greenberg, 1999). In fact, our previous work suggests that re-estimating the acoustic observation densities of single-path syllable models is not sufficient to account for the many different forms that syllable pronunciations can assume (Hämäläinen et al., 2007a).

In the early days of ASR based on HMMs, Lee (1989) proposed a multi-path topology for phone models, inspired by phonetic knowledge about assimilation and reduction processes. The longest path consisted of three states with self loops, whereas two shorter paths were aimed at modelling reduced pronunciations. Speech recognition experiments subsequently showed that a single-path model consisting of three states was sufficient to capture all the variation within a phone. However, for syllable models, which have to capture more complex pronunciation variation than phone models, more intricate topologies of the kind proposed by Lee might be advantageous. The problem of bootstrapping these more intricate models is the price we have to pay for more modelling power. In this study, we decided to use phonetic transcriptions to define the topologies and to initialise the model parameters of the parallel

paths of multi-path syllable models. More specifically, we used biphones and triphones underlying 'major, distinct transcription variants' (MDVs) for this purpose. Figure 2 presents an example of an MDV-based multi-path syllable model. In a way, re-estimated multi-path syllable models correspond to the nonsegmental multiple-entry representations proposed by Johnson (2004).

FIGURE 2 HERE

Figure 2: Multi-path model for the syllable /har/, with the three parallel paths initialised with the triphones underlying the 'major, distinct transcription variants' /ar/, /har/ and /ha/, respectively.

Many of the earlier studies on syllable models (Ganapathiraju et al., 2001; Jouvet and Messina, 2004; Sethy and Narayanan, 2003; Sethy et al., 2003) present speech recognition results without in-depth analysis of the aspects of pronunciation variation that the models are actually able to capture. The goal of this paper is to fill that gap. We aim to investigate the effects of within-syllable pronunciation variation and syllable context from the point of view of speech recognition performance. We attempt to interpret our findings in the context of segmental (explicit) versus nonsegmental (implicit) modelling of pronunciation variation. To reach our goal, we construct single-path and multi-path models for a set of 94 frequent 'target syllables'. We use these syllable models to represent monosyllabic words, constituent syllables of polysyllabic words, or both. In the final 'mixed-model' recognisers, the syllable models are combined with triphone models that cover the other syllables in a Dutch read speech recognition task. In addition, for a baseline, we build a word-internal triphone recogniser. To obtain insights into the factors under investigation, we study the evolution from untrained to retrained syllable models. First, we compare the speech recognition performance of the mixed-model recognisers with untrained and retrained syllable models with each other and with the performance of the baseline triphone recogniser. Second, we

analyse the word-level and sentence-level errors made by the most revealing mixed-model recognisers both before and after the Baum-Welch re-estimation.

This paper is further organised as follows. In Section 2, we describe the speech material used in the study, and discuss the issues concerning the selection of model topologies and parameter initialisation techniques. We also introduce the concept of MDVs, and describe their selection process. In Section 3, we detail the experimental set-up, including the acoustic model training. We present the results from the recognition experiments in Section 4, and analyse and discuss the speech recognition results in Section 5. We further discuss the issues at hand in Section 6, and suggest possible directions for future research in Section 7. In Section 8, we present our conclusions.

2. Method

2.1.Speech Material

We used read speech extracted from the Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN) (Oostdijk et al., 2002), consisting of novels read out loud for a library for the blind. We divided a total of 41 hours of speech into three non-overlapping sets comprising fragments from 303 speakers: a set for training the acoustic models, a development set for optimising the language model scaling factor, the word insertion penalty and the optimal number of Baum-Welch re-estimation rounds, and a test set for evaluating the acoustic models. Table 1 presents the main statistics of the speech material, and Table 2 the syllabic structure of the word tokens in the corpus.

Table 1: Main statistics of the speech material.

TABLE 1 HERE

Table 2: The syllabic structure of the word tokens in the corpus.

TABLE 2 HERE

A 6.5-hour subset of the training data contained manually verified broad phonetic transcriptions and word-level segmentations of the speech. We obtained a list of plausible transcription variants for all the syllables in the manually verified subset by aligning the manual phonetic transcriptions of word tokens with their canonical counterparts. For the alignment process, we used 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 and using a fixed penalty for deletions and insertions (Elffers et al., 2005). To ensure syllable-level alignment, we utilised the syllable boundaries that were available for the canonical transcriptions in the CGN lexicon and CELEX (Baayen et al., 1995) in the alignment process.

Using the transcription variants retrieved for the target syllables and canonical transcriptions for the rest of the syllables, we performed a forced alignment of the training data with 8-Gaussian triphones (see Section 3.3.1) to determine which transcription variants best represented the target syllables in the part of the corpus that only came with orthographic transcriptions. For instance, the canonical transcription of the bisyllabic word 'nadruk' ('emphasis') is /nadrYk/. As the first syllable /na/ belonged to the set of target syllables because of its high frequency, we fed the forced alignment process with all the four transcription variants observed in the manually verified subset (corresponding to the following sequences of biphones: /#-n+a n-a+#/, /#-n+@ n-@+#/, /#-n+A n-A+#/ and /#-N+a N-a+#/) and were, therefore, able to ascertain which variants acoustically best matched the relevant stretches of the speech signal. Since the second syllable /drYk/ did not belong to the set of target syllables, it was always labelled as the canonical sequence /#-d+r d-r+Y r-Y+k Y-

k+#/. To ensure that the complete training corpus was consistently handled in the same manner, we also applied the forced alignment procedure to the manually transcribed part of the data.

When building the single-path and multi-path mixed-model recognisers, we concentrated our modelling efforts on a set of 94 most frequent syllables found in the manually verified subset (Hämäläinen et al., 2007a). All of the target syllables appeared as part of polysyllabic words, and 71 of them also appeared as monosyllabic words. The target syllables covered 57% of all the syllable tokens in the training data, the least frequent of them occurring 850 times and the most frequent 35 000 times. 50% of all the target syllable tokens in the training data corresponded to monosyllabic words and, when modelled with context-independent syllable models, did not lose any context information as compared with the baseline word-internal triphone recogniser. An example of such a target syllable is /har/ (see Figures 1 and 2), which corresponds to the monosyllabic word 'haar' (the possessive pronoun 'her' or the noun 'hair'). 17% of the target syllable tokens occurred as the first syllable and 24% as the last syllable of a polysyllabic word. The last phone of the word-initial syllables lost right context information, whereas the first phone of the word-final syllables lost left context information. Examples of such cases are the target syllables /x@/ and /d@/, which appear, for instance, as the first and the last syllable of the words 'geleerd' (the past participle form of the verb 'to learn') and 'belde' (the singular imperfect form of the verb 'to call'), respectively. 9% of the target syllable tokens appeared word-internally and lost both left and right context information. An example of such a case is the target syllable /ni/, which appears, for example, as the third syllable of the word 'anonimiteit' ('anonymity'). The target syllables had an average of 8.7 transcription variants per syllable, with the actual number of variants differing from 1 to 27. Since the manually verified subset is representative of the whole corpus, we are

confident that the transcription variants that we retrieved cover all reasonable transcriptions of the target syllables.

Our corpus contained read speech. Even though read speech is not representative of all the problems that are typical of spontaneous speech (hesitations, restarts, repetitions etc.), the kinds of fundamental issues related to articulation that this paper addresses are present in *all* speech styles. In fact, using spontaneous speech would have added complexity into the recogniser that would have made it more difficult to isolate the effects of the kinds of articulatory issues we were interested in. An alternative for using syllable transcription variants derived from the manually verified subset of training data would have been to generate transcription variants using phonological rules for Dutch (e.g. Booij, 1999) and then perform a forced alignment with these transcription variants to determine which transcription variants best represented the target syllables in the training data. Yet, for our experiments, which were to test the validity of our method, we wanted to have as accurate transcription variants and as reliable information about their frequency as possible. We did not want to take the risk of omitting transcription variants or generating noise by using automatically derived transcription variants. Therefore, we decided to use the manually verified phonetic transcriptions available in CGN.

2.2.Selection of major, distinct transcription variants for the initialisation of multi-path syllable models

If the amount of data available for the re-estimation of the acoustic observation densities of single-path syllable models is already an issue (see Section 1), the situation is only more difficult for multi-path models. Therefore, the optimal initialisation of the parallel paths is of utmost importance. To accomplish this, we decided to initialise each path using the

parameters of the sequence of triphones that is most representative of the path in question. We obtained these representative sequences of triphones using the concept of 'major, distinct transcription variants' (MDVs). The identification of MDVs was guided by two principles. First, we wanted the MDVs to be as frequent as possible ('major'), while at the same time as different from each other as possible ('distinct'). Second, we had a preference for MDVs containing fewer symbols than the canonical variant. This preference stemmed from the high frequency of phone deletions reported in the literature (Greenberg, 1999; Johnson, 2004).

Except for the fact that one probably should not exceed the number of transcription variants observed amongst the manually verified phonetic transcriptions, it is not a priori evident how many different paths one should include in the topologies of multi-path syllable models. There are at least two criteria that should be taken into account:

(1) To reliably re-estimate the acoustic observation densities of the multi-path syllable models, a minimum number of training tokens is needed. A good estimate would be the minimum number of training tokens needed for the robust training of single-path syllable models multiplied by the number of the parallel paths in the multi-path syllable model.

(2) To add an extra path, it must be possible to initialise it with a sequence of triphones that guarantees a sufficiently large distance to the paths that are already present in the model.

To avoid an unnecessarily complex procedure, we decided to use all the transcription variants for building parallel paths for the syllables that only had up to three transcription variants (10 % of all the target syllables). For the syllables that had more than three transcription variants, we used the concept of MDVs to select the variants that best represented three maximally

different pronunciation variants. Three parallel paths per syllable appeared a good compromise between too little training data and too small a distance between the triphone sequences used to initialise the paths. Our assumption about the optimal number of paths could later be verified by carrying out a forced alignment of the training data with the syllable models; the majority of the paths were frequently entered (Hämäläinen et al, 2007b). In addition, removing the paths that were rarely used during the forced alignment showed that the recognition performance remained virtually unchanged (Hämäläinen et al, 2006).

We devised the following steps for selecting the optimal MDV triplet for each target syllable:

(1) Count the frequency of each transcription variant of the target syllable in the training data.

(2) Compute a matrix with articulatory distances between all transcription variant pairs for the target syllable. To compute the distances, we used the same feature-based algorithm as we did when aligning the manual and canonical transcriptions to find the transcription variants for the syllables (see Section 2.1).

(3) Compile a ranked list of transcription variant triplets, each variant of which optimally serves as a centroid of variant clusters, given the distances between and the frequencies of all the variants. The criterion for optimality is the overall distance of all variants to their closest centroid, multiplied with the frequency of the variant. This means that variants are more likely to be part of a high-ranking triplet if the variant is more frequent and/or more distinct from the other variants. For instance, the triplet /hAt/-/hat/-/At/ ranked the highest for the syllable /hAt/, whereas the triplet /Ad/-/jAt/-/jA/ ranked the lowest – mainly because of the low frequencies of the variants in question.

(4) Post-process the list produced in Step 3 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 triplet that contains at least one variant with at least one symbol less than the canonical. When none of the triplets satisfies the length criterion, select the highest-ranking triplet. The variants included in the selected triplet are the MDVs used in the initialisation of the HMM paths.

In practice, one of the MDVs for all of the target syllables was the canonical transcription itself. 85% of the bi- and triphonemic target syllables (81% of all the target syllables) had one or two MDVs with fewer phones than the canonical, whereas 39% of all the target syllables had one MDV with more phones than the canonical.

3. Experimental set-up

3.1.Feature extraction

We carried out the feature extraction at a frame rate of 10 ms using a 25-ms Hamming window and applied first order pre-emphasis to the signal using a coefficient of 0.97. For a total of 39 features, we calculated 12 Mel Frequency Cepstral Coefficients (MFCCs) and log-energy with first and second order derivatives. We applied channel normalisation using cepstral mean normalisation over complete recordings, and then chunked the recordings to sentence-length entities for creating the language model and carrying out the recognition experiments.

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 baseline triphone recogniser, this single pronunciation comprised a string of canonical phones from the CGN lexicon. In the case of the mixed-model recognisers, it consisted of the following:

a) syllable units,

b) canonical phones, or

c) a combination of a) and b).

To use the bisyllabic word 'wereld' ('world') as an example, the possible pronunciations were the following:

a) /we r@lt/,

b) /#-w+e w-e+r e-r+@ r-@+l @-l+t l-t+#/,

c_1) /we #-r+@ r-@+l @-l+t l-t+#/, or

c_2) /#-w+e w-e+# r@lt /.

The syllable /we/ belonged to the list of 94 target syllables, whereas the syllable /r@lt/ did not. Therefore, c_1) was the actual representation in the lexicon.

One of the issues to consider when building syllable models is ambisyllabic consonants, i.e. consonants at syllable boundaries that belong, in part, to both the preceding and the following syllable. Unlike Ganapathiraju et al. (2001), who assigned ambisyllabics to both syllables, we decided to assign them to the following syllable only. We had two main motivations to do so. First, one of the main issues that we wanted to address with the syllable models was reduction, which often manifests itself as durational reduction. Hence, we did not want to add any more states into the syllable models by assigning the ambisyllabics to both the preceding and the following syllable. Second, assigning the ambisyllabics to both syllables would have resulted in a larger set of syllable models. This would have inevitably resulted in a decrease in the amount of data available for training each syllable model. Therefore, our choice can be

seen as a trade-off between the (linguistic) accuracy of the models and the amount of the data available for training them.

We built a word-level bigram network for the task using the data in the training, test and development test sets. The purpose of this seemingly unconventional choice was to allow us to study changes in acoustic modelling only, without the risk of language modelling issues masking the effects. As a consequence of this choice, the out-of-vocabulary (OOV) rate was zero. The test set perplexity, computed on a per-sentence basis using HTK (Young et al., 2002), was 92.

3.3. Acoustic modelling

We used HTK (Young et al., 2002) as the speech recognition platform. Because of the large number of contexts that the target syllables appeared in, building context-dependent syllable models would have exploded the number of models in the recogniser. This would have necessitated the use of state tying between different syllable models and the parallel paths of these syllable models. As there is no straightforward way to implement this with HTK, we built context-independent single-path and multi-path syllable models for our mixed-model recognisers.

Both in terms of context modelling and the total number of states in the recognisers, a word-internal triphone recogniser was the most comparable conventional phone-based recogniser to compare the context-independent syllable models with. To facilitate the analysis of our results, we took a word-internal triphone recogniser as the starting point and took carefully controlled steps to build the experimental recognisers. First, we built an “impaired” triphone recogniser in which context information was removed at the boundaries of the target syllables

within polysyllabic words (see Section 3.3.2). Second, we constructed single-path mixed-model recognisers in which context-independent syllable models were included for the target syllables in monosyllabic or polysyllabic words only, or in both monosyllabic and polysyllabic words (see Section 3.3.3). Third, we repeated the exercise with multi-path syllable models (see Section 3.3.4).

To study the stepwise changes from untrained to retrained single-path and multi-path syllable models, we evaluated the performance of the single-path and multi-path mixed-model recognisers both before and after the Baum-Welch re-estimation. In addition, we analysed the word-level and sentence-level recognition errors of the single-path and multi-path mixed-model recognisers that could teach us the most about the different factors playing a role in pronunciation variation modelling with syllable models. We also compared the performance of the single-path and multi-path mixed-model recognisers with that of the baseline triphone recogniser. This section details the acoustic model training procedures used in building the recognisers.

3.3.1. Baseline triphone recogniser

We used a standard procedure with decision tree state tying to train the word-internal triphone recogniser. The procedure was based on asking yes/no questions about the left and right contexts of each triphone; the decision trees attempted to find the contexts that made the largest difference to the acoustics and that should, therefore, distinguish clusters. The questions at each node of the decision trees were chosen to locally maximise the likelihood of the training data given the final set of state tyings (Young et al., 2002).

We first trained initial 32-Gaussian monophones for 37 'native' Dutch phones using linear segmentation of canonical transcriptions within automatically generated word segmentations. After that, we used the monophones to perform a forced alignment of the training data, and bootstrapped the triphones using the resulting phone segmentations. When carrying out the state tying, the minimum occupancy count that we used for each cluster resulted in approximately 3 500 distinct triphones in the recogniser. Table 3 presents the recogniser complexity in terms of the total number of distinct states in the recogniser. We trained triphone recognisers with up to 128 Gaussians per state, and optimised the values for the language model scaling factor and the word insertion penalty. The 64-Gaussian triphone recogniser was the best performing triphone recogniser, and was therefore used as the baseline triphone recogniser.

Table 3: The complexities for the following recognisers: baseline triphone recogniser (TR), single-path mixed-model recogniser with the target syllables covered with syllable models in monosyllabic words (SPM), single-path mixed-model recogniser with the target syllables covered with syllable models in polysyllabic words (SPP), single-path mixed-model recogniser with the target syllables covered with syllable models in both monosyllabic and polysyllabic words (SP), multi-path mixed-model recogniser with the target syllables covered with syllable models in monosyllabic words (MPM), multi-path mixed-model recogniser with the target syllables covered with syllable models in polysyllabic words (MPP), and multi-path mixed-model recogniser with the target syllables covered with syllable models in both monosyllabic and polysyllabic words (MP). There was no state tying for syllable models. To facilitate a fair comparison, the complexity of the syllable models was estimated with the same tying ratio as that used in building the triphone models.

TABLE 3 HERE

3.3.2. Impaired triphone recogniser

Before building context-independent syllable models for the mixed-model recognisers, we wanted to test the effect of removing the same context information from the baseline triphone recogniser described in Section 3.3.1. In practice, this meant replacing the triphones at the

boundaries of the target syllables within polysyllabic words by biphones. For instance, the word 'behandeling' ('handling') was represented by the following string of triphones in the baseline triphone recogniser: /#-b+@ b-@+h @-h+A h-A+n A-n+d n-d+@ d-@+l @-l+I l-I+N I-N+#/. As the first syllable /b@/ and the third syllable /d@/ belonged to the set of 94 target syllables, they were to lose context in the mixed-model recognisers. This loss of context at the boundaries of these syllables was simulated by using the following pronunciation in the impaired triphone recogniser: /#-b+@ b-@+# #-h+A h-A+n A-n+# #-d+@ d-@+# #-l+I l-I+N I-N+#/. Whenever biphones needed for the impaired triphone recogniser did not exist in the baseline triphone recogniser, we synthesised them (i.e. tied them to existing triphones) using the decision trees described in Section 3.3.1 (Young et al., 2002).

We tested impaired triphone recognisers with up to 64 Gaussians per state. We carried out the tests both with optimised values for the language model scaling factor and the word insertion penalty, and with the values that were optimal for the baseline triphone recogniser.

3.3.3. Single-path mixed-model recognisers

When building the single-path mixed-model recognisers, we employed a procedure similar to that used in Hämäläinen et al. (2007a). We initialised the context-independent models for the target syllables by picking the initial syllable state parameters from the biphones and triphones corresponding to the canonical syllable transcriptions (see Figure 1). Some of the biphones necessary for building the syllable models did not exist in the baseline triphone recogniser. These unseen biphones were identical to the unseen biphones in the impaired triphone recogniser, and were tied to existing triphones in exactly the same way. To represent the syllables that were not covered with syllable models, we used the original triphones.

We built three types of single-path mixed-model recognisers:

a) A single-path mixed-model recogniser with the target syllables covered with syllable models in monosyllabic words (SPM). As the baseline triphone recogniser (see Section 3.3.1) did not contain cross-word context information, the untrained version of this type of mixed-model recogniser was essentially identical to it. The only difference was that the biphones and triphones constituting the monosyllabic words in question appeared as separate models in the case of the baseline triphone recogniser, whereas they were bound to the context-independent syllable models in the case of the mixed-model recogniser.

b) A single-path mixed-model recogniser with the target syllables covered with syllable models in polysyllabic words (SPP). As compared with the baseline triphone recogniser, the untrained version of this type of mixed-model recogniser had lost context at the boundaries of the target syllables within polysyllabic words. However, it was essentially identical to the impaired triphone recogniser (see Section 3.3.2). The only difference was that the biphones and triphones constituting the target syllables in the polysyllabic words appeared as separate models in the case of the impaired triphone recogniser, whereas they were bound to the context-independent syllable models in the case of the mixed-model recogniser.

c) A single-path mixed-model recogniser with the target syllables covered with syllable models in both monosyllabic and polysyllabic words (SP). As compared with the baseline triphone recogniser, the untrained version of this type of mixed-model recogniser had also lost context at the boundaries of the target syllables within polysyllabic words. However, it was essentially identical to the impaired triphone recogniser and the single-path mixed-model recogniser with the target syllables covered

with syllable models in polysyllabic words. The only difference was that some or all of the biphones and triphones constituting the target syllables in both the monosyllabic and the polysyllabic words appeared as separate models in the case of the impaired triphone recogniser and in the single-path mixed-model recogniser with the target syllable covered with syllable models in polysyllabic words.

We carried out recognition experiments on the development test set to define the optimal number of Baum-Welch re-estimation rounds for the mixed-model recognisers; one round of Baum-Welch re-estimation resulted in the best performance in all cases. In addition, we optimised the language model scaling factor and the word insertion penalty both before and after the retraining. We trained and tested single-path mixed-model recognisers with up to 64 Gaussians per state. The syllable models were initialised using biphones and triphones with the same number of Gaussians per state as in the final mixed-model recognisers. Table 3 presents the complexity of the single-path mixed-model recognisers in terms of the total number of states.

3.3.4. Multi-path mixed-model recognisers

We followed the steps described in Section 2.2 to select the MDVs for each of the 94 target syllables, and initialised the parallel paths of the corresponding context-independent multi-path models by picking the initial state parameters from the biphones and triphones corresponding to these MDVs (Hämäläinen et al., 2007a; Sethy and Narayanan, 2003). The previously unseen biphones were again synthesised using the decision trees described in Section 3.3.1. Before applying the Baum-Welch algorithm to capture within-syllable coarticulation and reduction effects, we combined the initialised paths into multi-path syllable models such as that shown in Figure 2. In practice, this meant that we did not assign specific

training tokens for the re-estimation of the model parameters of specific parallel paths. Instead, we left the Baum-Welch algorithm to take care of the weighted assignment of the training tokens during the re-estimation. In other words, the Baum-Welch algorithm used each training token to update the model parameters of each parallel path. In addition, the Baum-Welch algorithm updated the transition probabilities of entering the different parallel paths. As a consequence, in the final multi-path syllable models, the probability of entering the path associated with the most common pronunciation was the highest. We chose to use the Baum-Welch algorithm instead of Viterbi training in order to better model the gradual character of pronunciation variation phenomena (such as reduction). While the use of Viterbi training would have entailed the assumption that only one of the parallel paths is 'correct' for each training token, the Baum-Welch algorithm updated the model parameters of *each* parallel path using each training token. In practice, this means that the result of the Baum-Welch algorithm offers a better match between individual syllable tokens and the multi-path syllable models. The result of the Viterbi training does converge to the result of the Baum-Welch algorithm but for a very large number of training tokens only.

We built three types of multi-path mixed-model recognisers:

a) A multi-path mixed-model recogniser with the target syllables covered with syllable models in monosyllabic words (MPM). As the baseline triphone recogniser (see Section 3.3.1) did not contain cross-word context information, the fundamental difference between it and the untrained version of this type of mixed-model recogniser was that adding the parallel paths to the syllable models essentially translated into adding pronunciation variants for the monosyllabic words involved.
b) A multi-path mixed-model recogniser with the target syllables covered with syllable models in polysyllabic words (MPP). As compared with the baseline triphone

recogniser, the untrained version of this type of mixed-model recogniser had lost context at the boundaries of the target syllables within polysyllabic words. In addition, adding the parallel paths to the syllable models again meant adding pronunciation variants for the polysyllabic words involved.

c) A multi-path mixed-model recogniser with the target syllables covered with syllable models in both monosyllabic and polysyllabic words (MP). As compared with the baseline triphone recogniser, the untrained version of this type of mixed-model recogniser had lost context at the boundaries of the target syllables within polysyllabic words. In addition, adding the parallel paths to the syllable models meant adding pronunciation variants for both the monosyllabic and the polysyllabic words involved.

To define the optimal number of Baum-Welch re-estimation rounds for the mixed-model recognisers, we carried out recognition experiments on the development test set; one round of Baum-Welch re-estimation resulted in the best performance in all cases. Both before and after the retraining, we also optimised the language model scaling factor and the word insertion penalty. We trained and tested multi-path mixed-model recognisers with up to 64 Gaussians per state. The parallel paths of the syllable models were initialised using biphones and triphones with the same number of Gaussians per state as in the final mixed-model recognisers. Table 3 presents the complexity of the multi-path mixed-model recognisers in terms of the total number of states.

4. Speech recognition results

Figure 3 presents the most relevant speech recognition results in terms of word error rate (WER). 64 Gaussians per state resulted in the best recognition performance for all recogniser

types. The figure shows the performance of the single-path and multi-path mixed-model recognisers both before and after the Baum-Welch re-estimation for two conditions:

a) with the same language model scaling factor and word insertion penalty as used for the baseline triphone recogniser.

b) with the language model scaling factor and the word insertion penalty optimised for the best possible speech recognition performance.

Table 4 presents the corresponding numbers of insertion, deletion and substitution errors, as well as the corresponding recognition parameter values. Varying between 14 and 18, the language model scaling factor remained stable for all the experimental conditions. On the contrary, the behaviour of the word insertion penalty (modelled in HTK as a word entrance probability) is more interesting. The higher the value of this parameter, the more favourable it becomes to enter a word. In effect, high values of the word insertion penalty lead to word insertions, whereas low values result in word deletions. The fact that the word insertion penalty usually had to be decreased for optimal performance in the case of the recognisers with lost context information (ITR, SPP, SP, MP) and the recognisers with parallel paths (MPM, MP) suggests that the addition of multi-path syllable models into a recogniser affects the weighting between the acoustic and the linguistic models of the recogniser. Qualitatively, it is straightforward to understand this; the introduction of syllable models will, in general, improve the match of the affected words with the signal. Since this improvement only holds for a subset of the words in the lexicon, the entire word competition regime is skewed. Retuning the word entrance penalty and the language model scaling factor can, apparently, only partially compensate for this change.

FIGURE 3 HERE

Figure 3: WERs with a 95% confidence interval for the following recognisers with 64 Gaussians per state: baseline triphone recogniser (TR), impaired triphone recogniser (ITR), single-path mixed-model recogniser with the target syllables covered with syllable models in monosyllabic words (SPM), single-path mixed-model recogniser with the target syllables covered with syllable models in polysyllabic words (SPP), single-path mixed-model recogniser with the target syllables covered with syllable models in both monosyllabic and polysyllabic words (SP), multi-path mixed-model recogniser with the target syllables covered with syllable models in monosyllabic words (MPM), multi-path mixed-model recogniser with the target syllables covered with syllable models in polysyllabic words (MPP), and multi-path mixed-model recogniser with the target syllables covered with syllable models in both monosyllabic and polysyllabic words (MP). The subscript "def" indicates that the language model scaling factor (-s) was kept at 16 and that the word insertion penalty (-p) was kept at 25. The subscript "opt" indicates that the recognition parameters had been optimised for the best possible performance. The dark grey bars for the mixed-model recognisers represent the untrained and the light grey bars the retrained recognisers.

Table 4: The number of insertion, deletion and substitution errors corresponding to the WERs in Figure 3. The subscript "bt" refers to the untrained recognisers (see the dark grey bars in Figure 3) and the subscript "at" to the retrained recognisers (see the light grey bars in Figure 3). -s and -p are the corresponding language model scaling factors and word insertion penalties, respectively.

TABLE 4 HERE

Figure 3 and Table 4 show that, before the recognition parameter optimisation, the speech recognition results are identical for the baseline triphone recogniser and the untrained single-path mixed-model recogniser with the target syllables covered with syllable models in monosyllabic words ($SPM_{def,\ bt}$). Similarly, the results are identical for the impaired triphone recogniser (ITR_{def}) and both the untrained single-path mixed-model recogniser with the target syllables covered with syllable models in polysyllabic words ($SPP_{def,\ bt}$) and the untrained single-path mixed-model recogniser with the target syllables covered with syllable models in both monosyllabic and polysyllabic words ($SP_{def,\ bt}$). This proves that it does not make a difference for the speech recognition performance whether or not the biphones and triphones

constituting the target syllables are "loose", or bound to the context-independent syllable models before training the mixed-model recognisers further.

From the confidence intervals in Figure 3, one can see that most of the untrained single-path mixed-model recognisers ($SPM_{def,\ bt}$, $SPM_{opt,\ bt}$, $SP_{def,\ bt}$, $SP_{opt,\ bt}$) significantly outperformed the corresponding untrained multi-path mixed-model recognisers ($MPM_{def,\ bt}$, $MPM_{opt,\ bt}$, $MP_{def,\ bt}$, $MP_{opt,\ bt}$) both before and after the recognition parameter optimisation. The only exception was the untrained mixed-model recognisers with the target syllables covered with syllable models in polysyllabic words; the recognition results did not differ from each other significantly whether single-path ($SPP_{def,\ bt}$, $SPP_{opt,\ bt}$) or multi-path ($MPP_{def,\ bt}$, $MPP_{opt,\ bt}$) syllable models were used.

Before the recognition parameter optimisation, most of the re-trained single-path mixed-model recognisers ($SPM_{def,\ at}$, $SP_{def,\ at}$) again outperformed the corresponding re-trained multi-path mixed-model recognisers ($MPM_{def,\ at}$, $MP_{def,\ at}$). The only exception was the re-trained mixed-model recognisers with the target syllables covered with syllable models in polysyllabic words; the recognition results ($SPP_{def,\ at}$ and $MPP_{def,\ at}$) were identical. After the recognition parameter optimisation, the re-trained single-path mixed-model recogniser outperformed the re-trained multi-path mixed-model recogniser both in the case of the mixed-model recognisers with the target syllables covered with syllable models in monosyllabic words ($SPM_{opt,\ at}$ vs. $MPM_{opt,\ at}$), and in the case of the mixed-model recognisers with the target syllables covered with syllable models in both monosyllabic and polysyllabic words ($SP_{opt,\ at}$ vs. $MP_{opt,\ at}$). However, the difference in the recognition performance was significant only in the first case. In the case of the mixed-model recognisers with the target syllables

covered with syllable models in polysyllabic words, the recognition results ($SPP_{opt,\ at}$ and $MPP_{opt,\ at}$) were still identical after the recognition parameter optimisation.

As we can see from Figure 3, the retraining usually improved the performance of a recogniser significantly. The only exception was the single-path mixed-model recogniser with the target syllables covered with syllable models in monosyllabic words ($SPM_{def,\ bt}$ vs. $SPM_{def,\ at}$ and $SPM_{opt,\ bt}$ vs. $SPM_{opt,\ at}$). In this case, the re-training did improve the performance of the recogniser but this improvement was very small (0.1 percentage points). It is interesting to notice that the mixed-model recognisers that essentially started off as being identical to the baseline triphone recogniser ($SPM_{def,\ bt}$) and the impaired triphone recogniser ($SPP_{def,\ bt}$ and $SP_{def,\ bt}$) outperformed the corresponding triphone recognisers after the retraining. On the other hand, the recognition parameter optimisation affected the recognition results significantly only in the case of the untrained multi-path mixed-model recogniser with the target syllables covered with syllable models in monosyllabic words ($MPM_{def,\ bt}$ vs. $MPM_{opt,\ bt}$) and in the case of the untrained multi-path mixed-model recogniser with the target syllables covered with syllable models in both monosyllabic and polysyllabic words ($MP_{def,\ bt}$ vs. $MP_{opt,\ bt}$).

The best-performing recogniser was the re-trained single-path mixed-model recogniser with the target syllables covered with syllable models in monosyllabic words ($SPM_{opt,\ at}$). Except for the baseline triphone recogniser, it significantly outperformed all other types of recognisers. Even though the baseline triphone recogniser outperformed the re-trained multi-path mixed-model recogniser with the target syllables covered with syllable models in

monosyllabic words ($MPM_{opt,\ at}$), the difference in the recognition performance was not significant.

5. Discussion of the speech recognition results

The speech recognition results reported in Section 4 confirm our previous finding that the introduction of syllable models does not necessarily result in better speech recognition performance (Hämäläinen et al., 2007a). In the following subsections, we discuss the speech recognition results with respect to the different factors playing a role in pronunciation variation modelling with syllable models. These issues include the following: syllable context, lexical confusability, word-specific pronunciation variation, and long-span spectral and temporal dependencies in speech. We also discuss the effect of the Baum-Welch re-estimation in the context of the aforementioned factors.

5.1.Syllable context

Using context-independent syllable models in the untrained mixed-model recognisers essentially meant sacrificing some or all context information at the syllable boundaries in the case of syllables embedded in polysyllabic words (see Sections 3.3.3 and 3.3.4). From the recognition results, it immediately becomes clear that syllable context is the single most important factor in successful pronunciation variation modelling with syllable models. The effect of losing syllable context information is insulated in the case of the impaired triphone recogniser and the untrained single-path mixed-model recogniser with the target syllables covered with syllable models in polysyllabic words. This is because the loss of syllable context information at the boundaries of the target syllables within polysyllabic words is the only fundamental difference between the baseline triphone recogniser and these two

recognisers. In terms of recognition performance, this loss of syllable context information translated into a drastic 2.7 percentage point deterioration before the recognition parameter optimisation (ITR_{def}, $SPP_{def, bt}$) and a 2.4 percentage point deterioration after the recognition parameter optimisation (ITR_{opt}, $SPP_{opt, bt}$) as compared with the baseline triphone recogniser.

Apart from the recognition results, we can illustrate the effect of the lost syllable context information, as well as the impact of the retraining and the recognition parameter optimisation, using a detailed analysis of the word-level recognition errors made by the optimised single-path mixed-model recogniser with the target syllables covered with syllable models in polysyllabic words both before and after the retraining ($SPP_{opt, bt}$ and $SPP_{opt, at}$). For this analysis, we treated the recognition output of the baseline triphone recogniser as the reference transcriptions. This is because we wanted to show why the mixed-model recogniser performed worse than the baseline triphone recogniser and were, therefore, not so interested in the errors made by *both* the triphone recogniser and the mixed-model recognisers. We first compared the output of the optimised untrained mixed-model recogniser with the output of the baseline triphone recogniser. To analyse the effect of the retraining in the recognition output, we also compared the output of the optimised retrained mixed-model recogniser with the output of the baseline triphone recogniser. Using the output of the baseline triphone recogniser as the reference, 108 (10%) of all the 1122 'errors' made by the optimised untrained single-path mixed-model recogniser with the target syllables covered with syllable models in polysyllabic words ($SPP_{opt, bt}$) were insertions, 274 (24%) deletions and 740 (66%) substitutions[1]. Of the total of 668 errors made by the optimised retrained recogniser ($SPP_{opt, at}$), 107 (16%) were insertions, 99 (15%) deletions and 462 (69%) substitutions[1]. Therefore, substitutions were by far the most important type of errors from the WER point of view.

[1] As the output of the triphone recogniser was used as the reference in the analysis, these figures cannot straightforwardly be related to the figures in Table 4.

Figures 4 and 5 present the numbers of substitution errors before and after the retraining. As we can see from the figures, most of the substituted words contained syllable models both before and after the retraining. There are two reasons why one would expect most of the substitution errors to originate from monosyllabic words. First, monosyllabic words cover 65% of the corpus (see Table 2). Second, polysyllabic words generally exhibit a relatively low WER (Greenberg and Chang, 2000). However, Figure 4 shows that bisyllabic words containing syllable models were the most problematic type of words before the retraining. This finding supports the conclusion we were already able to make based on the speech recognition results; the loss of syllable context information at the boundaries of the target syllables within polysyllabic words is detrimental for the speech recognition performance. The fewer syllables the polysyllabic words have, the more serious the problem (see Figure 4).

FIGURE 4 HERE

Figure 4: The number of substitution errors for words with varying numbers of syllables in the optimised untrained single-path mixed-model recogniser with the target syllables covered with syllable models in polysyllabic words ($SPP_{opt, bt}$)[1]. The errors are shown separately for words that include one or more syllable models and for words that are entirely modelled as a sequence of triphones.

FIGURE 5 HERE

Figure 5: The number of substitution errors for words with varying numbers of syllables in the optimised retrained single-path mixed-model recogniser with the target syllables covered with syllable models in polysyllabic words ($SPP_{opt, at}$)[1]. The errors are shown separately for words that include one or more syllable models and for words that are entirely modelled as a sequence of triphones.

The retraining had the largest effect on polysyllabic words containing syllable models (see Figures 4 and 5). The number of substitution errors reduced as much as 50% (from 313 to 158 errors), and 51% (from 61 to 30 errors) for bisyllabic and trisyllabic words, respectively. The same figure was only 28% (from 263 to 189 errors) for monosyllabic words represented by syllable models. These figures suggest that the retraining was able to reintroduce some of the context information that was lost during the initialisation. In fact, the retraining and the recognition parameter optimisation resulted in a 1.7-percentage-point decrease in the WER (see Figure 3). Nevertheless, even after the retraining, the single-path mixed-model recogniser with the target syllables covered with syllable models in polysyllabic words yielded a significantly higher WER than the baseline triphone recogniser.

We looked further into the substitution errors to check for any potential error patterns, and were indeed able to find systematic errors in the case of polysyllabic words containing one or more syllable models. These errors illustrate exactly how the lost context information affects the recogniser output. There were two main types of systematic errors. First, we saw polysyllabic words with syllable models being substituted by words that were identical to the original word except for the deletion of a syllable (e.g. weg-ge-legd → ge-legd; ge-had → had). In some cases, the deleted syllable had been inserted into the sentence as a separate word. In other cases, it had been deleted altogether. Second, we saw polysyllabic words with syllable models being substituted by words corresponding to a part of the original word, rather than a syllable or several syllables of the original word (e.g. ja-ren → jaar). 161 errors exemplified these two types of substitution errors before the retraining. After the retraining, the same figure was 71. In other words, the retraining was able to reduce these systematic errors by 56%.

An example of a case in which a polysyllabic word with a syllable model had erroneously been substituted by two words, can be seen in the following sentence pair. The word ‘weggelegd’ (the past participle form of the verb ‘to lay aside’) had been substituted by the word ‘gelegd’ (the past participle form of the verb ‘to place’) but the word ‘weg’ (‘away’) had been inserted as a word on its own.

Baseline triphone recogniser:	die al is in de pap	weggelegd	voor ‘t zout en de specerijen
Mixed-model recogniser:	die al is weg in de pap	gelegd	voor ‘t zout en de specerijen

As the word ‘weggelegd’ was modelled with the model sequence /#-w+E w-E+# G@ #-l+E l-E+x E-x+t x-t+#/ (i.e. context information was lost between the last phone of the syllable ‘weg-’ and the first phone of the syllable ‘-ge-’ during the initialisation), this seemed to be a case of the lack of context information affecting the recogniser output. However, these types of errors are – of course – also related to the value of the word insertion penalty. As this particular error occurred both before and after the retraining, the retraining or the recognition parameter optimisation had not been able to correct it.

An example of a case in which a polysyllabic word with a syllable model had been substituted by a word that is identical to the original word except for the deletion of a syllable, and in which the deleted syllable had completely been deleted, can be seen in the following sentence pair. The pronominal adverb ‘erop’ had been substituted by the locative adverb ‘er’. The preposition ‘op’ (‘on’) had been deleted altogether.

Baseline triphone recogniser: wat stond erop

Mixed-model recogniser: wat stond er

The word ‘erop’ was modelled with the two syllable models /Er/ and /Op/, with context information lost at the syllable boundary. In this case, the sentence was correctly recognised after the retraining. This suggests that the retraining had reintroduced the lost context information. Considering the fact that ‘erop’ is a frequently occurring word, i.e. that both of the syllables frequently appeared in each other’s context in the training data, this is not surprising.

More often than the above type of cases, however, we saw polysyllabic words with syllable models being substituted by a word corresponding to a part of the original word. For example, before the retraining, the word ‘jaren’ (‘years’) was substituted by the word ‘jaar’ (‘year’) four times. The word ‘jaren’ was modelled with the two syllable models /ja/ and /r@/, whereas the word ‘jaar’ was modelled with the triphone sequence /#-j+a j-a+r a-r+#/. Similarly, the word ‘hadden’ (the plural imperfect form of the verb ‘to have’) was substituted by the word ‘had’ (the singular imperfect form of the same verb) four times before the retraining. The word ‘hadden’ was modelled with the model sequence /#-h+A h-A+# d@/, whereas the word ‘had’ was modelled with the syllable model /hAt/. These errors are related to resyllabification. The singular form ‘jaar’ corresponds to a syllable with a CVC structure, whereas the bisyllabic plural form ’jaren’ corresponds to a CV-CV structure. This raises the question whether all CV syllables are born equal. It might be that C_kV_i syllables resulting from $C_kV_iC_m$-V_a words, with V_a being an affix starting with a vowel, should not be clustered

with 'genuine' C_kV_i syllables. The fact that a large part of these errors disappeared in the retraining supports this hypothesis; the retraining was able to reintroduce some of the context information lost at the initialisation stage.

To summarise, our findings show that syllable context information is crucial for any attempt to model pronunciation variation with syllable models. From the point of view of human speech production and perception, syllables may have fewer interdependencies than phonemes. However, inter-syllable dependencies are clearly essential for automatic speech recognition.

5.2.Lexical confusability

Adding parallel paths to the syllable models essentially translates into adding pronunciation variants into the search space (see Section 3.3.4). 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 (e.g. Kessens et al., 2002). Similarly, the parallel paths of the untrained multi-path syllable models are obviously increasing the lexical confusability. Going from the baseline triphone recogniser to the untrained multi-path mixed-model recogniser with the target syllables covered with syllable models in monosyllabic words, the only fundamental difference was increasing the number of pronunciation variants for monosyllabic words in terms of parallel paths in the multi-path syllable models. Therefore, this type of mixed-model recogniser was the most appropriate recogniser to pinpoint the effect of the increased lexical confusability in the case of monosyllabic words. From the point of view of recognition performance, the increased confusability meant a 1.4 percentage point deterioration before the recognition parameter

optimisation ($MPM_{def,\ bt}$) and a 0.9 percentage point deterioration after the recognition parameter optimisation ($MPM_{opt,\ bt}$) as compared with the baseline triphone recogniser. It is interesting to notice that the increased lexical confusability deteriorated the recognition performance less than the loss of syllable context information in the case of polysyllabic words.

The significance of the lexical confusability issue in the case of monosyllabic words becomes clear when one considers the fact that 91% of the monosyllabic words represented with multi-path syllable models were function words. Function words typically carry less information than content words and are often pronounced in a highly reduced fashion (Bell et al., 2003; Greenberg, 1999; Jurafsky et al., 2001a; Pluymaekers et al., 2005; Van Son and Pols, 2003). Consequently, our initialisation approach produced short, easily confusable model paths particularly in the case of 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 on the word-level, and – when the language model could not assist in solving the problem – could have a direct impact on the WER.

It is interesting to notice that adding parallel paths to the syllable models in the case of polysyllabic words apparently does not cause problems with lexical confusability – nor does it improve the recognition performance. These conclusions can be drawn by comparing the recognition performance of the single-path mixed-model recogniser with the target syllables

covered with syllable models in polysyllabic words ($SPP_{def,\ bt}$, $SPP_{opt,\ bt}$, $SPP_{def,\ at}$ and $SPP_{opt,\ at}$) and the multi-path mixed-model recogniser with the target syllables covered with syllable models in polysyllabic words ($MPP_{def,\ bt}$, $MPP_{opt,\ bt}$, $MPP_{def,\ at}$ and $MPP_{opt,\ at}$, respectively), as well as the corresponding number of insertion, deletion and substitution errors. The comparable recognition results are virtually identical, and the numbers of errors – in particular, the number of substitution errors – are remarkably similar. In general, a word is less susceptible to recognition errors the more syllables it has (Greenberg and Chang, 2000; Hämäläinen et al., 2007a). It, therefore, appears that the other syllables and the language model are able to save the polysyllabic words from being misrecognised due to the increased number of pronunciation variants resulting from the addition of parallel paths in the multi-path syllable models.

To get further support for our hypothesis that initialising model paths with MDVs containing fewer symbols than the canonical variant was increasing the lexical confusability, we checked if the shorter paths were indeed contributing to misrecognitions more often than the other paths. To this end, we analysed the sentences with syllables modelled with multi-path syllable models in the case of the mixed-model recognisers. For these sentences, we calculated the total number of states visited during recognition by the baseline triphone recogniser and the optimised untrained and retrained multi-path mixed-model recognisers. We then checked how these numbers compared across the recognisers. The reason for us to calculate the total number of states on the sentence-level rather than on the word-level was that one speech recognition error typically causes recognition errors elsewhere in the sentence, as well. We carried out the analysis for four conditions:

a) for sentences that had been recognised correctly by both the baseline triphone recogniser and the optimised multi-path mixed-model recogniser.

b) for sentences that had been recognised correctly by the baseline triphone recogniser but incorrectly by the optimised multi-path mixed-model recogniser.

c) for sentences that had been recognised incorrectly by the baseline triphone recogniser but correctly by the optimised multi-path mixed-model recogniser.

d) for sentences that had been recognised incorrectly by both the baseline triphone recogniser and the optimised multi-path mixed-model recogniser.

Tables 5, 7 and 9 show the results of the analysis for the three types of multi-path mixed-model recognisers before the retraining, and Tables 6, 8 and 10 after the retraining. Condition b) is particularly revealing. Whenever the output of the multi-path mixed-model recogniser contained errors and the output of the baseline triphone recogniser did not, the total number of states visited by the mixed-model recogniser was smaller than the total number of states visited by the baseline triphone recogniser in most of the cases, both before and after retraining. On the contrary, when both recognisers were correct (condition a)), the total number of states visited was equal between the two recognisers in the vast majority of the cases. These results support our statement that paths shorter than the canonical cause misrecognitions. On the other hand, in particular condition a) shows that paths shorter (and longer) than the canonical can also be beneficial for the recognition results; their use often resulted in 100% recognition accuracy, too. Paths longer than the canonical were, however, the least helpful in the case of the multi-path mixed-model recogniser with the target syllables covered with syllable models in monosyllabic words (see condition a) in Tables 5 and 6). This is in line with the high reduction rates of monosyllabic words (Bell et al., 2003; Greenberg, 1999; Jurafsky et al., 2001a; Pluymaekers et al., 2005; Van Son and Pols, 2003).

Table 5: Comparison of the total number of states visited by the baseline triphone recogniser and the optimised untrained multi-path mixed-model recogniser with the target syllables covered with syllable models in monosyllabic words ($MPM_{opt,\ bt}$). The analysis only included sentences with syllables modelled with multi-path syllable models in the case of the mixed-model recognisers.

TABLE 5 HERE

Table 6: Comparison of the total number of states visited by the baseline triphone recogniser and the optimised retrained multi-path mixed-model recogniser with the target syllables covered with syllable models in monosyllabic words ($MPM_{opt,\ at}$). The analysis only included sentences with syllables modelled with multi-path syllable models in the case of the mixed-model recognisers.

TABLE 6 HERE

Table 7: Comparison of the total number of states visited by the baseline triphone recogniser and the optimised untrained multi-path mixed-model recogniser with the target syllables covered with syllable models in polysyllabic words ($MPP_{opt,\ bt}$). The analysis only included sentences with syllables modelled with multi-path syllable models in the case of the mixed-model recognisers.

TABLE 7 HERE

Table 8: Comparison of the total number of states visited by the baseline triphone recogniser and the optimised retrained multi-path mixed-model recogniser with the target syllables covered with syllable models in polysyllabic words ($MPP_{opt,\ at}$). The analysis only included sentences with syllables modelled with multi-path syllable models in the case of the mixed-model recognisers.

TABLE 8 HERE

Table 9: Comparison of the total number of states visited by the baseline triphone recogniser and the optimised untrained multi-path mixed-model recogniser with the target syllables covered with syllable models in both monosyllabic and polysyllabic words ($MP_{opt,\ bt}$). The analysis only included sentences with syllables modelled with multi-path syllable models in the case of the mixed-model recognisers.

TABLE 9 HERE

Table 10: Comparison of the total number of states visited by the baseline triphone recogniser and the optimised retrained multi-path mixed-model recogniser with the target syllables covered with syllable models in both monosyllabic and polysyllabic words ($MP_{opt,\ at}$). The analysis only included sentences with syllables modelled with multi-path syllable models in the case of the mixed-model recognisers.

TABLE 10 HERE

To summarise, while parallel paths make the syllable models acoustically more accurate (because the Gaussian mixtures along the parallel paths are able to capture the acoustic variation observed in the training data in much greater detail than a single path can do with the same number of Gaussians per state), they are increasing lexical confusability during recognition. Based on our findings, it may be particularly dangerous to add paths shorter than the canonical. This can, of course, also be explained by the well-known bias towards the use of shorter paths; the frame-state assignment is n-to-1 with n ≥1. So, while unreduced syllable tokens may be modelled with shorter state sequences, the short, reduced syllable tokens cannot be modelled by longer state sequences.

5.3. Word-specific pronunciation variation

Previous research on syllable models (Ganapathiraju et al., 2001; Hämäläinen et al., 2007a; Jouvet and Messina, 2004; Sethy and Narayanan, 2003; Sethy et al., 2003) does not discuss the appropriateness of using the same syllable models for syllables appearing in both monosyllabic and polysyllabic words. Based on the current recognition results, this might not, indeed, be an important issue in the case of single-path syllable models. The results gained with the single-path mixed-model recogniser with the target syllables covered with syllable models in both monosyllabic and polysyllabic words ($SP_{def,\ at}$ or $SP_{opt,\ at}$) can be explained by combining the results achieved with the single-path mixed-model recogniser with the target syllables covered with syllable models in monosyllabic words ($SPM_{def,\ at}$ or $SPM_{opt,\ at}$) and the single-path mixed-model recogniser with the target syllables covered with syllable models in polysyllabic words ($SPP_{def,\ at}$ or $SPP_{opt,\ at}$). However, using the same syllable models for syllables appearing in both monosyllabic and polysyllabic words does seem to be an issue in

the case of multi-path syllable models. There are at least two pieces of evidence pointing to this direction. First, one would not expect the performance of the multi-path mixed-model recogniser with the target syllables covered with syllable models in both monosyllabic and polysyllabic words ($MP_{def,\ at}$ or $MP_{opt,\ at}$) to be worse than the performances of both the multi-path mixed-model recogniser with the target syllables covered with syllable models in monosyllabic words ($MPM_{def,\ at}$ or $MPM_{opt,\ at}$) and the multi-path mixed-model recogniser with the target syllables covered with syllable models in polysyllabic words ($MPP_{def,\ at}$ or $MPP_{opt,\ at}$). Second, the fact that the optimal value for the word insertion penalty for the multi-path mixed-model recogniser with the target syllables covered with syllable models in both monosyllabic and polysyllabic words was so deviant from the other two multi-path mixed-model recognisers (see Table 4), suggests that it is difficult to find a word insertion penalty that would be suitable for both the monosyllabic and the polysyllabic words.

The importance of having different multi-path syllable models for canonically equivalent syllables appearing in monosyllabic and polysyllabic words makes sense intuitively. After all, the parallel paths are based on segmental variation. The segmental variation exhibited by the highly reduced monosyllabic function words is different from the segmental variation exhibited by a canonically equivalent syllable occurring in a polysyllabic word. Even if some of the segmental variants are the same, the probabilities of these variants are most likely to differ considerably between monosyllabic and polysyllabic words. The experiments carried out for this paper do not, however, allow us to draw any conclusions about the importance of more detailed information (e.g. which polysyllabic word the syllable appears in, which position the syllable appears in in the polysyllabic word) for the construction of multi-path syllable models.

5.4.Long-span spectral and temporal dependencies

In the literature (Ganapathiraju et al., 2001; Sethy and Narayanan, 2003; Sethy et al., 2003), the difficulty of capturing long-span spectral and temporal dependencies in speech with phoneme-length acoustic models has been cited as an important reason for using syllable models. According to Sethy et al. (2003), for instance, units of syllabic duration or longer are much more effective in capturing the cross-phone correlations and temporal dependencies than units of phonemic duration. In this subsection, we discuss this issue in more detail.

As explained in Section 3.3.3, the untrained version of the single-path mixed-model recogniser with the target syllables covered with syllable models in monosyllabic words ($SPM_{def,\ bt}$) was essentially identical to the baseline triphone recogniser. Therefore, this type of mixed-model recogniser is the most appropriate recogniser to pinpoint the effect of incorporating the long-span dependencies into the syllable models by means of retraining. As we can see in Figure 3, the retraining led into an insignificantly small improvement in the recognition performance. This indicates that, for a single-path system, coarticulation- and reduction-related spectral and temporal dependencies in speech that make a significant difference for speech recognition performance are already well covered by triphones – let alone context-dependent phone models with +/-2 phone context. Such long-span dependencies may, however, be slightly more important in the case of spontaneous speech. When compared with the performance of a phoneme-based recogniser, the absolute improvement that Sethy et al. (2003) obtained with mixed models on a particularly challenging database of spontaneous speech was 0.5%. However, some of their improvement can certainly be attributed to the fact that they used both a mixed syllabic-phonetic and a pure phonetic pronunciation variant for each word in the recognition lexicon. In any case, our results show that modelling syllable context is far more important for speech recognition

performance than modelling the long-span dependencies. The modelling of long-span spectral and temporal dependencies with syllable models may become more beneficial as the size of speech databases increases and as the number of syllables with a sufficient number of training tokens becomes larger.

6. General discussion

Thus far, explicit pronunciation variation modelling by adding pronunciation variants in the lexicon has made a disappointing contribution to improving speech recognition performance (Hain, 2005). Therefore, our research was focussed on the question whether implicit pronunciation variation modelling within the HMMs could yield better results. The problem of pronunciation variation is at the very heart of ASR since it is directly related to the question how observed continuous acoustic variation can successfully be modelled by a more discrete framework (e.g. distinct variants in the lexicon or distinct paths in an HMM). Of course, there are many different ways of attempting implicit modelling. The focus of the present study was on implicit modelling of long-span coarticulation and reduction effects with syllable-length acoustic models. More specifically, we studied a number of factors that may affect the performance of syllable-based recognisers.

First and foremost, we must conclude that implicit pronunciation variation modelling with syllable models does not per se lead to significant improvements in recognition performance as compared with explicit modelling with context-dependent phone models. In our experiments on TIMIT and a smaller set of read speech from CGN (Hämäläinen et al., 2007a), the performance of retrained single-path mixed-model recognisers with the target syllables covered with syllable models in both monosyllabic and polysyllabic words did not differ significantly from the performance of triphones. However, in the current study,

triphones (with a larger number of Gaussian mixtures) significantly outperformed a similar retrained single-path mixed-model recogniser. The performance of the baseline triphone recogniser was only reached and slightly improved upon by a mixed-model recogniser in which the most common monosyllabic words were covered with syllable models. Our results are comparable with other studies (Ganapathiraju et al., 2001; Jouvet and Messina, 2004; Sethy et al., 2003) in which single-path syllable models did not yield considerable improvements in recognition performance. In Hämäläinen et al. (2007a), we hypothesised that the lack of improvement in recognition performance was caused by the fact that the many different forms that syllable pronunciations can assume cannot be accounted for with a single path through the syllable model. We still believe that this is part of the reason for the disappointing recognition performance. However, our current study also shows that the loss of context information at some syllable boundaries puts the single-path mixed-model recognisers (as well as the multi-path mixed-model recognisers) with context-independent syllable models in polysyllabic words at a disadvantage as compared with a well-engineered triphone recogniser.

We expected that the richer topology of multi-path syllable models would be better at accounting for pronunciation variation than triphone models, or single-path syllable models that merely have their parameters adjusted on the basis of dedicated syllable tokens. In a way, untrained multi-path models initialised with MDVs re-introduce explicit pronunciation variation modelling. Such models correspond to the segmental multiple-entry models of auditory word recognition (Johnson, 2004). However, we assumed that re-estimating multi-path syllable models initialised with MDVs would 'specialise' the model paths to such an extent that a potential increase in lexical confusability would not be a problem. Such models would be in line with Johnson's nonsegmental multiple-entry models of auditory word

recognition. In reality, the Baum-Welch re-estimation turned out not to be as powerful as we had expected. The re-estimation may have adjusted the probabilities of entering the different parallel paths and taken us some distance from the symbolic level to the subsymbolic level but this was not enough to avoid the problem of lexical confusability. In fact, some of the retrained parallel paths were still closely related to the MDVs used to initialise them. In Hämäläinen et al. (2007b), we carried out a forced alignment of the training data with the multi-path mixed-model recogniser and analysed the training tokens assigned to each path of the syllable models. Our analysis showed that the token-to-path assignment was clearly related to the articulatory similarity – or dissimilarity – between the transcriptions of the training tokens and the MDVs used to initialise the parallel paths. In Hämäläinen et al. (2007c), on the other hand, we investigated the Kullback-Leibler distance (KLD) between the initial and the retrained model paths. It appeared that the KLDs between the initial and the retrained distributions for the states of the paths corresponding to the canonical transcriptions were relatively minor. The distances between the initial and the retrained paths for non-canonical paths were often (much) larger. The error analysis described in this paper showed that, to a large extent, the problem of lexical confusability could be attributed to parallel paths that were shorter than the canonical.

Properly trained parallel paths make syllable models acoustically more accurate because the Gaussian mixtures along the parallel paths are able to capture the acoustic variation observed in the training data in much greater detail than a single path can do with the same number of Gaussians per state. However, the greater acoustic accuracy comes at the cost of increased lexical confusability. To be able to benefit from the greater acoustic accuracy and to reduce the problem of the increased lexical confusability during recognition, the pronunciation variation modelled by the parallel paths should be linked to specific verbal contexts. After all,

some of the paths may represent pronunciation variants that only occur in certain words or – in particular in the case of monosyllabic function words – in certain cross-word contexts. However, the architecture of a conventional HMM decoder, such as HTK (Young et al., 2002), does not provide hooks for controlling which paths can be used with which words and contexts. One is left to do with the probabilities of words (and n-grams) as defined in the language model, and the "loose" transition probabilities of entering the different parallel paths of the syllable models that remain unchanged in all verbal contexts. Our speech recognition results with multi-path mixed-model recognisers show that this is not sufficient to achieve improved recognition performance.

The kinds of long-span coarticulation and reduction effects that we attempted to model are arguably more common in spontaneous speech than in read speech. As syllables are more stable than phones as basic units of speech (Greenberg, 1999), one might intuitively expect a greater gain from a syllable-based modelling approach in the case of spontaneous speech than in the case of read speech (Ganapathiraju et al., 2001). We do not, however, expect this to be the case in reality. This is because of the greater amount of variation in spontaneous speech. To model this variation, one would expect more parallel paths to be necessary. More parallel paths would, however, result in more confusion – as shown by our experimental results on read speech.

Based on the similarity between our and other researchers' (Ganapathiraju et al., 2001; Jouvet and Messina, 2004; Sethy et al., 2003) results with single-path syllable models, we also expect our results with multi-path syllable models to generalise to other tasks and to other languages of a similar syllabic composition. The approach may hold more promise in the case of languages that have much fewer syllables and a more constrained syllable structure (e.g.

Chinese). For such languages, it may be easier to build context-dependent multi-path syllable models.

HTK (Young et al., 2002) exemplifies a conventional HMM decoder. Therefore, one would expect our findings to generalise across all HMM-based recognisers. However, it is clear that a bigram language model is not the strongest possible language model. Using a higher-order language model would certainly help in the kind of scenario where the two Dutch definite articles 'de' and 'het' are confused with each other when there is an adjective between the article and the noun (see Section 5.2). Yet, a higher-order language model would be beneficial for all the different kinds of recognisers. Hence, even if the multi-path mixed-model recogniser had more to gain from a higher-order language model (because of the added confusability caused by the parallel paths), the effect of such a local improvement would be unlikely to fundamentally change our findings. When it comes to comparing WERs, one must also not forget that other types of recognisers with context-dependent phone models (e.g. context-dependent phone models with +/-2 phone context, context-dependent phone models with pronunciation variants in the lexicon) are known to outperform the type of baseline recogniser that we used. For our experimental set-up, a word-internal triphone recogniser with a single canonical pronunciation per word in the lexicon was the most suitable baseline recogniser. The goal of our experiments was not necessarily to look for the best performing recogniser. After all, it is not the reduction of WER alone that is important; for long-term development in the field, it is equally – if not more important – to really understand the issues that we are battling with (Bourlard et al., 1996). The experiments reported in this paper increase our understanding about the potential and the limitations of syllable-based models.

7. Directions for future research

In a nutshell, our results – supported by the results of others (Ganapathiraju et al., 2001; Jouvet and Messina, 2004; Sethy et al., 2003) – indicate that a successful approach to deal with pronunciation variation with syllable-length models must be based on a procedure that meets the following four conditions. First, one must account for the observed phonetic variation. Second, one must model syllable context information. Third, one must take the possible increase in lexical confusability into account if creating alternative model paths. Fourth, because of word-specific pronunciation variation, one must not use the same multi-path syllable models for both monosyllabic and polysyllabic words.

Even if there were no data sparseness issues when building multi-path syllable models, we would essentially be faced with two challenges: context modelling and lexical confusability. Jouvet and Messina (2004) employed a parameter sharing method that allowed them to build context-dependent syllable models. The improvements in recognition performance that they achieved with single-path syllable models were small and depended heavily on the recognition task: for telephone numbers, the performance even deteriorated. This may be an indication that the amount of training data they had available was not enough to capture all the relevant context effects. However, as the context modelling led to improvements in most of their tasks, one might expect a similar approach to be more fruitful in combination with a large amount of training data and properly initialised multi-path syllable models.

As the retrained parallel paths of the multi-path syllable models are still closely related to the MDVs used to initialise them (Hämäläinen et al., 2007b; Hämäläinen et al., 2007c), one might argue that we could alleviate the problem of lexical confusability by refining our MDV selection approach. Based on discriminative training methods (Lin and Yvon, 2007; Markov

and Nakamura, 2007) or existing methods to detect confusable words (Anguita et al., 2005; Roe and Riley, 1994), we could devise ways of avoiding MDVs that would result in overlapping pronunciations with other words in the lexicon. However, it is difficult to see how pronunciation variants could be added without increasing the confusability of the lexicon. Perhaps, the additional confusability should not be an insurmountable problem. After all, humans seem to be dealing with the problem with such ease that it often goes completely unnoticed. Staying within the probabilistic framework of mainstream ASR, the question then becomes how humans manage to obtain context-dependent local estimates of the prior probabilities of the words and their possible pronunciation variants. While it may be possible to embed single-path syllable models explicitly in the probabilistic decoding machinery of a speech recogniser, it is much less clear how the same could be accomplished with multi-path models. As explained in Section 6, in a conventional HMM decoder, the probabilities of the parallel paths can only be modelled as transition probabilities of entering the different parallel paths. These probabilities cannot directly be linked with the language model. One option would, of course, be to replace the non-emitting first and last states of the multi-path syllable models by three independent non-emitting states. Doing so would not only offer a solution for linking language model scores to pronunciation variants but also for specifying that a specific path is much more likely if the syllable occurs as part of a polysyllabic word. However, this would be a step back in the direction of the conventional multiple-entry representations.

MDV-based multi-path syllable models seem to suffer from the same kinds of problems as explicit pronunciation variation modelling in the recognition lexicon. It is difficult to see how other initialisation approaches could altogether avoid these problems. Still, we can maintain that straightforward left-to-right HMM topologies are not able to capture the relevant

pronunciation variation on the syllable-level. Therefore, we must conclude that multi-path syllable models, however they may be initialised and trained, may not be the way towards solving the pronunciation variation problem in ASR. Using the acoustic variation in speech as the basis for constructing parametric models of speech (Deng et al., 2006; Han et al., 2007; Zen et al., 2007) will not solve the context modelling problem either. It may well aggravate the problem because it is difficult to link bottom-up acoustic variation to the lexicon and the language model. Therefore, it may be necessary to altogether abandon parametric models and to move on to exemplar-based models (Aradilla et al., 2006; de Wachter, 2007), even if this approach will also need to come to grips with the proper integration of acoustic, lexical and linguistic probabilities.

8. Conclusions

The goal of our paper was to investigate the importance of within-syllable pronunciation variation and syllable context from the point of view of speech recognition performance. To this end, we constructed context-independent single-path and multi-path models for frequent syllables in a large vocabulary continuous speech recognition task. Our hypothesis was that the multi-path syllable models would be better at accounting for pronunciation variation than the single-path syllable models. We incorporated the single-path and multi-path syllable models into speech recognisers in which the other syllables in the task were covered with triphones. Comparing the recognition performance and recognition errors of the resulting mixed-model recognisers against the performance and errors of a baseline triphone recogniser allowed us to draw conclusions about the importance of the factors under investigation. Our study showed that the greater acoustic accuracy of multi-path syllable models comes at the cost of increased lexical confusability. This effect is particularly pronounced in the case of monosyllabic function words, which usually are some of the few syllables that have a

sufficient amount of training data available for the training of parallel paths. In fact, modelling within-syllable pronunciation variation with parallel paths in a conventional HMM decoder does more harm than good for the speech recognition performance. At least part of the reason is that the architecture of a conventional HMM decoder does not provide hooks for controlling which paths can be used with which words and with which cross-word contexts. Using the transition probabilities of entering the different parallel paths, which remain unchanged in all lexical contexts, obviously is not enough. In addition to highlighting the unfavourable imbalance between the greater acoustic accuracy of the multi-path syllable models and the lexical confusability caused by the parallel paths, our results showed the importance of context modelling at syllable boundaries. The main contribution of this paper, then, is to add to our understanding of speech modelling by providing insights into the complex issues that are of importance when modelling pronunciation variation with syllable models.

Acknowledgements

Annika Hämäläinen's research was carried out within the Interactive Multimodal Information eXtraction (IMIX) program, which is sponsored by the Netherlands Organisation for Scientific Research (NWO). Louis ten Bosch participates in the European FET project ACORNS (nr FP6-034362). The authors would like to thank two anonymous reviewers for their valuable ideas and suggestions.

References[2]

Anguita, J., Hernando, J., Peillon, S., Bramoulle, A., 2005. Detection of confusable words in automatic speech recognition. *IEEE Signal Processing Letters* 12(8), 585-588.

Aradilla, G., Vepa, J., Bourlard, H., 2006. Using Pitch as Prior Knowledge in Template-Based Speech Recognition, in: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '06)*, Toulouse, France.

Baayen, R.H., Piepenbrock, R., Gulikers, L., 1995. *The CELEX Lexical Database (Release 2)*, Linguistic Data Consortium, University of Pennsylvania, Philadelphia, Pennsylvania, USA.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., Gildea, D., 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *Journal of the Acoustical Society of America* 113, 1001–1024.

Booij, G., 1999. *The phonology of Dutch.* Oxford University Press, New York.

Bourlard, H., Hermansky, H., Morgan, N., 1996. Towards increasing speech recognition error rates. *Speech Communication* 18, 205-231.

Deng, L., Yu, D., Acero, A., 2006. Structured speech modeling. *IEEE Transactions on Audio, Speech and Language Processing (Special Issue on Rich Transcription)* 14(5), 1492-1504.

We have previously published the following conference papers related to the present study:

Hämäläinen, A., ten Bosch, L., Boves, L., 2006. Pronunciation variant –based multi-path HMMs for syllables, in: *Proceedings of the 9th International Conference on Spoken Language Processing (ICSLP '06)*, Pittsburgh, PA, USA.

Hämäläinen, A., ten Bosch, L., Boves, L., 2006. Multi-path syllable models based on phonetic knowledge, in: Aulanko, R., Wahlberg, L., Vainio, M. (Eds.), *Proceedings of the Phonetics Symposium 2006*, Publications of the Department of Speech Sciences, University of Helsinki, pp. 57-66.

Hämäläinen, A., ten Bosch, L., Boves, L., 2007. Modelling pronunciation variation using multi-path HMMs for syllables, in: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '07)*, Honolulu, HA, USA.

Hämäläinen, A., ten Bosch, L., Boves, L., 2007. Construction and analysis of multiple paths in syllable models, in: *Proceedings of the 10th European Conference on Speech Communication and Technology (EUROSPEECH '07)*, Antwerp, Belgium.

The above mentioned conference papers introduced the construction of MDV-based syllable models and analysed the models from different points of view. Many of the recognition results reported in this paper have not been published elsewhere. The error analyses and the conclusions presented in this paper are also new.

de Wachter, M., 2007. *Example based continuous speech recognition*, PhD thesis, University of Leuven, Leuven, Belgium.

Elffers, B., Van Bael, C., Strik, H., 2005. *ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcriptions*, Technical report, Radboud University Nijmegen, Nijmegen, The Netherlands.

Ganapathiraju, A., Hamaker, J., Picone, J., Ordowski, M., Doddington, G., 2001. Syllable-based large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing* 9(4), 358-366.

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

Greenberg, S., Chang, S., 2000. Linguistic dissection of switchboard-corpus automatic speech recognition systems, in: *Proceedings of Automatic Speech Recognition: Challenges for the New Millenium (ISCA ITRW ASR '00)*, pp. 195-202, Paris, France.

Hain, T., 2005. Implicit modelling of pronunciation variation in automatic speech recognition. *Speech Communication* 46(2), 171-188.

Han, Y., de Veth, J., Boves, L., 2007. Trajectory Clustering for Solving the Trajectory Folding Problem in Automatic Speech Recognition. *IEEE Transactions on Audio, Speech and Language Processing* 15(4), 1425-1434.

Hämäläinen, A., ten Bosch, L., Boves, L., 2006. Multi-path syllable models based on phonetic knowledge, in: Aulanko, R., Wahlberg, L., Vainio, M. (Eds.), *Proceedings of the Phonetics Symposium 2006*, Publications of the Department of Speech Sciences, University of Helsinki, pp. 57-66.

Hämäläinen, A., Boves, L., de Veth, J., ten Bosch, L., 2007a. On the Utility of Syllable-Based Acoustic Models for Pronunciation Variation Modelling. *EURASIP Journal on Audio, Speech, and Music Processing* 2007, Article ID 46460, 11 pages.

Hämäläinen, A., ten Bosch, L., Boves, L., 2007b. Modelling pronunciation variation using multi-path HMMs for syllables, in: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '07)*, Honolulu, HA, USA.

Hämäläinen, A., ten Bosch, L., Boves, L., 2007c. Construction and analysis of multiple paths in syllable models, in: *Proceedings of the 10th European Conference on Speech Communication and Technology (EUROSPEECH '07)*, Antwerp, Belgium.

Johnson, K., 2004. Massive reduction in conversational American English, in: Yoneyama, K., Maekawa, K. (Eds.), *Spontaneous Speech: Data and Analysis,* The National Institute for Japanese Language, Tokyo, pp. 29-54.

Jones, R.J., Downey, S., Mason, J.S., 1997. Continuous speech recognition using syllables, in: *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH '97)*, vol. 3, pp. 1171-1174, Rhodes, Greece.

Jouvet D., Messina, R., 2004. Context dependent "long units" for speech recognition, in: *Proceedings of the 8th International Conference on Spoken Language Processing (ICSLP '04)*, pp. 645-648, Jeju Island, Korea.

Jurafsky, D., Bell, A., Gregory, M., Raymond, W., 2001a. Probabilistic relations between words: Evidence from reduction in lexical production, in: Bybee, J., Hopper, P. (Eds.), *Frequency and the emergence of linguistic structure*, John Benjamins, Amsterdam, pp. 229-254.

Jurafsky, D., Ward, W., Jianping, Z., Herold, K., Xiuyang, Y., Sen, Z., 2001b. What kind of pronunciation variation is hard for triphones to model?, in: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '01)*, vol. 1, pp. 577-580, Salt Lake City, Utah, USA.

Kessens, J., Strik, H., Cucchiarini, C., 2002. Modeling pronunciation variation for ASR: Comparing criteria for rule selection, in: *Proceedings of ISCA Tutorial and Research*

Workshop on Pronunciation Modeling and Lexicon Adaptation for Spoken Language (PMLA '02), pp. 18-23, Estes Park, Colorado, USA.

Kessens, J., Cucchiarini, C., Strik, H., 2003. A data-driven method for modeling pronunciation variation. *Speech Communication*, 40(4):517-534.

Lee, K.-F., 1989. *Automatic speech recognition: The development of the SPHINX system*, Kluwer Academic Publishers, Boston.

Lin, S.S., Yvon, F., 2007. Optimization on decoding graphs by discriminative training, in: *Proceedings of the 10th European Conference on Speech Communication and Technology (EUROSPEECH '07)*, Antwerp, Belgium.

Markov, K., Nakamura, S., 2007. Never-ending learning with dynamic hidden Markov network, in: *Proceedings of the 10th European Conference on Speech Communication and Technology (EUROSPEECH '07)*, Antwerp, Belgium.

McAllaster, D., Gillick, L., 1999. Studies in acoustic training and language modeling using simulated speech data, in: *Proceedings of the 6th European Conference on Speech Communication and Technology (EUROSPEECH '99)*, pp. 1787-1790, Budapest, Hungary.

Oostdijk, N., Goedertier, W., Van Eynde, F., Boves, L., Martens, J.P., Moortgat, M., Baayen, H., 2002. Experiences from the Spoken Dutch Corpus Project, in: *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC '02)*, vol. 1, pp. 340-347, Las Palmas, Canary Islands, Spain.

Ostendorf, M., 1999. Moving beyond the 'beads-on-a-string' model of speech, in: *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '99)*, Keystone, Colorado, USA.

Plannerer, G., Ruske, B., 1992. Recognition of demisyllable based units using semicontinuous hidden Markov models, in: *Proceedings* of IEEE International Conference on

Acoustics, Speech, and Signal Processing *(ICASSP '92)*, vol. 1, pp. 581-584, San Francisco, California, USA.

Pluymaekers, M., Ernestus, M., Baayen, R.H., 2005. Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica* 62, 146-159.

Roe, D.B., Riley, M.D., 1994. Prediction of word confusabilities for speech recognition, in: *Proceedings of the 3rd International Conference on Spoken Language Processing (ICSLP '94)*, pp. 227-230, Yokohama, Japan.

Saraclar, M., Khudanpur, S., 2000. Pronunciation ambiguity vs pronunciation variability in speech recognition, in: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '00)*, vol. 3, pp. 1679-1682, Istanbul, Turkey.

Saraclar, M., Nock, H., Khudanpur, S., 2000. Pronunciation modeling by sharing Gaussian densities across phonetic models. Computer Speech and Language 14(2): 137-160.

Sethy, A., Narayanan, S., 2003. Split-lexicon based hierarchical recognition of speech using syllable and word level acoustic units, in: *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '03)*, vol. 1, pp. 772-776, Hong Kong.

Sethy, A., Ramabhadran, B., Narayanan, S., 2003. Improvements in ASR for the MALACH project using syllable-centric models, in: *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '03),* St. Thomas, Virgin Islands, USA.

Strik, H., Cucchiarini, C., 1999. Modeling pronunciation variation for ASR: A survey of the literature. *Speech Communication* 29, 225–246.

Van Son, R.J.J.H., Pols, L.C.W., 2003. Information structure and efficiency in speech production, in: *Proceedings of the 8th European Conference on Speech Communication and Technology (EUROSPEECH '03)*, pp. 769-772, Geneva, Switzerland.

Wells, J., 2000. *Longman Pronunciation Dictionary*, 2nd Edition. Pearson Education Limited, Harlow.

Wester, M., 2002. *Pronunciation variation modeling for Dutch automatic speech recognition*, PhD thesis, University of Nijmegen, Nijmegen, The Netherlands.

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

Zen, H., Tokuda, K., Kitamura, T., 2007. Reformulating the HMM as a trajectory model by imposing explicit relationships between static and dynamic feature vector sequences. *Computer Speech & Language* 21, 153-173.

Figure 1

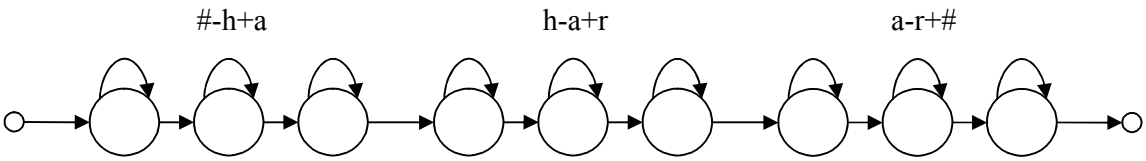


Figure 2

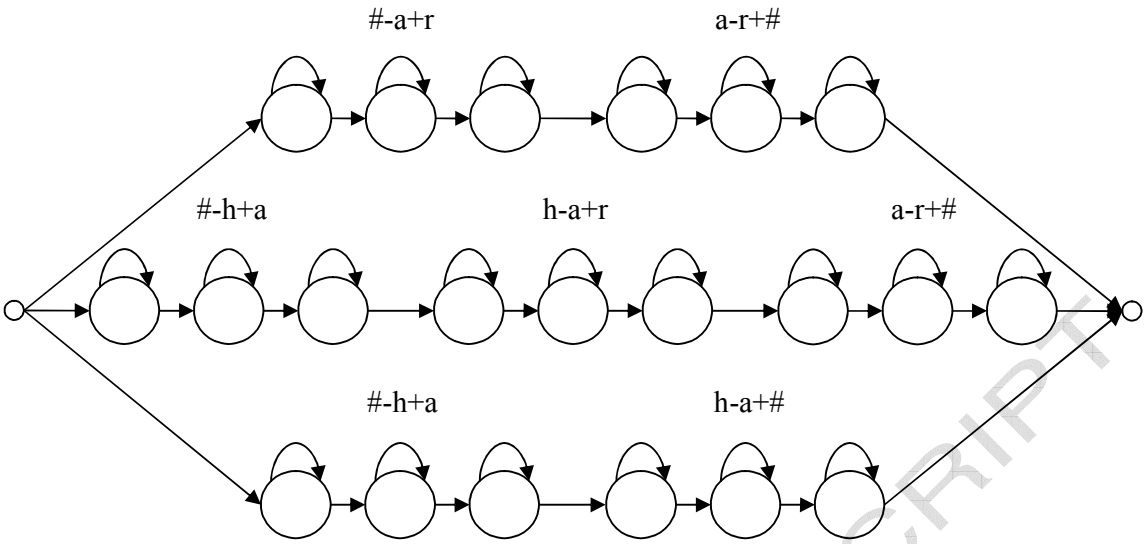


Figure 3

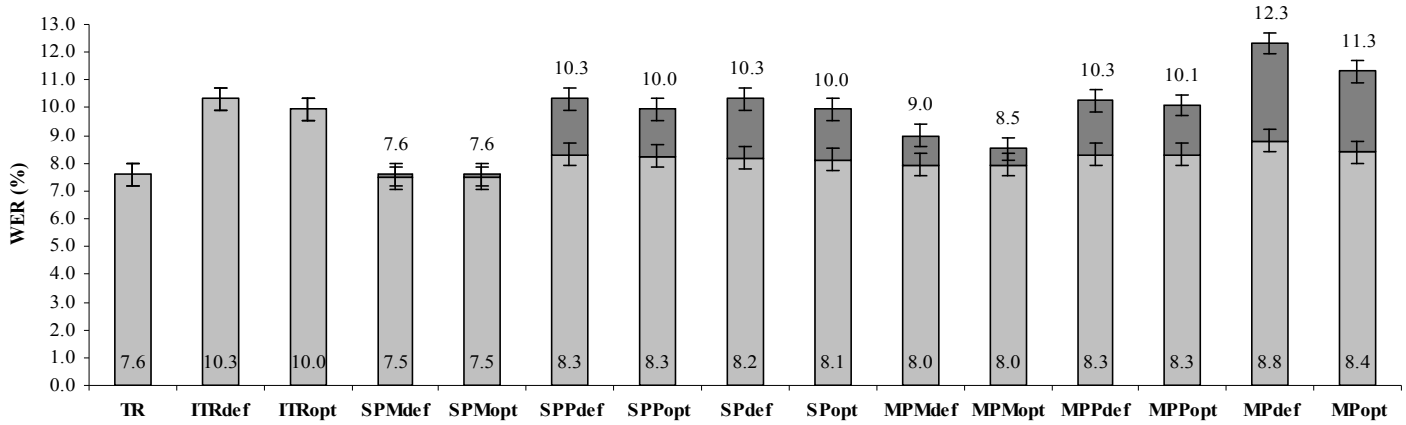


Figure 4

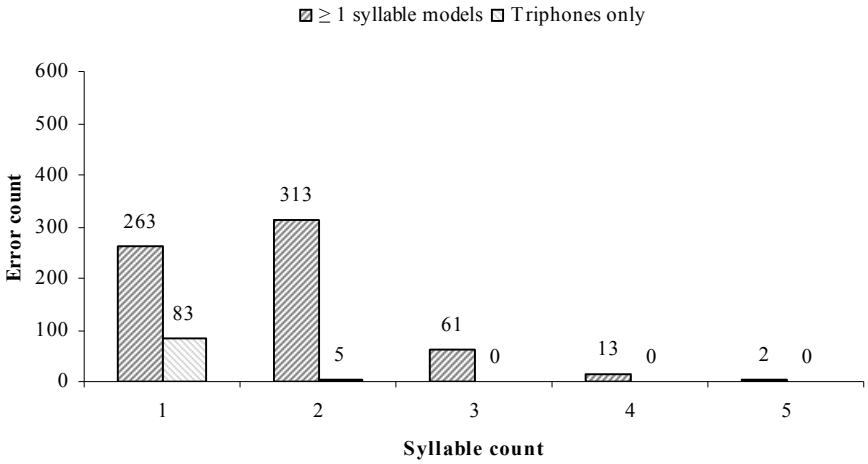


Figure 5

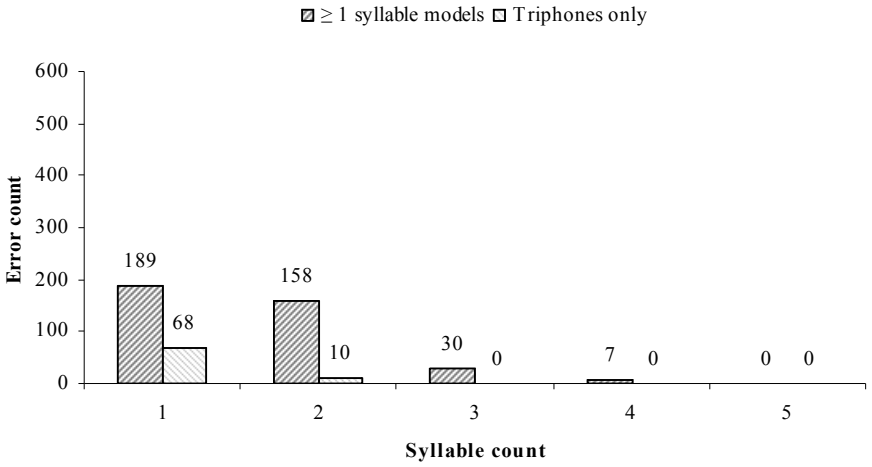


Table 1

	Train	Test	Dev. test
Word tokens	396 187	22 289	22 100
Word types	28 164	5 154	5 074
Syllable tokens	604 211	33 921	33 588
Syllable types	6 146	2 722	2 623
Duration (hh:mm:ss)	37:00:20	02:04:21	02:03:33

Number of syllables	Proportion (%)
1	65.0
2	22.5
3	8.7
4	3.0
≥ 5	0.9

	Total number of states
TR	1 535
SPM	1 605
SPP	1 621
SP	1 603
MPM	1 726
MPP	1 782
MP	1 764

Table 4

	-s	-p	Ins	Del	Subs
TR	16	25	163	350	1184
ITR_{def}	16	25	359	317	1626
ITR_{opt}	18	15	167	520	1534
$SPM_{def, bt}$	16	25	163	350	1184
$SPM_{def, at}$	16	25	168	299	1201
$SPM_{opt, bt}$	16	25	163	350	1184
$SPM_{opt, at}$	16	25	168	299	1201
$SPP_{def, bt}$	16	25	359	317	1626
$SPP_{def, at}$	16	25	238	310	1305
$SPP_{opt, bt}$	18	15	167	520	1534
$SPP_{opt, at}$	16	20	195	374	1271
$SP_{def, bt}$	16	25	359	317	1626
$SP_{def, at}$	16	25	234	290	1299
$SP_{opt, bt}$	18	15	167	520	1534
$SP_{opt, at}$	16	20	183	351	1280
$MPM_{def, bt}$	16	25	322	293	1391
$MPM_{def, at}$	16	25	277	254	1241
$MPM_{opt, bt}$	14	10	150	438	1312
$MPM_{opt, at}$	16	25	277	254	1241
$MPP_{def, bt}$	16	25	315	361	1609
$MPP_{def, at}$	16	25	239	317	1298
$MPP_{opt, bt}$	18	25	225	440	1583

$MPP_{opt, at}$	16	25	239	317	1298
$MP_{def, bt}$	16	25	523	298	1926
$MP_{def, at}$	16	25	336	255	1370
$MP_{opt, bt}$	16	5	161	657	1702
$MP_{opt, at}$	14	10	185	383	1302

Table 5

Number of States	Triphone correct; multi-path correct	Triphone correct; multi-path wrong	Triphone wrong; multi-path correct	Triphone wrong; multi-path wrong
Triphone = multi-path	731	38	23	437
Triphone > multi-path	113	69	17	268
Triphone < multi-path	36	23	12	132

Table 6

Number of States	Triphone correct; multi-path correct	Triphone correct; multi-path wrong	Triphone wrong; multi-path correct	Triphone wrong; multi-path wrong
Triphone = multi-path	692	23	21	445
Triphone > multi-path	162	24	18	189
Triphone < multi-path	67	45	36	179

Table 7

Number of States	Triphone correct; multi-path correct	Triphone correct; multi-path wrong	Triphone wrong; multi-path correct	Triphone wrong; multi-path wrong
Triphone = multi-path	472	38	5	332
Triphone > multi-path	125	73	14	245
Triphone < multi-path	180	45	14	198

Table 8

Number of States	Triphone correct; multi-path correct	Triphone correct; multi-path wrong	Triphone wrong; multi-path correct	Triphone wrong; multi-path wrong
Triphone = multi-path	440	12	8	315
Triphone > multi-path	155	32	17	179
Triphone < multi-path	264	26	26	262

Table 9

Number of States	Triphone correct; multi-path correct	Triphone correct; multi-path wrong	Triphone wrong; multi-path correct	Triphone wrong; multi-path wrong
Triphone = multi-path	523	55	13	282
Triphone > multi-path	160	153	22	378
Triphone < multi-path	156	47	7	189

Table 10

Number of States	Triphone correct; multi-path correct	Triphone correct; multi-path wrong	Triphone wrong; multi-path correct	Triphone wrong; multi-path wrong
Triphone = multi-path	500	14	19	298
Triphone > multi-path	213	67	27	269
Triphone < multi-path	246	56	28	250