Title: Automatic Phonetic Transcription of Large Speech Corpora

Authors: Christophe Van Bael, Lou Boves, Henk van den Heuvel, Helmer Strik

Affiliation: Centre for Language and Speech Technology (CLST), Radboud University Nijmegen, The Netherlands

Name and address for correspondence:
Name: Christophe Van Bael
Mail: Centre for Language and Speech Technology (CLST), Radboud University Nijmegen
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
Tel: + 31 24 361 29 08
Fax: + 31 24 361 29 07
E-mail: c.v.bael@let.ru.nl

Abstract:
Most large speech corpora are delivered with a lexicon that contains a canonical transcription of every word in the orthographic transcription. Such a lexicon can be used for generating a hypothetical 'canonical' phonetic transcription from the orthography. In addition, time and money permitting, some speech corpora are provided with a manually verified broad phonetic transcription of at least part of the material. Since the manual verification of phonetic transcriptions is time-consuming and expensive, we investigated whether existing automatic transcription procedures and combinations of such procedures can offer a quick and cheap alternative for the generation of phonetic transcriptions like the manually verified transcriptions delivered with large speech corpora. In our study, we used ten automatic transcription procedures to generate a broad phonetic transcription of well-prepared speech (read-aloud texts) and spontaneous speech (telephone dialogues) from the Spoken Dutch Corpus. The performance was assessed in terms of the number and the nature of the discrepancies between the emerging phonetic transcriptions and the corresponding manually verified phonetic transcriptions delivered with the Spoken Dutch Corpus. The resulting automatic transcriptions appeared to be comparable to the manually verified transcriptions.

Keywords: Large speech corpora, Automatic phonetic transcription, Transcription evaluation.

1. INTRODUCTION

In the last fifteen years we have witnessed the development of various large speech corpora. Well-known examples are TIMIT (1990), Switchboard (Godfrey et al., 1992), Verbmobil (Hess et al., 1995), the Spoken Dutch Corpus (Corpus Gesproken Nederlands - CGN, Oostdijk, 2002) and the Corpus of Spontaneous Japanese (Maekawa, 2003). The usability of such corpora largely depends on the availability of accurate annotations. It is probably fair to say that the lasting popularity of the not-so-big TIMIT corpus is due to the fact that is comes with very accurate phonetic labelling. Since broad phonetic transcriptions are often used and sometimes even required for diverse purposes such as lexical pronunciation variation modelling for automatic speech recognition (ASR - Strik, 2001), unit selection for speech synthesis (Mizutani and Kagoshima, 2005), automatic pronunciation training and assessment in Computer Assisted Language Learning (Neri et al., 2006; 2007) and general research on pronunciation variation (Riley et al., 1999), contemporary speech corpora are usually provided with a broad phonetic transcription of at least part of their material.

Almost all large speech corpora are provided with a phonemic lexicon that can be used to generate a hypothetical canonical phonetic representation of the material. In addition, time and money permitting, contemporary speech corpora are at least partially enriched with broad phonetic transcriptions with the help of human transcribers in order to ensure a more accurate representation of the material. Since the employment of human transcribers is known to be exceedingly time-consuming and expensive when they have to transcribe speech from scratch, it is common practice to provide human transcribers with an example transcription they have to verify on the basis of their own perception of the speech signal. Switchboard, Verbmobil and the Spoken Dutch Corpus are three corpora which, in addition to a canonical transcription of all their material, received a manually verified phonetic transcription of a limited subset of their data (Greenberg et al., 1996; Geumann et al., 1997; Goddijn & Binnenpoorte, 2003). The example transcription the transcribers of the Spoken Dutch Corpus were presented with already reflected the obligatory cross-

word assimilation and degemination processes of Dutch (Binnenpoorte & Cucchiarini, 2003). The modelling of these processes decreased the discrepancies between the original canonical example transcription and the actual speech signal, and as such it also reduced the number of required corrections and the time it took the transcribers to complete the transcription task. Despite them being quicker and therefore also less expensive than manual transcription procedures with human experts starting from scratch however, verification procedures also have their drawbacks.

It has been suggested that verifying example transcriptions may bias the resulting transcriptions towards the example transcriptions they are based upon (Binnenpoorte, 2006). In addition, the remaining costs are often still quite substantial. Demuynck et al. (2002) reported that their students needed 15 minutes to manually verify the transcription of one minute of public lectures, and 40 minutes for one minute of spontaneous speech. This explains why human transcribers verified an example transcription of 'only' one million words of the 9-million-word Spoken Dutch Corpus, and why 'only' four hours of Switchboard speech were phonetically transcribed as an afterthought. Still, despite these drawbacks, manually verified phonetic transcriptions are presently considered to be the best transcriptions one can feasibly obtain if large amounts of speech have to be transcribed. It is therefore worthwhile investigating whether the same transcription quality can be obtained by means of quicker and cheaper automatic transcription procedures. Because of their high transcription speed and their limited costs, automatic transcription procedures not only hold the promise of increasing transcription speed and reducing transcription costs, they even have the potential of transcribing corpora that are too large to be ever transcribed with the help of human transcribers.

Several studies already reported benefits of using automatic phonetic transcriptions (APTs) for the development of ASR systems (e.g. Riley et al., 1999; Saraçlar & Khundanpur, 2004; Tjalve & Huckvale, 2005; Wester, 2003; Yang & Martens, 2000) and speech synthesis systems (e.g. Bellegarda, 2005; Jande, 2005, Wang et al. 2005). However, since in these studies the transcriptions were used as mere tools for the development of specific speech applications, the procedures with which the transcriptions were generated were not evaluated in terms of their ability to approximate the quality of manually verified phonetic transcriptions. Therefore, our study was aimed at investigating whether existing automatic transcription procedures and combinations of such procedures can approximate manually verified phonetic

transcriptions and, consequently, whether they can offer a sound alternative to commonly used but nonetheless time-consuming and expensive verification procedures for the transcription of large speech corpora.

We assessed the quality of three well-known transcription procedures, two combinations of these procedures and extensions to these five procedures in terms of the resemblance of their transcriptions and a manually verified broad phonetic transcription of read speech and of spontaneous telephone dialogues from the Spoken Dutch Corpus. Since we aimed at approximating transcriptions that were made with a limited symbol set and that originated from canonical example transcriptions, it should be clear that our experiments were not aimed at comparing or improving the transcription procedures in terms of the accuracy with which they can describe the actual speech signal.

In order to ensure the applicability of the transcription procedures in contexts where only minimal resources are available, we optimised our procedures with limited resources and minimal human effort. Most procedures only required a standard continuous speech recogniser, an algorithm to align phonetic transcriptions, an orthographically transcribed corpus, a canonical lexicon and a manually verified phonetic transcription of a relatively small sample of the corpus. The manually verified phonetic transcription was required to tune the transcription procedures and to evaluate their performance. Some procedures also required software for the implementation of decision trees, and some (also) a list of phonological processes describing pronunciation variation in the language under investigation (Dutch). Expert human effort was limited to the compilation of such a list of phonological processes, and the aforementioned manual verification of an example transcription of a limited amount of speech.

This paper is organised as follows. In Section 2, we introduce the material and tools we used in our study. Section 3 sketches the various transcription procedures. Section 4 presents the evaluation of the emerging transcriptions. In Section 5, we discuss our results, and in Section 6, we formulate our conclusions.

2. MATERIAL AND TOOLS

2.1. Speech Material

We worked with speech material from the Spoken Dutch Corpus (Oostdijk, 2002). We considered speech of native speakers from the Netherlands only. In order not to base the assessment of the transcription procedures on the transcription of speech from one particular speech style, we chose to work with read speech as well as spontaneous telephone dialogues.

The read speech was recorded at 16 kHz (16-bit PCM) with high-quality table-top microphones for the compilation of a library for the blind. The telephone dialogues, comprising much more spontaneous speech, were recorded at 8 kHz (8-bit A-law) through a telephone platform. As part of the orthographic transcription process, the speech material was manually segmented into speech chunks of approximately 3 seconds each. The transcribers were instructed to put chunk boundaries in naturally occurring pauses. Only if speech stretched for substantially longer than 3 seconds without a silent pause, the transcribers were requested to put chunk boundaries between adjoining words with minimal cross-word co-articulation. We adhered to this chunk-level annotation. In order to be able to focus on phonetic transcription proper, we excluded speech chunks that, according to the orthographic transcription, contained non-speech, unintelligible speech, broken words and foreign speech. Chunks containing overlapping speech (in the telephone dialogues) were excluded as well.

The statistics of the data are presented in Table 1. We divided the data of each speech style into a training set, a development set and an evaluation set. To this end, we listed all speech chunks of all speakers, we randomised their ordering, and we extracted the subsets. This guaranteed mutually exclusive data sets with similar material. The resulting data sets of the two speech styles differ in size, but we preferred to work with all the material meeting our requirements rather than ignoring half of the read speech.

[Table 1]

2.2. Canonical Lexicon

Our canonical lexicon was a comprehensive multi-purpose in-house lexicon. It was compiled by merging various existing lexical resources such as CELEX (Baayen et al, 1995), RBN (ReferentieBestand Nederlands, 2005) and PAROLE (PAROLE lexicon, 2005). The pronunciation forms in the lexicon reflected the standard pronunciation of

words as they would be carefully pronounced in isolation according to the obligatory word-internal phonological processes of Dutch (Booij, 1999). Each word was represented by just one standard broad phonetic transcription. We ignored all information about syllabification and syllabic stress in order to ensure the applicability of the transcription procedures in research contexts where a lexicon with this kind of linguistic information is unavailable.

2.3. Reference Transcriptions (RTs)

We used the manually verified phonetic transcriptions of the Spoken Dutch Corpus as Reference Transcriptions (RTs) to tune (with the RTs of the development sets) and evaluate (by means of the RTs of the evaluation sets) the transcription procedures. The manually verified transcriptions of the Spoken Dutch Corpus were generated in three steps. First, the canonical representation of every word was selected from the lexicon. Subsequently, two cross-word phonological processes of Dutch, voice assimilation and degemination, were applied to the phones at word boundaries in order to decrease the discrepancies between the canonical transcription and the speech signal. The resulting transcriptions were finally verified and corrected by human transcribers. The transcribers acted according to a strict protocol instructing them to change the example transcription only if they were certain that it did not correspond to the speech signal (Binnenpoorte & Cucchiarini, 2003).

2.4. Continuous Speech Recogniser (CSR)

Except for the canonical transcriptions, all automatic phonetic transcriptions (APTs) were generated by means of a continuous speech recogniser (CSR) that was based on Hidden Markov Models and that was implemented with the HTK Toolkit (Young et al., 2001). Our CSR used 39 gender- and context independent, but speech style-specific acoustic models with 128 Gaussian mixture components per state (37 phone models, one model for silences of 30 ms or more and one model for the optional silence between words).

We trained our acoustic models in three stages with the canonical transcriptions of the training data (see Figure 1). First, we trained flat start acoustic models with 32 Gaussian mixture components in 41 iterations. Subsequently, we used these models to

obtain a more realistic segmentation of the speech material. We used this segmentation to bootstrap a new set of acoustic models, which we retrained (with 55 iterations) to models with 128 Gaussian mixture components per state. Experiments with the development sets of both speech styles showed that acoustic models with 128 mixture components yielded transcriptions that resembled the target transcriptions more closely than transcriptions that were generated with models with fewer mixture components per state.

[Figure 1]

2.5. Algorithm for Dynamic Alignment of Phonetic Transcriptions (ADAPT)

ADAPT (Elffers et al., 2005) is a dynamic programming algorithm designed to align two strings of phonetic symbols according to the articulatory distance between them. We used ADAPT to align phonetic transcriptions for the generation of lexical pronunciation variants for forced recognition (Section 3.1.2), and for the quality assessment of the automatic phonetic transcriptions through their alignment with the manually verified reference transcriptions (Section 3.2).

3. METHOD

We investigated the suitability of ten automatic transcription procedures for the phonetic transcription of large speech corpora. The transcription procedures are introduced in Section 3.1. In Section 3.2 we describe the evaluation procedure by means of which the automatic phonetic transcriptions and, consequently, the transcription procedures were assessed.

3.1. Generation of phonetic transcriptions with different transcription procedures

Figure 2 shows the ten transcription procedures by means of which our APTs were generated. We used three generic procedures (Section 3.1.1), two combinations of these procedures (Section 3.1.2), and five procedures in which we used decision trees to further tune the output of the aforementioned procedures towards the type of transcription we were trying to approximate (Section 3.1.3). Most of the procedures

required the tuning of several parameters to optimally approximate the RTs of the data in the development sets. The optimal parameter settings were subsequently used for the transcription of the data in the evaluation sets.

[Figure 2]

3.1.1. Generic transcription procedures

Lexicon lookup (canonical) transcription procedure

The canonical phonetic transcriptions (CAN-PTs) were generated through a lexicon lookup procedure. Cross-word processes were not modelled. In general, canonical transcriptions like these can be easily obtained, since many corpora are provided with an orthographic transcription and a canonical pronunciation lexicon comprising a broad phonetic transcription of the words in the orthographic transcription.

Data-driven transcription procedure

The data-driven phonetic transcriptions (DD-PTs) were based on the acoustic *data*. The DD-PTs were generated through constrained phone recognition; a CSR segmented and labelled the speech signal by means of its acoustic models and a phonotactic model. The phonotactic models (one for each speech style) were trained on the RTs of the development data.

Figure 3 shows the last three steps of the data-driven transcription procedure. The first step, the training of the phonotactic models, is not included in the figure. We trained bigram, trigram, four-gram, five-gram and six-gram models. Since the current version of HTK (v.3.2) only supports the use of unigram and bigram models in its first decoding pass, we used a bigram model in the first pass, and higher order n-gram models to rescore the resulting phone lattices (step 3). The final phonetic transcription of the data (step 4) was obtained with a four-gram model. Transcription experiments with the development data of both speech styles indicated that the use of four-gram models yielded transcriptions that resembled the RTs more closely than the bi-, tri-, penta- and hexagram phonotactic models.

[Figure 3]

Knowledge-based transcription procedure

ASR research often draws on the linguistic literature for the extraction of knowledge to generate lexical pronunciation variants for recognition (Kessens et al., 1999; Strik, 2001). Figure 4 illustrates the three-step procedure we used to generate knowledge-based phonetic transcriptions (KB-PTs).

[Figure 4]

We first compiled a list of 20 prominent phonological processes from the literature on the phonology of Dutch (Booij, 1999). We implemented these processes as context-dependent rewrite rules modelling both within-word and cross-word contexts in which phones from the CAN-PT could be deleted, inserted or substituted with other phones. Most of the processes identified by Booij (1999) could be described in terms of operations on phoneme symbols or articulatory features. However, some of the processes could only be described with information about the prosodic or syllabic structure of words. We reformulated most of these processes in terms of phonetic symbols and features, since we wanted to exclude non-segmental information from our experiments (see Section 2.2). We implemented the rules in a conservative manner in order to minimise the risk of over-generation. The resulting rule set comprised phonological rules describing progressive and regressive voice assimilation, nasal assimilation, t-deletion, n-deletion, r-deletion, schwa deletion, schwa epenthesis, palatalisation, degemination and more specific rules modelling pronunciation variation in high-frequency words (e.g. demonstratives) in Dutch. The reduction and the deletion of full vowels, two prominent phonological processes in Dutch, were not implemented since they could not be formulated without the explicit use of supra-segmental information.

In the second step of the procedure, we used the phonological rewrite rules to generate pronunciation variants from the CAN-PTs of the speech chunks. Note that it was necessary to apply the rules to the speech chunks rather than to the words in isolation, for cross-word processes could only be modelled if the neighbouring words were known. The rules only applied once, and their order of application was manually optimised. Analysis of the resulting pronunciation variants suggested that hardly any implausible variants were generated, and that no obvious variants were missing. It may well be, however, that two-level rules (Koskenniemi, 1983) or an iterative application of rewrite rules are needed for the generation of all plausible pronunciation variants in languages other than Dutch.

In the third step of the procedure, the pronunciation variants (including the original CAN-PTs) of each individual speech chunk were listed. Since the linguistic literature hardly ever provides accurate information on the frequency of phonological processes, and since trustworthy priors can only be learned from the analysis of a sufficiently large amount of manually verified transcriptions (the amount of manual transcriptions that is hardly every available), our knowledge-based pronunciation variants did not comprise prior probabilities. The optimal knowledge-based phonetic transcription (KB-PT) was identified through forced recognition.

3.1.2. Combinations of generic transcription procedures

After having generated the CAN-PTs, DD-PTs and KB-PTs, we combined these transcriptions to obtain new transcriptions. Chunk-level pronunciation variants were generated through the automatic alignment of two APTs at a time. Since the KB-PTs were based on the CAN-PTs, we only combined the CAN-PTs with the DD-PTs (CAN/DD-PT) and the KB-PTs with the DD-PTs (KB/DD-PT) to generate new pronunciation variants in addition to the original CAN-PTs, DD-PTs and KB-PTs. Figure 5 shows how new pronunciation variants were generated through the alignment of the phones in two different APTs. These pronunciation variants were listed, after which our CSR was forced to choose the best matching pronunciation variant for every chunk of words in the orthographic transcriptions. The three steps of this combined transcription procedure are illustrated in Figure 6.

[Figure 5] + [Figure 6]

We combined the APTs from the different transcription procedures to provide our CSR with additional linguistically plausible pronunciation variants for the words in the orthographic transcriptions. After all, canonical transcriptions do not model pronunciation variation, and our KB-PTs only modelled the pronunciation variation that was manually implemented in the form of phonological rewrite rules. The DD-PTs, however, were based on the speech signal. Therefore, they were potentially better at representing the actual speech signal, at the risk of being linguistically less plausible than the CAN-PTs and the KB-PTs. It was reasonable to expect that the combination of the different transcription procedures would reinforce the advantages and alleviate the disadvantages of the individual procedures.

3.1.3. Transcription procedures with decision trees

The use of data-driven transcription procedures can result in too many, too few or very unlikely lexical pronunciation variants (Wester, 2003). Therefore, ASR developers often use decision trees to reduce the number of unlikely pronunciation variants and to optimise the number of plausible pronunciations in recognition lexicons (Riley et al., 1999; Wester, 2003). Figure 7 illustrates our four-step procedure to improve the CAN-PTs, DD-PTs, KB-PTs, CAN/DD-PTs and KB/DD-PTs through the use of decision tree filtering. The decision trees were generated with the C4.5 algorithm (Quinlan, 1993), which is provided with the Weka package (Written & Frank, 2005), a collection of Java-based machine learning algorithms.

[Figure 7]

First, the APT (each of the aforementioned transcriptions individually) and the RT of the development data were aligned. Second, all the phones and their context phones in the APT were listed. We will call these phone sequences 'phonetic windows' for the sake of convenience. The size of these phonetic windows was limited to the target phone and its immediately left and right neighbours. Word boundaries were included as extra information in order to model pronunciation variation across word boundaries. The correspondences of the phonetic windows in the APT and the phones in the RT, and the frequencies of these correspondences were used to estimate:

$$P(\text{RT\_phone} \mid \text{APT\_phonetic\_window}) \quad (1)$$

i.e. the probability of a phone in the reference transcription given a particular phonetic window in the APT.

Figure 8 shows a simplified version of the decision tree trained for the phone /e/. The tree strongly predicts (because in the development data, in 12 out of 13 cases, it was the case) that a word-initial /e/ or an /e/ preceded by /@/ and followed by either /@/, /n/ or /j/ should be deleted. Based on the observations in the development data, in all other contexts, all /e/s in the APT should remain. The application of this knowledge is illustrated in the lower half of Figure 8.

[Figure 8]

In the third step of the procedure, the decision trees were used to generate plausible pronunciation variants for the APT of the unseen evaluation data. The decision trees were used to predict:

$$P\ (\text{pronunciation\_variants} \mid \text{APT\_phonetic\_window}) \tag{2}$$

i.e. the probability of a phone with optional pronunciation variants given a particular phonetic window in the APT. In our experiments, all phone variants with a probability lower than 0.1 were ignored. This reduced the number of pronunciation variants and, more importantly, it pruned unlikely pronunciation variants originating from idiosyncrasies in the original APT. The retained phone-level variants were combined to word-level variants. These variants were listed in a multiple pronunciation lexicon. Their probabilities were normalised so that the probabilities of all variants of a word added up to 1.

In the fourth and final step of the transcription procedure, our CSR selected the most likely pronunciation variant for every word in the orthographic transcription. The consecutive application of the decision trees to the CAN-PTs, DD-PTs, KB-PTs, CAN/DD-PTs and KB/DD-PTs resulted in new transcriptions hereafter referred to as $[\text{CAN-PTs}]_d$, $[\text{DD-PTs}]_d$, $[\text{KB-PTs}]_d$, $[\text{CAN/DD-PTs}]_d$ and $[\text{KB/DD-PTs}]_d$.

3.2. Evaluation of the phonetic transcriptions and the transcription procedures

The APTs of the data in the evaluation sets were evaluated in terms of their deviations from the manually verified RTs. We compared the transcriptions by means of ADAPT (Elffers et al., 2005). The disagreement metric was defined as:

$$\text{percentage disagreement} = \left(\frac{Sub + Del + Ins}{N} \right) * 100 \tag{3}$$

i.e. the sum of all phone substitutions (S*ub*), deletions (*Del*) and insertions (*Ins*) divided by the total number of phones in the reference transcription (*N*). Considering the aim of our research, a smaller deviation from the reference transcription indicated a 'better' transcription. A detailed analysis of the number and the nature of the deviations allowed us to systematically investigate the magnitude and the nature of

the improvements and deteriorations caused by the use of the different transcription procedures.

4. RESULTS

The figures in Table 2 show the disagreements between the APTs and the RTs of the evaluation data. From top to bottom and from left to right we see the disagreement scores (%dis) between the different APTs and the RTs of the read speech and the telephone dialogues. In addition, the statistics of the substitutions (sub), deletions (del) and insertions (ins) are presented in order to provide insight into the nature of the disagreements.

[Table 2]

The proportions of disagreements observed in the CAN-PTs and the KB-PTs differed significantly from each other for both speech styles ($p < .01$; we report *t*-tests throughout this article). However, the CAN-PT of the read speech was more similar to the RT than the KB-PT ($\Delta = 6.3\%$ rel.), while the opposite held for the telephone dialogues ($\Delta = 5.9\%$ rel.). In both speech styles, the proportion of substitutions was about equal in the CAN-PT and the KB-PT. Deletions made up only a very small proportion of the discrepancies, so the most important difference was in the insertions; the proportion of insertions was much higher in the telephone speech than in the read speech. The ten most frequent mismatches in the CAN-PTs and the KB-PTs of the two speech styles are presented in Tables 3 and 4, respectively. We observed many similar mismatches due to voiced/unvoiced classification of obstruents, as well as insertions of schwa and various consonants (in particular /r/, /t/ and /n/). Most substitutions and deletions (about 62-75% for the various transcriptions) occurred at word boundaries, but the absolute numbers in the KB-PTs were lower due to the cross-word pronunciation modelling inherent to the knowledge-based transcription procedure.

[Table 3] + [Table 4]

The disagreement scores obtained with the DD-PTs were much higher than the scores obtained with the CAN-PTs and the KB-PTs. This holds for both speech styles. Most discrepancies between the DD-PTs and the RTs were deletions and (a variety of) substitutions. In addition to consonant substitutions due to voicing, we observed

various consonant substitutions due to place of articulation, and vowel substitutions with schwa (and vice versa).

The proportions of disagreements in the CAN/DD-PTs and the KB/DD-PTs were lower than in the DD-PTs, but much higher than in the CAN-PTs and KB-PTs. Thus, the combination of the transcription procedures improved the DD-PTs, but deteriorated the CAN-PTs and KB-PTs. The CAN/DD-PTs and the KB/DD-PTs comprised twice as many substitutions as the CAN-PTs and the KB-PTs. Whereas the highly increased number of deletions in the CAN/DD-PT of the telephone dialogues (as compared to the CAN-PT) coincided with a - be it moderate - decrease of insertion errors, the CAN/DD-PT of the read speech showed even more insertions than the CAN-PT.

We used decision trees to narrow the gap between the ten aforementioned APTs (5 procedures x 2 speech styles) and the reference transcriptions. In nine out of ten cases, the use of decision trees improved the original transcriptions; only the $[\text{DD-PT}]_d$ of the telephone dialogues comprised more disagreements than the original DD-PT. The magnitude of the improvements differed substantially, though. The improvements were negligible for the DD-PTs, somewhat larger for the APTs that emerged from the combined procedures, and most outspoken for the CAN-PTs and the KB-PTs. This is quite remarkable, because one would expect the biggest improvement for the worst baseline. Our results show the opposite. For both speech styles, the $[\text{CAN-PT}]_d$ proved most similar to the RT. The $[\text{KB-PTs}]_d$ were slightly worse. The $[\text{CAN-PTs}]_d$ comprised on average 20.5% less mismatches with the RTs than the original CAN-PTs, which is a significant improvement at a 99% confidence level. Likewise, we observed on average 14.1% less mismatches in the $[\text{KB-PTs}]_d$ than in the original KB-PTs ($p < .01$).

5. DISCUSSION

5.1. Reflections on the evaluation procedure

We assessed our automatic phonetic transcriptions in terms of their resemblance to reference transcriptions that were based on example transcriptions. Previous studies have shown that the use of an example transcription for verification speeds up the transcription process (relative to manual transcription from scratch), but that it also

tempts human experts into adhering to the example transcription despite contradicting acoustic cues in the speech signal. Demuynck et al. (2004), for example, reported cases where human transcribers preferred not to change the example transcription in the presence of contradicting acoustic cues, and cases where transcribers left phones in the example transcription that could not be aligned with a specific portion of the speech signal.

This observation is important for our study, because it implies that our RTs may have been biased towards the canonical example transcriptions they were based upon. Considering that both the RTs and the KB-PTs were based on the CAN-PTs, it is reasonable to assume that the quality assessments of the CAN-PTs and the KB-PTs have been positively biased in our experiments. At the same time, the assessment of the DD-PTs may have been negatively biased, since these transcriptions were only based on the signal. Most probably, the transcribers' instruction to accept the example transcription as long as the acoustic evidence did not unequivocally suggest another transcription has contributed to the discrepancies between the DD-PTs and the RTs.

5.2. On the suitability of a low-cost transcription procedure for the automatic phonetic transcription of large speech corpora

5.2.1. Generic transcription procedures

Lexicon lookup (canonical) transcription procedure

The quality of the CAN-PT of the telephone dialogues (18% disagreement) was rather good as compared to human inter-labeller disagreement scores reported in the literature. Greenberg et al. (1996), for example, reported 25 to 20% disagreements between human transcriptions of American English telephone conversations, and Kipp et al. (1997) reported 21.2 to 17.4% inter-labeller disagreements between human transcriptions of German spontaneous speech. Binnenpoorte (2006), assessing the inter-labeller disagreements between manually verified phonetic transcriptions of spontaneous conversations in the Spoken Dutch Corpus, reported between 14 and 11.4% disagreements. The proportion of disagreements between the CAN-PT and the human RT (10.1% disagreement) of the read speech was still relatively high as compared to human inter-labeller disagreement scores reported in the literature. Kipp et al. (1996) reported 6.9 to 5.6% disagreements between human transcriptions of

German read speech, and Binnenpoorte (2006) reported 6.2 to 3.7% inter-labeller disagreements between manually verified transcriptions of Dutch read speech.

Considering the very low cost of CAN-PTs, and considering the similarities with previously published human inter-labeller disagreement scores, it appears that the production of CAN-PTs is a viable option in transcription projects in which limited resources are available. However, we still found a high proportion of substitutions and insertions at word boundaries. This is not surprising, because cross-word phonological processes were not accounted for in the CAN-PTs.

Data-driven transcription procedure

Constrained phone recognition proved suboptimal to approximate the manually verified phonetic transcriptions. The high number and the wide variety of substitutions suggest that the use of phonotactic models did not sufficiently tune our CSR towards the RTs. The high number of deletions implies that, in spite of extensive tuning of the phone insertion penalty, our CSR had too large a preference for transcriptions containing fewer symbols. Close inspection of the DD-PTs suggested that many deletions were systematic, but unlikely. Thus, it is not likely that the discrepancy between the DD-PTs and the RTs are fully due to a bias towards canonical representations of the human transcribers. Kessens & Strik (2004) observed that the use of shorter acoustic models for sounds like /@/ (e.g. two-state models that can be aligned to signal segments as short as 20 ms instead of the conventional three-state models that cover at least 30 ms of the speech signal) may reduce this tendency for deletions, but the diverse nature of the deletions in our results makes a substantial reduction of deletions through the mere use of shorter acoustic models rather unlikely.

Knowledge-based transcription procedure

The use of linguistic knowledge to model pronunciation variation at the lexical level improved the quality of the transcription of the telephone dialogues, but it deteriorated the transcription of the read speech. The availability of pronunciation variants is probably more beneficial for the transcription of spontaneous speech, since more spontaneous speech is often characterised by a larger degree of pronunciation variation (Goddijn & Binnenpoorte, 2003). Most probably, the CSR often preferred non-canonical variants for the transcription of the read speech, while the human

transcribers had a preference for the canonical example transcription, according to their instruction.

The knowledge-based multiple pronunciation lexicon of the telephone dialogues comprised on average 1.39 pronunciation variants per word, the lexicon of the read speech 1.47 variants per word. The higher average number of pronunciation variants in the read speech lexicon can be explained by the fact that the pronunciation variants of both speech styles were derived from the canonical transcriptions by applying a fixed set of rules. Since the words in the telephone dialogues were shorter than the words in the read speech (an average of 3.3 vs. 4.1 canonical phones per word in the telephone dialogues and the read speech), the canonical transcription of the telephone dialogues was less susceptible to the application of rewrite rules than the canonical transcription of the read speech.

In order to estimate the possible impact of the application of knowledge-based rewrite rules on the CAN-PTs, we computed the maximum and minimum accuracy that could be obtained with the knowledge-based recognition lexicons for read and spontaneous speech. For every chunk, every combination of the pronunciations of the words was aligned with the RT, and the highest and the lowest disagreement measures were retained. We found that the knowledge-based recognition lexicon of the telephone dialogues was able to provide KB-PTs of which 22.6 to only 13.2% of the phones differed from the RT. The knowledge-based lexicon of the read speech was able to provide KB-PTs of which 16.3 to only 7.4% of the phones differed from the RT. The eventual quality of the KB-PTs (17.3% and 10.9% disagreement for the telephone dialogues and the read speech, respectively) shows that there was still room for improvement; the acoustic models of our CSR often opted for suboptimal transcriptions.

5.2.2. Combinations of generic transcription procedures

The blend of data-driven pronunciation variants with canonical or knowledge-based variants into CAN/DD and KB/DD lexicons allowed our CSR to better approximate human transcription behaviour than through constrained phone recognition alone, but the combination of the procedures did not outperform the canonical lexicon lookup (CAN-PT) and the knowledge-based transcription procedure (KB-PT). The improvement with regard to the original DD-PTs must have been due to the fact that the CSR could now only select phoneme sequences from the multiple pronunciation

lexicons. This constituted a substantial bias in the direction of the RTs as compared to the constrained phone recognition through which the DD-PTs were generated. The fact that the CAN/DD and KB/DD transcriptions suffered from the addition of the signal-based pronunciation variants could be due to the added variants closer resembling the signal than the canonical representations did (and the representations derived by means of phonological rules), whereas the transcribers adhered to the canonical example transcriptions. We conclude that the mere combination of signal-based and canonical or knowledge-based lexical pronunciation variants was not effective for approximating the manually verified phonetic transcriptions.

5.2.3. Transcription procedures with decision trees

Contrary to our expectations, the $[\text{DD-PT}]_d$ of the telephone dialogues differed more from the RT (though not significantly more, $p > .1$) than the original DD-PT. The $[\text{DD-PT}]_d$ of the read speech was only slightly (again, not significantly, $p > .1$) better than the original DD-PT. The inability of the decision trees to tune the data-driven transcriptions towards the RTs was probably due to the high degree of confusability in the recognition lexicons in the absence of reliable estimates of prior probabilities. The recognition lexicon for the telephone dialogues had an average of 9.5 variants per word, and the lexicon for the read speech an average number of 3.5 variants per word.

Note that, contrary to the pronunciation variants in the knowledge-based recognition lexicons, the pronunciation variants in the $[\text{DD-PT}]_d$ lexicons were based on the speech signal rather than on the application of phonological rewrite rules on the CAN-PT. This resulted, in particular for the $[\text{DD-PTs}]_d$ of the more spontaneous telephone dialogues, in more discrepancies with the RTs, all of which were modelled in the decision trees. Even after pruning unlikely pronunciation variants from the decision trees, the decision trees apparently still comprised enough pronunciation variants to boost the average number of pronunciation variants per word in the recognition lexicons. From experience with ASR tasks it is known that an average number of 2.5 pronunciations per word is close to the optimum in terms of word error rate (Kessens et al., 2003). It was shown that the addition of more pronunciation variants to recognition lexicons increases the risk of lexical confusability. In our study, for the purpose of automatic phonetic transcription, the CSR had to choose between highly similar alternatives. Apparently, an average of 9.5 pronunciation

variants per word in the recognition lexicon for the telephone dialogues was too high, whereas an average of 3.5 variants in the lexicon for the read speech seemed tolerable, even though it was more than the optimum of 2.5 variants previously reported for ASR.

The small improvements obtained through the use of decision trees for the enhancement of the CAN/DD-PTs and the KB/DD-PTs, as well as the large improvements obtained through the use of decision trees for the enhancement of the CAN-PTs and the KB-PTs can be explained along the same line of reasoning. The numerous discrepancies between the CAN/DD-PTs and the KB/DD-PTs on the one hand and the RTs on the other hand yielded numerous pronunciation variants in the resulting recognition lexicons (though less than in the DD-PT lexicons). The higher similarity between the original $[\text{CAN-PTs}]_d$, the $[\text{KB-PTs}]_d$ and the RTs led to fewer branches in the decision trees and fewer pronunciation variants in the resulting recognition lexicons. As a consequence, the corresponding prior probabilities of the variants were intrinsically more robust than the probabilities in the data-driven lexicons comprising more pronunciation variants per word.

Recall that we did not implement vowel reduction and deletion for the generation of the KB-PTs, and that we based our KB-PTs on canonical transcriptions without using supra-segmental information. We investigated whether the disregard of this knowledge in our knowledge-based transcription procedure made a substantial contribution to the discrepancies between the KB-PTs (and consequently also the $[\text{KB-PTs}]_d$) and the RTs. This proved not to be the case; the missing vowel rules and the reformulation of the phonological processes did not hamper the pronunciation variation modelling in the knowledge-based transcriptions procedures to any substantial degree.

We obtained our best transcriptions by means of the procedure in which our fully canonical transcriptions were tuned towards the manually verified reference transcriptions by means of pronunciation variation modelling inspired by speech processes that were attested in the reference transcriptions. Apparently, learning intra-word and cross-word phonological processes from a small sample of real transcriptions works better than predicting the results of these processes from linguistic and phonetic knowledge. It remains to be explained why the KB-PTs were a less effective starting point for the learning process. We think that this is most likely due to a canonically-oriented bias in the RTs that was so strong that no other

point of departure could close the gap. Thus, in order to approximate manually verified transcriptions resulting from the auditory verification of close-to-canonical example transcriptions (like in the Spoken Dutch Corpus), it is worthwhile learning the most obvious differences between the canonical and the reference transcriptions through the use of decision trees. One should bear in mind, though, that a canonical point of departure may be suboptimal to approximate RTs that are not based on a (similar) example transcription. This is especially true for our signal-based APTs which were essentially ignorant of the canonical representation of the material.

5.3. What about the remaining discrepancies?

The number of remaining discrepancies in the $[\text{CAN-PTs}]_d$ of the telephone dialogues (14.6% disagreement) and the read speech (8.1% disagreement) was only slightly higher than human inter-labeller disagreement scores reported in the literature. Recall that Binnenpoorte (2006) reported human inter-labeller disagreements between 14 and 11.4% on transcriptions of Dutch spontaneous conversations, and between 6.2 and 3.7% disagreements on transcriptions of Dutch read speech from the Spoken Dutch Corpus. In the context of the figures reported in Binnenpoorte (2006), a closer look at the 20 most frequent dissimilarities distinguishing our $[\text{CAN-PTs}]_d$ from the human RTs shows a comparable number of insertions and deletions, and a set of substitutions in which the mismatches between voiced and voiceless phones were dominant (see Table 5).

[Table 5]

Similar disagreements were previously observed between different human transcribers who verified the same example transcription (Binnenpoorte et al., 2003). Therefore, we believe that our automatic transcription procedures have faced the same 'mission impossible' as humans when making broad phonetic transcriptions. The limited number of phonetic symbols available forces human transcribers and machines to classify auditory observations in a continuous space into discrete categories. For observations that are close to (hypothetical) category boundaries, forced choices inevitable cause a large proportion of disagreements. Fortunately, if for some application in which phonetic transcriptions must be used independent criteria can be formulated for classifying a fricative as voiced or unvoiced (to mention one of the most volatile phonetic differences in Dutch) it is probably quite easy to train an

acoustic classifier to re-label all fricatives in the corpus according to the new criteria. Most probably, such a re-labelling will be equally advantageous for manually verified broad phonetic transcriptions, for the same reason: they also involve classifications that may not fully adhere to the newly introduced criteria. Thus, we can conclude that we found a very quick, simple and cheap transcription procedure able to approximate manually verified phonetic transcriptions of a large speech corpus by training an automatic procedure on the basis of a relatively small set of data. Our procedure applied uniformly to well-prepared and spontaneous speech. It remains to be shown that the procedure is equally effective for manual transcriptions that are made in a way that is significantly different from the procedure used in the Spoken Dutch Corpus (and in most other large speech corpora, for that matter). However, the machine learning procedure on which our approach is based seems sufficiently general and powerful to approximate different types of transcriptions, as long as learning can be initialised from a starting point that is not too far from the eventual target.

6. CONCLUSIONS

The aim of our study was to investigate whether existing automatic transcription procedures and combinations of such procedures can approximate the quality of manually verified phonetic transcriptions of speech. If such procedures would be able to do so, we would have a quick and cheap alternative to deploying human experts for the generation of the type of transcription of large speech corpora. We used ten automatic transcription procedures to generate a phonetic transcription of well-prepared speech (read-aloud texts) and of spontaneous speech (telephone dialogues) from the Spoken Dutch Corpus. The resulting transcriptions were compared to the corresponding manually verified phonetic transcriptions from the Spoken Dutch Corpus.

Our results showed that, in order to approximate the quality of the manually verified phonetic transcriptions in the Spoken Dutch Corpus, one only needs an orthographic transcription, a canonical lexicon, a small sample of manually verified phonetic transcriptions, software for the implementation of decision trees and a standard continuous speech recogniser. Our study suggests that it is sufficient to verify the phonetic transcription of only a small portion of a corpus by hand in order to automatically generate similar transcriptions for the remainder of the corpus by

means of decision trees. The best point of departure for such an automatic procedure will probably depend on the procedure by means of which the manual reference transcriptions were obtained.

ACKNOWLEDGEMENT

The work of Christophe Van Bael was funded by the Speech Technology Foundation (Stichting Spraaktechnologie), Utrecht, The Netherlands. The authors would like to thank two anonymous reviewers for their useful suggestions.

7. REFERENCES

1. Baayen, R.H., Piepenbrock, R., Gulikers, L. (1995). *The CELEX Lexical Database (Release 2).* Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania, USA.
2. Bellegarda, J.R. (2005). Unsupervised, Language-independent Grapheme-to-phoneme Conversion by Latent Analogy. In: *Speech Communication*, vol. 46/2, pp. 140-152.
3. Binnenpoorte, D., Cucchiarini, C. (2003). Phonetic Transcription of Large Speech Corpora: How to Boost Efficiency without Affecting Quality. In: *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*, Barcelona, Spain, pp. 2981-2984.
4. Binnenpoorte, D., Goddijn S.M.A., Cucchiarini, C. (2003). How to Improve Human and Machine Transcriptions of Spontaneous Speech. In: *Proceedings of the ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, Tokyo, Japan, pp. 147-150.
5. Binnenpoorte, D. (2006). *Phonetic Transcriptions of Large Speech Corpora.* Ph.D. Dissertation, Radboud University Nijmegen, the Netherlands.
6. Booij, G. (1999). *The Phonology of Dutch*. Oxford University Press, New York.
7. CELEX Lexical Database (2005). [http://www.ru.nl/celex/].
8. Demuynck, K., Laureys, T., Gillis, S. (2002). Automatic Generation of Phonetic Transcriptions for Large Speech Corpora. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Denver, USA, pp. 333-336.

9. Demuynck. K., Laureys, T., Wambacq, P., Van Compernolle, D. (2004). Automatic Phonemic Labeling and Segmentation of Spoken Dutch. In: *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Lisbon, Portugal, pp. 61-64.
10. Elffers, B, Van Bael, C., Strik, H. (2005). *ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcriptions.* Internal report, Department of Language and Speech, Radboud University Nijmegen, the Netherlands. [http://lands.let.ru.nl/literature/elffers.2005.1.pdf].
11. Geumann, A., Oppermann, D., Schaeffler, F. (1997). *The Conventions for Phonetic Transcription and Segmentation of German Used for the Munich Verbmobil Corpus*. Verbmobil Memo 129-96, University of Munich, Germany.
12. Goddijn, S.M.A., Binnenpoorte, D. (2003). Assessing Manually Corrected Broad Phonetic Transcriptions in the Spoken Dutch Corpus. In: *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)*, Barcelona, Spain, pp. 1361-1364.
13. Godfrey, J., Holliman, E., McDaniel, J. (1992). SWITCHBOARD: Telephone Speech Corpus for Research and Development. In: *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, San Francisco, USA, pp. 737-740.
14. Greenberg, S., Hollenback, J., Ellis, D. (1996). Insights into Spoken Language Gleaned from Phonetic Transcription of the Switchboard Corpus. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Philadelphia, USA, pp. S24-27.
15. Hess, W., Kohler, K.J., Tillman, H.-G. (1995). The Phondat-Verbmobil Speech Corpus. In: *Proceedings of Eurospeech*, Madrid, Spain, pp. 863-866.
16. Jande, P.A. (2005). Inducing Decision Tree Pronunciation Variation Models from Annotated Speech Data. In: *Proceedings of Interspeech*, Lisbon, Portugal, pp. 1945-1948.
17. Kessens, J.M., Wester, M., Strik, H. (1999). Improving the Performance of a Dutch CSR by Modeling Within-word and Cross-word Pronunciation Variation. In: *Speech Communication*, vol. 29, pp. 193-207.
18. Kessens, J.M., Cucchiarini, C., Strik, H. (2003). A Data-driven Method for Modeling Pronunciation Variation. In: *Speech Communication*, vol. 40/4, pp.517-534.

19. Kessens, J.M., Strik, H. (2004). On Automatic Phonetic Transcription Quality: Lower Word Error Rates Do Not Guarantee Better Transcriptions. In: *Computer, Speech and Language*, vol. 18, pp. 123-141.
20. Kipp, A., Wesenick, M.-B., Schiel, F. (1996). Automatic Detection and Segmentation of Pronunciation Variants in German Speech Corpora. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP),* Philadelphia, USA, pp. 106-109.
21. Kipp, A., Wesenick, M.-B., Schiel F. (1997). Pronunciation Modelling applied to Automatic Segmentation of Spontaneous Speech. In: *Proceedings of Eurospeech*, Rhodes, Greece, pp. 1023-1026.
22. Koskenniemi, K. (1983). *Two-level Morphology: A General Computational Model of Word-form Recognition and Production*. Technical Report Publication No. 11, Dept. of General Linguistics, University of Helsinki, Finland.
23. Maekawa, K. (2003). Corpus of Spontaneous Japanese: Its Design and Evaluation. In: *Proceedings of the ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, Tokyo, Japan.
24. Mizutani, T. and Kagoshima, T. (2005). Concatenative Speech Synthesis Based on the Plural Unit Selection and Fusion Method. In: *IEICE Transactions on Information and Systems*, vol. E88-D/11, pp. 2565-2572
25. Neri, A., Cucchiarini, C., Strik, H. (2006). Selecting segmental errors in non-native Dutch for optimal pronunciation training. In: *International Review of Applied Linguistics*, vol. 44/4, pp. 357-404.
26. Neri, A., Cucchiarini, C., Strik, H. (2007). Pronunciation training in Dutch as a second language on the basis of automatic speech recognition. To appear in: *Stem, Spraak en Taakpathologie*
27. Oostdijk, N. (2002). The Design of the Spoken Dutch Corpus. In: Peters, P., Collins, P., Smith, A. (Eds.) *New Frontiers of Corpus Research*. Rodopi, Amsterdam, pp. 105-112.
28. PAROLE lexicon. (2005). [http://ww2.tst.inl.nl].
29. Quinlan, J. R. (1993). C4.5: *Programs for Machine Learning.* San Mateo: Morgan Kaufmann, USA.
30. Referentiebestand Nederlands (RBN). (2005). [http://ww2.tst.inl.nl].
31. Riley, M., Byrne, 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*, vol. 29, pp. 209-224.

32. Saraçlar, M., Khundanpur, S (2004). Pronunciation Change in Conversational Speech and its Implications for Automatic Speech Recognition. In: *Computer, Speech and Language*, vol. 18, pp. 375-395.
33. Strik, H. (2001). Pronunciation Adaptation at the Lexical Level. In: *Proceedings of the ISCA Tutorial and Research Workshop (ITRW) 'Adaptation Methods for Speech Recognition'*, Sophia-Antipolis, France, pp. 123-131.
34. TIMIT Acoustic-Phonetic Continuous Speech Corpus (1990). National Institute of Standards and Technology Speech Disc 1-1.1, NTIS Order No. PB91-505065, 1990.
35. Tjalve, M., Huckvale, M., (2005). Pronunciation Variation Modelling using Accent Features. In: *Proceedings of Interspeech*, Lisbon, Portugal, pp.1341-1344.
36. Wang, L., Zhao, Y., Chu, M., Soong, F., Cao, Z. (2005). Phonetic Transcription Verification with Generalised Posterior Probability. In: *Proceedings of Interspeech*, Lisbon, pp. 1949-1953.
37. Wester, M. (2003). Pronunciation Modeling for ASR - Knowledge-based and Data-derived Methods. In: *Computer Speech and Language*, vol. 17/1, pp. 69-85.
38. Witten, I.H., Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd Edition. Morgan Kaufmann, San Francisco, USA.
39. Yang, Q., Martens, J.-P., (2000). Data-driven Lexical Modelling of Pronunciation Variations for ASR. In: *Proceedings of the International Conference on Spoken Language Processing (ICSLP)*, Beijing, China, pp. 417-420.
40. Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., Woodland, P. (2001). The HTK Book (for HTK version 3.2), Cambridge University Engineering Department, UK.

TABLES

Speech style		Training sets	Development sets	Evaluation sets
Read speech	# word tokens	532,451	7,940	7,940
	hh:mm:ss	44:55:59	0:40:10	0:41:39
	# distinct speakers	561	126	126
Telephone dialogues	# word tokens	263,501	6,953	6,955
	hh:mm:ss	18:20:05	0:30:02	0:29:50
	# distinct speakers	344	92	91

Table 1: Statistics of the data sets.

Comparison with RT	Read speech				Telephone dialogues			
	Subs	Dels	Ins	%dis	Subs	Dels	Ins	%dis
CAN-PT	6.3	1.2	2.6	10.1	9.1	1.1	8.1	18.3
DD-PT	16.1	7.4	3.6	27.0	26.0	18.0	3.8	47.8
KB-PT	6.3	3.1	1.5	10.9	9.0	2.5	5.8	17.3
CAN/DD-PT	13.1	2.0	4.8	19.9	21.5	6.2	7.1	34.7
KB/ DD-PT	12.8	3.1	3.6	19.5	20.5	7.8	5.4	33.7
$[\text{CAN-PT}]_d$	4.8	1.6	1.7	8.1	7.1	3.3	4.2	14.6
$[\text{DD-PT}]_d$	15.7	7.4	3.5	26.7	26.0	18.6	3.8	48.3
$[\text{KB-PT}]_d$	5.0	3.2	1.2	9.4	7.1	3.5	4.2	14.8
$[\text{CAN/DD-PT}]_d$	12.0	2.3	4.3	18.5	20.1	7.2	5.5	32.8
$[\text{KB/ DD-PT}]_d$	11.6	3.1	3.1	17.8	19.3	9.4	4.5	33.1

Table 2: Evaluation of the transcription procedures. Fewer disagreements (%dis) indicate better transcriptions and therefore better transcription procedures.

Read speech						Telephone dialogues					
Substitutions		Deletions		Insertions		Substitutions		Deletions		Insertions	
RT	CAN-PT	RT	CAN-PT	RT	CAN-PT	RT	CAN-PT	RT	CAN-PT	RT	CAN-PT
f	v			-	@	f	v	n	-	-	r
s	z			-	r	s	z			-	h
d	t			-	t	@	E			-	n
x	G			-	n	d	t			-	t
g	k			-	d	x	G				

Table 3: 10 most frequent mismatches between the CAN-PTs and the RTs.

Read speech						Telephone dialogues					
Substitutions		Deletions		Insertions		Substitutions		Deletions		Insertions	
RT	KB-PT	RT	KB-PT	RT	KB-PT	RT	KB-PT	RT	KB-PT	RT	KB-PT
f	v	@	-	-	h	f	v	@	-	-	@
s	z	n	-			s	z			-	r
@	E	r	-			d	t			-	t
x	G					x	G			-	d
d	t									-	n
t	d										

Table 4: 10 most frequent mismatches between the KB-PTs and the RTs.

Read speech						Telephone dialogues					
Substitutions		Deletions		Insertions		Substitutions		Deletions		Insertions	
RT	$[CAN-PT]_d$	RT	$[CAN-PT]_d$	RT	$[CAN-PT]_d$	RT	$[CAN-PT]_d$	RT	$[CAN-PT]_d$	RT	$[CAN-PT]_d$
v	f	@	-	-	@	d	t	@	-	-	@
s	z	r	-	-	d	z	s	r	-	-	t
g	k	n	-	-	r	v	f	n	-	-	r
d	t	h	-	-	t	g	k	h	-	-	d
t	d			-	h	@	A			-	n
G	x			-	n	G	x			-	j
@	A					A	a				
z	s					t	d				
A	a					s	z				
@	a					f	v				

Table 5: 20 most frequent mismatches between the $[CAN-PTs]_d$ and the RTs.

FIGURES

!FOR A HIGH-QUALITY VERSION OF EACH PICTURE, PLEASE REFER TO THE ATTACHED FILE!

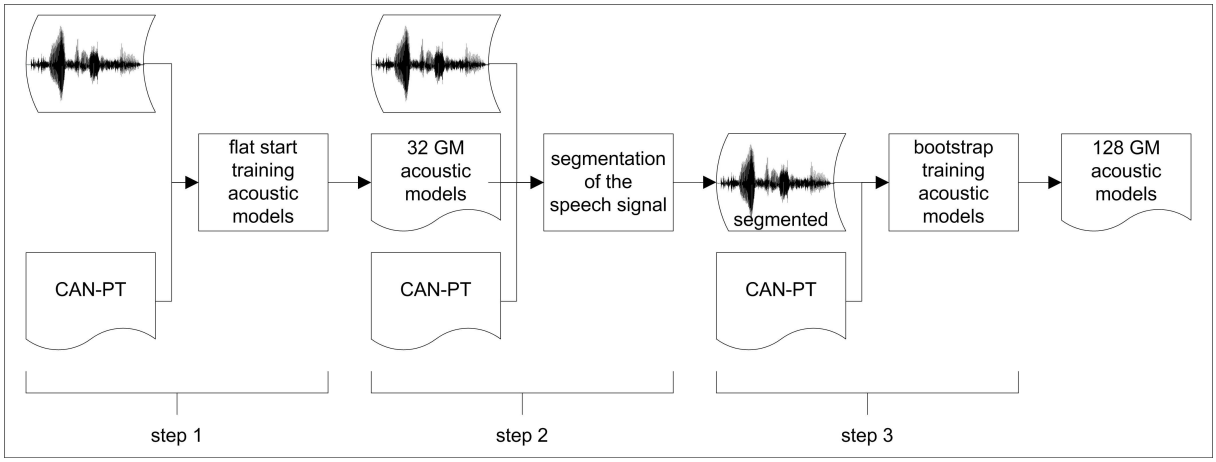


Figure 1: The procedure by means of which the acoustic models were trained.

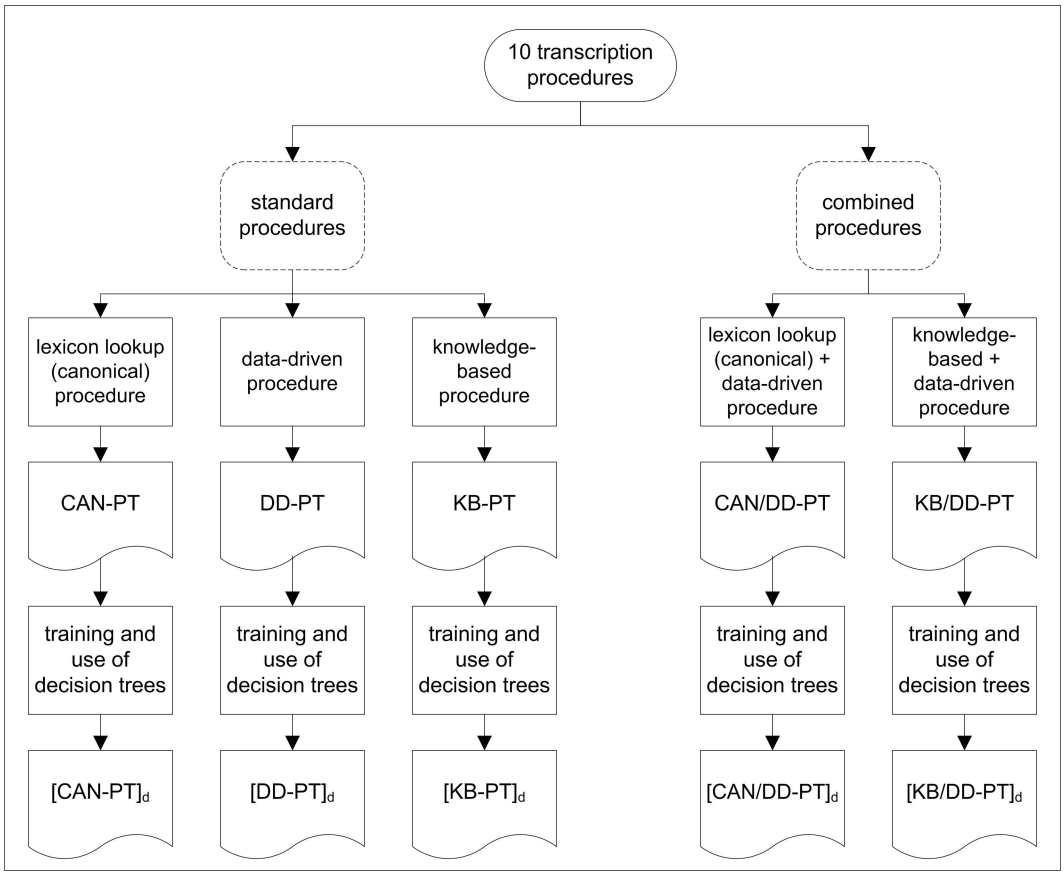


Figure 2: Overview of the ten investigated transcription procedures.

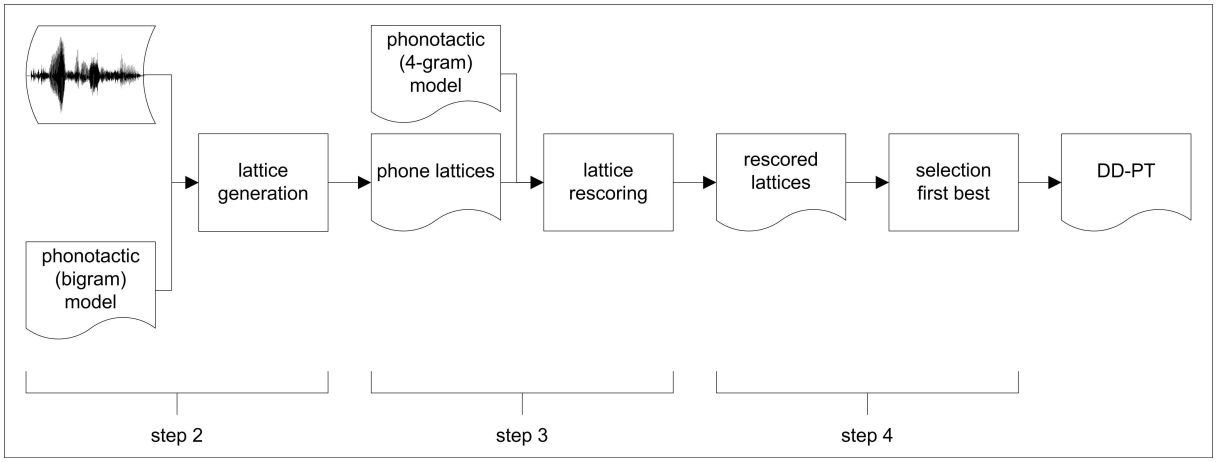


Figure 3: Data-driven phonetic transcription through constrained phone recognition (step 1 – the training of the phonotactic models – is not included).

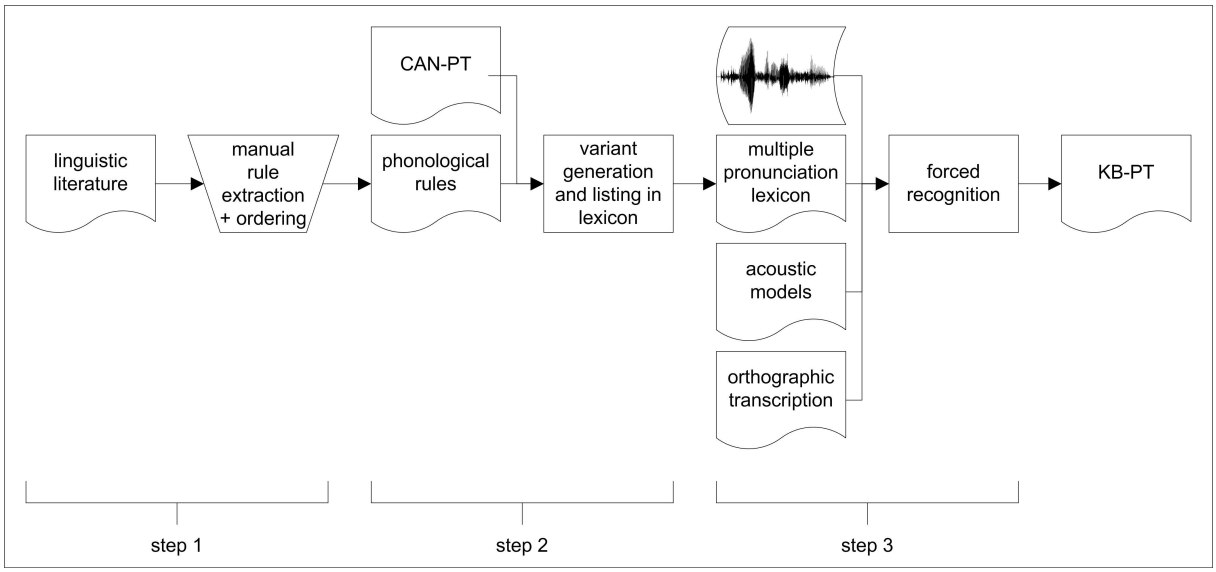


Figure 4: Knowledge-based phonetic transcription.

CAN-PT	A n	A t	d @	A p @ l t a r t
+				
DD-PT	A n	A t	d	A b @ l t a t
multiple pronunciation variants	A n	A t	d @	A p @ l t a r t
	A n	A t	d	A p @ l t a r t
	A n	A t	d @	A b @ l t a r t
	A n	A t	d	A b @ l t a r t
	A n	A t	d @	A p @ l t a t
	A n	A t	d	A p @ l t a t
	A n	A t	d @	A b @ l t a t
	A n	A t	d	A b @ l t a t

Figure 5: Generation of pronunciation variants through the alignment of two phonetic transcriptions.

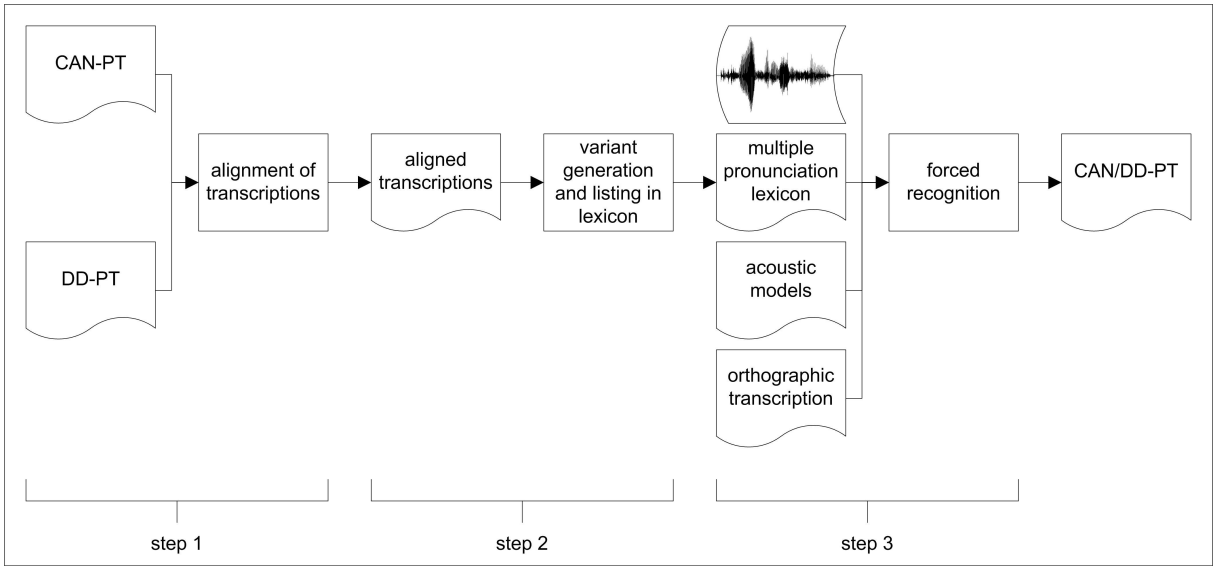


Figure 6: Combination of transcription procedures (in this case: CAN-PT and DD-PT).

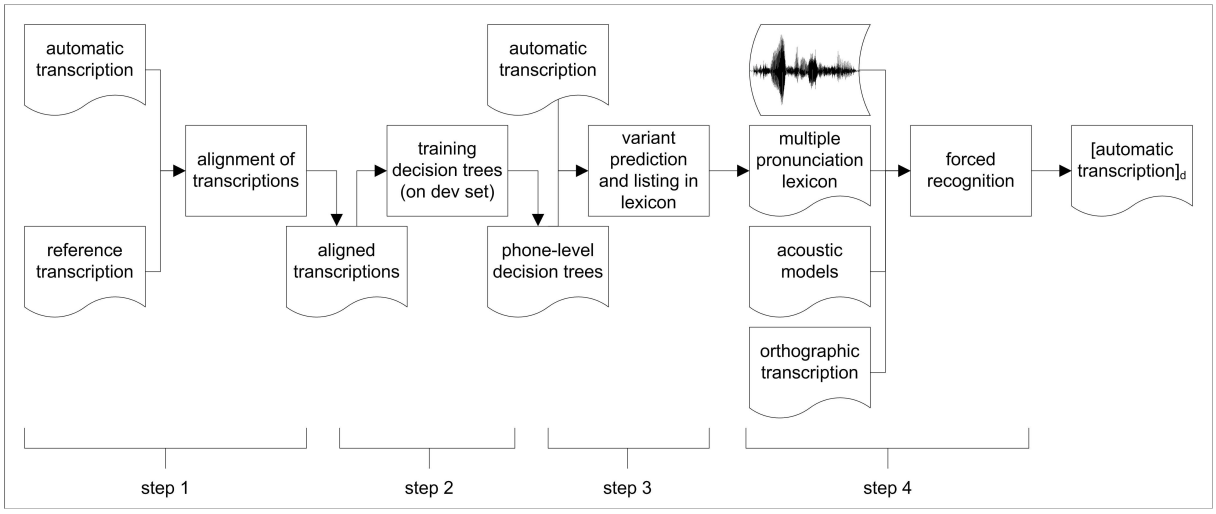
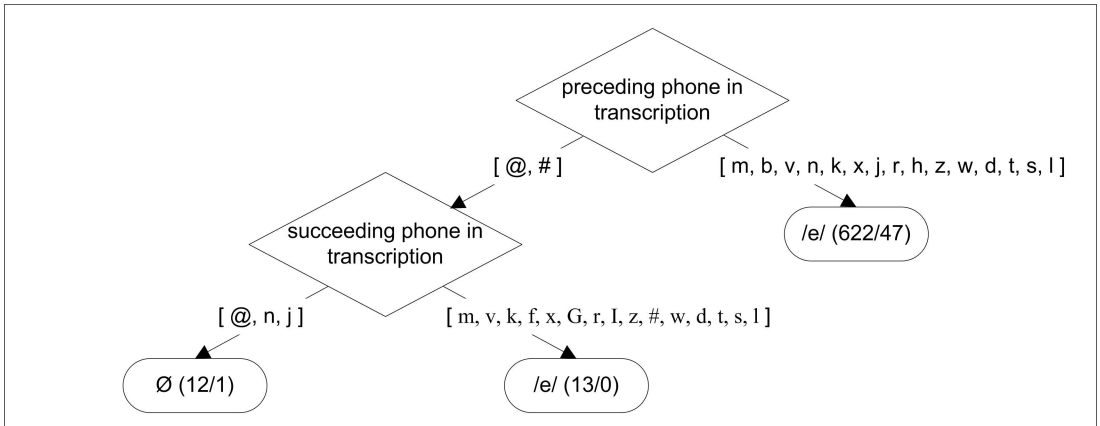


Figure 7: Automatic phonetic transcription with decision trees.



application of a trained decision tree for the phone /e/ :

orthography		canonical transcription	most likely transcription
meer	(more)	# m e r #	# m e r #
er één appel	(there one apple)	# E r # e n # A p @ l #	# E r # n # A p @ l #
eekhoorn	(squirrel)	# e k h o r n #	# e k h o r n #

Figure 8: Illustration and application of a decision tree for the phone /e/, given its left and right context phones (# = word boundary, Ø = deletion of /e/).