

MULTI-PATH SYLLABLE MODELS BASED ON PHONETIC KNOWLEDGE

Annika Hämmäläinen, Louis ten Bosch, Lou Boves
Radboud University Nijmegen, The Netherlands
{A.Hamalainen, L.tenBosch, L.Boves}@let.ru.nl

Abstract

Recent research suggests that syllable-length acoustic models might be more appropriate for pronunciation variation modelling than the context-dependent phones that conventional automatic speech recognisers use. In this paper, we compare the recognition performance of two types of recognisers: a conventional recogniser that only uses triphones, and an experimental recogniser that employs a mix of context-independent syllable models for a set of frequent syllables and sequences of triphones for the less frequent ones. The syllable models of the mixed-model recogniser are designed to consist of multiple HMM paths that are expected to capture major pronunciation variants. These paths are initialised using phonetic knowledge and re-estimated using a data-driven solution. When applied to 94 frequent syllables in a 37-hour corpus of Dutch read speech, the multi-path mixed-model recogniser outperforms a much more complex triphone recogniser.

Keywords: automatic speech recognition, pronunciation variation, multi-path HMM, syllable

1 Introduction

Conventional large-vocabulary continuous speech recognisers use context-dependent phone models, such as triphones, to model the elementary acoustic units of 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 the robust training of acoustic models when reasonable amounts of training data are available and when state tying methods are used to deal with triphones with insufficient training data. When using triphones, one must assume speech to be representable as a sequence of discrete phonemes ('beads on a string') that can only be substituted, inserted or deleted to account for pronunciation variation (Ostendorf 1999). Given this assumption, pronunciation variation should be possible to account for at the level of the phonetic transcriptions in the recognition lexicon (perhaps together with the introduction of a limited number of extra triphones that might occur in reduced syllables). Modelling pronunciation variation by adding transcription variants in the recognition lexicon has, however, met with limited success because of the resulting increase in lexical confusability (Kessens 2002, Wester 2002). Furthermore, while triphones are able to capture short-span contextual effects such as phoneme substitution and reduction (Jurafsky et al. 2001), the long-span spectral and temporal dependencies introduced by coarticulation are not easy to capture in models with such a limited duration (Ganapathiraju et al. 2001).

To alleviate the problems of the 'beads on a string' representation of speech, the use of longer-length acoustic models has been proposed. In particular, syllable-based acoustic models have been suggested as an alternative to triphones (Ganapathiraju et al. 2001, Hämmäläinen et al. 2005, Jones et al. 1997, Messina & Juvet 2004, Sethy &

Narayanan 2003, Sethy et al. 2003). The most important challenge of using syllable models is the inevitable sparseness of data in the model training; infrequent syllables do not have sufficient data available for reliable model parameter estimation, and there are no straightforward fallback methods similar to backing off to generalised triphones, diphones and monophones when there is not enough data to train all conceivable triphones. The solutions suggested to the data sparsity problem are two-fold. First, syllable models are only trained for frequent ‘target syllables’ for which at least around 100–150 training tokens are available; the syllables with fewer than the minimum required number of training tokens are modelled as sequences of triphones (Ganapathiraju et al. 2001, Sethy & Narayanan 2003). Second, to enable reliable training of the syllable models with such a limited number of training tokens, the model parameters are initialised with the parameters of the triphones underlying the canonical transcriptions of the syllables in question (Sethy & Narayanan 2003, Hämäläinen et al. 2005). An example of this kind of single-path syllable model initialised with triphones is presented in Figure 1.

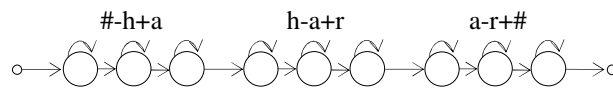


Figure 1. Single-path model for the syllable /har/, with the single path through the model initialised with the triphones underlying the canonical transcription. The symbol “#” in the triphone notation denotes the start or end of a syllable

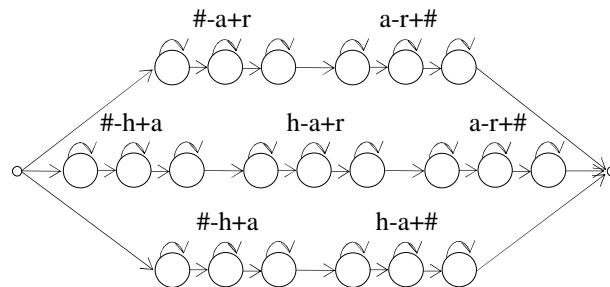


Figure 2. Multi-path model for the syllable /har/, with the three parallel paths initialised with the triphones underlying the MDVs /ar/, /har/ and /ha/, respectively

Previous experiments have shown that re-estimating the acoustic observation densities of single-path syllable models initialised with triphones does indeed appear to capture at least some of the coarticulation-related variation. However, it seems that this is not sufficient to account for the many different forms that syllable pronunciations can assume (Hämäläinen et al. 2005). Greenberg (1999) – amongst other authors – has shown that, while syllables are seldom deleted completely, they do display considerable variation in the identity and number of phonetic symbols that best reflect their pronunciation. At the same time, it is clear that a substantial part of the variation defies modelling in the form of different sequences of symbols, i.e. modelling at the level of

the phonetic transcriptions in the recognition lexicon (Ostendorf 1999). We believe that pronunciation variation could be modelled better by using syllable models with parallel paths that represent ‘major, distinct transcription variants’ (hereafter MDVs), and re-estimating these parallel paths to better capture the dynamic nature of articulation. An example of a multi-path syllable model is shown in Figure 2.

In this paper, we construct multi-path models for frequent syllables using a combination of knowledge-based and data-driven methods. The knowledge-based part of our approach uses manually verified broad phonetic transcriptions of the target syllables for selecting MDVs, and for initialising the observation densities of the parallel paths aimed at modelling these MDVs. The data-driven part amounts to us leaving the training entirely to the Baum-Welch algorithm, instead of predefining which training tokens to use for re-estimating the model parameters of each parallel path. We use a mixed-model recognition scheme in which syllable models for 94 frequent syllables are combined with triphone models that cover the less frequent syllables in a Dutch read speech recognition task. We investigate whether multi-path syllable models improve recognition performance as compared with a conventional triphone recogniser.

This paper is further organised as follows. The speech material used in the study is described in Section 2. The concept of MDVs and the selection of the MDVs used for the initialisation of the parallel paths are discussed in Section 3. The experimental set-up, including the acoustic model training, is detailed in Section 4. Results from the recognition experiments are presented and discussed in Section 5. Finally, the conclusions are formulated in Section 6.

2 Speech material

The speech material used in this study was read speech extracted from the Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN), which – amongst other types of annotations – contains manually verified orthographic transcriptions for all of the data (Oostdijk et al. 2002). The data were divided into three sets comprising non-overlapping fragments of all 303 speakers: a set for training the acoustic models, a development set for optimising the language model scaling factor and word insertion penalty, and a test set for evaluating the acoustic models. Details of the data are presented in Table 1.

Table 1. Main statistics of the speech material

Statistic	Train	Development	Test
# Word tokens	396,187	22,100	22,289
# Speakers	303	303	303
Duration (hh:mm:ss)	37:00:20	02:03:33	02:04:21

A 60,600-word subset of the training data containing manually verified broad phonetic transcriptions and word-level segmentations was used to retrieve transcription variants for syllables. In this study, a set of 37 phone labels was used. A list of plausible transcription variants for all the syllables in the manually verified subset was obtained by aligning the manual phonetic transcriptions of word tokens with their canonical counterparts using 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 distance for

deletions and insertions (Elffers et al. 2005). To ensure syllable-level alignment, the syllable boundaries that were available for the canonical transcriptions were utilised in the alignment process.

When building the multi-path mixed-model recogniser, we concentrated our modelling efforts on a set of 94 most frequent syllables in the manually verified subset (Hämäläinen et al. 2005). Using the transcription variants retrieved for these target syllables and canonical transcriptions for the rest of the syllables, a forced alignment of the training data was performed with 8-Gaussian triphones to determine which 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, the forced alignment process was fed with all the four transcription variants observed in the manually verified subset (corresponding to the following sequences of triphones: “#-n+a n-a+#”, “#-n+@ n-@+#”, “#-n+A n-A+#” and “#-N+a N-a+#”) and was, therefore, able to ascertain which variants acoustically best matched the relevant stretches of the speech signal. This labelling process was, of course, applied to all instances of the syllable /na/ in the training data. The second syllable /drYk/ did not belong to the set of target syllables and was only allowed to be labelled as the canonical sequence of triphones “#-d+r d-r+Y r-Y+k Y-k+#”. To ensure that the complete training corpus was handled in the same manner, the forced alignment procedure was also applied to the manually transcribed part of the data. Comparing the proportions of the different transcription variants of the target syllables in the manually verified and the automatically transcribed data confirmed the reliability of the automatic transcription procedure.

3 Selection of major, distinct transcription variants

If the amount of data available for the re-estimation of the acoustic observation densities of single-path syllable models is already an issue, the situation is only more difficult for multi-path models. Therefore, the optimal initialisation of the parallel paths is of the 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. These representative sequences of triphones were obtained by means of the so-called ‘major, distinct transcription variants’. The selection of MDVs was guided by two principles. First, we wanted to keep the canonical variant as one of the MDVs, unless a different transcription variant was more frequent in the training corpus. Second, we had a preference for MDVs containing fewer symbols than the canonical variant. This preference stemmed from an analysis of syllable durations obtained by using a single-path mixed-model recogniser to perform a forced alignment of the CGN data used in (Hämäläinen et al. 2005). Figure 3 shows the duration distribution histogram for the 40 most common CV syllables. Since the HMM topologies consisted of three states per each phoneme in the canonical syllable transcriptions, and each state had to be aligned with at least one acoustic frame corresponding to a 10 ms interval in the signal, the shortest possible duration for the CV syllables was 60 ms. The large number of syllables with durations corresponding to 6 or 7 frames suggests that the standard three states per underlying phoneme topology may have been too long. Furthermore, although multi-path models derived using trajectory clustering resulted in a significant improvement in recognition performance in (Han et al. 2006), we concluded that the equal length of the parallel paths was hindering the performance gain.

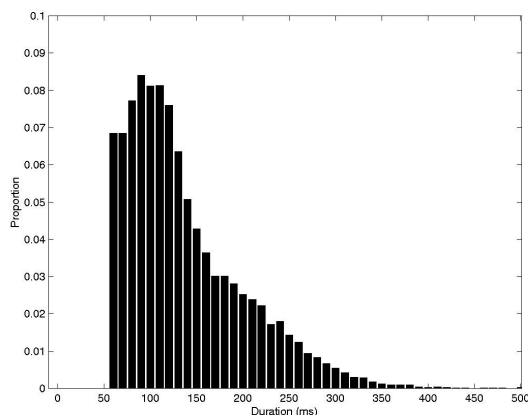


Figure 3. Duration distribution for 40 most common CV syllables established through a forced alignment with single-path syllable models containing three states per each underlying phoneme in the canonical transcription

The set of 94 target syllables used in this study covered 57% of all the syllable tokens in the training data, the least frequent of them, /ti/, occurring 850 times and the most frequent, /d@/, occurring 35,000 times. The syllable /d@/ is highly frequent, because it occurs as the definite article “de” (feminine, masculine, singular and plural), and as part of many polysyllabic words. The target syllables had an average of 8.7 transcription variants per syllable. The actual number of variants differed considerably: the syllable /mu/ only had one variant, whereas the syllable /hEt/ had 27 variants. The syllable /mu/ corresponds to the monosyllabic word “moe” (‘tired’, or exceptionally ‘mother’), and can also occur as part of several polysyllabic words (in some of which it is the result of the re-syllabification of morphemes that contain one or more coda consonants). In any case, /mu/ always carries word stress and does not, therefore, display a lot of variation in its pronunciation. The syllable /hEt/ appears as the definite article “het” (neuter, singular), and as part of a small number of polysyllabic words. The large number of transcription variants is due to the many ways in which the unstressed article can engage in clitisation processes. 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. This could be estimated to 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 three parallel paths for all syllables that had at least three transcription variants. For the syllables that had more than three transcription variants, we used the concept of MDVs to select the variants that best represented three truly different pronunciation variants. Three parallel paths per syllable appeared the best compromise between too little training data and too small a distance between the triphone sequences used to initialise the paths.

We devised the following steps for selecting the optimal combination of three MDVs for each target syllable:

- (1) Compute articulatory distances between all transcription variant pairs for the target syllable. To compute the distances, we used the same algorithm as we did when aligning the manual and canonical transcriptions to find the transcription variants for the syllables (see Section 2).
- (2) Weighing the transcription variants by their frequency of occurrence, compile a ranked list of transcription variant combinations the constituent variants of which are articulatorily maximally different from each other. For instance, the combination /hAt/-/hat/-/At/ ranked the highest for the syllable /hAt/, whereas the combination /Ad/-/jAt/-/jA/ ranked the lowest, mainly because of the low frequencies of the variants in question.
- (3) Post-process the list produced in Step 2 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 transcription variant combination that contains at least one variant with at least one symbol less than the canonical. When none of the variant combinations satisfies the length criterion, select the highest-ranking variant combination. The variants included in the selected transcription variant combination are the MDVs used in the initialisation of the HMM paths.

4 Experimental set-up

4.1 Feature extraction

Feature extraction of the speech material was carried out at a frame rate of 10 ms using a 25-ms Hamming window and a pre-emphasis factor of 0.97. 12 Mel Frequency Cepstral Coefficients (MFCCs) and log-energy with corresponding first and second order time derivatives were calculated, for a total of 39 features. Channel normalisation was applied using cepstral mean normalisation over complete recordings, which were then chunked to sentence-length entities for the purpose of further processing.

4.2 Lexicon and language model

In order to study possible improvements due to changes in acoustic modelling only, without the risk of language modelling issues masking the effects, out-of-vocabulary words were not allowed in the task. In effect, the recognition lexicon and word-level bigram network were built using all orthographic words in the training and test sets. The recognition lexicon consisted of a single pronunciation for each word. In the case of the triphone recogniser, the pronunciation for each word consisted of a string of canonical phones from the CGN lexicon. In the case of the mixed-model recognisers, it consisted

of a) syllable units b) canonical phones, or c) a combination of a) and b). The vocabulary comprised about 29,700 words, and the test set perplexity, computed on a per-sentence basis, was 92.

4.3 Acoustic modelling

Speech recognition experiments were designed to test whether a mixed-model recogniser with multi-path models for the target syllables would outperform a conventional triphone recogniser. This section details the acoustic model training procedures used in building these recognisers.

A standard procedure with decision tree state tying was used to train the triphone recogniser (Young et al. 2002). Initial 32-Gaussian monophones were trained using linear segmentation of canonical transcriptions within automatically generated word segmentations. The monophones were used to perform a forced alignment of the training data; triphones were then bootstrapped using the resulting phone segmentations. Triphone recognisers with up to 64 Gaussian mixtures per state were trained and tested.

The steps described in Section 3 were followed to select the optimal combinations of three MDVs for each of the 94 target syllables. The parallel paths of the context-free multi-path models for the target syllables were initialised by picking the initial state parameters from the triphones corresponding to these MDV combinations (Sethy & Narayanan 2003, Hämäläinen et al. 2005). Before applying the Baum-Welch algorithm to capture within-syllable co-articulation effects, the initialised paths were combined into multi-path 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, but left the training entirely to the Baum-Welch algorithm. Triphones were used to represent the syllables that did not belong to the set of 94 target syllables. The mix of syllable and triphone models underwent four passes of Baum-Welch re-estimation. Multi-path mixed-model recognisers with up to 16 Gaussian mixtures per state were trained and tested. The parallel paths of the syllable models were initialised using triphone models with the same number of Gaussian mixtures per state as in the final mixed-model recogniser.

5 Results and discussion

An analysis of the MDV combinations used in building the multi-path models for the target syllables showed that the canonical transcription was always included. In fact, the canonical transcription was the most frequent transcription for all of the 94 syllables. 85% of the bi- and tri-phonemic target syllables (81% of all the target syllables) had one or two MDVs with fewer phones than the canonical. Somewhat surprisingly, 39% of all the target syllables had one MDV with more phones than the canonical. The long paths could be attributed to phenomena such as syllable-initial glide insertion and syllable-initial epenthetic schwa-insertion. Glide insertion at word boundaries is a very frequent phonological process in Dutch; it may take place when a word that ends in a vowel is followed by a word that starts with a vowel. Epenthesis is a manifestation of articulatory complexity. For example, the word “werken” (‘to work’) is often pronounced /wEr@k@/, while the canonical pronunciation is /wErk@/. In these cases, the syllable alignment procedure (see Section 2) assigned the epenthetic schwa to the second syllable, despite the fact that this resulted in a syllable structure that violates the conventional phonotactic restrictions of the Dutch language. We could have, of course, decided to consider the transcription variant with the epenthetic schwa as a tri-syllabic

word (/wE-r@-k@/), but this would also have been problematic. First, the syllable /wE/ would also have violated phonotactic constraints: short vowels such as /E/ cannot normally occur in a syllable-final position. Second, it is unclear which canonical syllable the syllable /r@/ should have been considered a variant of. All in all, we came to the conclusion that accepting bi-vocalic syllables such as /@k@/ was less problematic.

In Table 2, the speech recognition results and the recogniser complexities measured in terms of the total number of Gaussians are presented for the most relevant recognisers: the 16-Gaussian triphone recogniser, the 32-Gaussian triphone recogniser (best performing triphones), and the 16-Gaussian multi-path mixed-model recogniser (best performing multi-path mixed-model recogniser). The complexity of the syllable models was estimated with the same tying ratio as that used in building the triphone models. Even with the loss of context information at some syllable boundaries, the 16-Gaussian multi-path mixed-model recogniser outperformed both the 16-Gaussian triphone recogniser and the much more complex 32-Gaussian triphone recogniser. In the latter case, the reduction in WER was not significant, but the result does suggest that using multi-path models for frequent syllables is a more effective way of increasing modelling power than just increasing the number of Gaussians per state in triphones. In effect, the multi-path syllable models add prior knowledge about structure, whereas the triphone models only add detail in terms of straightforward statistics of an unstructured population.

Table 2. Word error rates with a 95% confidence interval, and the total number of Gaussians in the recognisers

Recogniser	WER (%)	# Gaussians
16-Gaussian triphone	10.3 ± 0.4	24,560
32-Gaussian triphone	10.1 ± 0.4	49,120
16-Gaussian multi-path mixed-model	9.9 ± 0.4	28,224

We emphasise the contrast between phonetic knowledge and data, because the use of knowledge in computational models is usually far from straightforward. In the early days of automatic speech recognition based on HMMs, Kai-Fu Lee (1989) proposed quite an elaborate topology for phone models. This topology was inspired by phonetic knowledge about assimilation and reduction processes, and comprised three parallel paths. The longest path consisted of three states with self loops, whereas the two shorter paths were aimed at modelling reduced pronunciations of phonemes. 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, when dealing with longer-span speech segments with an inherently larger amount of variation, using more than one parallel path is an obvious proposition. The problem of bootstrapping these more intricate models is the price we have to pay for more modelling power. Practice has shown that, in case of many computational models, the estimation of the optimal model topology from observed data is more difficult than the estimation of the associated model parameters. Therefore, we decided to utilise phonetic knowledge in determining the syllable model topologies and initialising the model parameters, and to fine-tune the parameters within these predefined topologies using data.

The improved recognition performance suggests that important variation is indeed accounted for by the parallel paths. However, we acknowledge that our approach is oversimplified. We chose to train up to three parallel paths per target syllable; in practice, the optimal combination of three MDVs was used when constructing the parallel paths for 90% of the syllables, whereas all transcription variants were used for the remaining 10% that only had one to three transcription variants. This may not have always been the best choice, as more paths might have been needed for some syllables and fewer might have sufficed for others – in particular, those that might have had up to three somewhat similar transcription variants. To estimate the number of potentially redundant paths, we used the multi-path mixed-model recogniser to perform a forced alignment of the training data, and checked the proportion of syllable tokens captured by each parallel path. We considered a path redundant if fewer than 5% of all syllable tokens were assigned to it. 31% of the syllables contained one redundant path, and one syllable, /wAt/, had two redundant paths. As so many syllables had potentially redundant paths, we decided to train another 16-Gaussian mixed-model recogniser in which the paths capturing fewer than 5% of all syllable tokens were removed. The resulting recogniser reached a WER of 9.8% – a recognition result that is not significantly better than that of the original 16-Gaussian multi-path mixed-model recogniser, but does show a trend towards better performance. Therefore, refining our approach when it comes to the optimal number and type of MDVs used in the initialisation of the parallel paths seems like a worthwhile direction for future research. As an obvious starting point for this work, we will carry out an in-depth analysis of the variation captured by the parallel paths of the current multi-path models, and compare the recognition errors made by the 32-Gaussian triphone recogniser on the one hand and the 16-Gaussian multi-path mixed-model recogniser on the other hand.

6 Conclusions

In this paper, we constructed multi-path models for frequent syllables. The approach we used combined knowledge-based and data-driven techniques by using phonetic knowledge to initialise the parallel paths of the syllable models, and by subsequently leaving the further training entirely to the Baum-Welch algorithm. In essence, the approach provides a solution for initialising parallel paths of different lengths. Experiments with a mixed-model recogniser with 16 Gaussians per state suggested that multi-path syllable models capture important effects of pronunciation variation. Even though the reduction in WER was not significant, the multi-path mixed-model recogniser outperformed a much more complex 32-Gaussian triphone recogniser. This suggests that, beyond a certain number of Gaussians per state, adapting model topologies is a more effective way of increasing modelling power than just increasing the number of Gaussians per state in triphones.

7 Acknowledgements

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

References

- Elffers, Bram, Van Bael, Christophe and Strik, Helmer (2005). *ADAPT: Algorithm for dynamic alignment of phonetic transcriptions*. Internal report, Department of Language and Speech, Radboud University Nijmegen, The Netherlands.
- Ganapathiraju, Aravind, Hamaker, Jonathan, Ordowski, Mark, Doddington, George and Picone, Joseph (2001). Syllable-based large vocabulary continuous speech recognition. *IEEE Transactions on Speech and Audio Processing*, Vol. 9(4), 358–366.
- Greenberg, Steven (1999). Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. *Speech Communication* 29, 159–176.
- Han, Yan, Hämäläinen, Annika and Boves, Lou (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* (Vol. I, pp. 1169–1172). Toulouse, France.
- Hämäläinen, Annika, Boves, Lou and de Veth, Johan (2005). Syllable-length acoustic units in large-vocabulary continuous speech recognition. In G. Kokkinakis, N. Fakotakis, E. Dermatas and R. Potapova (Eds.), *Proceedings of the 10th International Conference Speech and Computer* (pp. 499–502). Patras, Greece.
- Jones, Rhys James, Downey, Simon and Mason, John S. (1997). Continuous speech recognition using syllables. In *Proceedings of the 5th European Conference on Speech Communication and Technology* (Vol. 3, pp. 1171–1174). Rhodes, Greece.
- Jurafsky, Daniel, Ward, Wayne, Jianping, Zhang, Herold, Keith, Xiuyang, Yu and Sen Zhang (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* (Vol. I, pp. 577–580). Salt Lake City, UT, USA.
- Kessens, Judith (2002). *On automatic transcription and modeling of Dutch pronunciation variation for automatic speech recognition*. PhD thesis, University of Nijmegen, The Netherlands.
- Lee, Kai-Fu (1989). *Automatic Speech Recognition – The Development of the SPHINX-System*. Kluwer Academic Publishers, Boston.
- Messina, Ronaldo and Jouviet, Denis (2004). Context-dependent long units for speech recognition. In *Proceedings of the International Conference on Spoken Language Processing*, (pp. 645–648). Jeju Island, Korea.
- Oostdijk, Nelleke, Goedertier, Wim, Van Eynde, Frank, Boves, Lou, Martens, Jean-Pierre, Moortgat, Michael and Baayen, Harald (2002). Experiences from the Spoken Dutch Corpus project. In M. González Rodríguez and C. Paz Suarez Araujo (Eds.), *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (Vol. 1, pp. 340–347). Las Palmas de Gran Canaria, Spain.
- Ostendorf, Mari (1999). Moving beyond the ‘beads-on-a-string’ model of speech. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*. Keystone, CO, USA.
- Sethy, Abhinav and Narayanan, Shrikanth (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*, (Vol. 1, pp. 772–776). Hong Kong.
- Sethy, Abhinav, Ramabhadran, Bhuvana and Narayanan, Shrikanth (2003). Improvements in ASR for the MALACH project using syllable-centric models. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop*. St. Thomas, US Virgin Islands.
- Wester, Mirjam (2002). *Pronunciation variation modeling for Dutch automatic speech recognition*. PhD thesis, University of Nijmegen, The Netherlands.
- Young, Steve, Evermann, Gunnar, Hain, Thomas, Kershaw, Dan, Moore, Gareth, Odell, Julian, Ollason, Dave, Povey, Dan, Valtchev, Valtcho and Woodland, Phil (2002). *The HTK book (for HTK version 3.2.1)*. Cambridge University, UK.