# Improving Automatic Phonetic Transcription of Spontaneous Speech through Variant-Based Pronunciation Variation Modelling

**Diana Binnenpoorte, Catia Cucchiarini, Helmer Strik and Lou Boves**

Department of Language and Speech, University of Nijmegen
Erasmusplein 1, Nijmegen, The Netherlands
{D.Binnenpoorte,C.Cucchiarini,W.Strik,L.Boves}@let.kun.nl

**Abstract**

In this paper we present an experiment aimed at improving automatic phonetic transcription of Dutch spontaneous speech through a variant-based method of pronunciation variation modelling. For spontaneous speech, the literature does not always provide enough rules to describe its characteristic phonological processes. Therefore, other methods should be applied to model pronunciation variation for automatic phonetic transcription. We show that a large amount of manually transcribed phonetic data is an extremely useful source for collecting pronunciation variants and their prior probabilities. From the results we can conclude that the adopted method is indeed suitable for improving automatic transcription of spontaneous speech, and that further improvements can be obtained by combining this method with rule-based methods of pronunciation variation modelling.

## Introduction

Annotated large speech databases are a rich resource for various linguistic studies. Manual annotation of speech signals is very time-consuming and costly. Especially phonetic transcriptions are known to be extremely labour intensive and therefore expensive. Recourse to automatic techniques would partly solve this problem. Although in the last decades considerable progress has been made in the field of speech recognition technologies, still a continuous speech recognizer (CSR) performs better on read speech than on conversational, spontaneous speech. This does not only apply to automatic speech recognition, but also to automatic transcription of speech (Cucchiarini and Strik, 2003).

However, many real-life situations in which ASR techniques can be applied concern spontaneous speech rather than read speech, which therefore constitutes a very good reason for trying to improve CSR performance on spontaneous speech. Since in this process automatic phonetic transcription has an important role to play, there are good reasons too for improving CSR performance on automatic phonetic transcription of speech data. This topic will be the focus of the present paper.

The fact that CSR performance on automatic transcription is systematically lower for spontaneous speech than for read speech can be explained in two different ways. The first explanation is that spontaneous speech is intrinsically more difficult to transcribe than read speech. The alternative explanation is that we are much better at modelling read speech than spontaneous speech, because the bulk of the knowledge accumulated so far in speech research does concern carefully pronounced laboratory speech, which is more similar to read speech than to spontaneous speech. The third possibility is a combination of the previous two: spontaneous speech is intrinsically more difficult to transcribe than read speech, but the discrepancy in CSR performance on automatic transcription of read and spontaneous speech can be reduced by better modelling spontaneous speech.

Although we believe that spontaneous speech might somehow be more difficult to transcribe for both humans and machine, we are convinced that the current levels of CSR performance on automatic transcription of spontaneous speech can be improved to a certain extent through better modelling. In particular, current approaches to automatic transcription have made little use of the spontaneous speech corpora that are now becoming available for various languages, and which appear to be invaluable sources of information for various purposes, among which pronunciation variation modelling. In this paper we will show how automatic transcription of spontaneous speech can be improved by modelling some of the variation that characterizes this type of speech in a way that was not feasible until large spontaneous speech corpora became available: variant-based pronunciation variation modelling as opposed to rule-based pronunciation variation modelling.

In the remainder of this paper we go more deeply into the adopted method, then we present the results after which a discussion is presented together with the conclusions.

## Experiment

In the following section we first describe the method of the experiment, followed by a description of the speech material we used, how the automatic phonetic transcription is created based on a lexicon containing pronunciation variants, how a reference transcription of a small test corpus is made and finally how the latter was used to determine the quality of the automatically generated phonetic transcription.

### Method

One way of obtaining automatic phonetic transcriptions is by having a speech recognizer in forced recognition mode select the variant that best matches the acoustic signal from a list of pronunciation variants contained in the lexicon. These variants can be generated in different ways (for an overview, see Strik & Cucchiarini, 1999). A very common method consists in generating the variants by means of rewrite rules that are either obtained from the literature or are extracted from speech data. A second option consists in extracting the variants directly from a large speech corpus (enumerated). The advantage of the first method, which we will call rule-based, is that the rules can be applied to all words in the lexicon, whereas in the second approach, which we will call variant-based, only variants that are found in the corpus can be included

in the lexicon. However, the variant-based approach has the advantage that it allows modelling of word-specific phenomena that cannot otherwise be captured by rules. Especially in spontaneous speech it often happens that highly frequent words undergo extreme reduction processes that can delete even up to complete syllables. Until recently, variant-based modelling could not be applied to Dutch, because we did not have an adequate corpus. Since we are now fortunate to have a large corpus of transcribed Dutch spontaneous speech, the Spoken Dutch Corpus (Oostdijk, 2000), we decided to study the effect of this type of pronunciation modelling on automatic transcription. In this experiment we limited ourselves to modelling frequently found words that are known to be enormously reduced in spontaneous speech.

## Material

The speech material used in this experiment is divided into two parts, one for the extraction of variants to be added to the lexicon, the other for testing the performance of automatic transcription. The material was selected from the Spoken Dutch Corpus. We selected all the spontaneous material that had a manual phonetic transcription. This material consists of telephone conversations and dialogues (and multi-logues) that were recorded in home environments, using one central (stereo) microphone and a minidisk recorder. The different recording conditions of these two speech types result in different acoustic qualities. Nonetheless, we chose to use both types of spontaneous material because of the extemporaneous character of the speech that is almost the same in both conditions.
In Table 1 the most important statistics of the data are summarized: the total duration of the speech material in hh:mm:ss, the number of words, the number of unique words and the average number of pronunciation variants per word.

| | duration | # words | # unique | av. vars |
|---|---|---|---|---|
| TRAIN | 24:26:07 | 304502 | 14113 | 21,5 |
| TEST | 0:13:04 | 2850 | 826 | 9,9 |

Table 1: Statistics of both train and test set

In total 7620 words in the training set were found with only one pronunciation, most of which are proper names, infrequent inflections of verbs and broken words (start-repairs). The forty most frequent words cover 50% of all the words in the training set and most of these are short (monosyllabic) function words and first person inflections of the verbs ‘zijn’ (to be) and ‘hebben’ (to have). Multisyllabic function words, such as ‘natuurlijk’ (of course), ‘helemaal’ (totally), ‘eigenlijk’ (actually) and ‘allemaal’ (all), are also very frequent and can be found in the top hundred of most frequent words.

## Lexicon training set

The broad phonetic transcriptions were obtained by having trained transcribers verify and possibly correct an optimized automatically generated phonetic transcription. Then, in a second round, the resulting transcriptions were verified and corrected, if needed, by another transcriber. Besides this manual phonetic transcription and the original orthographic transcription, the training material is also manually time-aligned to the speech signal on word level. Thus, every orthographic entity is unambiguously linked to a phonetic transcription.
All the word types in the training set are collected together with their transcription and sorted on frequency. Then a prior probability for each pronunciation variant is calculated given the frequency of occurrence of its orthographic counterpart in the training material. The list created this way contains all possible pronunciations of the words found in the training set and their probability of occurrence. 80 orthographic words from the test set did not occur in the training set. For these words a unique canonical phonetic transcription was obtained, by consulting the general CGN lexicon and these transcriptions were assigned a prior probability of 1. Furthermore, 65 words in the test set only occurred once in the training set and were assigned the observed pronunciation variant in the lexicon.

## Automatically generated transcription - AGT

We used a CSR (Strik, et al, 1996) in forced recognition mode to choose the most likely pronunciation variant from the lexicon given a class-based language model and the acoustics of the speech signal. The acoustic models are continuous density hidden Markov models with 32 Gaussians per state trained on phonetically rich sentences uttered through a telephone. We converted the wide band material of the test set, the recordings in the home environments, to telephone bandwidth in order to avoid the mismatch between the acoustic properties of the models and the test data.
For each utterance in the test set a pronunciation lexicon is extracted from the training lexicon, where each word in the utterance has all the pronunciation variants as they were found in the training material.
The language model was a class-based bigram model. The prior probabilities of the pronunciation variants of a word are captured in the unigram part. Here, the classes, or categories, are the words in the utterance; the transitions between words are modelled by the class bigram (Brown et al, 1992).
The result of the forced recognition is a sequence of the pronunciation variants of the words in the utterance that best matches the speech signal, the AGT.

## Reference transcription – RT

A reference transcription (RT) can serve as a benchmark against which other transcriptions, in this case an AGT, can be validated. A consensus transcription is probably the best possible approximation of the ‘true’ transcription (Shriberg, 1991).
Two phonetically trained and experienced listeners were asked to make a consensus transcription of the speech material in the test set. They transcribed from scratch and had to agree on each symbol in the transcription. They used the same symbol set as was used for the AGT. This led to a broad phonetic consensus transcription, which will serve as the RT in this experiment.

## Alignment

A dynamic programming algorithm was used to make an alignment between the AGT and the RT in order to determine the agreement between the former and the

latter. The program provides the number of substitutions, deletions and insertions on phoneme level. Each of these errors is assigned a weighting, which is used as a distance measure during the alignment procedure. The weightings are calculated in terms of articulatory features, such as place and manner of articulation, voice, lip rounding, length, etc. The results of the alignment show in what respects the AGT differs from the RT.

## Results

### Phone error rates

In the first row in Table 2 the results of the alignment between the AGT and the RT are shown in percentages of substitutions, deletions and insertions on phoneme level. The total percentage disagreement (last column) is the phone error rate (PER). In order to put the data in perspective, the second row gives the result that was obtained by modelling frequent phonological processes by means of rules for the same data (Binnenpoorte & Cucchiarini, 2003). Finally, in the last row the percentage disagreement on phoneme level between a simple concatenation of canonical forms and the RT for the same material is displayed.

| % | SUB | DEL | INS | TOTAL |
|---|---|---|---|---|
| AGT | 10.01 | 7.22 | 4.50 | 21.73 |
| STATIC | 10.37 | 1.57 | 11.83 | 23.77 |
| CANON | 12.50 | 2.00 | 12.87 | 27.37 |

Table 2: Quantitative results of alignment between AGT and RT and previously found results.

In Binnenpoorte et al. (2003) four trained transcribers were asked to transcribe a part of the spontaneous speech material as contained in the test set. When comparing their transcriptions with the corresponding part in the RT we found total PERs ranging from 13.4% to 15.7%, with inter-transcriber agreement ranging from 85.7% to 94.9% (where the latter figure relates to agreement found by comparing the transcription of the first transcriber with the correction of that first transcription by a second transcriber). Although the data set of the human transcription differs from the AGT, the results obtained in this experiment surpass the best AGT performance in previous experiments. Still the AGT does not come close to human performance yet, which is not surprising if we consider that in this experiment we only applied the variant-based method.

### Analysis of PERs

Closer inspection of the output of the alignment between the AGT and RT reveals that for all substitutions, deletions and insertions a relatively small number of phonetic processes cover more than half of the errors. To illustrate, the 13 most frequent substitutions (8.3% of all the substitution types) are responsible for 50% of the substitution errors. In case of the deletions, 50% of the errors can be accounted for by only 4 deletion types (11,4% of total), and also the 4 most frequent insertions (12.1% of total) are responsible for 50% of the insertion errors.

The most frequent substitutions are confusions between phonemes that only differ in one articulatory feature, see Table 3, primarily related to the feature voice (in fricatives and plosives) and length (in vowels). In addition, confusions between any vowel and schwa are also frequent. Most deletions are related to /@/, /r/, /d/ and /n/. Finally, for insertions we found that most of the errors are due to insertion of /@/, /n/, /r/ and /t/.

| SUBSTITUTIONS | | DELETIONS | | INSERTIONS | |
|---|---|---|---|---|---|
| # | phones | # | phone | # | phone |
| 51 | G,x | 117 | @ | 68 | @ |
| 50 | s,z | 68 | r | 60 | n |
| 46 | d,t | 63 | d | 39 | r |
| 41 | A,@ | 47 | n | 29 | t |
| 37 | f,v | 39 | t | 24 | j |

Table 3: Top five of substitutions, deletions and insertions in dataset containing 8063 phonemes

## Discussion

The data in Table 2 show that our attempt to optimise automatic phonetic transcription by means of a lexicon with pronunciation variants observed in a large manually transcribed corpus has been successful. The improvements are mainly the result of fewer insertions, which means that the CSR has chosen variants in which reduction of specific phonemes was modelled. On the other hand, the number of deletions has risen enormously. We believe that many –but not all- of the remaining discrepancies between our APT and RT are due to inherent limitations of the HMM recogniser used as a transcription tool. The 117 /@/ deletions can illustrate this: The topology of the acoustic models in our CSR requires that phonemes span at least 30 ms to be detected. It seems that the two expert listeners had a lower durational threshold for /@/. We believe that we see similar problems with the other frequent insertions and deletions. Dutch has a substantial number of frequent unstressed syllables with a vowel followed by /r/ or /n/. In all these cases the acoustic basis for the detection of the individual phonemes in the canonical representation is rather weak, especially in spontaneous speech. More often than not, the presence of one ‘sound’ is fully encoded in the phonetic details of its neighbours. Phoneticians are able to reach a high degree of agreement on the segmental transcription of these syllables (cf. the agreement data in Goddijn & Binnenpoorte, 2003), but this is probably due to a common interpretation of these acoustic complexes, biased by the fact that they understand the words and therefore can rely on knowledge of the underlying canonical form. However, a phone-based HMM system is fundamentally unable to reproduce this behaviour.

The most frequent substitutions that remain in our approach are related to the feature “voice”. Due to the fact that the lexicon only contained observed pronunciation variants, we may have missed a number of realistic variants, especially in words that are not among the most frequent. Also, our approach may not be the best solution for cross-word voice assimilation, a process that is known to be quite important (Binnenpoorte & Cucchiarini, 2003). However, also in this case we think that the HMM system is partially to be blamed. Especially for fricatives “voice” has quite an uncertain status. As a consequence, it is

virtually impossible to train HMMs that can tell the voiced and unvoiced cognates apart. To approximate human-like performance in voiced-unvoiced distinction we will need a two stage procedures that operates on the segmentation of the HMM system, and that applies independent acoustic evidence for the classification.

In this paper, we adopted a variant-based approach to generate pronunciation variants. We put all observed variants in the lexicon. A disadvantage of this approach is that only 'seen' variants of a word can be modelled. For words that did not occur in the corpus from which the variants were derived, the lexicon will contain only the canonical form. In our case, 1.4% of the total number of discrepancies between APT and RT originates from the 80 'unseen' words. To obtain pronunciation variants for these and other less frequent words we can use the manually annotated corpus for the extraction of rules. This can be achieved by comparing the manually transcribed data with canonical transcriptions of that same data to generalize over all differences given a certain context (Wester, 2003; Scharenborg & Boves, 2002; Riley et al, 1999).

The combination of rewrite rules together with prior probabilities of pronunciation variants could be especially promising for multiword expressions. These are frequently used expressions in everyday language, such as institutionalized phrases. Most of the time the individual words of a multiword expression are pronounced with much more reduction in the multiword construction than in other not so frequent constructions. Multiword expression should therefore be considered as one entity in the same way as 'normal' words.

## General discussion

In this paper we have shown that automatic phonetic transcription of spontaneous speech can be improved to a certain extent by modelling pronunciation variation through a variant-based method which could not be applied before a large corpus of spontaneous speech became available for Dutch. It's clear that the more transcribed data are available, the better spontaneous speech can be modelled, which, in turn, means that the APT can be improved such that more transcriptions can become available at lower costs.

In spite of this enhancement in performance, there is still much room for improvement to obtain performance levels that much more resemble those obtained for read speech. However, this is not surprising if we consider that in this experiment only the variant-based method of pronunciation variation modelling was applied, thus neglecting the modelling of other processes that, as we know, are best addressed through the rule-based method. The challenge will now be to find the optimal combination of these two methods which provides the best performance levels. This will be the focus of our research in the near future.

## Conclusions

Based on the results of the experiment reported on in this paper we can conclude that the adopted technique of modelling real-life pronunciation variants does improve automatic phonetic transcription quality, but is still not sufficient to resemble human phonetic transcriptions. A combination of variant-based and rule-based methods will probably offer the best solution.


## Acknowledgements

We would like to thank J. Sturm and O. Scharenborg for their contributions to this research.


## References


Cucchiarini, C. and Strik, H. (2003). Automatic phonetic transcription: An overview. In Proceedings of 15th ICPhS, Barcelona, Spain (pp. 347-350).

Strik, H. and Cucchiarini, C. (1999). Modeling pronunciation variation for ASR: A survey of the literature. In Special Issue of Speech Communication on 'Modeling Pronunciation Variation for Automatic Speech Recognition', Vol. 29, No. 2 – 4, (pp. 225–246).

Oostdijk, N. (2000). The Spoken Dutch Corpus: Overview and first Evaluation. In Proceedings LREC2000, Athens, Greece, (pp. 887-893).

Strik, H., Russel, A., van den Heuvel, H., Cucchiarini, C. and Boves, L. (1996). A spoken dialogue system for the Dutch public transport information service. In International Journal Speech Technology 2 (2), (pp. 119-129).

Brown, P., Della Pietra, V., deSouza, P., Lai, J., Mercer, R. (1992). Class-based n-gram Models of Natural Language. In Computational Linguistics, 18 (4), (pp. 467-480).

Shriberg, L. D. and Lof, L. (1991). Reliability studies in broad and narrow phonetic transcription. In Clinical Linguistics and Phonetics, 5, (pp. 225-279).

Binnenpoorte, D. and Cucchiarini, C. (2003), Phonetic Transcription of Large Speech Corpora: How to boost efficiency without affecting quality. In In Proceedings of 15th ICPhS, Barcelona, Spain (pp. 2981-2984).

Binnenpoorte, D., Goddijn, S. and Cucchiarini, C. (2003). How to Improve Human and Machine Transcriptions of Spontaneous Speech. In Proceedings of ISCA IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR), Tokyo, Japan, (pp. 147–150).

Goddijn, S. and Binnenpoorte, D. (2003). Assessing Manually Corrected Broad Phonetic Transcriptions in the Spoken Dutch Corpus. In Proceedings of 15th ICPhS, Barcelona, Spain (pp. 2981 – 2984).

Wester, M. (2003). Pronunciation modeling for ASR -- knowledge-based and data-derived methods. In Computer Speech and Language 2003, 17 (pp. 69-85)

Scharenborg, O. and Boves, L. (2002). Pronunciation Variation Modelling in a Model of Human Word Recognition. In Proceedings of Workshop on Pronunciation Modeling and Lexicon Adaptation, Estes Park, USA, (pp. 65-70).

Riley, M., Byrene, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraclar, M., Wooters, C., Zavaliagkos, G. (1999). Stochastic pronunciation modelling from hand-labelled phonetic corpora. In Speech Communication 29, (pp. 209-224).