Radboud Repository

Radboud University Nijmegen

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is an author's version which may differ from the publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/75065

Please be advised that this information was generated on 2017-12-06 and may be subject to change.

PRONUNCIATION VARIATION MODELLING IN A MODEL OF HUMAN WORD RECOGNITION

Odette Scharenborg and Lou Boves

A^{2}RT, Department of Language and Speech
University of Nijmegen, The Netherlands
{O.Scharenborg,L.Boves}@let.kun.nl

ABSTRACT

Due to pronunciation variation, many insertions and deletions of phones occur in spontaneous speech. The psycholinguistic model of human speech recognition Shortlist is not well able to deal with phone insertions and deletions and is therefore not well suited for dealing with real-life input. The research presented in this paper explains how Shortlist can benefit from pronunciation variation modelling in dealing with real-life input.

Pronunciation variation was modelled by including variants into the lexicon of Shortlist. A series of experiments was carried out to find the optimal acoustic model set for transcribing the training material that was used as basis for the generation of the variants.

The Shortlist experiments clearly showed that Shortlist benefits from pronunciation variation modelling. However, the performance of Shortlist stays far behind the performance of other, more conventional speech recognisers.

1. INTRODUCTION

In spontaneous speech, many insertions, deletions, and substitutions of phones occur (cf. [1]). It is common knowledge that the performance of an automatic speech recogniser (ASR) degrades if this pronunciation variation is not properly accounted for. But not only automatic speech recognisers have to deal with pronunciation variation. In this paper, we will explain how we expect the psycholinguistic model of human speech recognition Shortlist [2] to benefit from pronunciation variation modelling in order to deal with real-life input.

Shortlist is a model of human speech recognition that is able to account for a wide range of results from psycholinguistic experiments related to word recognition. In its present implementation, the recognition process in Shortlist is a two-stage process. In the first stage, an exhaustive lexical search yields a shortlist of (typically) maximally 30 word candidates that are roughly consistent with the phonemic input. This search is repeated from scratch for each phoneme in the input when an utterance is processed in a strict left-to-right manner. The activation of the words in the shortlist is determined by their degree of fit with the phonemic input. If a phoneme in a word matches the input, the word activation is increased; for each mismatching phoneme the word activation is reduced. In the second stage, the activated words in the shortlists compete with each other by means of their initial activation and the inhibition of other words in the list. During the processing of the phonemic input, the activation as a function of time of each word in the shortlist can be observed. These activation functions are used for the simulation and explanation of human speech recognition. In this research however, we do not deal with the simulation process of Shortlist; we use Shortlist solely as a speech recogniser. To this end, we only use the output of Shortlist after all phonemes are processed. We take the word with the highest activation as the recognised word. Note that if a word is not included in the shortlist generated in the first stage, it can never be recognised as a result of the competition process.

Despite its explanatory power, the current version of Shortlist suffers from two unrealistic simplifications. The first is that the input should consist of a single string of phoneme symbols instead of an acoustic signal. Not only Shortlist suffers from this simplification, virtually all psycholinguistic models start from a discrete segmental representation instead of the acoustic signal; and therefore, cover only parts of the human speech recognition process. To enable Shortlist to deal with an acoustic signal as input and thus cover a larger part of the human speech recognition process, we developed an automatic phone recogniser (APR) that converts the acoustic signal into a discrete segmental representation. This APR is used as an acoustic front-end for Shortlist. The second simplification in Shortlist is that the current implementation makes it difficult to deal with insertions and deletions. So, in the present implementation there is a large premium for inputs with a number of phones identical to the number of phones in the lexical representation of that word in the internal lexicon of Shortlist. Consequently, the current version of Shortlist has difficulty in dealing with input consisting of 'accurate' phonetic transcriptions of the acoustic signal, even if these are made by expert phoneticians, because transcriptions that closely represent acoustic signals of 'normal' speech tend to contain insertion, deletion, and substitution 'errors' compared to the canonical phonemic representations of the words in the internal lexicon of Shortlist.

The discrepancy between accurate phonetic transcriptions on the one hand, and the preference for unique canonical representations of words in linguistic theory on the other, forms a fundamental problem that ASR as well as (Psycho-) linguistics need to deal with. 'Normal' listening is listening for content/meaning, and for this aim a phonetic representation that is closer to canonical lexical representations than to the idiosyncratic acoustic signal might be preferred. At the same time it is clear that every realistic lexicon should be able to account for some degree of pronunciation variation. The final goal of the research reported in this paper is to improve our

understanding of the optimal trade-off between phonetic accuracy and the requirements of symbolic processing.

Shortlist's problem concerning insertions and deletions can be approached from two directions. One may try to find the optimal balance between generating an input phoneme string that is close to the signal and generating an input phoneme string that contains the least possible number of insertions and deletions. Alternatively, Shortlist could be adapted so that it is capable of dealing with accurate phonetic transcriptions. In this paper, we investigate the second approach by modelling pronunciation variation in the lexicon of Shortlist. This paper describes the generation of the pronunciation variants and the effect of adding pronunciation variants to the lexicon of Shortlist on the performance of Shortlist as a recogniser.

The research presented in [3] has shown that to obtain the best automatic phonetic transcriptions, the APR used to create the transcriptions should be optimised on this specific task. In our study, this optimisation takes place in two steps. First, before the pronunciation variants can be generated, a proper set of acoustic models for the automatic phonetic transcription of the training material should be found. Section 3 describes the three methods we investigate to find the best possible acoustic models for the APR on the task. The transcriptions generated by the best performing model set are then used to generate the pronunciation variants in a procedure that is described in Section 4. The pronunciation variants are then used to improve the transcriptions of the training and test material (cf. Section 5). With these improved transcriptions the final acoustic models are trained that are used to generate the final phonetic transcription of the test corpus. By optimising the transcriptions of the training material, we try to derive cleaner acoustic models that describe the acoustic signal more closely. This is the second step in the optimisation of the acoustic models. The pronunciation variants derived in the first step are also used to extend the lexicon of Shortlist, to make it more suitable for the processing of 'real speech'. The results of the experiments with the Shortlist lexicon containing pronunciation variants are presented in Section 6.

In Section 2, the corpora, lexicons, and language models that are used throughout all experiments will be described. Finally, Section 7 will present our conclusions.

2. MATERIAL

2.1. Corpora

For training and testing the APR, data from the Dutch Directory Assistance Corpus (DDAC) were used [4]. The material to train the acoustic models of the APR comprises 24,559 utterances. The total duration of the speech frames is 4 hours and 40 minutes. Each utterance consists of one Dutch city name or 'ik weet het niet' ('I don't know') pronounced in isolation, although audible hesitations like 'eh' were allowed. All utterances were recorded through the telephone. The reason for this strict selection is that we wanted to obtain the cleanest possible phone models with a training corpus for which initially only canonical transcriptions of the words were available. Including longer utterances in the training corpus would probably yield more contaminated models, since longer utterances contain more variation, resulting in a larger mismatch between the acoustic signal and the phonetic transcription.

A second set of utterances used for training the APR consists of 42,101 short utterances of the Dutch Polyphone database (SHUTT) [5]. The total duration of the speech is 9 hours and 32 minutes. The Polyphone short utterances consist of no more than three words per utterance and contain various types of items, e.g. digits, ZIP codes, times, application words and city names recorded through the telephone. The speaking styles are read and extemporaneous speech.

The independent test set consists of 10,643 utterances from the DDAC corpus with a total number of 11,890 words. These utterances may also contain disfluencies and connected speech responses like 'haarlem noordholland' (i.e., a city name plus the name of a province).

2.2. Lexicons and language models

The DDAC training lexicon consists of 2,392 entries: 2,381 city names and alternative expressions for some city names, as well as 8 entries for (alternative expressions of) 'ik weet het niet' ('I don't know'). Also entries for hesitation sounds ('eh' and 'ehm') and noise were present. For each entry in the lexicon a unique canonical phonemic representation was available.

The SHUTT training lexicon consists of 10,242 distinct words including 22 spelled letters. Also an entry for noise was present. For each entry in the lexicon a unique canonical phonemic representation was available.

The standard lexicon of Shortlist consists of all entries of the DDAC training lexicon extended with the names of the Dutch provinces and the words 'de', 'plaatsnaam', and 'is' ('the', 'city name', 'is', respectively). This makes a total of 2,404 entries.

The uni- and bi-gram language models (LMs) used in the APR experiments were trained on the phonetic transcriptions of the DDAC training utterances. So, the 'words' in these LMs were in fact phones.

3. OPTIMISING THE ACOUSTIC MODELS

The optimisation of the acoustic models is a two-step procedure. In the first step, described in this section, we determine the number of Gaussians per state (G/s) and the units to be modelled. Furthermore, we determine the training corpus that should be used to train the phone models. In the second step, we try to further improve the acoustic models by modelling pronunciation variation at the level of the transcriptions of the training corpus (Section 5). The reason for this two-step approach is the need to limit the number of conditions in the experiments to a reasonable maximum.

The APR was optimised using the following procedure. The test set was split into 4 subsets, after which the optimal values of the language model factor (LMF) and the insertion penalty (IP) were determined on each of the 4 subsets and tested on the remaining 3 subsets on which they were not optimised. IP determines the trade-off between insertions and deletions; LMF determines the weighting of the acoustic and the language model contribution to the total probability of a phoneme. The values of LMF and IP obtained from the tuning set that produced the lowest Phone Error Rate (PER) for the 4 test sets were regarded as optimal. The PER is defined as:

$$PER = \frac{(S + I + D)}{N} \quad (1)$$

with *S, I, D,* and *N* the number of substitutions, insertions and deletions, and total number of phones in the canonical phonetic transcriptions of the test corpus, respectively.

Concerning the optimal number of G/s to be modelled, all model sets were trained with maximally 32, 64, and 128 G/s to examine the effect of varying the maximum number of G/s.

3.1. Filled pauses

In our search to what units to model, we concentrated on the schwa (/ə/). The schwa pronounced in a word and the schwa pronounced in a filled pause ('eh' or 'ehm') can be treated as two instantiations of the same phone and can accordingly both be mapped on the same model of the /ə/. This is referred to as '-FP' in the remainder of this paper. Secondly, it is also possible to treat them as separate sounds. In this case, we trained a separate model for filled pauses in order to minimise the contamination of the phone model of the /ə/ ('+FP'). Since there were too few instantiations of 'ehm' to train two separate models, only one model was trained for both 'eh' and 'ehm'.

In the -FP condition, the lexicon in the APR only contains the 36 trained phones and an additional model for noise and silence. In the +FP condition, the new model for filled pauses was also added to the APR recognition lexicon and the LM was retrained accordingly.

3.2. Extra training material

The third parameter we investigated is the material for training the acoustic models. The amount of DDAC training material is not very large. By using more training material it is assumed that the performance of the APR will improve. To increase the amount of training data, the training corpus is expanded with the SHUTT corpus.

In the remainder of this paper, the acoustic model sets only trained on the DDAC material will be indicated as 'DDAC'. The acoustic model sets also trained on the Polyphone short utterances will be indicated as '+SHUTT'.

For the APR experiments, the Phicos recognition system was used [6]. 36 context independent phone models, 1 silence model, and 1 noise model were trained on the two training corpora: DDAC and +SHUTT. Of course, in the case of +FP, one extra model for filled pauses was trained. Each phone model and the noise (and filled pause) model consist of 3 pairs of 2 identical states, one of which can be skipped. The silence model consists of 1 state.

3.3. Results and discussion

Figure 1 shows the results of the recognition experiments with the various model sets. The PERs in Figure 1 are the average PERs calculated after the optimisation procedure; the PER obtained on the subset on which it was optimised was not taken into account.

Figure 1 clearly shows that adding training material significantly reduces the performance. A possible explanation for this decrease is that the utterances in the added material are not optimally matched to the requirements of the test set. It is not yet clear whether the mismatch is primarily acoustic (i.e. due to other types of background noise, handsets, transmission lines) or rather 'phonetic' in nature (for example due to the fact that the large majority of the Polyphone short utterances are read, while all DDAC utterances are extemporaneous).

Furthermore, it is clear in Figure 1 that the +FP model sets outperform the -FP model sets. This confirms our intuition that schwas in filled pauses and words are acoustically different, and thus that the /ə/ model is less contaminated in the +FP condition.

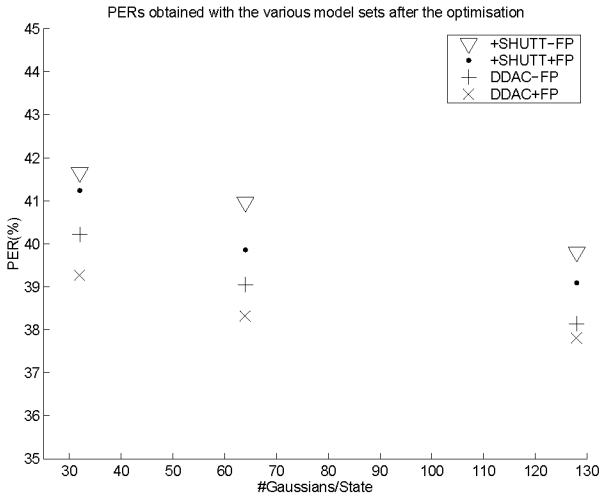

Figure 1: The PER plotted against the (maximum) number of G/s for the four types of model sets: DDAC+FP, DDAC-FP, +SHUTT+FP, and +SHUTT-FP.

Finally, increasing the number of maximally trained G/s significantly improves the recognition performance. The best performing model set is the DDAC+FP model set with maximally 128 G/s (DDAC+FP:128). However, the difference with DDAC-FP:128 is not significant. The fact that the difference in performance between DDAC+FP:128 and DDAC-FP:128 is not significant indicates that the extra Gaussians in the /ə/ model will model the filled pauses. Since the difference in PER between DDAC+FP:*xx* and DDAC-FP:*xx* is significant for all models with fewer G/s, the transcription produced by the DDAC+FP:128 model set will be used for the generation of the pronunciation variants.

4. GENERATING THE PRONUNCIATION VARIANTS

There are several ways to obtain pronunciation variants. These can roughly be divided in knowledge-based, e.g. using linguistic rules [7], and data-derived, e.g. using decision trees described for example in [8]. See for an overview of the various methods [9]. In this research, we used the data-derived decision tree (d-tree) approach, mainly because we were not sure that we could derive linguistic rules that have sufficient coverage -without too much over-generation- for the extemporaneous speech in the DDAC test corpus.

In the case of data-derived pronunciation variation modelling, an information source is needed. We used the training utterances phonetically transcribed by the APR using the DDAC+FP:128 model set. These sequences of phones output by the APR can be regarded as pronunciation variants.

Potential over- and under-generation of variants is also a problem with data-derived approaches [10]. In order to generalise from observed variants and at the same time keep the number of pronunciation variants manageable, the d-tree approach is used to smooth the output of the phone recognition before creating the lexicon.

The d-trees were built using the *Weka* d-tree tools [11], also described in [12]. In building the d-trees, only the left and right neighbours of the phones were used as input to the d-tree algorithm. For each phone a d-tree was built. d-trees predict pronunciation variants on the basis of an alignment between the reference phonetic transcription of the training utterance and the transcription obtained with the phone recognition. Before building the d-trees, the data were pruned by removing very low frequency contexts. If the combination of the phone in focus with its left of right neighbour occurred less than five times in the data, this combination was not taken into account when building the d-tree.

Next, using the distributions in the d-trees, a Finite State Grammar (FSG) was built for each training utterance. A second type of pruning occurred during the creation of these FSGs: transitions with a probability lower than 0.01 (default value) were not allowed. An FSG can be considered to represent all potential pronunciation variants for that utterance. Furthermore, the FSGs attach a probability to all pronunciation variants they produce, which is an estimate of the prior probability of a variant on the basis of the training material. These priors are later used to decide which pronunciation variants to include in the lexicon of Shortlist. The pronunciation variants that are created consist (mainly) of fewer phones than the canonical transcriptions.

5. IMPROVING THE TRANSCRIPTIONS AND THE ACOUSTIC MODELS

Previous research on automatic continuous speech recognition has shown that the best results are obtained when pronunciation variation is modelled on three levels, viz. at the level of the lexicon, the language model and the acoustic models (cf. [7]). For the phonetic transcription task, we use an automatic phone recogniser instead of a continuous speech recogniser for words; therefore, our lexicon consists of phones instead of words. This makes it harder – though not impossible – to model pronunciation variation at the level of the lexicon. Therefore, we only model pronunciation variation at the level of the acoustic models. To this end, the transcriptions of the training material were improved (see Section 5.1), after which new acoustic models were trained (see Section 5.2). Furthermore, we retrained the LM on the 'improved' transcriptions of the training material to improve the model of the phonotactic constraints found in the training material.

5.1. Improving the transcriptions

The transcriptions of the training material were improved using the following procedure. A lexicon was created consisting of the most likely pronunciation variants. In order to select the most likely pronunciation variants, all variants produced by the FSGs were ranked according to their prior probability. The pronunciation variants with a prior probability above the pre-set threshold of 0.063 were then added to the lexicon. This resulted in an average of 77 variants per word. Subsequently, an ASR trained on the canonical transcription of the training corpus, running in forced recognition mode, was used to find the most likely variant for each word in the training corpus from the lexicon with the new pronunciation variants. In the last step, the phonemic transcriptions of the corpus were updated by replacing the canonical transcriptions by the most likely variant as found by the forced recognition.

In order to obtain transcriptions of the test material that more closely match the acoustic signal, the same procedure was followed for the test material.

5.2. Retraining of the acoustic models

These 'improved' transcriptions of the training material were then used to train new, and presumably cleaner acoustic models. The new acoustic models were trained on the 'improved' transcriptions starting from a linear segmentation (NEW).

The maximum number of Gaussians trained per state was 128. Also, a separate model for filled pauses was trained.

5.3. Results and discussion

Figure 2 shows the PERs of the model sets described in Section 3 measured against the improved transcriptions of the test material. The number of phones in the test material was 83,614. As can be observed, the PER for the baseline system DDAC+FP:128 drops from 37.81% when measured against the canonical transcriptions (see Figure 1) to 34.73% when measured against the improved transcriptions. This decrease in PER is observed for all model sets. Whether this gain is caused by a tuning of the transcriptions to the idiosyncrasies of the recogniser or by the fact that the new transcriptions are really closer to the acoustic signal – and thus closer to what a phonetician would have transcribed – is as yet unclear. However, inspection of the pronunciation variants that were used to improve the transcription of the train and test corpora suggests that all added variants are reasonable. Therefore, we are confident that the updated transcriptions are indeed closer to acoustic signals than the canonical representations that yield a PER of about 38%.

To examine whether the improvement of the transcriptions of the training material resulted in improved acoustic models, a recognition experiment with the new trained model set was conducted on the test material. The lexicon used in the recognition experiment contained all 36 phones, and additional models for silence, noise, and filled pauses; the LM used was trained on the improved transcriptions of the training material. The APR was optimised using the procedure described in Section 3.

Table 1 shows the PER of the new model set. The PER was measured against the improved transcriptions of the test material. As a reference, the PER obtained for the baseline system DDAC+FP:128 on the improved transcriptions is also given. Table 1 shows that the performance of the model set trained on the improved transcriptions is indeed slightly better than the performance of the baseline, but the gain is quite small. This is in accordance with results found in other experiments. A possible reason for this only small gain is that the models are too much tuned to the training data.

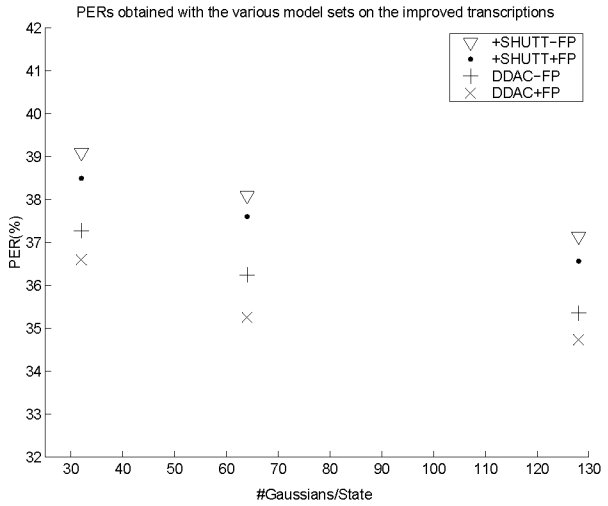

Figure 2: The PER obtained on the optimised transcriptions of the test set plotted against the (maximum) number of G/s for the four types of model sets: DDAC+FP, DDAC-FP, +SHUTT+FP, and +SHUTT-FP.

Model set	PER (%)
DDAC+FP:128	34.73
NEW	34.42

Table 1: PERs for baseline and NEW on the test set.

6. SHORTLIST EXPERIMENTS

In order to improve Shortlist's ability to deal with real-life speech, its lexicon should be adapted to include pronunciation variants next to the canonical transcriptions. The pronunciation variants to be added to Shortlist' s lexicon were selected from the lexicon used for improving the transcriptions of the training material. Several independent ASR studies (cf. [10],[13]) have shown that it is difficult to determine how many and which pronunciation variants to include in the lexicon in order to obtain the best result. One of the most important findings of these studies is that including a high number of variants will cause a deterioration of the recognition performance unless the prior probabilities of the variants can be used in the recognition process. If no prior probabilities are known, the research in e.g. [13] shows that an average of 2.5 variants per word gives the best performance on a continuous speech recognition task. Although the task we present to Shortlist is more like an isolated word recognition task than a continuous speech recognition task, we followed this finding, since it is impossible to include prior probabilities of the variants in Shortlist. A threshold was determined on the prior probabilities such that adding all variants with a higher probability to the Shortlist lexicon resulted in an average of 2.5 variants per word.

As already pointed out in the introduction, Shortlist is a simulator of human speech recognition. In order to make correct simulations, it is necessary that Shortlist recognises the correct words. In this experiment, we present Shortlist input containing phoneme sequences that do not match exactly with phonemic representations in the lexicon, but that are assumed to be a close symbolic representation of acoustic speech signals. It is interesting to know to what extent Shortlist is able to recognise the 'scrambled' input strings correctly.

As input, Shortlist was given the phonetic transcriptions of the test utterances obtained with the phone recognition with the NEW model set. Two experiments were carried out. In the first experiment, the original lexicon (no pronunciation variants) was used; in the second experiment the lexicon expanded with pronunciation variants was used. The performance was measured in terms of percentage of utterances for which the correct word did not have the highest activation, which means that the correct word was not recognised: the Word Error Rate (WER).

6.1. Results and discussion

With the original lexicon the WER obtained with Shortlist was 64.45%. The WER decreased to 48.2% with the expanded lexicon. Thus, adding pronunciation variants to the lexicon of Shortlist greatly improves its performance as a speech recogniser. However, the performance is still far from what a conventional automatic speech recogniser would obtain on the same test set. For instance, WERs in the ASR systems described in [14] range from 10 to 15%.

The question that immediately arises is why Shortlist is such a bad recogniser. In order to answer this question, we have to look at the current implementation of Shortlist. As already pointed out, if a word is not included in the shortlist generated in the first stage of the model, it can never be recognised. An analysis of the output of Shortlist reveals that there are few cases where the correct word is in the shortlist, but that a competitor receives a higher final activation. Thus, the problem appears to be in the initial selection phase. Adding pronunciation variants with (mainly) fewer phonemes improves the selection, but it is not sufficient.

The next question that arises is whether the performance of the APR is highly inaccurate, or whether Shortlist is not able to deal with accurate transcriptions. The current implementation of the first stage, the search, in Shortlist does not use sophisticated weighting of substitutions, insertions and deletions. Consequently, Shortlist has difficulties making a correct match between the phone string with 'errors' and the phonetic representation of the words in the internal lexicon. Furthermore, it is well-known that there is a considerable variety in the transcriptions created by human transcribers. In [15], it was shown that automatically generated transcriptions of read speech are very similar to manual phonetic transcriptions created by expert phoneticians. In the research reported here, the APR processed extemporaneous speech, which is likely to be more difficult to transcribe – both for machines and humans. Although the performance of the APR is far from perfect most of the transcriptions in the output of the recogniser are sensible and intelligible. Thus, it seems safe to assume that even if the transcriptions of the material used in our experiments had been created by human transcribers, the performance of Shortlist would not have been much better. So, the problem of the inadequate performance of Shortlist is most probably not due to a poor performance of the APR, but rather to the current implementation of the first stage of the model. To put this assumption to the test we are currently working on a more sophisticated version of the first stage of Shortlist, that should be better able to deal with transcription 'errors'.

7. CONCLUSIONS

In this research, we showed that modelling pronunciation variation improves the ability of the psycholinguistic model of human speech recognition Shortlist to deal with real-life input. We modelled pronunciation variation by including variants into the lexicon of Shortlist.

A series of experiments was carried out to find the optimal model set for transcribing the training material that was used as basis for the generation of the variants. The best model set for the transcription task was trained with maximally 128 Gaussians per state and had a separate model for filled pauses. Furthermore, a new acoustic model set was trained on the 'improved' transcriptions of the training material. This showed a slight improvement compared to the baseline system.

Adding pronunciation variants to Shortlist's lexicon increased its performance with 16.25% absolute, but still the error rate is high compared to state-of-the-art ASR systems. The disappointing performance of Shortlist as a recogniser is most probably due to the current implementation of the search in the first stage of the model. In order for Shortlist to better deal with real-life input, its implementation should be improved. Therefore, we are currently working on a more sophisticated version of the first stage of Shortlist.

8. ACKNOWLEDGEMENT

For converting the d-trees into FSGs, software developed at the International Computer Science Institute (ICSI), Berkeley, USA, was used. The authors like to thank Eric Fosler-Lussier for making available the software and Mirjam Wester for her help with the Weka-tools and ICSI-software.

9. REFERENCES

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

[2] Norris, D., "Shortlist: a connectionist model of continuous speech recognition", *Cognition 52*, 189-234, 1994.

[3] Kessens, J.M., Strik, H., "Lower WERs do not guarantee better transcriptions", *Proceedings Eurospeech*, pp. 1721-1724, 2001.

[4] Sturm, J., Kamperman, H., Boves, L., den Os, E., "Impact of speaking style and speaking task on acoustic models", *Proceedings ICSLP*, pp. 361-364, 2000.

[5] den Os, E.A., Boogaart, T.I., Boves, L., Klabbers, E., "The Dutch Polyphone Corpus", *Proceedings Eurospeech*, pp. 825-828, 1995.

[6] Steinbiss, V., Ney, H., Haeb-Umbach, R., Tran, B.-H., Essen, U., Kneser, R., Oerder, M., Meier, H.-G., Aubert, X., Dugast, C., Geller, D., "The Philips research system for large-vocabulary continuous-speech recognition", *Proceedings Eurospeech*, pp. 2125-2128, 1993.

[7] Kessens, J., Wester, M., Strik, H., "Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation", *Speech Communication 29*, 193-207, 1999.

[8] Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., Zavaliagkos, G., "Stochastic pronunciation modelling from hand-labelled phonetic corpora", *Speech Communication 29*, 209-224, 1999.

[9] Strik, H., Cucchiarini, C., "Modeling pronunciation variation for ASR: A survey of the literature", Speech Communication 27, 225-246, 1999.

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

[11] Weka-3 Machine Learning software in Java, http://www.cs.waikato.ac.nz/ml/weka/index.html.

[12] Witten, I., Frank, E., "Data Mining, practical machine learning tools and techniques with Java implementations", *Morgan Kaufmann Publishers*, 2000.

[13] Yang, Q., Martens, J.-P., "On the importance of exception and cross-word rules for the data-driven creation of lexica for ASR", *Proceedings 11th ProRisc Workshop*, pp. 589-593, 2000.

[14] Bouwman, G., Boves, L., "Using information on lexical stress for utterance verification", *Proceedings of ITRW on Prosody in ASRU*, pp. 29-34, 2001.

[15] Cucchiarini, C., Binnenpoorte, D.M., Goddijn, S.M.A., "Phonetic Transcriptions in the Spoken Dutch Corpus: how to combine efficiency and good transcription quality", *Proceedings Eurospeech*, pp. 1679-1682, 2001.