Hindawi Publishing Corporation
EURASIP Journal on Audio, Speech, and Music Processing
Volume 2007, Article ID 46460, 11 pages
doi:10.1155/2007/46460

Research Article

On the Utility of Syllable-Based Acoustic Models for Pronunciation Variation Modelling

Annika Hämäläinen, Lou Boves, Johan de Veth, and Louis ten Bosch

Centre for Language and Speech Technology (CLST), Faculty of Arts, Radboud University Nijmegen, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

Received 6 December 2006; Accepted 18 May 2007

Recommended by Jim Glass

Recent research on the TIMIT corpus suggests that longer-length acoustic models are more appropriate for pronunciation variation modelling than the context-dependent phones that conventional automatic speech recognisers use. However, the impressive speech recognition results obtained with longer-length models on TIMIT remain to be reproduced on other corpora. To understand the conditions in which longer-length acoustic models result in considerable improvements in recognition performance, we carry out recognition experiments on both TIMIT and the Spoken Dutch Corpus and analyse the differences between the two sets of results. We establish that the details of the procedure used for initialising the longer-length models have a substantial effect on the speech recognition results. When initialised appropriately, longer-length acoustic models that borrow their topology from a sequence of triphones cannot capture the pronunciation variation phenomena that hinder recognition performance the most.

1. INTRODUCTION

Conventional large-vocabulary continuous speech recognisers use context-dependent phone models, such as triphones, to model speech. Apart from their capability of modelling (some) contextual effects, the main advantage of triphones is that the fixed number of phonemes in a given language guarantees their robust training when reasonable amounts of training data are available and when state tying methods are used to deal with infrequent triphones. When using triphones, one must assume that speech can be represented as a sequence of discrete phonemes (beads on a string) that can only be substituted, inserted, or deleted to account for pronunciation variation [1]. Given this assumption, it should be possible to account for pronunciation variation at the level of the phonetic transcriptions in the recognition lexicon. Modelling pronunciation variation by adding transcription variants in the lexicon has, however, met with limited success, in part because of the resulting increase in lexical confusability [2]. Furthermore, while triphones are able to capture short-span contextual effects such as phoneme substitution and reduction [3], there are complexities in speech that triphones cannot capture. Coarticulation effects typically have a time span that exceeds that of the left and right neighbouring phones. The corresponding long-span spectral and temporal dependencies are not easy to capture with the limited window of triphones [4]. This is the case even if the feature vectors implicitly encode some degree of long-span coarticulation effects thanks to the addition of, for example, deltas and delta-deltas, or the use of augmented features and LDA. In an interesting study with simulated data, McAllaster and Gillick [5] showed that recognition accuracy decreases dramatically if the sequence of HMM models that is used to generate speech frames is derived from accurate phonetic transcriptions of Switchboard utterances, rather than from sequences of phonetic symbols in a sentence-independent multipronunciation lexicon. At the surface level, this implies that the recognition accuracy drops substantially if the state sequence licensed by the lexicon is not identical to the state sequence that corresponds to the best possible segmental approximation of the actual pronunciation. At a deeper level, this suggests that triphones fail to capture at least some relevant effects of long-span coarticulation. Ultimately, then, we must conclude that a representation of speech in terms of a sequence of discrete symbols is not fully adequate.

To alleviate the problems of the “beads on a string” representation of speech, several authors propose using longer-length acoustic models [4, 6–12]. These word or subword

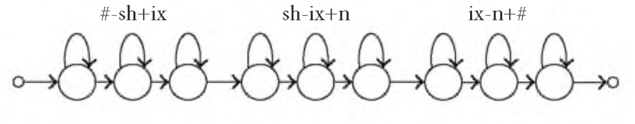

Figure 1: Syllable model for the syllable /sh ix n/. The model states are initialised with the triphones underlying the canonical syllable transcription [8]. The phones before the minus sign and after the plus sign in the triphone notation denote the left and right context in which the context-dependent phones have been trained. The hashes denote the boundaries of the context-independent syllable model.

models are expected to capture the relevant detail, possibly at the cost of phonetic interpretation and segmentation. Syllable models are probably the most commonly suggested longer-length models [4, 6–12]. Support for their use comes from studies of human speech production and perception [13, 14], and the relative stability of syllables as a speech unit. The stability of syllables is illustrated by Greenberg in [15] finding that the syllable deletion rate of spontaneous speech is as low as 1%, as compared with the 12% deletion rate of phones.

The most important challenge of using longer-length acoustic models in large-vocabulary continuous speech recognition is the inevitable sparseness of training data in the model training. As the speech units become longer, the number of infrequent units with insufficient acoustic data for reliable model parameter estimation increases. If the units are words, the number of infrequent units may be unbounded. Many languages—for instance, English and Dutch—also have several thousands of syllables, some of which will have very low-frequency counts in a reasonably sized training corpus. Furthermore, as the speech units comprise more phones, increasingly complex types of articulatory variation must be accounted for.

The solutions suggested for the data sparsity problem are two-fold. First, longer-length models with a sufficient amount of training data are used in combination with context-dependent phone models [4, 8–12]. In other words, context-dependent phone models are backed off to when a given longer-length speech unit does not occur frequently enough for reliable model parameter estimation. Second, to ensure that a much smaller amount of training data is sufficient, the longer-length models are cleverly initialised [8–10]. Sethy and Narayanan [8], for instance, suggest initialising the longer-length models with the parameters of the triphones underlying the canonical transcription of the longer-length speech units (see Figure 1). Subsequent Baum-Welch reestimation is expected to incorporate the spectral and temporal dependencies of speech into the initialised models by adjusting the means and covariances of the Gaussian components of the mixtures associated with the HMM states of the longer-length models.

Several research groups have published promising, but somewhat contradictory, results with longer-length acoustic models [4, 8–12]. Sethy and Narayanan [8] used the above described mixed-model recognition scheme, combining context-independent word and syllable models with triphones. They reported a 62% relative reduction in word error rate (WER) on TIMIT [16], a database of carefully read, and annotated American English. We adopted their method for our research, repeating the recognition experiments on TIMIT and, in addition, carrying out similar experiments on a corpus of Dutch read speech equipped with a coarser annotation. As was the case with other studies [4, 9, 10], the improvements we gained [11, 12] on both corpora were more modest than those that Sethy and Narayanan obtained. Part of the discrepancy between Sethy and Narayanan's impressive improvements and the much more equivocal results of others [4, 9–12] may be due to the surprisingly high baseline WER (26%) Sethy and Narayanan report. We did, however, also find much larger improvements on TIMIT than on the Dutch corpus. The goal of the current study is to shed light on the reasons for the varying results obtained on different corpora. By doing so, we show what is necessary for the successful modelling of pronunciation variation with longer-length acoustic models.

To achieve the goal of this paper, we carry out and compare speech recognition experiments with a mixed-model recogniser and a conventional triphone recogniser. We do this for both TIMIT and the Dutch read speech corpus, carefully minimising the differences between the two corpora and analysing the remaining (intrinsic) differences. Most importantly, we compare results obtained using two sets of triphone models: one trained with manual (or manually verified) transcriptions and the other with canonical transcriptions. By doing so, we investigate the claim that properly initialised and retrained longer-length acoustic models capture a significant amount of pronunciation variation.

Both TIMIT and the Dutch corpus are read speech corpora. As a consequence, they are not representative of all the problems that are typical of spontaneous conversational speech (hesitations, restarts, repetitions, etc.). However, the kinds of fundamental issues related to articulation that this paper addresses are present in all speech styles.

2. SPEECH MATERIAL

2.1. TIMIT

The DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus [16] is a database comprising a total of 6300 read sentences—ten sentences read by 630 speakers that represent eight major dialects of American English. Seventy percent of the speakers are males and 30% are females.

Two of the sentences for each speaker are identical, and are intended to delineate the dialectal variability of the speakers. We excluded these two sentences from model training and evaluation. Five of the sentences for each speaker originate from a set of 450 phonetically compact sentences, so that seven different speakers speak each of the 450 sentences. The remaining three sentences for each speaker are unique for the different speakers.

The TIMIT data are subdivided into a training set, and two test sets that the TIMIT documentation refers to as the complete test set and the core test set. No sentence or speaker

Table 1: The syllabic structure of the word tokens in TIMIT and CGN.

No. of Syllables	TIMIT/Proportion (%)	CGN/Proportion (%)
1	63.1	62.2
2	22.7	22.6
3	9.3	9.9
4	3.5	3.9
≥ 5	1.4	1.4

Table 2: Proportions of the different types of syllable tokens in TIMIT and CGN.

Type	TIMIT/Proportion (%)	CGN/Proportion (%)
CV	31.6	38.0
CVC	23.8	31.4
VC	10.1	12.6
V	7.3	2.2
CVCC	6.1	5.9
CCV	5.9	3.4
CCVC	4.5	3.4
Other	10.7	3.1

Table 3: TIMIT phone mappings. The remaining phonetic labels of the original set were not changed.

Original label	New label
dx	d
q	—
jh	d z
ch	t sh
zh	z y
em	m
en	n
eng	ng
nx	n
hv	hh
el	l
ih	ix
aw	aa uw
oy	ao ix
ux	uw
er	axr
ax-h	ax

appears in both the training set and the test sets. We used the training set, which comprises 462 speakers and 3696 sentences, for training the acoustic models. The complete test set contains 168 speakers and 1344 sentences, the core test set being a subset of the complete test set and containing 24 speakers and 192 sentences. We used the core test set as the development test set—that is, for optimising the language model scaling factor, the word insertion penalty, and the minimum number of training tokens required for the further training of a longer-length model (see Section 3.3.2). To ensure nonoverlapping test and development test sets, we created the test set by removing the core test set material from the complete test set. We used this test set, which comprised 144 speakers and 1152 sentences, for evaluating the acoustic models.

We intended to build longer-length models for words and syllables for which a sufficient amount of training data was available. To understand the relation between words and syllables, we analysed the syllabic structure of the words in the corpus. The statistics in the second column of Table 1 show that the large majority of all word tokens were monosyllabic. For these words, there was no difference between word and syllable models. In fact, no multisyllabic words occurred often enough in the training data to warrant the training of multisyllabic word models. Hence, the difference between word and syllable models becomes redundant, and we will hereafter refer to the longer-length models as syllable models. According to Greenberg [15], pronunciation variation affects syllable codas and—although to a lesser extent—nuclei more than syllable onsets. To estimate the proportion of syllable tokens that were potentially sensitive to large deviations from their canonical representation, we examined the structure of the syllables in the TIMIT database (see the second column of Table 2). If one considers all consonants after the vowel as coda phonemes, 53.7% of the syllable tokens had coda consonants, and were therefore potentially subject to a considerable amount of pronunciation variation.

TIMIT is manually labelled and includes manually verified phone and word segmentations. For consistency with the experiments on the corpus of Dutch read speech (see Section 2.2), we reduced the original set of phonetic labels to a set of 35 phone labels, as shown in Table 3. To determine the best possible phone mapping, we considered the frequency counts and durations of the original phones, as well as their acoustic similarity with each other. Most importantly, we merged closures with the following bursts and mapped closures appearing on their own to the corresponding bursts. Using the revised set of phone labels, the average number of pronunciation variants per syllable was 2.4. The corresponding numbers of phone substitutions, deletions, and insertions in syllables were 18040, 7617, and 1596.

2.2. CGN

The Spoken Dutch Corpus (Corpus Gesproken Nederlands, CGN) [17] is a database of contemporary standard Dutch spoken by adults in The Netherlands and Belgium. It contains nearly 9 million words (800 hours of speech), of which approximately two thirds originate from The Netherlands and one third from Belgium. All of the data are transcribed orthographically, lemmatised (i.e., grouped into categories of related word forms identified by a headword), and enriched with part-of-speech information, whereas more advanced transcriptions and annotations are available for a core set of the corpus.

For this study, we used read speech from the core set; these data originate from the Dutch library for the blind. To make the CGN data more comparable with the carefully

Table 4: CGN phone mapping. The remaining phonetic labels of the original set were not changed.

Original label	New label
g	k
S	s j
Z	z j
J	n j
E:	E
Y:	Y
O:	O
E~	E
A~	A
O~	O
Y~	Y

spoken TIMIT data, we excluded sentences with tagged particularities, such as incomprehensible words, nonspeech sounds, foreign words, incomplete words, and slips of the tongue from our experiments. The exclusions left us with 5401 sentences uttered by 125 speakers, of which 44% were males and 56% were females. TIMIT contains some repeated sentences; it therefore has higher frequency counts of individual words and syllables, as well as more homogeneous word contexts. Thus, we carried out the subdivision of the CGN data into the training set and the two test sets in a controlled way aimed at maximising the similarity between the training set and the test set on the one hand, and the training set and the development test set on the other hand. First, we created 1000 possible data set divisions by randomly assigning 75% of the sentences spoken by each speaker to the training set and 12.5% to each of the test sets. Second, for each of the three data sets, we calculated the probabilities of word unigrams, bigrams, and trigrams appearing 30 times or more in the set of 5401 sentences. Finally, we computed Kullback-Leibler distances (KLD) [18] between the training set and the two test sets using the above unigram, bigram, and trigram probability distributions. We made each KLD symmetric by calculating it in both directions and taking the average (KLD(p1, p2) = KLD(p2, p1)). The overall KLD-based measure used in evaluating the similarity between the data sets was a weighted sum of the KLDs for the unigram probabilities, the bigram probabilities, and the trigram probabilities. As the final data set division, we chose the division with the lowest overall KLD-based measure.

The final optimised training set comprised 125 speakers and 4027 sentences, whereas the final test sets contained 125 speakers and 687 sentences each. The third column of Table 1 shows how much data was covered by words with different numbers of syllables. As Table 1 illustrates, the word structure of CGN was highly similar to that of TIMIT. The third column of Table 2 illustrates the proportions of the different types of syllable tokens in CGN. CGN had slightly more CV and CVC syllables than TIMIT, but fewer V syllables.

The CGN data comprised manually verified (broad) phonetic and word labels, as well as manually verified word-level segmentations. Only 35 of the original 46 phonetic labels occurred frequently enough for the robust training of triphones. The remaining phones were mapped to the 35 phones, as shown in Table 4. After reducing the number of phonetic labels, the average number of pronunciation variants per syllable was 1.8. The corresponding numbers of phone substitutions, deletions and insertions in syllables were 16358, 6755, and 2875, respectively. Compared with TIMIT, the average number of pronunciation variants, as well as the number of substitutions and deletions, was lower. These numerical differences reflect the differences between the transcription protocols of the two corpora. The TIMIT transcriptions were made from scratch, whereas the CGN transcription protocol was based on the verification of a canonical phonemic transcription. In fact, the CGN transcribers changed the canonical transcription if, and only if, the speaker had realised a clearly different pronunciation variant. As a consequence, the CGN transcribers were probably more biased towards the canonical forms than the TIMIT transcribers; hence, the difference between the manual transcriptions and the canonical representations in CGN is smaller than that in TIMIT.

2.3. *Differences between TIMIT and CGN*

Regardless of our efforts to minimise the differences between TIMIT and CGN, there are some intrinsic differences between them. First and foremost, the two corpora represent two distinct—albeit Germanic—languages. Second, TIMIT contains carefully spoken examples of manually designed or selected sentences, whereas CGN comprises sections of books that the speakers read aloud and, in the case of fiction, sometimes also acted out. Due to the differing characters of the two corpora—and regardless of the optimised data set division of the CGN material—TIMIT contains higher frequency counts of individual words and syllables, and more homogeneous word contexts. Because of this, we chose the CGN training and development data sets to be larger than those for TIMIT. A larger training set guaranteed a similar number of syllables with sufficient training data for training syllable models, and a larger development test set ensured that the corresponding syllables occurred frequently enough for determining the minimum number of training tokens for the models. An additional intrinsic difference between the corpora is that TIMIT comprises five times as many speakers as CGN. Due to the relatively small number of CGN speakers, we included speech from all of the speakers in all of the data sets, whereas the TIMIT speakers do not overlap between the different data sets. All in all, each corpus has some characteristics that make the recognition task easier, and others that make it more difficult, as compared with the other corpus. However, we are confident that the effect of these characteristics does not interfere with our interpretation of the results.

3. EXPERIMENTAL SETUP

3.1. *Feature extraction*

Feature extraction was carried out at a frame rate of 10 milliseconds using a 25-millisecond Hamming window.

First-order preemphasis was applied to the signal using a coefficient of 0.97. 12 Mel frequency cepstral coefficients and log-energy with first, and second-order time derivatives were calculated for a total of 39 features. Channel normalisation was applied using cepstral mean normalisation over individual sentences for TIMIT and complete recordings (with a mean duration of 3.5 minutes) for CGN. Feature extraction was performed using HTK [19].

3.2. *Lexica and language models*

The vocabulary consisted of 6100 words for TIMIT and 10535 words for CGN. Apart from nine homographs in TIMIT and five homographs in CGN, each of which had two pronunciations, the recognition lexica comprised a single, canonical pronunciation per word. We did not distinguish homophones from each other. The language models were word-level bigram networks. The test set perplexity, computed on a persentence basis using HTK [19], was 16 for TIMIT and 46 for CGN. These numbers reflect the inherent differences between the corpora and the recognition tasks.

3.3. *Building the speech recognisers*

In preparation for building a mixed-model recogniser that employed context-independent syllable models and triphones, we built and tested two recognisers: a triphone and a syllable-model recogniser. The performance of the triphone recogniser determined the baseline performance for each recognition task.

3.3.1. *Triphone recogniser*

A standard procedure with decision tree state tying was used for training the word-internal triphones. The procedure was based on asking questions about the left and right contexts of each triphone; the decision tree attempted to find the contexts that made the largest difference to the acoustics and that should, therefore, distinguish clusters [19]. First, monophones with 32 Gaussians per state were trained. The manual (or manually verified) phonetic labels and linear segmentation within the manually verified word segmentations were used for bootstrapping the monophones. Then, the monophones were used for performing a sentence-level forced alignment between the manual transcriptions and the training data; the triphones were bootstrapped using the resulting phone segmentations. When carrying out the state tying, the minimum occupancy count that we used for each cluster resulted in about 4000 distinct physical states in the recogniser. We trained and tested these *"manual triphones"* with up to 32 Gaussians per state.

3.3.2. *Syllable-model recogniser*

The first step of implementing the syllable-model recogniser was to create a recognition lexicon with word pronunciations consisting of syllables. In this lexicon, syllables were represented in terms of the underlying canonical phoneme sequences. For instance, the word "action" in TIMIT was now represented as the syllable models ae_k and sh_ix_n.

To create the syllable lexicon, we had to syllabify the canonical pronunciations of words. In the case of TIMIT, we used the tsylb2 syllabification software available from NIST [20]. tsylb2 is based on rules that define possible syllable-initial and syllable-final consonant clusters, as well as prohibited syllable-initial consonant clusters [21]. The syllabification software produces a maximum of three alternative syllable clusters as output. Whenever several alternatives were available, we used the alternative based on the maximum onset principle (MOP); the syllable onset comprised as many consonants as possible. In the case of CGN, we used the syllabification available in the CGN lexicon and the CELEX lexical database [22]. As in the case of TIMIT, the syllabification of the words adhered to MOP.

After building the syllable lexicon, we initialised the context-independent syllable models with the 8-Gaussian triphone models corresponding to the underlying (canonical) phonemes of the syllables. Reverting to the example word "action" represented as the syllable models ae_k and sh_ix_n, we carried out the initialisation as follows. States 1–3 and 4–6 of the model ae_k were initialised with the state parameters of the 8-Gaussian triphones #-ae+k and ae-k+#, and states 1–3, 4–6, and 7–9 of the model sh_ix_n with the state parameters of the 8-Gaussian triphones #-sh+ix, sh-ix+n, and ix-n+# (see Figure 1). In order to incorporate the spectral and temporal dependencies in the speech, the syllable models with sufficient training data were then trained further using four rounds of Baum-Welch reestimation. To determine the minimum number of training tokens necessary for reliably estimating the model parameters, we built a large number of model sets, starting with a minimum of 20 training tokens per syllable, and increasing the threshold in steps of 20. After each round, we tested the resulting recogniser on the development test set. We continued this process until the WER on the development set stopped decreasing. Eventually, the syllable-model recogniser for TIMIT comprised 3472 syllable models, of which those 43 syllables with a frequency of 160 or higher were trained further. These syllables covered 31% of all the syllable tokens in the training data. The syllable-model recogniser for CGN consisted of 3885 syllable models, the minimum frequency for further training being 130 tokens and resulting in the further training of 94 syllables. These syllables covered 41% of all the syllable tokens in the training data. Syllable models with insufficient training data consisted of a concatenation of the original 8-Gaussian triphone models.

3.3.3. *Mixed-model recogniser*

We derived the lexicon for the mixed-model recogniser from the syllable lexicon by keeping the further-trained syllables from the syllable-model recogniser and expanding all other syllables to triphones. In effect, the pronunciations in the lexicon consisted of the following:

(a) syllables,
(b) canonical phones, or

(c) a combination of (a) and (b).

To use the word "action" as an example, the possible pronunciations were the following:

(a) /ae_k sh_ix_n/,
(b) /#-ae+k ae-k+sh k-sh+ix sh-ix+n ix-n+#/,
(c) /#-ae+k ae-k+# sh_ix_n/, or /ae_k #-sh+ix sh-ix+n ix-n+#/.

The syllable frequencies determined that the actual representation in the lexicon was /#-ae+k ae-k+# sh_ix_n/.

The initial models of the mixed-model recogniser originated from the syllable-model recogniser and the 8-Gaussian triphone recogniser. Four subsequent passes of Baum-Welch reestimation were used to train the mixture of models further. The difference between the syllable-model and mixed-model recognisers was that the triphones underlying the syllables with insufficient training data for further training were concatenated into syllable models in the syllable-model recogniser, whereas they remained free in the mixed-model recogniser. In practice, the triphones whose frequency exceeded the experimentally determined minimum number of training tokens for further training were also trained further in the mixed-model recogniser. The minimum frequency for further training was 20 in the case of TIMIT and 40 in the case of CGN. In the case of TIMIT, the mixed-model recogniser comprised 43 syllable models and 5515 triphones. The mixed-model recogniser for CGN consisted of 94 syllable models and 6366 triphones.

4. SPEECH RECOGNITION RESULTS

Figures 2 and 3 show the recognition results for TIMIT and CGN. We trained and tested manual triphones with up to 32 Gaussian mixtures per state; we only present the results for the triphones with 8 Gaussian mixtures per state, as they performed the best for both corpora. The use of longer-length acoustic models in both the syllable-model and the mixed-model recognisers resulted in statistically significant gains in the recognition performance (using a significance test for a binomial random variable), as compared with the performance of the triphone recognisers. However, the performance of the syllable-model and of the mixed-model recognisers did not significantly differ from each other. In the case of TIMIT, the relative reduction in WER achieved by going from triphones to a mixed-model recogniser was 28%. For CGN, the figure was a more modest 18%. Overall, the results for CGN were slightly worse than those for TIMIT. This can, however, be explained by the large difference in the test set perplexities (see Section 3.2).

The second and third columns of Tables 5 and 6 present the TIMIT and CGN WERs as a function of syllable count when using the triphone and mixed-model recognisers. The effect of the number of syllables is prominent: the probability of ASR errors in the case of monosyllabic words is more than five times the probability of errors in the case of polysyllabic words. This confirms what has been observed in previous ASR research: the more syllables a word has, the less susceptible it is to recognition errors. This can be explained by the fact that a large proportion of monosyllabic words are function words that tend to be unstressed and (heavily) reduced. Polysyllabic words, on the other hand, are more likely to be content words that are less prone to heavy reductions.

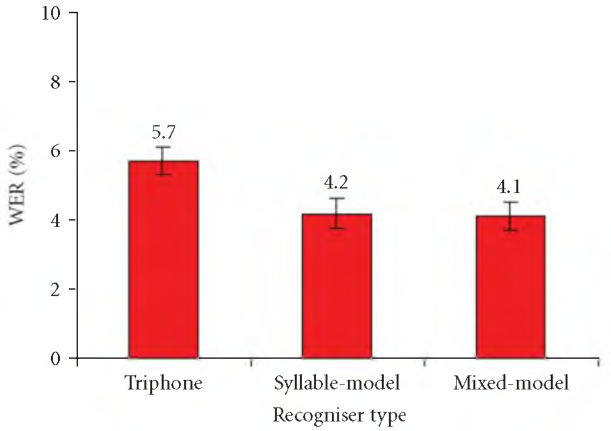

Figure 2: TIMIT WERs, at the 95% confidence level, when using manual triphones.

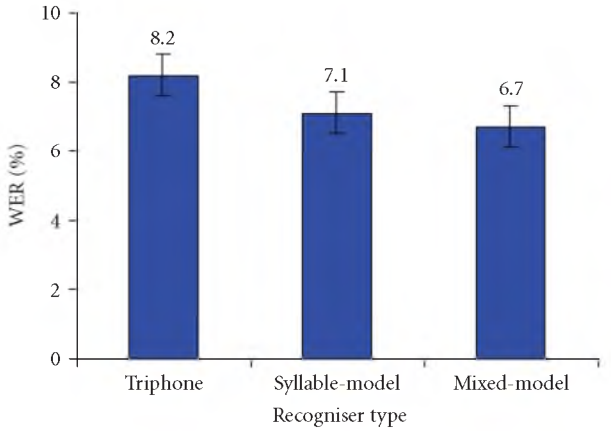

Figure 3: CGN WERs, at the 95% confidence level, when using manual triphones.

The fourth columns of Tables 5 and 6 show the percentage change in the WERs when going from the triphones to the mixed-model recognisers. For TIMIT, the introduction of syllable models results in a 50% reduction in WER in the case of bisyllabic and trisyllabic words. For CGN, the situation is different. The WER does decrease for bisyllabic words, but only by 11%. The WER for trisyllabic words remains unchanged. We believe that this is due to a larger proportion of bisyllabic and trisyllabic words with syllable deletions in CGN. Going from triphones to syllable models without adapting the lexical representations will obviously not help if complete syllables are deleted.

Table 5: TIMIT WERs and percentage change as a function of syllable count when using the triphone and mixed-model recognisers based on manual triphones.

No. of Syllables	Triphone/WER (%)	Mixed-model/WER (%)	Change (%)
1	4.8	3.6	−25
2	0.6	0.3	−50
3	0.2	0.1	−50
4	0.1	0	−100
≥ 5	0	0	±0

Table 6: CGN WERs and percentage change as a function of syllable count when using the triphone and mixed-model recognisers based on manual triphones.

No. of Syllables	Triphone/WER (%)	Mixed-model/WER (%)	Change (%)
1	7.1	5.7	−20
2	0.9	0.8	−11
3	0.2	0.2	±0
4	0.1	0	−100
≥ 5	0	0	±0

5. ANALYSING THE DIFFERENCES

The 28% and 18% relative reductions in WER that we achieved fall short of the 62% relative reduction in WER that Sethy and Narayanan [8] present. Other studies have also used syllable models with varying success. The absolute improvement in recognition accuracy that Sethy et al. [9] obtained with mixed-models was only 0.5%, although the comparison with the Sethy and Narayanan study might not be fair for at least two reasons. First, Sethy et al. used a cross-word left-context phone recogniser, the performance of which is undoubtedly more difficult to improve upon than that of a word-internal context-dependent phone recogniser. Second, their recognition task was particularly challenging with a large amount of disfluencies, heavy accents, age-related coarticulation, language switching, and emotional speech. On the other hand, however, the best performance was achieved using a dual pronunciation recogniser in which each word had both a mixed syllabic-phonetic and a pure phonetic pronunciation variant in the recognition lexicon. Even though Jouvet and Messina [10] employed a parameter sharing method that allowed them to build context-dependent syllable models, the gains from including longer-length acoustic models were small and depended heavily on the recognition task: for telephone numbers, the performance even decreased. In any case, it appears that the improvements on TIMIT, as reported by Sethy and Narayanan and ourselves, are the largest.

Obviously, using syllable models only improves recognition performance in certain conditions. To understand what these conditions are, we carried out a detailed analysis of the differences between the TIMIT and CGN experiments. First, we examined the possible effects of linguistic and phonetic differences between the two corpora. Second, since it is only reasonable to expect improvements in recognition performance if the acoustic models differ between the recognisers, we investigated the differences between the retrained syllable models and the triphones used to initialise them.

5.1. *Structure of the corpora*

In our experiments, we only manipulated the acoustic models, keeping the language models constant. As a consequence, any changes in the WERs are dependent on the so-called acoustic perplexity (or confusability) of the tasks [23]. One should expect a larger gain from better acoustic modelling if the task is acoustically more difficult. The proportion of monosyllabic and polysyllabic words in the test sets provides a coarse approximation of the acoustic perplexity of a recognition task. Table 1, as well as Tables 5 and 6, suggest that TIMIT and CGN do not substantially differ in terms of acoustic perplexity.

Another difference that might affect the recognition results is that the speakers in the TIMIT training and test sets do not overlap, whereas the CGN speakers appear in all three data sets. One might argue that long-span articulatory dependencies are speaker-dependent. Therefore, one would expect syllable models to lead to a larger improvement in the case of CGN, and not vice versa. So, this difference certainly does not explain the discrepancy in the recognition performance.

Articulation rate is known to be a factor that affects the performance of automatic speech recognisers. Thus, we wanted to know whether the articulation rates of TIMIT and CGN differed. We defined the articulation rate as the number of canonical phones per second of speech. The rates were 12.8 phones/s for TIMIT and 13.1 phones/s for CGN, a difference that seems far too small to have an impact.

We also checked for other differences between the corpora, such as the number of pronunciation variants and the

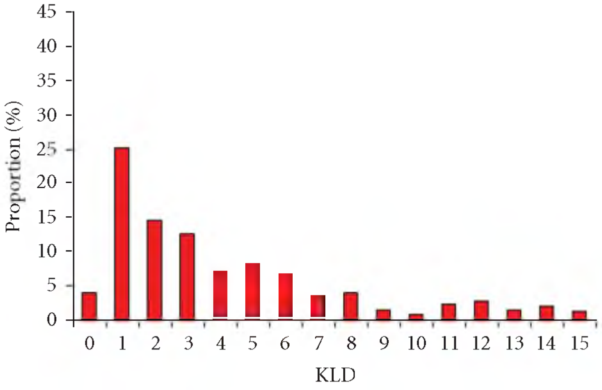

Figure 4: KLD distributions for the states of retrained syllable models for TIMIT when using manual triphones.

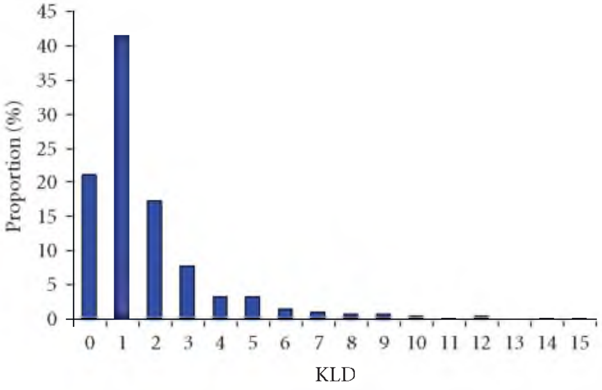

Figure 5: KLD distributions for the states of retrained syllable models for CGN when using manual triphones.

durations of syllables. However, we were not able to identify any linguistic or phonetic properties of the corpora that could possibly explain the differences in the performance gain.

5.2. *Effect of further training*

To investigate what happens when syllable models are trained further from the sequences of triphones used for initialising them, we calculated the distances between the probability density functions (pdfs) of the HMM states of the retrained syllable models and the pdfs of the corresponding states of the initialised syllable models in terms of the Kullback-Leibler distance (KLD) [18]. Figures 4 and 5 illustrate the KLD distributions for TIMIT and CGN. The distributions differ from each other substantially, the KLDs generally being higher in the case of TIMIT. This implies that the further training affected the TIMIT models more than the CGN models. Given the greater impact of the longer-length models on the recognition performance, this is what one would expect.

There were two possible reasons for the larger impact of the further training on the TIMIT models. Either the boundaries of the syllable models with the largest KLDs had shifted substantially, or the effect was due to the switch from the manually labelled phones to the retrained canonical representations of the syllable models. Since syllable segmentations obtained through forced alignment did not show major differences, we pursued the issue of potential discrepancies between manual and canonical transcriptions. To that end, we performed additional speech recognition experiments, in which triphones were trained using the canonical transcriptions of the uttered words. These *"canonical triphones"* were then used for building the syllable-model and mixed-model recognisers.

In the case of TIMIT, the mixed-model recogniser based on canonical triphones contained 86 syllable models that had been trained further within the syllable-model recogniser using a minimum of 100 tokens. The corresponding syllables covered 42% of all the syllable tokens in the training data. The mixed-model recogniser for CGN comprised 89 syllable models trained further using a minimum of 140 tokens, and the corresponding syllables covered 56% of all the syllable tokens in the training data. Further Baum-Welch reestimation was not necessary for the mixture of triphones and syllable models; tests on the development test set showed that training the mixture of models further would not lead to improvements in the recognition performance. This was different from the syllable models initialised with the manual triphones; tests on the development test set showed that the mixture of models should be trained further for optimal performance. With hindsight, this is not surprising. As a result of the retraining, the syllable models initialised in the two different ways became very similar to each other. However, the syllable models that were initialised with the manual triphones were acoustically further away from this final "state" than the syllable models that were initialised with the canonical triphones and, therefore, needed more reestimation rounds to conform to it.

Figures 6 and 7 present the results for TIMIT and CGN. The best performing triphones had 8 Gaussian mixtures per state in the case of TIMIT and 16 Gaussian mixtures per state in the case of CGN. Surprising as it may seem, the results obtained with the canonical triphones substantially outperformed the results achieved with the manual triphones (see Figures 2 and 3). In fact, the canonical triphones even outperformed the original mixed-model recognisers (see Figures 2 and 3). The performances of the mixed-model recognisers containing syllable models trained with the two differently trained sets of triphones did not differ significantly at the 95% confidence level. In addition, the performance of the canonical triphones was similar to that of the new mixed-model recognisers. Smaller KLDs between the initial and the retrained syllable models (see Figures 8 and 9) reflected the lack of improvement in the recognition performance. Evidently, only a few syllable models benefited from the further training, leaving the overall effect on the recognition performance negligible. These results are in line with results from other studies [4, 9, 10], in which improvements achieved

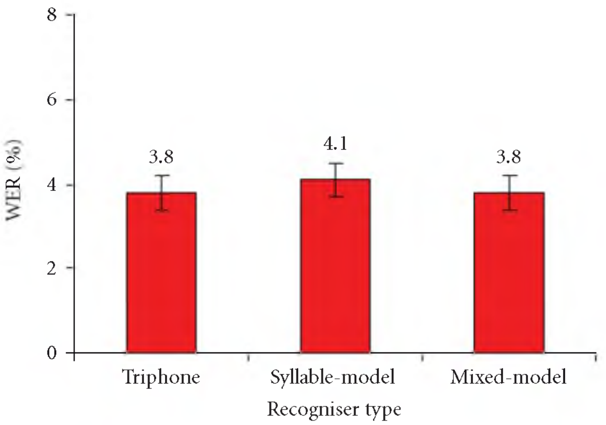

Figure 6: TIMIT WERs, at the 95% confidence level, when using canonical triphones.

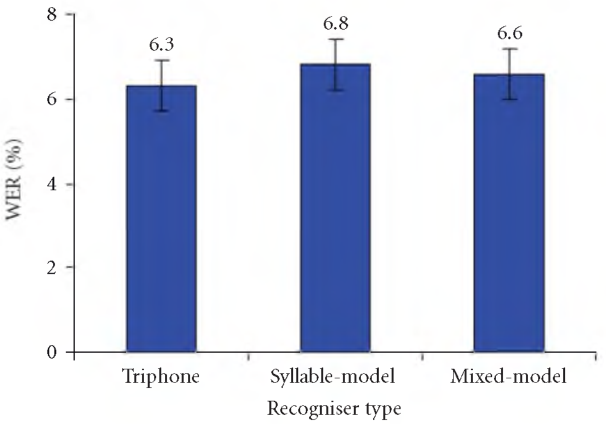

Figure 7: CGN WERs, at the 95% confidence level, when using canonical triphones.

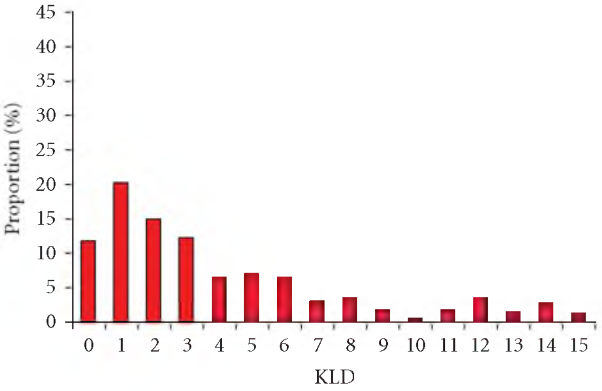

Figure 8: KLD distributions for the states of retrained syllable models for TIMIT when using canonical triphones.

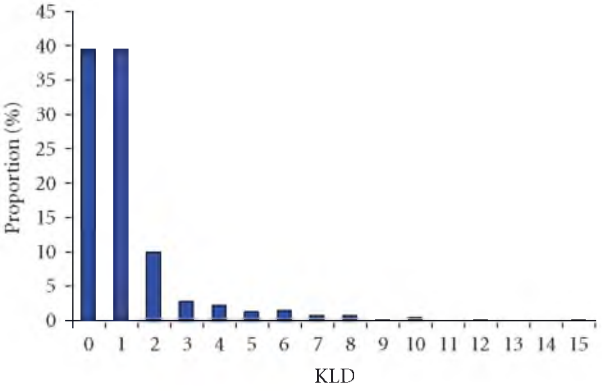

Figure 9: KLD distributions for the states of retrained syllable models for CGN when using canonical triphones.

with longer-length acoustic models are small, and deteriorations also occur.

The second and third columns of Tables 7 and 8 present the TIMIT and CGN WERs as a function of syllable count when using the triphone and mixed-model recognisers. As in the case of the experiments with manual triphones (see Tables 5 and 6), the probability of errors was considerably higher for monosyllabic words than for polysyllabic words. The fourth columns of the tables show the percentage change in the WERs when going from the triphones to the mixed-model recognisers. The data suggest that the introduction of syllable models might deteriorate the recognition performance in particular in the case of bisyllabic words. This may be due to the context-independency of the syllable models and the resulting loss of left or right context information at the syllable boundary. As words tend to get easier to recognise as they get longer (see Section 5.1), the words with more than two syllables do not seem to suffer from this effect.

The most probable explanation for the finding that the canonical triphones outperform the manual triphones is the mismatch between the representations of speech during training and testing. While careful manual transcriptions yield more accurate acoustic models, the advantage of these models can only be reaped if the recognition lexicon contains a corresponding level of information about the pronunciation variation present in the speech [24]. Thus, at least part, if not all, of the performance gain obtained with retrained syllable models in the first set of experiments (and probably also in Sethy and Narayanan's work [8]) resulted from the reduction of the mismatch between the representations of speech during training and testing. Because the manual transcriptions in CGN were closer to the canonical transcriptions than those in TIMIT (see Section 2.2), the mismatch was smaller for CGN. This also explains why the impact of the syllable models was smaller for CGN.

6. DISCUSSION

So far, explicit pronunciation variation modelling has made a disappointing contribution to improving speech recognition performance [25]. There are many different ways to attempt implicit modelling. To avoid the increased lexical

Table 7: TIMIT WERs and percentage change as a function of syllable count when using the triphone and mixed-model recognisers based on canonical triphones.

No. of Syllables	Triphone/WER (%)	Mixed-model/WER (%)	Change (%)
1	3.2	3.2	±0
2	0.4	0.5	+25
3	0.1	0.1	±0
4	0	0	±0
≥ 5	0	0	±0

Table 8: CGN WERs and percentage change as a function of syllable count when using the triphone and mixed-model recognisers based on canonical triphones.

No. of Syllables	Triphone/WER (%)	Mixed-model/WER (%)	Change (%)
1	5.4	5.6	+4
2	0.6	0.8	+33
3	0.2	0.2	±0
4	0.1	0	−100
≥ 5	0	0	±0

confusability of a multiple pronunciation lexicon, Hain [25] focused on finding a single optimal phonetic transcription for each word in the lexicon. Our study confirms that a single pronunciation that is consistently used both during training and during recognition is to be preferred over multiple pronunciations derived from careful phonetic transcriptions. This is in line with McAllaster and Gillick's [5] findings, which also suggest that consistency between—potentially inaccurate—symbolic representations used in training and recognition is to be preferred over accurate representations in the training phase if these cannot be carried over to the recognition phase.

The focus of the present study was on implicit modelling of long-span coarticulation effects by using syllable-length models instead of the context-dependent phones that conventional automatic speech recognisers use. We expected Baum-Welch reestimation of these models to capture phonetic detail that cannot be accounted for by means of explicit pronunciation variation modelling at the level of phonetic transcriptions in the recognition lexicon. Because of the changes we observed between the initial and the retrained syllable models (see Figures 8 and 9), we do believe that retraining the observation densities incorporates coarticulation effects into the longer-length models. However, the corresponding recognition results (see Figures 6 and 7) show that this is not sufficient for capturing the most important effects of pronunciation variation at the syllable level. Greenberg [15], amongst other authors, has shown that while syllables are seldom deleted completely, they do display considerable variation in the identity and number of the phonetic symbols that best reflect their pronunciation. Greenberg and Chang [26] showed that there is a clear relation between recognition accuracy and the degree to which the acoustic and lexical models reflect the actual pronunciation. Not surprisingly, the match (or mismatch) between the knowledge captured in the models on the one hand and the actual articulation is dependent on linguistic (e.g., prosody, context) as well as nonlinguistic (e.g., speaker identity, speaking rate) factors. Sun and Deng [27] tried to model the variation in terms of articulatory features that are allowed to overlap in time and change asynchronously. Their recognition results on TIMIT are much worse than what we obtained with a more conventional approach.

We believe that the aforementioned problems are caused by the fact that part of the variation in speech (e.g., phone deletions and insertions) results in very different trajectories in the acoustic parameter space. These differently shaped trajectories are not easy to model with observation densities if the model topology is identical for all variants. We believe that pronunciation variation could be modelled better by using syllable models with parallel paths that represent different pronunciation variants, and by reestimating these parallel paths to better incorporate the dynamic nature of articulation. Therefore, our future research will focus on strategies for developing multipath model topologies for syllables.

7. CONCLUSIONS

This paper contrasted recognition results obtained using longer-length acoustic models for Dutch read speech from a library for the blind with recognition results achieved on American English read speech from TIMIT. The topologies and model parameters of the longer-length models were initialised by concatenating the triphone models underlying their canonical transcriptions. The initialised models were then trained further to incorporate the spectral and temporal dependencies in speech into the models. When using manually labelled speech to train the triphones, mixed-model recognisers comprising syllable-length and phoneme-length models substantially outperformed them. At first sight, these results seemed to corroborate the claim that properly initialised and retrained longer-length acoustic models capture a significant amount of pronunciation variation. However, detailed analyses showed that the effect of training syllable-sized models further is negligible if canonical representations of the syllables are initialised with triphones trained with the canonical transcriptions of the training corpus. Therefore, we conclude that single-path syllable models that borrow their topology from a sequence of triphones cannot capture the pronunciation variation phenomena that hinder recognition performance the most.

ACKNOWLEDGMENTS

The research was carried out within the Interactive Multimodal Information eXtraction (IMIX) program sponsored by The Netherlands Organisation for Scientific Research (NWO).

REFERENCES

[1] M. Ostendorf, "Moving beyond the 'beads-on-a-string' model of speech," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '99)*, pp. 79–84, Keystone, Colo, USA, December 1999.

[2] J. M. Kessens, C. Cucchiarini, and H. Strik, "A data-driven method for modeling pronunciation variation," *Speech Communication*, vol. 40, no. 4, pp. 517–534, 2003.

[3] D. Jurafsky, W. Ward, Z. Banping, K. Herold, Y. Xiuyang, and Z. Sen, "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, Utah, USA, May 2001.

[4] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski, and G. R. Doddington, "Syllable-based large vocabulary continuous speech recognition," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 4, pp. 358–366, 2001.

[5] D. McAllaster and L. Gillick, "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, September 1999.

[6] B. Plannerer and G. Ruske, "Recognition of demisyllable based units using semicontinuous hidden Markov models Plannerer," in *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP '92)*, vol. 1, pp. 581–584, San Francisco, Calif, USA, March 1992.

[7] R. J. Jones, S. Downey, and J. S. Mason, "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, September 1997.

[8] A. Sethy and S. Narayanan, "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. 772–775, Hong Kong, April 2003.

[9] A. Sethy, B. Ramabhadran, and S. Narayanan, "Improvements in English ASR for the MALACH project using syllable-centric models," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU '03)*, pp. 129–134, St. Thomas, Virgin Islands, USA, November-December 2003.

[10] D. Jouvet and R. Messina, "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, October 2004.

[11] A. Hämäläinen, J. de Veth, and L. Boves, "Longer-length acoustic units for continuous speech recognition," in *Proceedings of European Signal Processing Conference (EUSIPCO '05)*, Antalya, Turkey, September 2005.

[12] A. Hämäläinen, L. Boves, and J. de Veth, "Syllable-length acoustic units in large-vocabulary continuous speech recognition," in *Proceedings of the 10th International Conference on Speech and Computer (SPECOM '05)*, pp. 499–502, Patras, Greece, October 2005.

[13] N. O. Schiller, A. S. Meyer, and W. J. M. Levelt, "The syllabic structure of spoken words: evidence from the syllabification of intervocalic consonants," *Language and Speech*, vol. 40, no. 2, pp. 103–140, 1997.

[14] C. Pallier, "Phonemes and syllables in speech perception: size of attentional focus in French," in *Proceedings of the 5th European Conference on Speech Communication and Technology (EUROSPEECH '97)*, pp. 2159–2162, Rhodes, Greece, September 1997.

[15] S. Greenberg, "Speaking in shorthand—a syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, no. 2, pp. 159–176, 1999.

[16] "TIMIT acoustic-phonetic continuous speech corpus," NTIS Order PB91-505065, National Institute of Standards and Technology, Gaithersburg, Md, USA, 1990, Speech Disc 1-1.1.

[17] N. Oostdijk, W. Goedertier, F. Van Eynde, et al., "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, May 2002.

[18] S. Kullback and R. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[19] S. Young, G. Evermann, T. Hain, et al., *The HTK Book (for HTK Version 3.2.1)*, Cambridge University, Cambridge, UK, 2002.

[20] W. M. Fisher, "tsylb2-1.1 syllabification software," August 1996, http://www.nist.gov/speech/tools/index.htm.

[21] D. Kahn, *Syllable-based generalisations in English phonology*, Ph.D. thesis, Indiana University Linguistics Club, Bloomington, Ind, USA, 1976.

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

[23] H. Printz and P. Olsen, "Theory and practice of acoustic confusability," in *Proceedings of Automatic Speech Recognition: Challenges for the New Millenium (ISCA ITRW ASR '00)*, pp. 77–84, Paris, France, September 2000.

[24] M. Wester, *Pronunciation variation modeling for Dutch automatic speech recognition*, Ph.D. thesis, University of Nijmegen, Nijmegen, The Netherlands, 2002.

[25] T. Hain, "Implicit modelling of pronunciation variation in automatic speech recognition," *Speech Communication*, vol. 46, no. 2, pp. 171–188, 2005.

[26] S. Greenberg and S. Chang, "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, September 2000.

[27] J. Sun and L. Deng, "An overlapping-feature-based phonological model incorporating linguistic constraints: applications to speech recognition," *Journal of the Acoustical Society of America*, vol. 111, no. 2, pp. 1086–1101, 2002.