VARIATION IN SPEECH

Describing Continuous Phenomena with Discrete Representations

Cover photo and design: Annika Hämäläinen
Printed and bound by Ipskamp Drukkers B.V., Nijmegen

ISBN: 978-90-9024276-7

VARIATION IN SPEECH

Describing Continuous Phenomena with Discrete Representations

Een wetenschappelijke proeve
op het gebied van de Letteren

Proefschrift

ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. mr. S.C.J.J. Kortmann,
volgens besluit van het College van Decanen
in het openbaar te verdedigen op vrijdag 4 december 2009
om 13.00 uur precies

door

Kaisa Annika Hämäläinen

geboren op 6 december 1975
te Rovaniemi, Finland.

Promotor

Prof. dr. L. Boves

Copromotor

Dr. L. ten Bosch

Manuscriptcommissie

Prof. dr. D. van Leeuwen

Prof. dr. J.-P. Martens, Universiteit Gent

Dr. R. van Son, Universiteit van Amsterdam

Acknowledgements

This thesis would never have seen the light of day without the help and support from a number of people and organisations that I would now like to thank.

In particular, I would like to thank my *promotor,* Lou Boves, and my *begeleider,* Louis ten Bosch, for sharing their ideas and expertise, and for their help and guidance throughout my project. Whenever I was convinced that my negative speech recognition results were good for nothing, Lou would come up with a way of turning them into publications with a positive slant. Working with someone as easy-going and enthusiastic as Louis was an absolute privilege. Both of my supervisors were very supportive when I decided to take five months off to work in industry, which is something that I am very grateful for as, without that work experience, I might not be working in industry now.

Thanks are also due to other colleagues and ex-colleagues from the Department of Language and Speech. I started my project under the *begeleiding* of Johan de Veth, during whose era the all-important technical basis for my experiments was built. Arek Nagorski, Christophe Van Bael, Diana Binnenpoorte, Eric Sanders, Michele Gubian and Yan Han were always willing to share their scripts, their programming tricks, and their knowledge on data and software. Bert Cranen and Hans Adamse were just an email away even when computer problems struck in the evenings or at weekends. Hella Cranen-Jooren was always happy to help with practical issues. Gies Bouwman helped me to get started with designing the cover of this thesis. Many colleagues also gave feedback on my papers, posters and presentations.

Furthermore, I would like to thank Mark Pluymaekers, Mirjam Ernestus and Harald Baayen for sharing their experimental data, and Mirjam Ernestus for her input when we started the research for the last article included in this thesis. I am also grateful to the anonymous reviewers of my papers and the members of my manuscript committee for their comments and time. Thanks are also due to Mark Rawlings-Smith who proofread my English texts on several occasions.

As for financial support, I am indebted to the following organisations: Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Walter Ahlströmin säätiö, Stichting Spraaktechnologie, International Speech Communication Association, and Insinööriliitto. Without this support, I would not have been able to carry out my research, or to attend as many conferences and workshops as I did.

Finally, a special thank you goes to my friends for providing moral support and much-needed distraction from my work, and to my family for their love and support over the years.

Contents

I INTRODUCTION

Introduction

Automatic Speech Recognition (ASR) is the technology aimed at correctly identifying the sequence of words which corresponds to a given stretch of an acoustic speech signal. The basic components of a speech recogniser are shown in the block diagram of Figure 1. At a high level, ASR consists of two steps: (a) feature extraction, i.e., the extraction of the relevant spectro-temporal details from the acoustic signal, and (b) the word search, in which the best matching word sequence is searched for.

The conventional form of feature extraction converts the speech signal into a sequence of so-called acoustic feature vectors, each representing the spectral properties in a short stretch of the signal. This representation of speech is used as input for the word search. Three different knowledge bases are essential for the word search: the lexicon, the acoustic models and the language model. The lexicon describes the set of possible words (orthographies) in the speech recognition task in terms of shorter speech units, such as phonemes. These speech units correspond to acoustic models, which model the acoustic properties of the speech units. The conventional approach to acoustic modelling is to use Hidden Markov Models (HMMs), which deal with the speech signal as a sequence of acoustic feature vectors, and allow each speech unit to be modelled in terms of the statistical properties of these acoustic feature vectors. The goal of the language model is to guide the word search to look for word sequences that are plausible in the language of the recognition task; the language model estimates the prior probability of a certain sequence of words. The acoustic models and the language model have to be trained on speech and text corpora before a recognition task can be performed. Together, the lexicon, the acoustic models and the language model 'model' the acoustic realisations of all possible sentences in the language. For a thorough introduction to ASR, the reader is referred to Holmes & Holmes (2001), and Rabiner & Juang (1993).

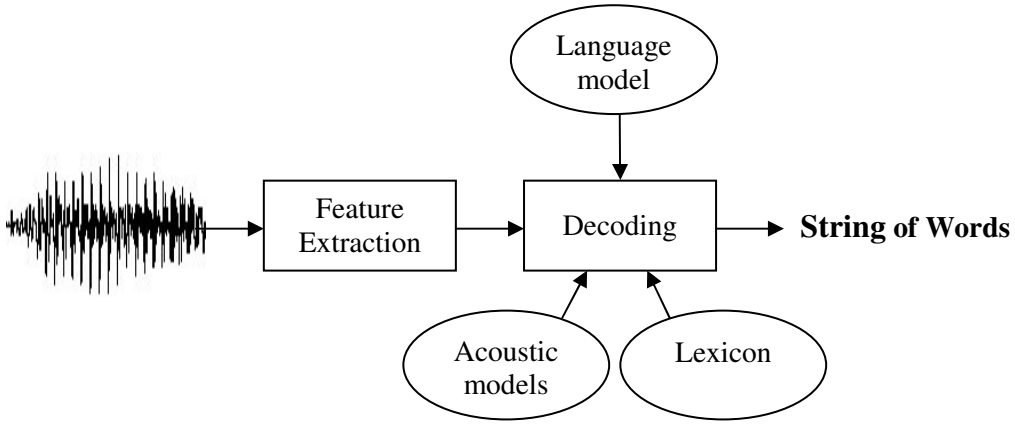

Figure 1: The basic components of a speech recogniser.

The algorithms developed for ASR can be used with two different goals in mind. One of these goals is to strive for better recognition accuracy on any given recognition task. As testified by the publications in the field, this is the goal of the majority of ASR studies. The other goal is very different from the first goal, and involves using ASR-based algorithms as a speech analysis tool – for example, to unravel phonetic details of the speech signal. In this case, ASR-based algorithms are used to locate and/or identify specific acoustic/phonetic events in the speech signal.

The body of this thesis consists of three articles. These three articles describe studies addressing both of the aforementioned goals. The first two articles are related to the first goal; they detail approaches to improve upon the performance of a conventional HMM-based speech recogniser by specifically focussing on *pronunciation variation*, one of the most important causes of speech recognition errors. The third article, on the other hand, describes a study related to the second goal; it explores the possibility of using ASR to analyse a specific type of pronunciation variation, namely acoustic reduction, in the speech signal.

As all of the articles included in this thesis are related to pronunciation variation, we continue by introducing the phenomenon of pronunciation variation in more detail. We then explain why conventional automatic speech recognisers have problems dealing with pronunciation variation, and propose an alternative approach that is investigated in the first two articles. We go on to suggest an ASR-based method to analyse acoustic reduction in the speech signal, as described in more detail in the third article. We end this introduction with an overview of the articles included in this thesis.

1 A brief introduction to pronunciation variation

Pronunciation variation is the phenomenon of words never being pronounced in exactly the same way by different (between-speaker variation) or even by the same (within-speaker variation) speakers. Between-speaker variation is caused by differences in vocal tract geometry, age, gender, regional and social accent, voice quality, and so on. The degree of within-speaker variation is affected by factors such as speaking style, speaking rate, state of health or emotional state of the speaker, allophonic variation, and suprasegmental features. (Wester, 2002.)

From the point of view of this thesis, allophonic variation is the most interesting type of pronunciation variation (even though different types of variation cannot be teased apart very easily). Allophones are variants of a particular phoneme; allophonic variation affects individual segments of a word due to context, such as surrounding sounds and syllable structure. Phonemes can often be related to their allophones in terms of phonological rules. Examples of phonological rules relating phonemes to their allophones are assimilation, deletion, epenthesis and reduction. Many of these phonological rules model coarticulation – a change in a segment caused by the movement of the articulators in the preceding or following segments. Variation in the degree of coarticulation is one of the causes of pronunciation variation. Assimilation, for instance, is the change of one sound segment into another because of the influence of

neighbouring sound segments on its articulation. For instance, the underlying unvoiced /k/ at the end of the first part of the Dutch compound 'zakdoek' ('handkerchief') changes into the voiced /g/ because of the voiced /d/ at the beginning of the second part of the compound. So, instead of the 'expected' pronunciation /zAkduk/, the canonical pronunciation of the word 'zakdoek' is /zagduk/. Deletion is quite common particularly in the case of faster or more casual speech; an example is the deletion of the final /t/ in the Dutch word 'niet' ('no'). Epenthesis is the insertion of one or more sounds in the middle of a word. For instance, the Dutch word 'werken' ('to work') is often pronounces /wEr@k@n/, while the canonical pronunciation is /wErk@n/. As an example of reduction, vowel reduction is one of the most important phonological processes. As a result of vowel reduction, many vowels in unstressed syllables are realised as 'reduced' vowels, the most common of which is schwa. While phonemes can often be related to their allophones in terms of phonological rules, this is virtually impossible in the case of 'massive reduction' (Johnson, 2004); the phonetic realisation of a word may involve a large deviation from the canonical pronunciation such that whole syllables are lost and/or a large proportion of the phones are changed. (Jurafsky & Martin, 2000; Ladefoged, 2001.)

2 Focus on acoustic modelling

2.1 Pronunciation variation and phoneme-based acoustic models

Pronunciation variation is known to cause problems for large-vocabulary continuous speech recognition, which has traditionally viewed speech as a sequence of *discrete* phonemes ('beads on a string'; Ostendorf, 1999) and modelled each individual phoneme by an HMM. To illustrate these problems, let us take an example from the Spoken Dutch Corpus (Oostdijk et al., 2002). In Figure 2, the canonical and actual (manually verified) pronunciations of the word sequence 'maar jij was er niet bij' ('but you were not there') are presented in terms of phonemic transcriptions. As we can see, the word-final consonants /r/, /r/ and /t/ of the words 'maar' ('but'), 'er' ('there') and 'niet' ('not') have been deleted, and the /s/ of the word 'was' ('were') has changed quality from the voiceless /s/ to the voiced /z/ because of the following vowel (another example of assimilation).

```
m a r j E+ w A s E r n i t b E+
m a   j E+ w A z E   n i   b E+
```

Figure 2: Canonical and actual pronunciations of the word sequence 'maar jij was er niet bij' (using the Spoken Dutch Corpus phone set).

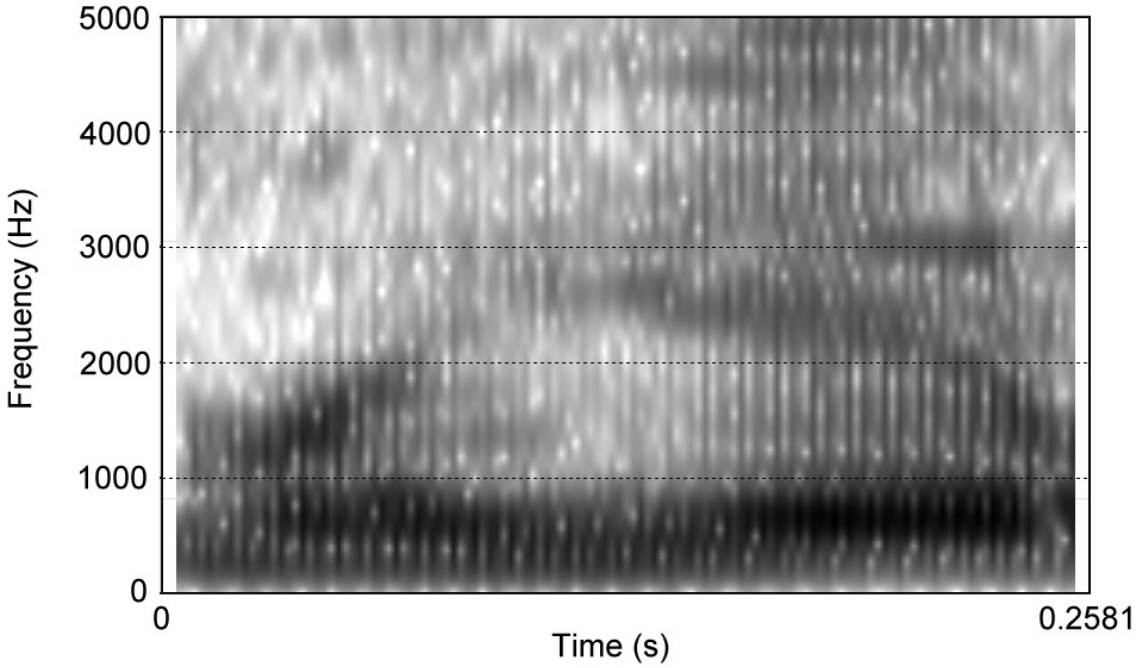

Figure 3: Spectrogram of the word sequence 'maar jij'.

Figure 3 shows a spectrogram representation of the word sequence 'maar jij' ('but you') from our example. Let us consider the spectrogram from the point of view of manual phonetic segmentation. What Figure 3 essentially illustrates is that speech is a sequence of spectro-temporal patterns that gradually transform from one to another. Even though we know that the phoneme sequence present is /majE+/, it is very difficult to accurately locate the phonemes and their boundaries in the spectrogram. This is because the transitions from one phoneme to another are particularly prone to phenomena such coarticulation. It is even possible to get the impression that a speaker has realised a specific phoneme, without there being a signal interval that can unambiguously be assigned to that phoneme. In these cases the 'underlying' presence of that phoneme is reflected in coarticulation effects (e.g., glottalisation) on the neighbouring phonemes (Jurafsky & Martin, 2000; Ostendorf, 1999). Even if the phoneme sequence /majE+/ could be segmented precisely, it would still differ from the phoneme sequence expected based on the canonical transcriptions of the words 'maar' (/mar/) and 'jij' (/jE+/).

The above-mentioned difficulties of manual phonetic segmentation illustrate why viewing speech as a sequence of discrete phonemes is also problematic for ASR. After all, such a view of speech only allows pronunciation variation to be represented in terms of phoneme-level substitutions, deletions and insertions – without the possibility of modelling the kinds of long-span spectro-temporal patterns illustrated by Figure 3 (Ostendorf, 1999). As a matter of fact, the traditional approach to pronunciation variation modelling involves listing multiple alternative phonemic representations of words in the recognition lexicon, with phonemes substituted,

deleted and inserted with respect to the canonical pronunciation. For instance, the pronunciation variation in the Dutch word 'bijvoorbeeld' ('for example') might be modelled by including both the canonical pronunciation /bE+vorbelt/ and the non-canonical pronunciation /b@vorb@lt/ in the lexicon. This approach has, however, met with limited success because of the resulting increase in lexical confusability (Kessens et al., 2003).

Other attempts to model pronunciation variation with phoneme-based acoustic models have, for instance, involved modifying the topologies of the phoneme-based models. In conventional automatic speech recognisers, phoneme-based HMMs typically have three emitting Markov states; three emitting states per model is a practical choice that has proved to result in reasonable speech recognition performance. Three emitting states is, however, not at all imperative. In the past, researchers have experimented with different kinds of model topologies in an attempt to model phoneme-level pronunciation variation, such as reduction and assimilation processes. Examples of these include phoneme-based models with skip states (e.g., Bakis, 1976) and phoneme-based models with parallel paths through the model (e.g., Lee, 1989).

2.2 From phoneme-based to longer-length acoustic models

How could pronunciation variation successfully be modelled in large-vocabulary continuous speech recognition? One of the possible ways is to use longer-length acoustic models. Since part of the variation is context-dependent, the inclusion of context *in* the model will simplify the description of variation *within* the model. The first two articles included in this thesis explore the idea of using longer-length acoustic models that would have phonemic variation and long-span spectro-temporal patterns inherently embedded into them. The motivations for the idea of longer-length acoustic models are both linguistic and technical. The linguistic motivation comes from human speech production and perception. While most humans have no difficulty recounting exactly which words or syllables have been uttered, they often struggle doing so in the case of phonemes. The technical motivation comes from the successful use of word models and so-called word-specific phoneme models (Odell & Durrani, 2006). Word models have successfully been used in applications with limited vocabularies – such as digit recognition and command and control applications. A limited vocabulary allows the use of longer-length models (even word models) without the risk of data sparseness issues; in the case of a limited vocabulary, the training corpus can easily contain enough instances of each longer-length speech unit. Similarly, word-specific phoneme models – i.e., phoneme-based models trained in specific lexical contexts – have successfully been used in commercial speech recognisers. Unlike phoneme-based models that are trained using instances of the phoneme in several different lexical contexts, longer-length acoustic models and word-specific phoneme models capture variation in a specific lexical context. They would therefore seem to be acoustically more accurate. It is, however, evident that there is a trade-off when using longer-length acoustic models: as the stretches of speech to be modelled are longer, longer-length acoustic models

allow more accurate modelling of the segments involved but, at the same time, must capture more variation. When trained without clever data sharing, longer-length acoustic models require more training data than phoneme-based models (Sethy & Narayanan, 2003). As the speech units become longer, the number of units with little or no acoustic data available for model parameter estimation will increase. If the speech units are words, there is an unbounded increase in the number of possible units. This so-called 'data sparseness' is an important issue when considering the use of longer-length acoustic models for large-vocabulary continuous speech recognition.

The use of longer-length acoustic models in large-vocabulary continuous speech recognition is not a novel idea. Syllable-based acoustic models have been suggested and even successfully used for some Asian tone languages, such as Mandarin Chinese (Hon et al., 1994) and Thai (Wutiwiwatchai & Furui, 2007). As for European languages, Ganapathiraju et al. (2001), Jouvet & Messina (2004), Sethy & Narayanan (2003) and Sethy et al. (2003), report speech recognition results with syllable-length acoustic models for English and French. Plannerer & Ruske (1992), on the other hand, suggest using demi-syllable-based models for German, while Jouvet & Messina (2004) also experiment with automatically derived longer-length acoustic models for French. Many recognition results reported in the aforementioned studies on European languages show promise. However, the results range from deterioration (e.g., Jouvet & Messina, 2004) to seemingly enormous improvements (Sethy & Narayanan, 2003) in speech recognition performance as compared with conventional phoneme-based speech recognisers. This makes it quite obvious that the use of longer-length acoustic models does not necessarily lead to improved speech recognition performance. Yet, earlier studies present speech recognition results without in-depth analysis on the aspects of pronunciation variation that the longer-length models are actually able to capture.

In fact, our main criticism on the earlier studies is the above-mentioned lack of in-depth analysis of the speech recognition results. Like Bourlard et al. (1996), we would like to support the view that it is not improvement in speech recognition performance 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. The speech recognition community can learn a lot from carefully analysed results. The first two articles included in this thesis try, in their part, to fill this gap in the literature. In the studies described in these articles, we carry out speech recognition experiments with syllable-length acoustic models and analyse our results in ways that shed light on the reasons behind the changes observed in the recognition performance. This way, we are able to increase our understanding of the complex issues playing a role in pronunciation variation modelling with syllable-length models. The insights we provide should be of help for future research on pronunciation variation modelling.

2.3 Using acoustic models for analysing acoustic reduction

Reduction basically originates in articulation: the trajectories taken by the moving articulators approximate the 'canonical' targets, without reaching all of them. These articulatory shortcuts have an impact on the acoustic signal. Traditionally used measures of acoustic reduction include, for instance, the duration of speech segments, the formant values of vowels, and the centre of gravity of spectra. However, the relation between articulatory and acoustic reduction is so complex that simple acoustic measures are unlikely to capture all the variation. In any case, all scholars investigating reduction agree that the phenomenon of reduction must be interpreted as deviation from a canonical pronunciation. Reduction is then manifest in the deviation between an observed acoustic token (e.g., a particular instance of a phoneme or a word) and the canonical model of the token (e.g., an average acoustic representation of a carefully pronounced token of the phoneme or the word).

Automatic speech recognisers might be able to provide estimates of the degree of reduction in a particular stretch of a speech signal. During the so-called forced alignment process, a speech recogniser is given a particular sequence (or several particular sequences) of acoustic models to align with a particular stretch of a speech signal. The recogniser then returns the log-likelihoods ('acoustic scores') of those acoustic models given the speech signal. Forced alignment is commonly used to estimate the actual pronunciation of words in an utterance: given a set of possible phonemic transcriptions of the words and a set of phoneme-based models corresponding to those phonemic transcriptions, the speech recogniser returns the sequence of phonemes that best matches the speech signal in terms of the acoustic scores. Should the speech recogniser only be given one possible phonemic transcription per word (or any other speech unit), the total acoustic score for each instance of a given word would express how well the signal matches that single phonemic transcription. Should that single transcription be a canonical transcription, the total acoustic score would express how deviant the signal is from the canonical transcription. Therefore, the total acoustic score might be able to serve as an estimate of the degree of reduction in the word (or some other speech unit). The third article included in this thesis describes a study in which the idea of using acoustic scores as a measure of the degree of acoustic reduction in the speech signal is explored for four Dutch affixes.

3 Overview

This section provides a short overview of the three articles included in the body of this thesis. The thesis concludes with a summary of the main results and conclusions of the articles, as well as suggestions for future research.

Article 1: On the utility of syllable-based acoustic models for pronunciation variation modelling

This study aims at increasing our understanding of the conditions in which syllable-length acoustic models result in improvements in the performance of large-vocabulary continuous

speech recognisers. The motivation for the study comes from the relatively large improvements in recognition performance that Sethy & Narayanan (2003) report with syllable-length acoustic models with a single path through the model, as compared with the more modest performance that other studies (Ganapathiraju et al., 2001; Jouvet & Messina, 2004; Sethy et al., 2003) report. To answer our research question, we replicate Sethy & Narayanan's (2003) speech recognition experiments, carry out similar experiments on a Dutch speech corpus, and analyse the differences between the two sets of results. In particular, we focus on the role of the procedure used to initialise the syllable-length acoustic models. Sethy & Narayanan (2003) proposed initialising the syllable-length acoustic models using a sequence of phoneme-based acoustic models corresponding to the canonical transcriptions of the syllables in question. We establish that the details of the procedure used for training these phoneme-based models have a substantial effect on the speech recognition results. Training the phoneme-based models using manual(ly verified) transcriptions of the training data but including canonical transcriptions in the recognition lexicon causes a mismatch that has a negative impact on the baseline speech recognition results. Consequently, the improvement obtained with syllable-length acoustic models seems much larger than it is in reality.

Article 2: Modelling pronunciation variation with single-path and multi-path syllable models: Issues to consider

This study aims to investigate the importance of modelling within-syllable pronunciation variation and syllable context when using syllable-length acoustic models for large-vocabulary continuous speech recognition. In an attempt to model within-syllable pronunciation variation more accurately, the study introduces a method for adding parallel paths to syllable-length acoustic models. The motivation for adding parallel paths comes from analysing the number of pronunciation variants per syllable. For instance, the Dutch syllable /hEt/ corresponding to the Dutch definite article 'het' has 27 different pronunciation variants in the part of the Spoken Dutch Corpus used for the experiments. Therefore, it is intuitively difficult to believe that a single path through the syllable model would be sufficient to capture all the possible pronunciation variation. To reach our goal, we construct context-independent single-path and multi-path syllable models and use these syllable models to represent monosyllabic words, constituent syllables of polysyllabic words, or both. We then compare the recognition performance of the different recognisers with each other and with the recognition performance of a conventional phoneme-based recogniser, and analyse the word-level and sentence-level errors made by the recognisers that are the most revealing when it comes to the factors under investigation. Both the phoneme-based models and the single-path syllable models outperform multi-path syllable models. The error analyses show that the most important factors affecting the recognition performance are syllable context and lexical confusability. Furthermore, the 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.

Article 3: Analysis of acoustic reduction using spectral similarity measures

This study introduces a measure of spectral reduction, which is based on the log-likelihoods ('acoustic scores') returned by an automatic speech recogniser aligning a particular sequence of phoneme-based models with a particular stretch of a speech signal. Using data for four Dutch affixes from a large database of face-to-face conversations, it builds upon an earlier study examining the effects of lexical frequency on durational reduction in spoken Dutch (Pluymaekers et al., 2005), and investigates whether the proposed measure of reduction could either replace or add to duration as a measure of reduction. The results suggest that spectral reduction scores capture other aspects of reduction than duration. While duration can – to a moderate degree – be predicted by a number of linguistically motivated variables (such as word frequency, segmental context, and speech rate), spectral reduction scores cannot. This may be due to the fact that spectral reduction is inherently a multidimensional phenomenon. However, at the same time, spectral reduction scores are able to predict a substantial amount of the variation in duration that the linguistically motivated variables do not account for.

References

Bakis, R. (1976). Continuous speech word recognition via centi-second acoustic states, in: *Proceedings of the 91st Meeting of the Acoustical Society of America*, Washington, D.C., USA.

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

Davis, S., Mermelstein, P. (1980). "Comparison of parametric representation for monosyllable word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 28(4), 357-366.

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

Holmes, J., Holmes, W. (2001). *Speech synthesis and recognition* (Taylor & Francis, London and New York).

Hon, H.-W., Yuan, B., Chow, Y.-L., Narayanan, S., Lee, K.-F. (1994). Towards large-vocabulary Mandarin Chinese speech recognition, in: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Adelaide, Australia, Vol. 1, pp. 545-548.

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

Jurafsky, D., Martin, J.H. (2000). *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition* (Prentice Hall, Upper Saddle River, New Jersey).

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

Ladefoged, P. (2001). *A course in phonetics* (Harcourt College Publishers, Orlando).

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

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

Odell, J., Durrani, S. (2006). *Word-specific acoustic models in a speech recognition system*. US Patent 7062436.

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 International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Canary Islands, Spain, Vol. 1, pp. 340-347.

Ostendorf, M. (1999). Moving beyond the 'beads-on-a-string' model of speech, in: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Keystone, CO, USA, pp. 79-84.

Pluymaekers, M., Ernestus, M., Baayen, R.H. (2005). "Lexical frequency and acoustic reduction in spoken Dutch," *Journal of the Acoustical Society of America*, 118, 2561-2569.

Rabiner, L.R., Juang, B.H. (1993). *Fundamentals of speech recognition* (Prentice Hall, Englewood Cliffs, New Jersey).

Plannerer, G., Ruske, B. (1992). Recognition of demisyllable based units using semicontinous hidden Markov models, in: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, San Francisco, CA, USA, Vol. 1, pp. 581-584.

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

Sethy, A., Ramabhadran, B., Narayanan, S. (2003). Improvements in ASR for the MALACH project using syllable-centric models, in: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, St. Thomas, U.S. Virgin Islands, USA, pp. 129-134.

Wester, M. (2002). *Pronunciation variation modelling for Dutch automatic speech recognition* (University of Nijmegen, The Netherlands).

Wutiwiwatchai, C., Furui, S. (2007). "Thai speech processing technology: A review," *Speech Communication*, 49, 8-27.

II ARTICLES

A. Hämäläinen, L. Boves, J. de Veth, and L. ten Bosch (2007). “On the utility of syllable-based acoustic models for pronunciation variation modelling,” *EURASIP Journal on Audio, Speech, and Music Processing*, Article ID 46460, 11 pages, doi:10.1155/2007/46460. (Reformatted.)

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

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 & 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 multi-pronunciation 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 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's [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 & 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 re-estimation 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.

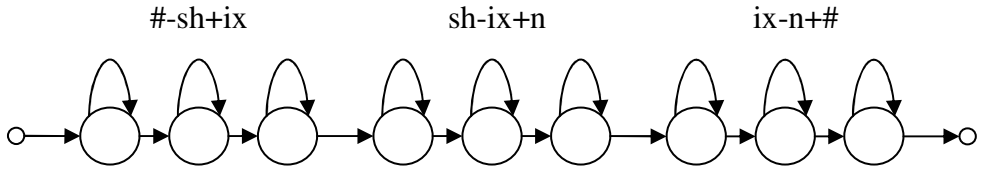

Figure 1: Syllable model for the syllable /sh ix n/. The model states are initialised with the triphones underlying the canonical syllable transcription [8]. The phones before the minus sign and after the plus sign in the triphone notation denote the left and right context in which the context-dependent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

Several research groups have published promising, but somewhat contradictory results with longer-length acoustic models [4, 8-12]. Sethy & 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 & Narayanan obtained. Part of the discrepancy between Sethy & 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 & 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(ly verified) transcriptions and the other with canonical transcriptions. By doing so, we investigate the claim that properly initialised and re-trained 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. 70% 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 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 non-overlapping 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 2nd 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 2nd 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 Appendix A. 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.

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

# 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

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 spoken TIMIT data, we excluded sentences with tagged particularities, such as incomprehensible words, non-speech 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 male and 56% were female. 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 3rd column of Table 1 shows how much of the 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 3rd 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 Appendix B. 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, were 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 set-up

3.1 Feature extraction

Feature extraction was carried out at a frame rate of 10 ms using a 25-ms Hamming window. First order pre-emphasis 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 per-sentence 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(ly 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 re-estimation. 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+#/, and
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 re-estimation were used to train the mixture of models further. The difference between the syllable-model and mixed-model recogniser 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 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).

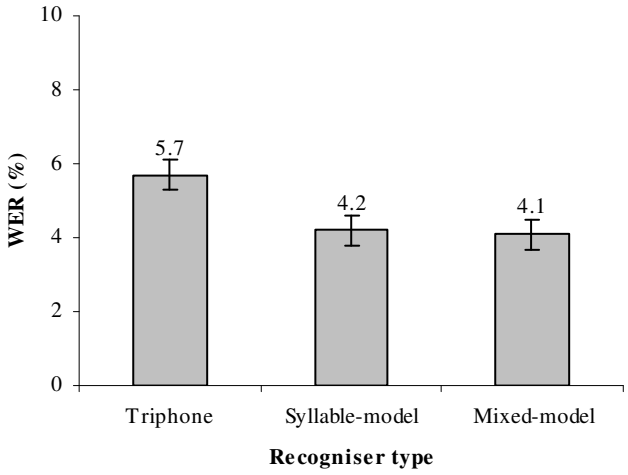

Figure 2: TIMIT WERs, at the 95% confidence level, when using manual triphones.

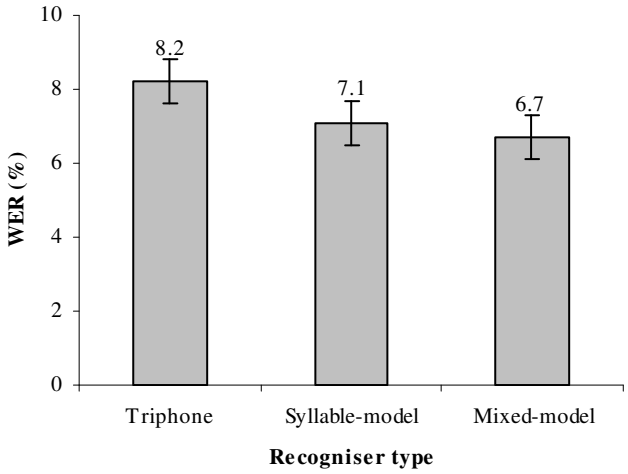

Figure 3: CGN WERs, at the 95% confidence level, when using manual triphones.

Table 3: TIMIT WERs and percentage change as a function of syllable count when using the triphone and mixed-model recognisers based on manual triphones.

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

# 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

The 2nd and 3rd columns of Tables 3 and 4 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.

The 4th columns of Tables 3 and 4 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.

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 & 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 & 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 Messina & Jouvet [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 & 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 re-trained 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 3 and 4 – 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 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 re-trained 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.

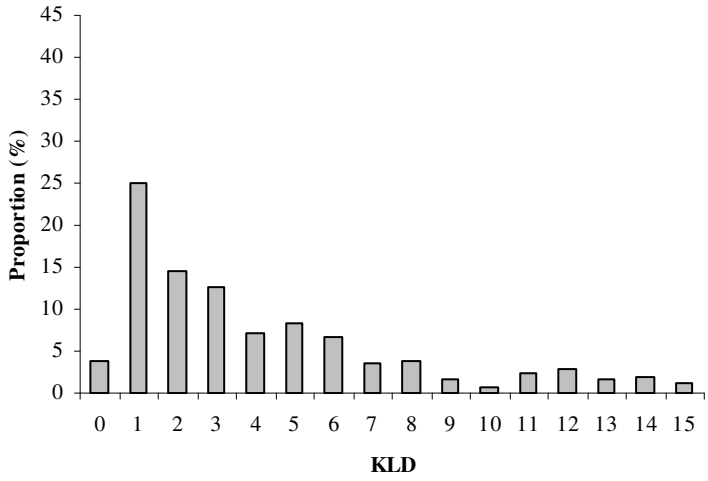

Figure 4: KLD distributions for the states of re-trained syllable models for TIMIT when using manual triphones.

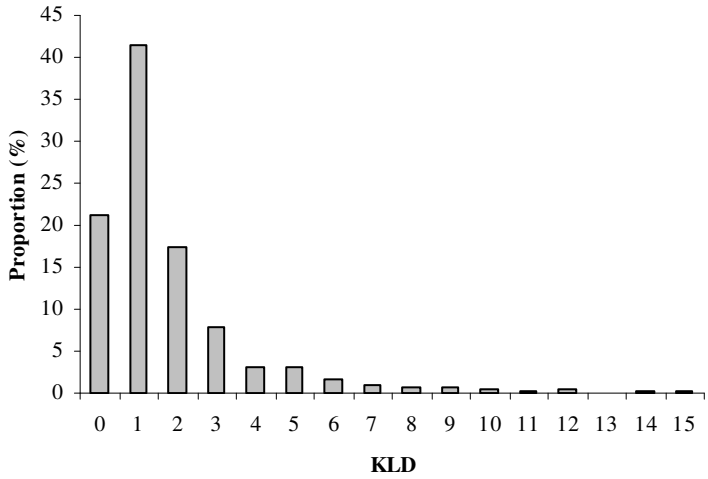

Figure 5: KLD distributions for the states of re-trained syllable models for CGN when using manual triphones.

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 re-trained 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 re-estimation 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 re-training, 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 re-estimation 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 re-trained 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 with longer-length acoustic models are small, and deteriorations also occur.

The 2nd and 3rd 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. As in the case of the experiments with manual triphones (see Tables 3 and 4), the probability of errors was considerably higher for monosyllabic words than for polysyllabic words. The 4th 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 re-trained syllable models in the first set of experiments (and probably also in Sethy & 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.

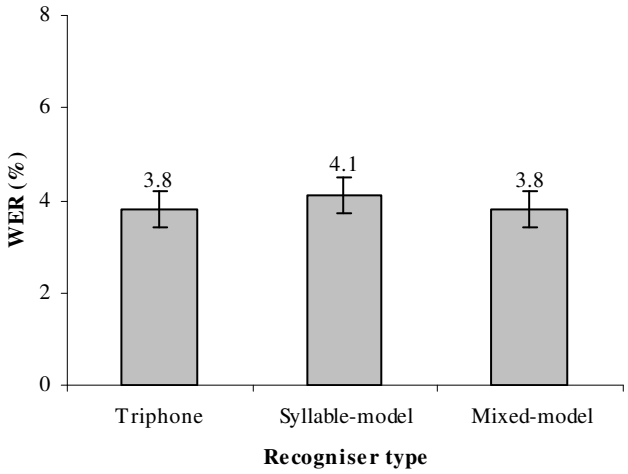

Figure 6: TIMIT WERs, at the 95% confidence level, when using canonical triphones.

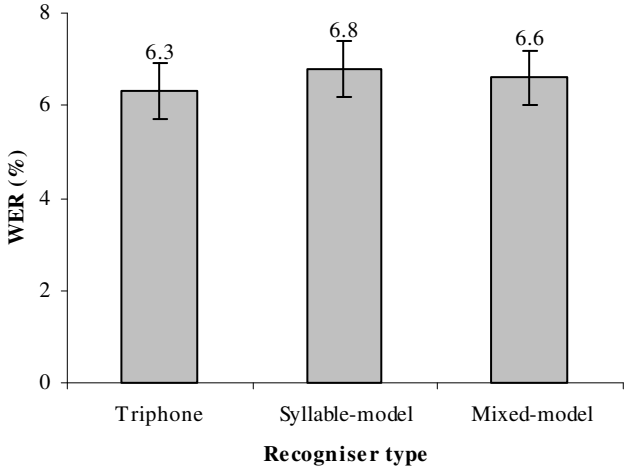

Figure 7: CGN WERs, at the 95% confidence level, when using canonical triphones.

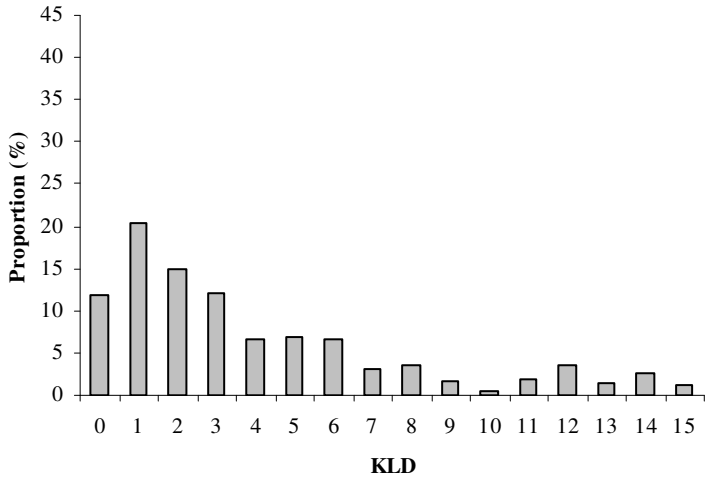

Figure 8: KLD distributions for the states of re-trained syllable models for TIMIT when using canonical triphones.

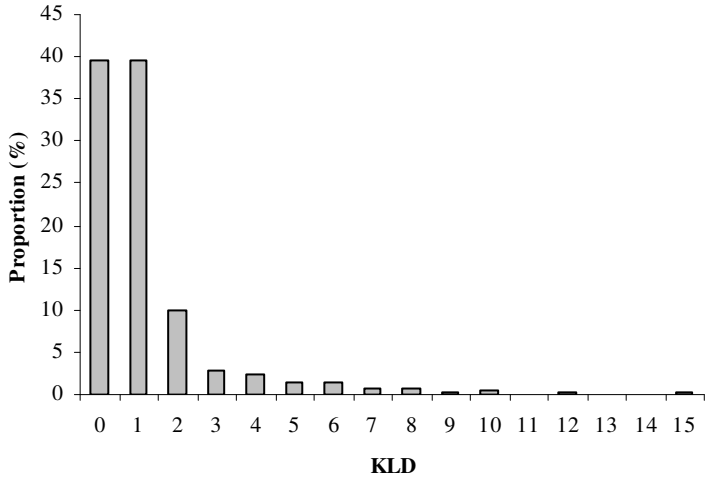

Figure 9: KLD distributions for the states of re-trained syllable models for CGN when using canonical triphones.

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

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

# 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

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 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 & 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 re-estimation 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 re-trained 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 & 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 (mis)match 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 non-linguistic (e.g., speaker identity, speaking rate) factors. Sun & 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 (for instance, 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 re-estimating these parallel paths to better incorporate the dynamic nature of articulation. Therefore, our future research will focus on strategies for developing multi-path 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 re-trained 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.

References

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

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

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

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

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

[6] Plannerer, G., Ruske, B. (1992). Recognition of demisyllable based units using semicontinous hidden Markov models, in: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, San Francisco, CA, USA, Vol. 1, pp. 581 -584.

[7] Jones, R.J., Downey, S., Mason, J.S. (1997). Continuous speech recognition using syllables, in: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Rhodes, Greece, pp. 1171-1174.

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

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

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

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

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

[13] Schiller, N.O., Meyer, A.S., Levelt, W.J.M. (1997). "The syllabic structure of spoken words: Evidence from the syllabification of intervocalic consonants," *Language & Speech*, 40, 103-140.

[14] Pallier, C. (1997). Phonemes and syllables in speech perception: size of attentional focus in French, in: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Rhodes, Greece, pp. 2159-2162.

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

[16] *TIMIT Acoustic-Phonetic Continuous Speech Corpus* (1990). (National Institute of Standards and Technology Speech Disc 1-1.1, NTIS Order No. PB91-505065).

[17] 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 International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Canary Islands, Spain, Vol. 1, pp. 340-347.

[18] Kullback, S., Leibler, R. (1951). "On information and sufficiency," *Annals of Mathematical Statistics*, 22, 79-86.

[19] 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, UK).

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

[21] Kahn, D. (1976). *Syllable-based generalisations in English phonology* (Indiana University Linguistics Club, Bloomington, IN, USA).

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

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

[24] Wester, M. (2002). *Pronunciation variation modelling for Dutch automatic speech recognition* (University of Nijmegen, The Netherlands).

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

[26] 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)*, Paris, France, pp. 195-202.

[27] Sun, J., Deng, L. (2002). “An overlapping-feature-based phonological model incorporating linguistic constraints: Applications to speech recognition,” *Journal of the Acoustical Society of America,* 111(2), 1086-1101.

Appendix A: 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

Appendix B: CGN phone mappings. 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

A. Hämäläinen, L. ten Bosch, and L. Boves (2009). “Modelling pronunciation variation with single-path and multi-path syllable models: Issues to consider,” *Speech Communication*, 51, 130-150. (Reformatted.)

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

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.

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.

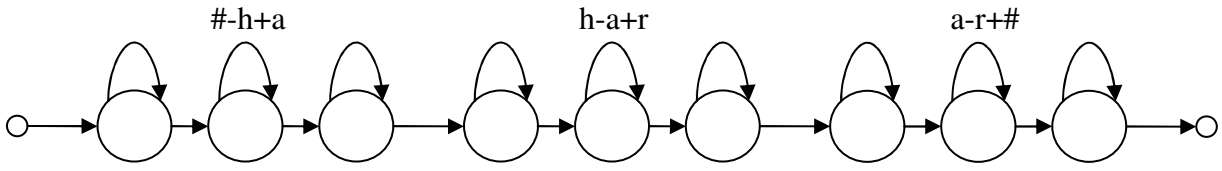

Figure 1: Single-path model for the syllable /har/, with the single path through the model initialised with the biphones and triphones underlying the canonical syllable transcription (Hämäläinen et al., 2007a; Sethy and Narayanan, 2003). The phones before the minus sign and after the plus sign in the 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.

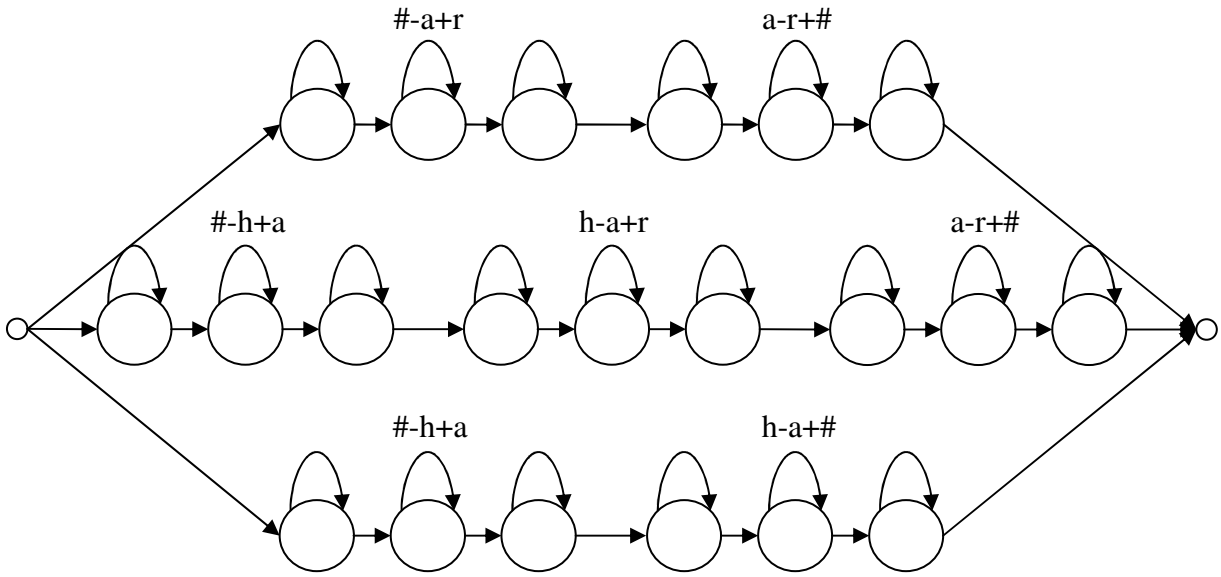

Figure 2: Multi-path model for the syllable /har/, with the three parallel paths initialised with the triphones underlying the 'major, distinct transcription variants' /ar/, /har/ and /ha/, respectively.

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).

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.

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.

Table 1: Main statistics of the speech material.

	Train	Test	Dev. test
Word tokens	396187	22289	22100
Word types	28164	5154	5074
Syllable tokens	604211	33921	33588
Syllable types	6146	2722	2623
Duration (hh:mm:ss)	37:00:20	02:04:21	02:03:33

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

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

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 35000 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 29700 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 3500 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.

	Total number of states
TR	1535
SPM	1605
SPP	1621
SP	1603
MPM	1726
MPP	1782
MP	1764

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.

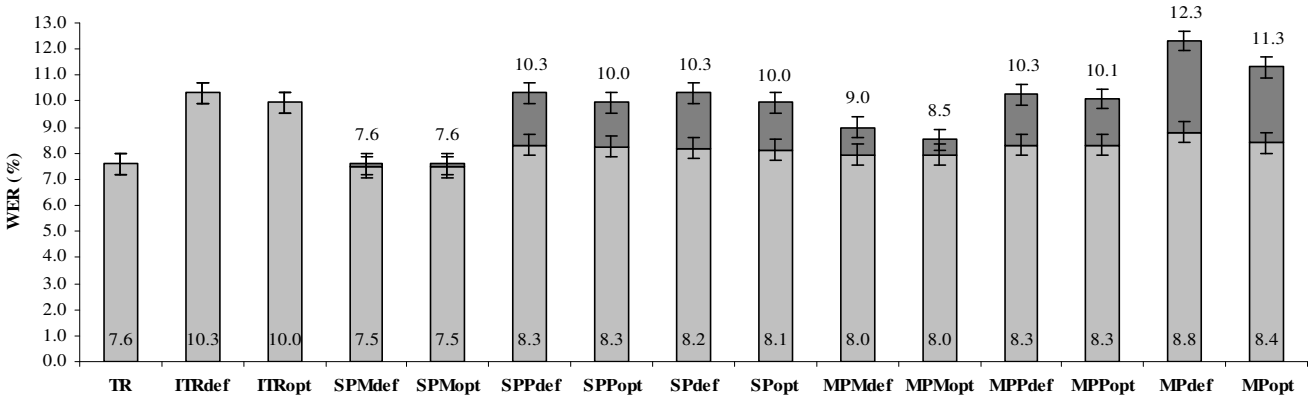

Figure 3: WERs with a 95% confidence interval for the following recognisers with 64 Gaussians per state: baseline triphone recogniser (TR), impaired triphone recogniser (ITR), single-path mixed-model recogniser with the target syllables covered with syllable models in monosyllabic words (SPM), single-path 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.

	-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

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[*]. 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[*]. Therefore, substitutions were by far the most important type of errors from the WER point of view.

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).

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.

[*] 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.

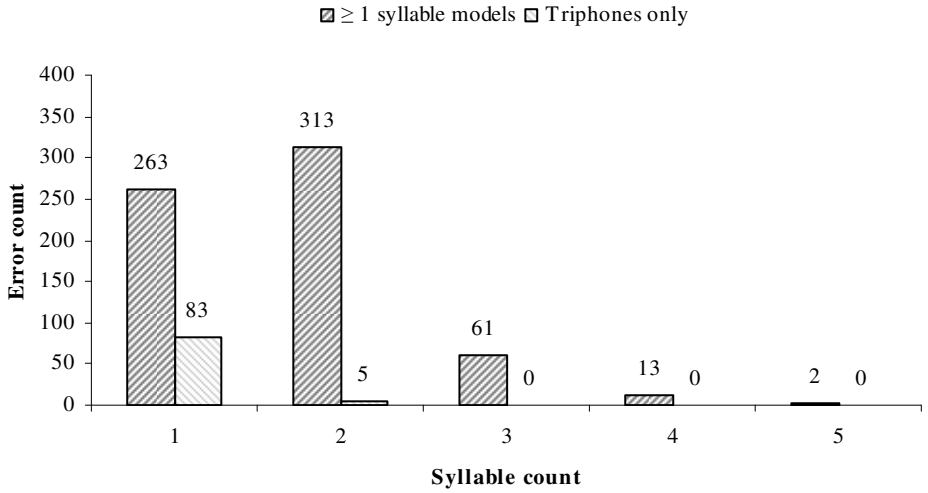

Figure 4: The number of substitution errors for words with varying numbers of syllables in the optimised untrained single-path mixed-model recogniser with the target syllables covered with syllable models in polysyllabic words ($SPP_{opt,\ bt}$). The errors are shown separately for words that include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

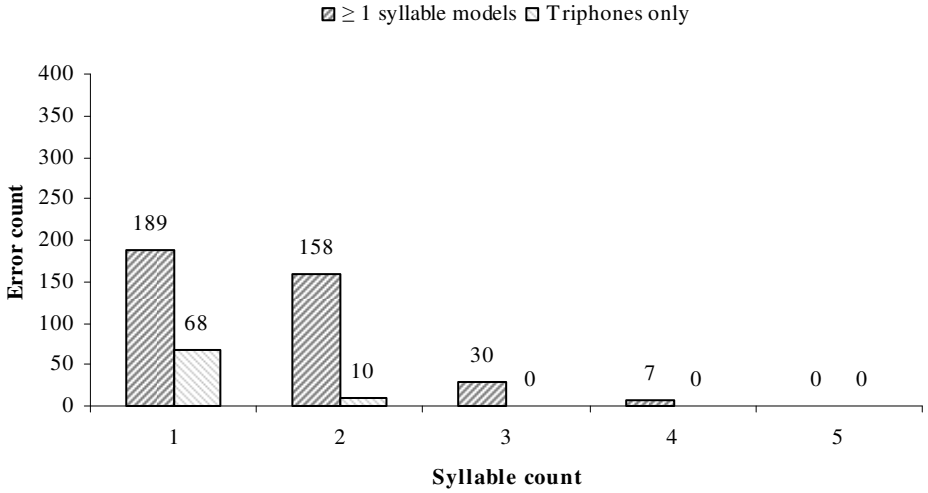

Figure 5: The number of substitution errors for words with varying numbers of syllables in the optimised retrained single-path mixed-model recogniser with the target syllables covered with syllable models in polysyllabic words ($SPP_{opt,\ at}$). The errors are shown separately for words that include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

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	weggelegd	voor 't zout en de specerijen in de pap
Mixed-model recogniser:	die al is	weg gelegd	voor 't zout en de specerijen in de pap

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.

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: 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.

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: 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.

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: 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.

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: 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.

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: 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.

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

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 $n \geq 1$. 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.

References

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 the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France.

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

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

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

Ganapathiraju, A., Hamaker, J., Ordowski, M., Doddington, G., Picone, J. (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)*, Paris, France, pp. 195-202.

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

Han, Y., Veth, J.M. de & 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: *Proceedings of the Phonetics Symposium*, Helsinki, Finland, 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*, Article ID 46460, 11 pages, doi:10.1155/2007/46460.

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

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

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

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

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

Jurafsky, D., Bell, A., Gregory, M., Raymond, W. (2001a). Probabilistic relations between words: Evidence from reduction in lexical production, in: *Frequency and the emergence of linguistic structure*, edited by J. Bybee and P. Hopper (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 the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Salt Lake City, UT, USA, Vol. 1, pp. 577-580.

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)*, Estes Park, CO, USA, pp. 18-23.

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 European Conference on Speech Communication and Technology (Eurospeech)*, Antwerp, Belgium.

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

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

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 International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Canary Islands, Spain, Vol. 1, pp. 340-347.

Ostendorf, M. (1999). Moving beyond the 'beads-on-a-string' model of speech, in: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Keystone, CO, USA, pp. 79-84.

Plannerer, G., Ruske, B. (1992). Recognition of demisyllable based units using semicontinous hidden Markov models, in: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, San Francisco, CA, USA, Vol. 1, pp. 581 -584.

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 International Conference on Spoken Language Processing (ICSLP)*, Yokohama, Japan, pp. 227-230.

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

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 the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Hong Kong, Vol. 1, pp. 772-776.

Sethy, A., Ramabhadran, B., Narayanan, S. (2003). Improvements in ASR for the MALACH project using syllable-centric models, in: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, St. Thomas, U.S. Virgin Islands, USA, pp. 129-134.

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 European Conference on Speech Communication and Technology (Eurospeech)*, Geneva, Switzerland, pp. 769-772.

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

Wester, M. (2002). *Pronunciation variation modelling for Dutch automatic speech recognition* (University of 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, 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.

A. Hämäläinen, M. Gubian, L. ten Bosch, and L. Boves. (in press). “Analysis of acoustic reduction using spectral similarity measures,” *Journal of the Acoustical Society of America*.

Analysis of acoustic reduction using spectral similarity measures

Annika Hämäläinen, Michele Gubian, 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

Articulatory and acoustic reduction can manifest itself in the temporal and spectral domains. This study introduces a measure of spectral reduction, which is based on the speech decoding techniques commonly used in automatic speech recognisers. Using data for four frequent Dutch affixes from a large corpus of spontaneous face-to-face conversations, it builds on an earlier study examining the effects of lexical frequency on durational reduction in spoken Dutch (Pluymaekers et al., 2005), and compares the proposed measure of spectral reduction with duration as a measure of reduction. The results suggest that the spectral reduction scores capture other aspects of reduction than duration. While duration can – albeit to a moderate degree – be predicted by a number of linguistically motivated variables (such as word frequency, segmental context, and speech rate), the spectral reduction scores cannot. This suggests that the spectral reduction scores capture information that is not directly accounted for by the linguistically motivated variables. The results also show that the spectral reduction scores are able to predict a substantial amount of the variation in duration that the linguistically motivated variables do not account for.

1 Introduction

It has long been known that words in normal speech – in particular in spontaneous speech – are frequently pronounced in a more reduced form than their canonical phonetic transcriptions would suggest (e.g., Ernestus, 2000; Ernestus et al., 2006; Jespersen, 1922; Lindblom, 1963; Pluymaekers et al., 2005; Zipf, 1929). Weak forms of reduction may become manifest in the acoustic signal as shortened segments with flatter spectral envelopes, while strong reduction may result in the deletion of phonemes or whole syllables (Greenberg, 1999; Johnson, 2004). It has been hypothesised that the degree of reduction could be explained by the amount of information carried by the word in question. This has resulted in competing theories, such as the Smooth Signal Redundancy Hypothesis (Aylett & Turk, 2004), the Probabilistic Redundancy Hypothesis (Jurafsky et al., 2001), and the Speech Efficiency Hypothesis (van Son & Pols, 2003). Different theories seem to invoke different cognitive and physiological processes, such as the compression of motor routines as a result of practice (Bybee, 2001), as well as adaptation to the needs of the listener (e.g., Jurafsky et al., 2001). All theories aim to explain reduction phenomena that are manifest in both the temporal and spectral domain.

It has, however, proved difficult to design experiments for investigating the causes of reduction in detail. This is because it is difficult to exert enough experimental control for a fair comparison of reduction when words do not only differ in frequency, but also in their intrinsic phonemic and morphological complexity, such as the number and type of phonemes they consist of (Pluymaekers et al., 2005). To avoid these difficulties, Pluymaekers et al. (2005) investigated

reduction by focusing on affixes, i.e., on morphemes that can occur in a large number of different words with varying frequencies. More specifically, they studied the role of various linguistically motivated predictors (e.g., word frequency, speech rate, and the age and regional origin of the speaker) in explaining reduction observed in syllable-sized affixes.

Pluymaekers et al. (2005) chose to use a correlate of reduction that is relatively easy to measure in the acoustic speech signal: duration. They showed that regression models based on linguistically motivated variables could, at best, predict moderate proportions of variance in duration (i.e. the dependent variable). Reduction is, however, known to manifest itself in many different ways, and duration only reflects part of the reduction phenomenon. Therefore, it is worthwhile investigating other indexes of reduction in the acoustic speech signal, as well. Because of the relation between the gestures of the articulators and the spectrum of the resulting speech signal, *spectral* reduction measures are particularly interesting. In this paper, we propose an automatically derived measure of spectral reduction and test it using the same data as Pluymaekers et al. (2005). The resulting spectral reduction scores reflect the reduction phenomenon in a different way than duration does. In this paper, we therefore investigate the relation between the newly developed spectral reduction measure, the duration-based reduction measure, and the linguistically motivated context variables employed by Pluymaekers et al. (2005).

Scholars agree that reduction must be interpreted as the deviation of an observed pronunciation from some reference pronunciation. Since speech production involves multiple articulators and results in acoustic trajectories in a high-dimensional acoustic space, deviation from a reference pronunciation can take place along several different dimensions. Using duration as the only measure of reduction would leave open the option of reduction being limited to a time compression of otherwise 'unreduced' articulatory gestures. However, most studies on reduction imply that, in addition to being shorter, the articulatory gestures are simplified. This 'simplification' should manifest itself in the spectral structure of the signals. A spectral measure of reduction captures the deviation between an actual trajectory and a 'reference trajectory' in the acoustic space. In our case, this is the deviation between an observed acoustic token (e.g., a particular instance of an affix) and the reference model of the token. Coarticulation is a pervasive phenomenon in speech, and its effects could be interpreted as just another, unavoidable manifestation of reduction. We should point out that our definition of reduction also holds in the presence of coarticulation effects; spectral reduction can always be interpreted as the deviation between the observed and the reference trajectories in the acoustic space, with the reference trajectories including coarticulation effects. The goal of this paper is to investigate whether duration and spectral reduction are overlapping or complementary indices of the underlying articulatory 'simplification' in the case of syllable-sized affixes. To that end, we carry out experiments using the same data as Pluymaekers et al. (2005).

This paper is further organised as follows. In Section 2, we introduce our approach to quantifying spectral reduction. We recapitulate the speech material in Section 3, and describe the statistical variables used in this study in Section 4. We discuss the design and results of our

first experiment in Section 5, and do the same for a follow-up experiment in Section 6. Finally, we discuss the results and suggest directions for future research in Section 7, and conclude the paper in Section 8.

2 Quantifying spectral reduction

The question how to quantify spectral reduction can be made more precise by asking how to quantify the amount of (dis)similarity between the reduced and reference realisations of a speech unit – in our case, syllable-sized affixes. To that end, speech decoding and alignment techniques developed for automatic speech recognition (ASR), provide powerful tools. Speech recognisers based on Hidden Markov Models (HMMs) are able to provide estimates of the degree of (dis)similarity between a particular stretch of speech signal and a model of the acoustics of the corresponding speech unit(s) (e.g., phoneme(s), syllable(s) or word(s)) derived from some corpus of training data. One such estimate is the log-likelihoods (usually referred to as *'acoustic scores'*) that HMM-based speech recognisers compute as a by-product of forced alignment. Forced alignment is a technique in which a speech signal is aligned with a predefined sequence of acoustic models associated with speech units (e.g., phonemes, syllables or words). The output of the alignment is a score for the goodness of the fit between the speech signal and the models, usually in combination with a corresponding segmentation. Forced alignment can also be used for estimating the best transcription for a word token: if a word is represented by more than one phonemic transcription in the recogniser lexicon, the forced alignment procedure is able to select the most likely one. The result of the forced alignment then depends on the available phonemic transcriptions ('candidate transcriptions') of the word in the lexicon and the quality of the acoustic models corresponding to these phonemic transcriptions. Should the recogniser lexicon only contain *one* possible transcription per word, the acoustic score for each token of that word would express how well the signal matches that single transcription. Should that single transcription be a canonical transcription (which is the closest we can get to a reference pronunciation of the word), the total acoustic score would express how well the signal matches the reference. Below, we argue why the acoustic scores obtained from forced alignment with a sequence of HMMs corresponding to a canonical transcription are viable estimators of spectral deviation and, consequently, viable estimators of spectral reduction. By using just a single scalar to represent the distance between the models and the actual acoustic signals, we clearly lose information about the details (temporal and spectral) of the deviation between the token and the model. However, the spectral reduction measure obtained in this way provides information that reflects the deviation from the reference in the articulatory and acoustic space better than a plain duration measure can do. Both measures reflect differences between acoustic trajectories, but focus on different kinds of differences between these trajectories.

The rationale underlying our approach to computing spectral reduction is as follows. Suppose $X=\{x_1,x_2,\ldots,x_N\}$ is a sequence of observed acoustic feature vectors, and $S=\{s_1,s_2,\ldots,s_K\}$ is the sequence of HMM states used in the forced alignment between the

speech signal and the corresponding acoustic models. The alignment procedure returns the log-likelihood $\log P(X \mid S)$ defined by Equation 1:

$$\log P(X \mid S) = \log \prod_{n,j} P_e(x_n \mid s_j) \prod_{j,i} P_t(s_j \mid s_i) \tag{1}$$

in which P_e and P_t denote the emission and transition probabilities and the (n, j) and (j, i) pairs are uniquely determined by the alignment path (the indices i and j specify the indices of the states, and n specifies the frame index, along the path resulting from the alignment.

To justify that Equation 1 leads to a viable estimator of acoustic reduction, please notice that, for a single feature vector x and an HMM state s, the distance (dissimilarity) between the feature vector and the HMM state can be written as

$$d^2 = -\log(P_e^M(x \mid s)) - \log(P_t(s \mid s_{previous})) \tag{2}$$

where P_e^M denotes the emission probability modelled by a mixture of M Gaussians. To obtain a measure of dissimilarity between a vector sequence and an acoustic model represented by a sequence of HMM states, the dissimilarity scores d^2 along the best path through the trellis must be accumulated:

$$D = \sum_{n=1}^{N} d_n^2 = -\sum_{n,j} \log(P_e^M(x_n \mid s_j)) - \sum_{i,j} \log(P_t(s_j \mid s_i)) \tag{3}$$

In this expression, the sum over $\log(P_e)$ represents the spectral distance between the token and the models, while the sum over $\log(P_t)$ represents the total scores associated with the state-to-state transition probabilities. The dissimilarity score D depends on the duration of the speech segment (represented by the number of frames in the sequence of input frames). To be able to compare the results of Equation 3 *across* tokens of different duration, we obtain an *average frame-to-state dissimilarity* by normalising the score D for the number of frames:

$$D_{norm} = \frac{-\sum_{n,j} \log(P_e^M(x_n \mid s_j)) - \sum_{i,j} \log(P_t(s_j \mid s_i))}{N} \tag{4}$$

Equation 4 is the expression used in this paper to compute the final spectral reduction scores.

In this study, we use the Hidden Markov Model Toolkit (HTK) (Young et al., 2002), which actually outputs *similarity scores* instead of dissimilarity scores. Therefore, we use $-D_{norm}$ from Equation 4 as the spectral reduction score in this paper.

3 Speech material

We re-used the affix data that were selected and measured by Pluymaekers et al. (2005). These data originate from spontaneous face-to-face conversations between speakers of Dutch (as spoken in the Netherlands) in the Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN) (Oostdijk et al., 2002).

We investigated the prefixes *ge-*, *ver-* and *ont-*, and the suffix *-lijk*. *Ge-* is commonly used to create the perfect participle in Dutch (e.g., *gespeculeerd* (the perfect participle form of the verb 'to speculate')), and can also appear as a nominal or a verbal prefix (e.g., *gebak* ('cake(s)'); *gebeuren* ('to happen')). However, we only investigated the participial instances of *ge-*. *Ver-* and *ont-* are verbalising prefixes expressing change of state (e.g., *verplaatsen* ('to move')) and reversal or inchoation (e.g., *onteigenen* ('to disown')). The suffix *-lijk* appears in adverbs and adjectives (e.g., *natuurlijk* ('natural(ly)'); *eigenlijk* ('actual(ly)')). The canonical phonetic transcriptions (using SAMPA (Speech Assessment Methods Phonetic Alphabet)) of the four affixes are /x@/, /v@r/, /Ont/, and /l@k/, respectively. (Pluymaekers et al., 2005.)

Pluymaekers et al. (2005) provide a detailed description of the selection of the affix tokens that were analysed. To summarise, they selected one token for each word type containing a target affix. As word types, they did not only consider words belonging to different lemmas but also different word forms of the same lemma (e.g., the sample for the affix *ont-* included both *ontwikkelt* 'develops' and *ontwikkelde* 'developed'). The recordings contained the complete utterances in which the affixes were embedded. Table 1 presents an overview of the affix samples used in the study.

Table 1: The number of tokens, the number of speakers, the maximum number of tokens uttered by each speaker and the broad phonetic transcriptions of the uttered tokens for each affix.

Affix	#Tokens	#Speakers	Max(#Tokens/ Speaker)	Phonetic transcriptions
ge-	427	132	12	/x@/, /x/, /G@/, /G/
ver-	137	80	8	/v@r/, /v@/, /vEr/, /vr/, /v/, /f@r/, /f@/, /f/
ont-	101	63	4	/Ont/, /Ond/, /Omp/, /Od/, /Om/, /On/, /Ot/, /@nd/, /@nt/, /@n/, /@t/
-lijk	157	87	6	/l@k/, /l@g/, /lEk/, /lIk/, /lYk/, /l@/, /lk/, /@k/, /@/, /g/, /k/

4 Statistical variables

The statistical variables we used in this study included the spectral reduction scores, which we used both as a dependent variable and as a predictor; duration, which we used as a dependent variable, and the linguistically motivated variables from Pluymaekers et al. (2005), which we used as predictors. In this section, we describe these variables in more detail.

4.1 Spectral reduction scores

We obtained the spectral reduction scores by carrying out forced alignment on the stretches of speech that Pluymaekers et al. (2005) had manually labelled as the target affixes. When carrying out the forced alignment, we used a single sequence of HMM states for each affix. This sequence was formed by concatenating the triphone models underlying the canonical transcription of the affix in question.

As the model topology for the triphone models, we used standard three-state left-to-right HMMs with no state skips allowed. We carried out feature extraction of the affix data and of the data used for training the triphone models at a frame rate of 5 ms using a 25-ms Hamming window and applied first order pre-emphasis to the signal using a coefficient of 0.97. Using the 'default' frame rate of 10 ms in combination with the chosen model topology would have required the *ge-* tokens to have a minimum duration of 60 ms (i.e. two phone models times three states per model, at least one frame per state) and the *ver-*, *ont-,* and *-lijk* tokens to have a minimum duration of 90 ms to allow alignment. Reducing the frame rate to 5 ms allowed us to obtain acoustic scores for the vast majority of the very short affix tokens as well. 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 the complete recordings.

We carried out forced alignment using two different sets of triphone models. The first set of triphones (*'manual triphones'*) comprised 8-Gaussian HMMs trained with the *manually verified* transcriptions of the read speech in the core set of CGN. The training data contained 45,172 orthographic word tokens (4 h 51 min 27 s of speech). The second set of triphones (*'canonical triphones'*) comprised 64-Gaussian HMMs trained with *canonical* transcriptions of a much larger part of the read speech data in CGN. The training data contained 396,187 orthographic word tokens (37 h 20 s of speech). The (standard) triphone training procedure is described in Hämäläinen et al. (2007) for the manual triphones, and in Hämäläinen et al. (2009) for the canonical triphones. For this study, we carried out state tying such that both sets of triphones had about 3400 physically distinct triphones. While the amount of training data and the number of Gaussian mixtures were different for the two sets, the number of data points (frames) used to define each diagonal-covariance Gaussian after tying was almost equal.

The reason to use triphones trained on *read speech* was that we wanted to base the spectral reduction scores on the dissimilarity between an individual affix token and a maximally unreduced form of the affix. Such maximally unreduced form can be considered maximally

similar to the canonical pronunciation of the affix. Triphones trained on carefully read speech provided us with a reference that was as unreduced as possible. The triphones trained with manually verified transcriptions were arguably the 'cleanest' models in this sense. However, as manually verified transcriptions are not always available in speech corpora because of their expensiveness, we also tested triphones trained with canonical transcriptions of read speech.

Unlike Pluymaekers et al. (2005), who also fitted models to predict the durations of the individual segments of the affixes, we only carried out statistical analyses on the affix level. This is because the acoustic scores obtained for individual segments using forced alignment are not necessarily meaningful due to differences between manual and automatic segmentation. The acoustic scores that the forced alignment process computes for each affix are sums of the acoustic scores of the constituent triphones. In addition to the acoustic scores, the alignment process provides a segmentation of the triphones. However, this automatic segmentation of the triphones might differ considerably from the manual segmentation of the corresponding phonemes. This is because the speech recogniser is forced to align the speech signal with the full sequence of constituent triphones and because the minimum duration of each triphone is 15 ms (with a frame rate of 5 ms and three emitting states per triphone). In the case of very short or deleted phonemes, the recogniser uses parts of the previous or the following phoneme to satisfy the minimum length criterion. This renders the acoustic scores for the individual segments of the affixes potentially meaningless.

4.2 Duration

For all target words, Pluymaekers et al. (2005) measured the duration of the affix and the durations of the individual segments in the affix in milliseconds. They placed the segment boundaries where they found clear formant transitions in the spectrogram supported by visible changes in the waveform pattern.

4.3 Linguistically motivated control variables

We took over the linguistically motivated control variables investigated by Pluymaekers et al. (2005). These include both probabilistic and non-probabilistic variables. The probabilistic variables comprise word frequency; the number of times the target word, or a word from the same inflectional paradigm had occurred earlier in the conversation; the number of times the target affix had occurred earlier in the conversation; mutual information; and word-stem ratio. The non-probabilistic variables include the rate of speech; the gender, age and regional origin of the speaker; the location of the target word in the utterance (utterance-initial/utterance-final); the presence of disfluencies directly before and after the target word; the segment following the affix (consonant/vowel); the number of consonants in the onset of the stem of the prefixed word (onset complexity); and the absence of segments in the affix. Pluymaekers et al. (2005) describe the motivations for using the above-listed control variables, and detail the ways they obtained their values.

5 Experiment 1

In Experiment 1, we investigated whether our spectral reduction scores capture the same information about acoustic reduction as duration. To achieve our goal, we repeated the experiments described by Pluymaekers et al. (2005) with the spectral reduction scores as the dependent variable (instead of duration) and using the same linguistically motivated variables as the predictors. We experimented with the spectral reduction scores based on both the manual triphones and the canonical triphones as the dependent variable ('*the manual score models*' and '*the canonical score models*', respectively). If our spectral reduction scores and duration (the duration models, referred to as '*Pluymaekers models*' in the remainder of this paper) captured essentially the same information about reduction, the models for the different dependent variables should be very similar.

For the results to be comparable across the three models, we first removed the 1-3 tokens per affix for which we were not able to generate acoustic scores because of their exceptionally short duration. We then determined the outlier tokens for the different models and removed them from all of the models (i.e. the final data sets used for the analyses were the same). Following Pluymaekers et al., we used leverage and Cook's distance values to determine the outliers. The resulting sets of affixes were slightly different from the selection used by Pluymaekers et al. (2005). Therefore, in order to allow a fair comparison, we recomputed the models for duration with the same data as used for the spectral reduction scores[†].

In other words, we fitted three different linear multiple regression models to the data. The Pluymaekers model had affix duration as the response variable, while the manual score model had the spectral reduction scores based on the manual triphones as the response variable, and the canonical score model had the spectral reduction scores based on the canonical triphones as the response variable. Eight data points were removed from each of the models for *ge-* because they were outliers for the Pluymaekers model, the manual score model and/or the canonical score model. For the same reason, seven data points were removed from the models for *ver-* and *ont-*. For *-lijk*, seven data points were removed from the models for words in non-final position, whereas six data points were removed from the models for words in final position. Table 2 summarises the results of Experiment 1 by presenting the amount of variance explained (R^2) by the three different models fitted for the different affixes. It becomes immediately clear from Table 2 that the spectral reduction scores cannot properly be predicted by the linguistically motivated variables. This would seem to suggest that the hypothesis of spectral reduction and duration representing the same information about reduction does not hold true. We return to this finding in Section 7.

[†] This explains the slight differences between our 'Pluymaekers models' and the numbers in he original paper.

Table 2: The amount of variance explained (R^2) by the Pluymaekers model, the manual score model and the canonical score model in Experiment 1.

Affix	Pluymaekers model	Manual score model	Canonical score model
ge-	.09	.04	.03
ver-	.10	.02	.01
ont-	.22	.04	.04
-lijk / Non-final	.13	.01	.01
-lijk / Final	.45	.02	.01

6 Experiment 2

Considering the results of Experiment 1, Experiment 2 was designed to test the hypothesis that reduction is a complex phenomenon of which temporal and spectral reduction measures each deal with different and incomplete aspects. Given this hypothesis, it would be unlikely that these two measures would capture exactly the same aspects of reduction. The second experiment, therefore, aimed to investigate the extent to which the more complex spectral reduction measure can help to *explain* duration as a measure of reduction over and above the contribution of the linguistically motivated variables (cf. Section 1). Again, we first fitted the statistical models described by Pluymaekers et al. (2005) ('*the Pluymaekers models*'). We then extended the Pluymaekers models with the spectral reduction scores based on both the manual triphones and the canonical triphones as another predictor ('*the manual score models*' and '*the canonical score models*', respectively). For the results to be comparable across the different models, we again excluded the very short affix tokens, determined the outlier tokens, and removed them from the data sets. Because the data set used for Experiment 2 was a bit larger than the data set used for Experiment 1 (in Experiment 1, we had to remove outliers for when duration *and* the spectral reduction scores were the dependent variable), the results we report with the Pluymaekers models also differ somewhat from the ones reported for Experiment 1.

We used least squares regression for the statistical analyses in this study. The proportion of variance accounted for by a model is expressed by the coefficient R^2. The signs of the reported $\hat{\beta}$ coefficients indicate whether there is a positive or a negative correlation between a predictor (independent) variable and the response (dependent) variable (for a more elaborate explanation of multiple regression models, see Izenman (2008, Chapter 5)). Before embarking on model building, we checked the distributions of the continuous variables (duration and the spectral reduction scores) for deviations of normality that would necessitate some kind of transformation of the data. No such transformation appeared to be necessary.

In other words, we used the duration of the prefix as the response variable and fitted three different linear multiple regression models to the data for each of the prefixes *ge-*, *ver-* and *ont-*: the Pluymaekers model, the manual score model and the canonical score model. In the case of the suffix *-lijk*, we followed Pluymaekers et al. (2005) by carrying out the analysis separately for

suffix tokens originating from words in non-final and final position. The number of data points removed as outliers was six for *ge-*, four for *ver-*, three for *ont-*, four for *-lijk* in the case of words in non-final position (114 observations), and five for *-lijk* in the case of words in final position (43 observations). The next four subsections present and discuss our results. To evaluate the significance of our results, we report the outcome of *t*-tests (*t*-statistics) for each response variable. The p-value is the probability of obtaining a statistical result (in this case, the result of a *t*-test) at least as extreme as the one that was actually observed, assuming that the null hypothesis ('the response variable is *not* significant') is true.

6.1 ge-

For the Pluymaekers model, we found the following effects: frequency ($\hat{\beta}$ = –3.5, $t(417) = -2.65$, $p < 0.01$), onset complexity ($\hat{\beta}$ = –6.7, $t(417) = -1.88$, $p < 0.1$), and speech rate ($\hat{\beta}$ = –8.3, $t(417) = -5.56$, $p < 0.0001$). The amount of variance (R^2) explained by this model was 9%. For the manual score model, we found the following effects: frequency ($\hat{\beta}$ = –4.1, $t(416) = -3.28$, $p < 0.01$), onset complexity ($\hat{\beta}$ = –3.3, $t(416) = -0.98$, $p \approx 0.33$), speech rate ($\hat{\beta}$ = –7.3, $t(416) = -5.16$, $p < 0.0001$), and manual score ($\hat{\beta}$ = 3.1, $t(416) = 7.49$, $p < 0.0001$). The R^2 of this model was 20%. For the canonical score model, we found the following effects: frequency ($\hat{\beta}$ = –4.0, $t(416) = -3.21$, $p < 0.01$), onset complexity ($\hat{\beta}$ = –3.5, $t(416) = -1.04$, $p \approx 0.30$), speech rate ($\hat{\beta}$ = –7.6, $t(416) = -5.34$, $p < 0.0001$), and canonical score ($\hat{\beta}$ = 3.1, $t(416) = 7.13$, $p < 0.0001$). The R^2 of this model was 19%. Words with a higher frequency had shorter realisations of *ge-*. The prefix was also shorter if the speech rate was high, or if the prefix was followed by a large number of consonants (onset complexity). The prefix was longer if the manual score or the canonical score was high.

Unlike in the Pluymaekers model, onset complexity was not significant as a predictor in the manual score model or in the canonical score model. In the Pluymaekers model, onset complexity was only significant at the 0.1 level, so the additional predictors may actually have turned it insignificant in the manual score model and in the canonical score model. Because the most complex onsets all start with a fricative, it may also be that onset complexity lost its significance because the spectral reduction scores account for its effect by capturing onset-specific coarticulation.

The observed effects of manual score and canonical score went in the expected direction. The shorter, i.e. the more reduced, the token, the worse one would expect it to match the sequence of models corresponding to the canonical transcriptions and the lower one would expect the score to be.

An analysis of variance showed that both the manual score model ($F(1, 416) = 56.13$, $p < 0.0001$) and the canonical score model ($F(1, 416) = 50.85$, $p < 0.0001$) differed from the Pluymaekers model significantly. (The *F*-statistic used in an analysis of variance is similar to the

t-statistic described earlier in this section, and the p-value is interpreted the same way as in the case of *t*-tests.) There was virtually no difference in the R^2 of the manual score model and the canonical score model.

6.2 ver-

For the Pluymaekers model, we found the following effects: onset complexity ($\hat{\beta} = -16.8$, $t(130) = -3.09$, $p < 0.01$) and the year of birth ($\hat{\beta} = -0.5$, $t(130) = -2.49$, $p < 0.05$). The R^2 of this model was 12%. For the manual score model, there were significant main effects of onset complexity ($\hat{\beta} = -17.4$, $t(129) = -3.38$, $p < 0.001$), the year of birth ($\hat{\beta} = -0.5$, $t(129) = -2.55$, $p < 0.05$), and manual score ($\hat{\beta} = 2.3$, $t(129) = 4.08$, $p < 0.0001$). The R^2 of this model was 22%. For the canonical score model, there were significant main effects of onset complexity ($\hat{\beta} = -17.4$, $t(129) = -3.34$, $p < 0.01$), the year of birth ($\hat{\beta} = -0.5$, $t(129) = -2.51$, $p < 0.05$), and canonical score ($\hat{\beta} = 2.2$, $t(129) = 3.54$, $p < 0.001$). The R^2 of this model was 20%. Younger speakers produced shorter prefixes. The prefix was also shorter if the number of consonants in the onset of the stem was high, or if the manual score or the canonical score was low.

An analysis of variance showed that both the manual score model ($F(1, 129) = 16.62$, $p < 0.0001$) and the canonical score model ($F(1, 129) = 12.56$, $p < 0.001$) differed from the Pluymaekers model significantly. The manual score model and the canonical score model did not, however, differ from each other much. Unlike in the case of *ge-*, onset complexity (which was significant at the 0.01 level in the Pluymaekers model) was not overridden by the spectral reduction scores. Apart from the fact that onset complexity was a more robust variable to begin with, it may well be that cross-syllable coarticulation is weaker and less systematic for the closed syllable /v@r/ than for the open syllable /x@/.

6.3 ont-

For the Pluymaekers model, there were significant main effects of the interaction between frequency and speech rate ($\hat{\beta} = -3.1$, $t(94) = -3.66$, $p < 0.001$), the interaction between frequency and the year of birth ($\hat{\beta} = 0.3$, $t(94) = 3.24$, $p < 0.01$), and the year of birth ($\hat{\beta} = -1.4$, $t(94) = -5.06$, $p < 0.0001$). The R^2 of this model was 25%. For the manual score model, there were significant main effects of the interaction between frequency and speech rate ($\hat{\beta} = -2.9$, $t(93) = -3.38$, $p < 0.01$), the interaction between frequency and the year of birth ($\hat{\beta} = 0.3$, $t(93) = 3.03$, $p < 0.01$), the year of birth ($\hat{\beta} = -1.4$, $t(93) = -4.96$, $p < 0.0001$), and manual score ($\hat{\beta} = 1.1$, $t(93) = 1.24$, $p \approx 0.22$). The R^2 of this model was 26%. For the canonical score model, there were significant main effects of the interaction between frequency and

speech rate ($\hat{\beta} = -3.0$, $t(93) = -3.43$, $p < 0.001$), the interaction between frequency and the year of birth ($\hat{\beta} = 0.3$, $t(93) = 3.06$, $p < 0.01$), the year of birth ($\hat{\beta} = -1.4$, $t(93) = -4.99$, $p < 0.0001$), and canonical score ($\hat{\beta} = 0.8$, $t(93) = 0.98$, $p \approx 0.33$). The R^2 of this model was 26%. Younger speakers produced shorter prefixes. The prefix was also shorter if the manual score or the canonical score was low.

An analysis of variance showed that neither the manual score model ($F(1, 93) = 1.53$, $p \approx 0.22$) nor the canonical score model ($F(1, 93) = 0.95$, $p \approx 0.33$) differed from the Pluymaekers model significantly. The manual score model and the canonical score model did not differ from each other either. It is unclear why spectral reduction was not a significant predictor for /Ont/. It could be that the degree of nasalisation in the vowel varies independently from reduction proper. It could also be that the variance induced by uncontrolled factors, such as between-speaker differences, limits the maximum proportion of variance that can be explained with the variables in the model.

6.4 -lijk

In the case of words in non-final position, there were significant main effects of frequency ($\hat{\beta} = -7.0$, $t(107) = -3.48$, $p < 0.001$) and the year of birth ($\hat{\beta} = -0.8$, $t(107) = -3.45$, $p < 0.001$) for the Pluymaekers model. The R^2 of this model was 19%. For the manual score model, there were significant main effects of frequency ($\hat{\beta} = -6.8$, $t(106) = -3.45$, $p < 0.001$), the year of birth ($\hat{\beta} = -0.8$, $t(106) = -3.63$, $p < 0.001$), and manual score ($\hat{\beta} = 1.9$, $t(106) = 2.20$, $p < 0.05$). The R^2 of this model was 22%. For the canonical score model, there were significant main effects of frequency ($\hat{\beta} = -6.9$, $t(106) = -3.46$, $p < 0.001$), the year of birth ($\hat{\beta} = -0.8$, $t(106) = -3.60$, $p < 0.001$), and canonical score ($\hat{\beta} = 1.6$, $t(106) = 1.89$, $p < 0.1$). The R^2 of this model was 21%. Words with a higher frequency had shorter realisations of *-lijk*. The prefix was also shorter if the speakers were young, or if the manual score or the canonical score was low.

In the case of words in final position, there were significant main effects of the presence of the plosive ($\hat{\beta} = 144.9$, $t(35) = -3.32$, $p < 0.01$) and speech rate ($\hat{\beta} = -32.8$, $t(35) = -3.92$, $p < 0.001$) for the Pluymaekers model. The R^2 of this model was 45%. For the manual score model, there were significant main effects of the presence of the plosive ($\hat{\beta} = 154.9$, $t(34) = -3.65$, $p < 0.001$), speech rate ($\hat{\beta} = -29.0$, $t(34) = -3.48$, $p < 0.01$), and manual score ($\hat{\beta} = 6.4$, $t(34) = 1.88$, $p < 0.01$). The R^2 of this model was 50%. For the canonical score model, there were significant main effects of the presence of the plosive ($\hat{\beta} = 157.1$, $t(34) = -3.69$, $p < 0.001$), speech rate ($\hat{\beta} = -29.9$, $t(34) = -3.65$, $p < 0.001$), and canonical score ($\hat{\beta} = 6.5$,

$t(34) = 1.89$, $p < 0.1$). The R^2 of this model was 50%. The prefix was shorter if the speech rate was high, the plosive was absent, or if the manual score or the canonical score was low.

For the words in non-final position, an analysis of variance showed that both the manual score model ($F(1, 106) = 4.84$, $p < 0.05$) and the canonical score model ($F(1, 106) = 3.55$, $p < 0.1$) differed from the Pluymaekers model significantly. Also for the words in final position, an analysis of variance showed that both the manual score model ($F(1, 34) = 3.53$, $p < 0.1$) and the canonical score model ($F(1, 34) = 3.57$, $p < 0.1$) differed from the Pluymaekers model significantly. Again, there was virtually no difference between the manual and canonical score models in either case. It is interesting to note that spectral reduction does not subtract from the predictive power of the categorical variable 'plosive present'. This should not be taken to mean that the absence or presence of /k/ does not affect the spectral reduction scores. Rather, these results are due to the mechanics of the model fit: if two or more predictors explain the same part of the variance, the most powerful variable will take it all – only leaving the residuals for its competitors. Thus, it seems that the categorical absence or presence of /k/ is a stronger predictor of the duration of the suffix than the spectral reduction scores.

7 General discussion

In this study, we investigated the use of log-likelihoods (normalised for duration) from an HMM-based forced alignment procedure as a correlate of acoustic reduction in the speech signal as an alternative for, or as an addition to duration as a correlate of reduction. We referred to these normalised log-likelihood values as spectral reduction scores. The results of our study suggest that the spectral reduction scores capture different aspects of reduction than duration – at least in the sense that the spectral reduction scores cannot be explained by the same linguistically motivated variables as duration. However, they do explain part of the duration variance unaccounted for by the linguistically motivated variables for three of the four Dutch affixes under investigation: *ge-*, *ver-*, and *-lijk*. This is supported by the finding that, for these affixes, the spectral reduction scores only weakly correlate with the durations of the affixes (the correlation between duration and the canonical scores is 0.33 (R^2 = 0.11) for *ge-*, 0.29 (R^2 = 0.08) for *ver-*, 0.12 (R^2 = 0.01) for ont-, 0.10 (R^2 = 0.01) for non-final *-lijk*, and 0.34 (R^2 = 0.12) for final *-lijk* without any outliers removed). Except for final *-lijk*, the increase in the proportion of variance in the multiple regression models explained by the spectral reduction measures is close to the R^2 for the bivariate correlation between spectral reduction and duration. This corroborates the conclusion that our measure of spectral reduction is largely orthogonal to the linguistic measures. At the same time, it is interesting to note that all correlations between spectral reduction and duration predict that shorter tokens correspond with larger spectral reduction. Since our spectral reduction measure is normalised for duration, this suggests that reduction is not limited to time compression, but that there is an additional effect on articulatory 'simplification'.

In our first experiment, we tried to predict the spectral reduction scores of the affixes using the linguistically motivated variables from Pluymaekers et al. (2005). None of the 'linguistic' models that we fitted explained more than 4% of the variance in the data. Considering the fact that duration *can* (partially) be predicted using the said linguistically motivated variables, and the fact that there is a weak correlation between duration and the spectral reduction scores, this finding is rather interesting. There are at least two potential explanations for it. First, it may be difficult for linguistically motivated variables to predict the spectral reduction scores because the latter are based on a complex measure that combines spectral and time-warp differences in the acoustic space into a single number (as opposed to 'duration', which is rather a simple, one-dimensional correlate of reduction (see Section 1)). Second, the spectral reduction scores are subject to token-by-token variation due to a large number of uncontrolled factors, such as speaker identity and phonetic context from the preceding and following morphemes. This may have added 'noise' to the spectral reduction scores. The same holds for duration but the variance contributed by the uncontrolled variables can again be expected to be smaller because of duration being a simpler correlate of reduction. While random variation should not affect the outcome of linear regression models if the number of observations is very high, the number of observations may have been an issue for all models except for *ge-*, which had more than 420 observations (see Table 1). Then again, in the case of *ge-*, the impact of the first phoneme of the following morpheme may have been particularly strong because the affix ends with a vowel.

As one can see from Equation 4, the distance between an observed token of an affix and the maximally unreduced pronunciation not only depends on the properties of the token itself, but also on the representation of the unreduced reference. We defined the reference as the sequence of the triphones underlying the canonical phonetic transcription of the affix. We investigated triphones trained with both manual(ly verified) and canonical transcriptions of read speech. The spectral reduction scores obtained using the two sets of triphones were almost identical (the correlation coefficients between the manual and the canonical scores were 0.98 for *ge-*, *ont-*, and *-lijk*, and 0.93 for *ver-*). However, it must be pointed out that both sets of acoustic models were based on the same type of training data. In other words, the distance from the canonical transcription is not a purely 'linguistic' measure; it is actually the distance from the training data.

Our spectral reduction measure is susceptible to the well-known trajectory folding problem (Han et al., 2007); different tokens taking different trajectories through the acoustic space may end up with identical log-likelihoods, even if their trajectories make very different auditory impressions. This is yet another reason why it may not be appropriate to map multidimensional acoustic reduction to a real number. While it is difficult to imagine how reduction could be described in terms other than deviation from some reference, it is not obvious that there is one unique reference or one correct way of defining it. In this paper, we used context- and speaker-independent statistical models as the reference. This implies the assumption that all effects of context, speech style, regional background, gender, age, etc. are accounted for by the models. As we have seen, this assumption may not be warranted. Including

'context' and 'speaker' as random factors in the regression models might be one way around this problem. However, this would require a data set that is orders of magnitude larger than the data set we had available for our research. Similarly, building a mixed model would not be possible with the amount of data that we had.

If we blame the failure to model spectral reduction on the inherent uncontrolled variation in the scores, the question arises what makes duration a measure of reduction that is so much easier to model. We believe that the answer lies in duration being less sensitive to factors such as phonetic context and speaker identity than the trajectories in the spectral space. In addition, while spectral reduction is a result of a trajectory in a multi-dimensional space, duration is inherently a scalar variable.

In passing, it may be interesting to note that the relation between the 'predictability of a linguistic unit' and its duration in a spoken utterance is not as clear-cut as one might think. In a recent study, Kuperman et al. (2007) found that infixes in Dutch (/@/, /@n/, or /s/ connecting two nouns that together form a compound) are *longer* if they are more predictable from the nouns that make up the compound. This finding is explained as a tendency to gloss over sounds of which the speaker is not very confident that they should be there.

Both in this study and in the paper of Pluymaekers et al. (2005), the proportion of variance in the affix durations that could be explained by the linguistically motivated variables ranged from the low $R^2 = 0.09$ for *ge-* to the high $R^2 = 0.45$ for *-lijk* in final position; the R^2 values for *ver-* and *-lijk* in non-final position were almost as low as the value for *ge-*, while the value for *ont-* ($R^2 = 0.25$) was in the middle. The original paper does not offer an explanation for the wide range of explained variance, and we are not in the position to offer a convincing explanation either. For *ge-*, *ver-* and non-final *-lijk* – i.e., for the affixes with a low R^2 in the Pluymaekers model – spectral reduction scores raised the proportion of explained variance to about 20%. For *ont-*, spectral reduction scores were unable to increase to proportion of explained variance much. We speculate this to be due to the effect of the nasal that is likely to cause substantial variance in the spectral reduction measure (over and above the variance introduced by deletions of /t/ and/or /n/).

Because extending the linguistically motivated variables with the spectral reduction scores as predictors increases R^2 for almost all models, one might ask if a similar effect would hold for models that predict spectral reduction scores with the combination of linguistically motivated variables and duration. This appears not to be the case; the explained variance for such models is much lower than the explained variance for models predicting duration with the combination of linguistically motivated variables and spectral reduction scores. Although this may seem surprising, it is an effect that is frequently encountered in regression studies that involve more than two variables (Langford et al., 2001).

In this study, we opted for a measure of spectral reduction that does not rely on the descriptive concepts of acoustic phonetics (e.g., formant frequencies). By doing so, we may seem to ignore previous research on the acoustic reduction of vowels (van Bergem, 1995) and consonants (van Son & Pols, 1999) in Dutch. However, we argue that an approach along the

lines of conventional acoustic phonetics is not feasible for capturing spectral reduction in the four affixes under investigation. Three of the affixes have a schwa in their canonical transcription; this raises the question how one could represent vowel reduction in terms of formant frequencies. Furthermore, the formant values of the /O/ in the prefix *ont-* may be affected both by the final phonemes in the preceding word and by spectral reduction of the affix proper; the potentially disturbing effects of the nasal have already been alluded to. As for consonant reduction, a representation in terms of formant frequencies is inherently questionable; the formant concept only applies with strong restrictions. Moreover, formants in the consonants occurring in spontaneous conversations defy any attempt at automatic measurement. Finally, known reduction measures from acoustic phonetics would only apply to individual phonemes in an affix, leaving us with the problem of incorporating these phoneme-based measures into a measure of acoustic reduction on the affix level.

8 Conclusions

In this study, we proposed a measure of spectral reduction that might either replace or add to duration as a measure of reduction in speech. It appeared that the proposed spectral reduction scores capture other aspects of reduction than duration: while duration can – to a moderate degree – be predicted by a number of linguistically motivated variables, spectral reduction scores cannot. At the same time, spectral reduction scores are able to predict a substantial amount of the variation in duration that the linguistically motivated variables do not account for. We discussed why spectral reduction measures are difficult to express in the form of a scalar. It appears that powerful models of spectral reduction require modelling techniques that can handle factors such as phonetic context, speaker, and speaking style as random variables. This will only be possible when very large corpora are available.

References

Aylett, M., Turk, A. (2004). "The smooth signal redundancy hypothesis: a functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech," *Language and Speech*, 47, 31-56.

Bybee, J.L. (2001). *Phonology and language use* (Cambridge University Press, UK).

Ernestus, M. (2000). *Voice assimilation and segment reduction in casual Dutch: A corpus-based study of the phonology-phonetics interface* (LOT, Utrecht, The Netherlands).

Ernestus, M., Lahey, M., Verhees, F., Baayen, R.H. (2006). "Lexical frequency and voice assimilation," *Journal of the Acoustical Society of America*, 120, 1040-1051.

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

Han, Y., Veth, J.M. de & 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., Boves, L., de Veth, J., ten Bosch, L. (2007). "On the utility of syllable-based acoustic models for pronunciation variation modelling," *EURASIP Journal on Audio, Speech, and Music Processing*, Article ID 46460, 11 pages, doi:10.1155/2007/46460.

Hämäläinen, A., ten Bosch, L., Boves, L. (2009). "Modelling pronunciation variation with single-path and multi-path syllable models: Issues to consider," *Speech Communication*, 51, 130-150.

Izenman, A.J. (2008). *Modern Multivariate Statistical Techniques: Regression, Classification and Manifold Learning* (Springer, New York).

Jespersen, O. (1922). *Language: its nature, development and origin* (George Allen & Unwin Ltd, London).

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

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

Kuperman, V., Pluymaekers, M., Ernestus, M., Baayen, R.H. (2007). "Morphological predictability and acoustic salience of interfixes in Dutch compounds," *Journal of the Acoustical Society of America*, 121, 2261-2271.

Langford, E., Schwertman, N., Owens, M. (2001). "Is the property of being positively correlated transitive?" *American Statistician*, 55(4), 322-325.

Lindblom, B. (1963). "Spectrographic study of vowel reduction," *Journal of the Acoustical Society of America*, 35(11), 1773-1781.

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 International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Canary Islands, Spain, Vol. 1, pp. 340-347.

Pluymaekers, M., Ernestus, M., Baayen, R.H. (2005). "Lexical frequency and acoustic reduction in spoken Dutch," *Journal of the Acoustical Society of America*, 118, 2561-2569.

van Bergem, D. (1995). *Acoustic and lexical vowel reduction* (IFOTT, University of Amsterdam, The Netherlands).

van Son, R.J.J.H., Pols, L.C.W. (1999). "An acoustic description of consonant reduction," *Speech Communication*, 28, 125-140.

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

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, UK).

Zipf, G. (1929). "Relative frequency as a determinant of phonetic change," *Harvard Studies in Classical Philology*, 15, 1-95.

III SUMMARY & CONCLUSIONS

Summary & conclusions

The common thread in the articles included in this thesis is that they all have to do with both pronunciation variation and the speech decoding methods deeply rooted in automatic speech recognition (ASR) technology. The first two articles concentrate on pronunciation variation modelling for the purpose of ASR using syllable-length acoustic models, while the third article focuses on the possibility of using techniques from ASR to analyse acoustic reduction, i.e., a specific form of pronunciation variation. We conclude this thesis by summarising the main findings of the three articles and suggesting directions for future research. Section 1 covers the first two articles, whilst Section 2 looks at the third article.

1 Syllable-length acoustic models

The first two articles included in this thesis investigated the use of syllable-length acoustic models for pronunciation variation modelling in large-vocabulary continuous speech recognition. The research was based on the hypothesis that syllable models would lead to improved speech recognition performance because they model variation in more specific lexical contexts than phone models, which are usually trained using instances of the phonemes in several different lexical contexts, and because they have long-span spectro-temporal patterns inherently embedded into them (Ganapathiraju et al., 2001). The need to investigate alternative model topologies for large-vocabulary continuous speech recognition stems from the difficulties of modelling pronunciation variation with the conventional context-dependent phone models (e.g., Kessens, 2002; Ostendorf, 1999).

Earlier studies on syllable models for large-vocabulary continuous speech recognition present inconsistent results, ranging from deterioration (e.g., Jouvet & Messina, 2004) to seemingly enormous improvements (Sethy & Narayanan, 2003) in speech recognition performance. The majority of the results (e.g., Ganapathiraju et al., 2001; Sethy et al., 2003) do not, however, differ much from the results obtained with conventional phoneme-based recognisers. The recognition results obtained with syllable models have usually been presented without in-depth analysis on the aspects of pronunciation variation that the syllable models are actually able to capture. Therefore, the research reported in the first two articles included in this thesis was intended to advance our understanding about the issues playing a role in pronunciation variation modelling with syllable models. The first article concentrated on the initialisation of the syllable models, while the second article examined the importance of modelling within-syllable pronunciation variation and syllable context when using syllable models. The next two subsections summarise the main findings from these two articles.

1.1 Insights into the initialisation of syllable-length acoustic models

The first article took a critical look at the above-mentioned Sethy & Narayanan (2003) paper. In that paper, Sethy & Narayanan vastly improve upon the performance of a conventional triphone recogniser by using a so-called mixed-model recogniser with context-independent models for the most common monosyllabic words and syllables in their task, and triphones for the rest of the words and syllables. They attribute the improvement to the careful selection of the syllables to be modelled with syllable models and to the method they introduce for initialising the syllable models. The initialisation method uses the triphones underlying the canonical transcriptions of the syllables to initialise the model topologies and parameters of the corresponding syllable models. Subsequent Baum-Welch re-estimation is expected to incorporate the long-span spectral and temporal dependencies in speech into the models.

For our article, we replicated Sethy & Narayanan's speech recognition experiments on the TIMIT database (1990) and carried out similar experiments on a comparably sized part of the Spoken Dutch Corpus (Oostdijk et al., 2002). We trained two different sets of triphones, which we then used to obtain baseline recognition results and to initialise the syllable models used in the mixed-model recognisers: 1) 'manual triphones' trained using the manual(ly verified) phonetic labels available in the two corpora, and 2) 'canonical triphones' trained using the canonical transcriptions of the words in the corpora. The mixed-model recognisers substantially outperformed the triphone recognisers only in the case of manual triphones. In the case of canonical triphones, there were no significant differences between the performance of the triphone recognisers and the mixed-model recognisers. By training two different sets of triphones and by analysing what happens when syllable models are trained further from the sequences of triphones used for initialising them, we were able to show that the apparent improvement in the recognition performance reported by Sethy & Narayanan was only due to the mismatch between the representations of speech during training and testing. In the case of the baseline triphone recognisers with manual triphones, the triphones were trained using manual(ly verified) transcriptions, whereas the recognition lexicon contained canonical transcriptions (as typically is the case in ASR). While training with careful manual transcriptions yields 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 (cf. Wester, 2002). The mismatch had little effect on the mixed-model recognisers because the initialisation method is based on canonical transcriptions and because the subsequent Baum-Welch re-estimation ensures that syllable models initialised with manual triphones actually become very similar to syllable models initialised with canonical triphones. In other words, the baseline recognisers with manual triphones had a disadvantage that led to seeming improvements in the recognition performance when using the mixed-model recognisers.

1.2 Findings on the importance of syllable context and within-syllable pronunciation variation

The experiments described above made it obvious that appropriately initialised context-independent 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. Therefore, by carrying out the experiments described in the second article included this thesis, we wanted to understand the relative importance of different types of pronunciation variation phenomena from the point of view of speech recognition performance. In particular, we were interested in the role of syllable context information and within-syllable pronunciation variation. As the number of pronunciation variants per syllable can be large, we suspected that it might not be possible to model within-syllable pronunciation variation in terms of the Gaussian mixtures of a single-path syllable model only – but that changes to the model topology would be necessary. In an attempt to model within-syllable pronunciation variation more accurately, we introduced a method for building multi-path syllable models. We then constructed mixed-model recognisers with context-independent single-path and multi-path syllable models used to represent monosyllabic words, constituent syllables of polysyllabic words, or both. To obtain insights into the factors under investigation, we compared the recognition performance of the different recognisers with each other and with the recognition performance of a baseline triphone recogniser, and analysed the word-level and sentence-level errors made by the recognisers that were the most revealing with respect to our goal. The error analyses showed that the most important factors affecting the recognition performance are syllable context and the lexical confusability caused by the additional paths in the syllable models. Furthermore, the recognition results suggested that the greater acoustic modelling accuracy of the multi-path syllable models is of use only if the information about the syllable-level pronunciation variation can be linked with the word-level information in the language model.

1.3 Suggestions for future research

The research covered in the two articles discussed above tried to come to grips with pronunciation variation using a combination of phonetic and linguistic techniques, and relatively small training corpora. With hindsight, we must ask ourselves whether this was the right direction to pursue at this point in time. Based on our findings, the most promising way forward with syllable models seems to be context-dependent syllable models. Building context-dependent syllable models would, however, require at least two things: 1) a method for sharing the model parameters of the contextual units and, potentially, of parts of similar syllables, and 2) large amounts of training data. Jouvet & Messina (2004) introduced a parameter sharing method that led to improvements in recognition performance on most of their French speech recognition tasks. For models with eight Gaussian mixtures per state, the improvement was 31% on a test set with isolated digits, 22% on a test set with isolated words, and 1% on test sets with city names and digit pairs from 00 to 99. Similarly, Wu & Wu (2007) have published promising

preliminary results on context-dependent syllable models on Mandarin Chinese, reporting a 5% improvement on a small test set with 15 minutes of speech. Jouvet & Messina (2004) had a training corpus with 300 hours of speech, while Wu & Wu's (2007) training corpus contained 360 hours of speech. These results also suggest that, together with ever-increasing amounts of training data, context-dependent syllable models could indeed help us in lowering word error rates. At least initially, such syllable models might be the most profitable in the case of languages with a simpler syllable structure and, consequently, a lower syllable count than the languages that we carried out experiments on in this thesis (Dutch and English). Such languages include, for instance, some Asian tone languages (e.g., Mandarin Chinese).

2 Using acoustic models for analysing acoustic reduction

2.1 Experimenting with an ASR-based measure of acoustic reduction

The third article included in this thesis investigated the usefulness of ASR-based speech decoding methods for analysing acoustic reduction. It was based on the hypothesis that the total log-likelihood ('acoustic score') obtained by carrying out a forced alignment of a canonical sequence of triphones for a specific speech unit (e.g., a word) against the corresponding stretch of speech signal would carry information about the degree of reduction in the speech unit. When normalised for duration, the acoustic score would provide a measure of spectral reduction. Such an automatically derived measure of spectral reduction would be of interest to the speech community because of the comparative ease of obtaining it, and because of the elimination of the inconsistencies typical of human measurements. Also, to be able to explain reduction, we probably need a better understanding of the underlying articulation. Making comprehensive articulatory measurements of spontaneous speech is virtually impossible. Neither is it possible to uniquely recover articulation from the acoustic speech signal. However, if it were possible to link acoustic reduction to speech models (such as HMMs), we might be able to start creating links to the underlying articulation.

Our article described a study in which the idea of using normalised acoustic scores as a measure of spectral reduction in the speech signal was explored for four Dutch affixes from a large corpus of face-to-face conversations. It built upon an earlier study examining the effects of lexical frequency on durational reduction in spoken Dutch (Pluymaekers et al., 2005), and investigated whether the proposed measure of reduction could either replace or add to duration as a measure of reduction. We used read speech for training the triphones that were used for the forced alignment to make sure that the acoustic scores would be based on the dissimilarity between an individual affix token and a maximally unreduced form of the affix. During the research described in the first article included in this thesis, we had come to the conclusion that 'manual triphones', which are trained using the manual(ly verified) phonetic transcriptions of the words in the training data, are not a priori useful for the purpose of conventional ASR (see Section 1.1). As they are acoustically more accurate than 'canonical triphones', which are

trained using canonical transcriptions, we hypothesised that they may be useful for certain types of speech analysis. Therefore, we trained both manual triphones and canonical triphones to investigate how important the additional accuracy of manual triphones would be for our spectral reduction scores. The results suggested that the spectral reduction scores capture other aspects of reduction than duration. While duration can – to a moderate degree – be predicted by a number of linguistically motivated variables (such as word frequency, segmental context, and speech rate), the spectral reduction scores cannot. This may be due to the fact that spectral reduction is inherently a multidimensional phenomenon. However, at the same time, the spectral reduction scores are able to predict a substantial amount of the variation in duration that the linguistically motivated variables do not account for. The difference between manual and canonical triphones proved insignificant for the results.

2.2 Suggestions for future research

Reduction must be interpreted as deviation from some canonical reference pronunciation. This raises the fundamental question how such a canonical reference pronunciation should be defined. In our research, we decided to use context- and speaker-independent statistical models (triphones trained on speech read by speakers that were different from the speakers in the face-to-face conversations from which the affix tokens were selected) to define the canonical reference pronunciations. This decision implies the assumption that all effects of context, speech style, regional background, gender, age, etc. are accounted for in the models. However, it appears that this assumption may not be warranted. Therefore, taking systematic context and speaker effects into account would be an interesting direction for future research. There are essentially two ways to do this. One could either train (tied) context-dependent affix models, or include ‘context’ and ‘speaker’ as random factors in the regression models. However, as in the case of context-dependent syllable models for large-vocabulary continuous speech recognition (see Section 1.3), both options would require a much larger amounts of data than what was available for our research.

Samenvatting (Summary in Dutch)

De kern van dit proefschrift bestaat uit drie wetenschappelijke artikelen. De rode draad die de artikelen met elkaar verbindt is dat alle drie te maken hebben met zowel uitspraakvariatie als met methoden voor spraakverwerking die gebruikt worden in automatische spraakherkenning. De eerste twee artikelen gaan over het modelleren van uitspraakvariatie voor automatische spraakherkenning door middel van akoestische modellen van syllaben (syllabemodellen), terwijl het derde artikel zich toespitst op het gebruiken van de methodes van automatische spraakherkenning voor het analyseren van akoestische reductie. Deze reductie kan gezien worden als één van de specifieke gevolgen van uitspraakvariatie. Hieronder geven we een korte samenvatting van de drie artikelen in dit proefschrift.

Het doel van het eerste artikel is het vergroten van ons inzicht in de condities waarin syllabemodellen betere herkenningsresultaten op kunnen leveren dan foneemmodellen voor automatische spraakherkenningstaken met een groot vocabulaire. Een belangrijk motief voor de opzet van het onderzoek was dat de herkenningsresultaten met syllabemodellen die in de literatuur gerapporteerd worden grote verschillen vertonen tussen verschillende spraakcorpora. Vooral Sethy & Narayanan (2003) rapporteren een veel grotere verbetering in herkenningsprestatie dan andere onderzoekers. Om hun resultaten beter te begrijpen hebben we hun experimenten herhaald en aangevuld met vergelijkbare experimenten met een Nederlands spraakcorpus. Vervolgens hebben we de verschillen tussen de resultaten van de twee sets van experimenten geanalyseerd. Het onderzoek richt zich vooral op de rol van de procedure die is gebruikt voor het initialiseren van de syllabemodellen; de syllabemodellen zijn geïnitialiseerd met de foneemmodellen van de canonische transcripties van de syllaben. De resultaten lieten zien dat de details van de procedure een substantieel effect op de herkenningsresultaten hebben. Het trainen van foneemmodellen op basis van handmatig gemaakte transcripties van het trainingscorpus in combinatie met het gebruiken van canonische transcripties in het lexicon van de spraakherkenner veroorzaakt een mismatch tussen de training en de test die een negatief effect heeft op de herkenningsresultaten van de foneemgebaseerde spraakherkenner. Die mismatch verdwijnt, als foneemmodellen gebruikt worden als startpunt voor het trainen van syllabemodellen. De grote verbetering die Sethy & Narayanan lieten zien was dus in feite een gevolg van een fout in de door hun gebruikte procedure. Als die fout gecorrigeerd wordt, blijkt dat de winst die met eenvoudige syllabemodellen geboekt kan worden relatief klein is.

Het doel van het tweede artikel is het onderzoeken van het belang van het modelleren van uitspraakvariatie binnen een syllabe en van het belang van fonetische context bij het gebruik van syllabemodellen voor spraakherkenners met een groot vocabulaire. Om syllabe-interne uitspraakvariatie accuraat te modelleren hebben we een methode ontwikkeld voor het toevoegen van parallelle paden in syllabemodellen. De motivatie voor het toevoegen van parallelle paden kwam voort uit de resultaten van een analyse van het aantal uitspraakvarianten per syllabe. We geven als voorbeeld de Nederlandse syllabe /hEt/. Deze syllabe (van het woord 'het'), heeft 27 verschillende uitspraakvarianten in het corpus dat voor de experimenten gebruikt werd. Door dat

grote aantal en door de grote verschillen tussen sommige varianten is het moeilijk voor te stellen dat één pad in het model genoeg zou zijn voor het modelleren van alle mogelijke uitspraakvariatie in de syllabe. Daarom hebben we context-onafhankelijke syllabemodellen met één en met meerdere paden in het model gemaakt en deze modellen gebruikt voor het representeren van monosyllabische woorden, de relevante syllaben van polysyllabische woorden of allebei. Om inzicht te krijgen in het belang van het modelleren van syllabe-interne uitspraakvariatie en fonetische context voor de herkenningsresultaten hebben we de herkenningsprestatie van een aantal verschillende spraakherkenners met elkaar vergeleken. We hebben in het bijzonder de fouten van de spraakherkenners op woord- en zinsniveau geanalyseerd. Zowel de foneemmodellen als de syllabemodellen met één pad waren beter dan syllabemodellen met parallelle paden. De foutanalyses toonden aan dat de belangrijkste factoren die de kwaliteit van de spraakherkenners bepaalden de syllabecontext en lexicale verwarbaarheid zijn. Verder gaven de herkenningsresultaten aan dat de voordelen van de syllabemodellen met parallelle paden alleen kunnen worden benut als de informatie over de uitspraakvariatie op syllabeniveau gerelateerd kan worden met de informatie op woordniveau in het taalmodel.

Het derde artikel introduceert een maat van spectrale reductie die gebaseerd is op de log-likelihoods (‘akoestische scores’) die worden opgeleverd door een spraakherkenner als een spraaksignaal wordt opgelijnd met een rij fonemen. Deze studie bouwt voort op een eerdere studie van Pluymaekers et al. (2005) en onderzoekt of een maat voor reductie die volledig gebaseerd is op de duur van fonemen en syllaben vervangen zou kunnen worden door de ons voorgestelde spectrale reductiemaat. De resultaten laten zien dat de spectrale reductiemaat andere (en meer) informatie geeft dat de duurmaat. Duur kan tot op zekere hoogte worden voorspeld door een aantal linguïstisch gemotiveerde variabelen (zoals woordfrequentie, context en spraaktempo), maar de spectrale reductiemaat is daarvoor te ingewikkeld. We geloven dat dit komt doordat spectrale reductie een inherent multidimensionaal fenomeen is, en complexer dan de een-dimensionale duurmaat. Ons interessantste resultaat is dat onze scores voor spectrale reductie een substantiële hoeveelheid van de variatie in duur, die niet door de linguïstische variabelen wordt gemodelleerd, kunnen voorspellen (en niet andersom).

De drie artikelen in dit proefschrift laten zien hoe complex het modelleren van spraak is. Ze tonen hoe het gebruik van geavanceerde technieken zoals automatische spraakherkenning kan leiden tot betere wetenschappelijke inzichten over de manier waarop spraak en uitspraakvariatie kan worden beschreven.

References

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

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

Kessens, J. (2002). *Making a difference: On automatic transcription and modeling of Dutch pronunciation variation for automatic speech recognition* (University of Nijmegen, The Netherlands).

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 International Conference on Language Resources and Evaluation (LREC)*, Las Palmas, Canary Islands, Spain, Vol. 1, pp. 340-347.

Ostendorf, M. (1999). Moving beyond the 'beads-on-a-string' model of speech, in: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Keystone, CO, USA, pp. 79-84.

Pluymaekers, M., Ernestus, M., Baayen, R.H. (2005). "Lexical frequency and acoustic reduction in spoken Dutch," *Journal of the Acoustical Society of America*, 118, 2561-2569.

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

Sethy, A., Ramabhadran, B., Narayanan, S. (2003). Improvements in ASR for the MALACH project using syllable-centric models, in: *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, St. Thomas, U.S. Virgin Islands, USA, pp. 129-134.

TIMIT Acoustic-Phonetic Continuous Speech Corpus (1990). National Institute of Standards and Technology Speech Disc 1-1.1, NTIS Order No. PB91-505065.

Wester, M. (2002). *Pronunciation variation modelling for Dutch automatic speech recognition* (University of Nijmegen, The Netherlands).

Wu, H., Wu, X. (2007). Context-dependent syllable acoustic model for continuous Chinese speech recognition, in: *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, Antwerp, Belgium.

Curriculum vitae

Annika Hämäläinen was born in Rovaniemi, Finland, on 6 December 1975. She graduated with a Bachelor's degree in Electrical Engineering from the Rovaniemi University of Applied Sciences in 2000, and with a Master's degree in Speech and Language Processing from the University of Edinburgh in Scotland in 2002. In January 2004, Annika started working on her PhD project at the Centre for Language and Speech Technology at Radboud University Nijmegen in the Netherlands. This thesis describes the research conducted during that project.

In addition to her studies, Annika has undertaken periods of employment in industry. Her previous work experience in the area of speech technology includes two work placements at the Advanced Speech Technology Unit at British Telecommunications in Ipswich, England, in 2001 and in 2002, and a position as a TTS language specialist for Finnish at Nuance Communications in Ghent, Belgium, in 2007. Since 2002, she has also worked as a freelance language consultant, delivering Finnish and English projects for several international customers. Today, Annika is employed as a TTS language consultant at Loquendo in Turin, Italy.

List of publications

Journal articles

Hämäläinen, A., Gubian, M., ten Bosch, L., Boves, L. (in press). "Analysis of acoustic reduction using spectral similarity measures," *Journal of the Acoustical Society of America*.

Hämäläinen, A., ten Bosch, L., Boves, L. (2009). "Modelling pronunciation variation with single-path and multi-path syllable models: Issues to consider," *Speech Communication*, 51, 130-150.

Hämäläinen, A., Boves, L., de Veth, J., ten Bosch, L. (2007). "On the utility of syllable-based acoustic models for pronunciation variation modelling," *EURASIP Journal on Audio, Speech, and Music Processing*, Article ID 46460, 11 pages, doi:10.1155/2007/46460.

Conference papers

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

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

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

Hämäläinen, A., ten Bosch, L., Boves, L. (2006). Multi-path syllable models based on phonetic knowledge, in: *Proceedings of the Phonetics Symposium*, Helsinki, Finland, pp. 57-66.

Hämäläinen, A., Han, Y., Boves, L., ten Bosch, L. (2006). Whither linguistic interpretation of acoustic pronunciation variation, in: *Proceedings of the Speech Recognition and Intrinsic Variation Workshop (SRIV)*, Toulouse, France.

Han, Y., Hämäläinen, A., Boves, L. (2006). Trajectory clustering of syllable-length acoustic models for continuous speech recognition, in: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France.

ten Bosch, L., Hämäläinen, A., Scharenborg, O., Boves, L. (2006). Acoustic scores and symbolic mismatch penalties in phone lattices, in: *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Toulouse, France.

Hämäläinen, A., Boves, L., de Veth, J. (2005). Syllable-length acoustic units in large-vocabulary continuous speech recognition, in: *Proceedings of the International Conference on Speech and Computer (SPECOM)*, Patras, Greece, pp. 499-502.

Hämäläinen, A., de Veth, J., Boves, L. (2005). Longer-length acoustic units for continuous speech recognition, in: *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Antalya, Turkey.