Automatic Phonetic Transcription of Large Speech Corpora

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

Centre for Language and Speech Technology (CLST)
Radboud University Nijmegen, the Netherlands
[c.v.bael,l.boves,h.v.d.heuvel,w.strik]@let.ru.nl

Abstract

This study is aimed at investigating whether automatic phonetic transcription procedures can approximate manual transcriptions typically delivered with contemporary large speech corpora. To this end, ten automatic procedures were used to generate a broad phonetic transcription of well-prepared speech (read-aloud texts) and spontaneous speech (telephone dialogues) from the Spoken Dutch Corpus. The resulting transcriptions were compared to manually verified phonetic transcriptions from the same corpus.

Most transcription procedures were based on lexical pronunciation variation modelling. The use of signal-based pronunciation variants prevented the approximation of the manually verified phonetic transcriptions. The use of knowledge-based pronunciation variants did not give optimal results either. A canonical transcription that, through the use of decision trees and a small sample of manually verified phonetic transcriptions, was modelled towards the target transcription, performed best. The number and the nature of the remaining disagreements with the reference transcriptions compared to inter-labeller disagreements reported in the literature.

1. Introduction

In the last decades we have witnessed the development of large multi-purpose speech corpora such as TIMIT (1990), Switchboard (Godfrey et al., 1992), Verbmobil (Hess et al., 1995), the Spoken Dutch Corpus (Oostdijk, 2002) and the Corpus of Spontaneous Japanese (Maekawa, 2003). In particular a good phonetic transcription increases the value of such corpora for scientific research and for the development of applications such as automatic speech recognition (ASR).

For some purposes (e.g. basic ASR development), a canonical phonetic representation of speech can be sufficient (Van Bael et al., 2006). However, for other purposes, such as linguistic research, a more accurate annotation of the signal is needed. For this reason, some corpora come with a manual transcription of the data (Hess et al., 1995; Greenberg et al., 1996; Oostdijk, 2002).

Despite efforts to improve the workflow of human experts, however, the human transcription process remains tedious and expensive (Cucchiarini, 1993). This explains why 'only' 4 hours of Switchboard speech were phonetically transcribed as an afterthought, and why the phonetic transcription of 'only' 1 million words of the 9-million-word Spoken Dutch Corpus was manually verified. Both for Switchboard and the Spoken Dutch Corpus, transcription costs were restricted by presenting trained students with an example transcription. The students were asked to verify this transcription rather than to transcribe from scratch (Greenberg et al. 1996; Goddijn & Binnenpoorte, 2003). Although such a check-and-correct procedure is very attractive in terms of cost reduction, it has been suggested that it may bias the resulting transcriptions towards the example transcription (Binnenpoorte, 2006). In addition, the costs involved in such a procedure are still quite substantial. Demuynck et al. (2002) reported that the manual verification process took 15 minutes for one minute of speech recorded in formal lectures and 40 minutes for one minute of spontaneous speech.

Several studies already reported the benefits of automatic phonetic transcriptions for ASR (e.g. Riley, 1999; Yang & Martens, 2000; Wester, 2003; Saraçlar & Khundanpur, 2004; Tjalve & Huckvale, 2005) and for speech synthesis (e.g. Bellegarda, 2005; Jande, 2005, Wang et al. 2005). In these studies, the phonetic transcriptions were used as tools to improve the performance of a specific system. Hence, they were not evaluated in terms of their similarity with manually verified broad phonetic transcriptions. Only a small number of studies evaluated automatic phonetic transcriptions in terms of their resemblance to manual transcriptions (e.g. Wesenick, & Kipp, 1996; Kipp, et al. 1997; Demuynck et al. 2004). These studies, however, reported the use and evaluation of only one or a limited number of similar procedures at a time. To our knowledge, no study has compared the performance of established automatic transcription procedures in terms of their ability to approximate manual transcriptions. We are also not aware of attempts to study the potential synergy of the combinatory use of existing transcription procedures.

The aim of this paper is to compare the performance of existing transcription procedures and to investigate whether combinations of these procedures lead to a better performance so that it will eventually be possible to minimise (or even eliminate) human labour in the phonetic transcription of large speech corpora, without reducing the quality of the transcriptions. Since transcriptions in large speech corpora are often designed to suit multiple purposes, our transcriptions are also intended to be multi-applicable rather than particularly suitable for one specific application such as ASR. Therefore, we will evaluate the transcriptions in terms of their similarity to a reference transcription, rather than in terms of a particular speech application. Because we want to approximate manually verified transcriptions, we will also discuss the characteristics of manual phonetic transcriptions obtained through verification of example transcriptions. Most of the procedures discussed in this article require a continuous speech recogniser to select the best fitting lexical pronunciation variant. The major difference between these procedures is the manner in which the lexical pronunciation variants were generated.

In order to ensure the applicability of the transcription procedure in situations where only limited resources are available, all procedures are designed to minimise human effort. Most procedures are based on the use of a standard continuous speech recogniser, an algorithm to align phonetic transcriptions, an orthographically transcribed

corpus, a lexicon with a canonical transcription of all words, and a manually verified transcription of a relatively small sample of the corpus. The manual transcriptions are required to tune the automatic transcription procedures and to evaluate their performance. Some procedures also require a list of phonological processes describing pronunciation variation in the language at hand. Human intervention and labour, if required at all, is limited to the compilation of such a list of phonological processes.

This paper is organised as follows. In Section 2, we introduce the corpus material used in our study. Section 3 sketches the various transcription procedures. Section 4 presents the validation of the corresponding transcriptions. In Section 5 the results are discussed, and in Section 6 general conclusions are formulated.

2. Material

2.1. Speech Material

The speech material was extracted from the Northern Dutch part of the Spoken Dutch Corpus (Oostdijk, 2002). In order not to restrict our study to one particular speech style, we selected read speech (RS) as well as spontaneous telephone dialogues (TD).

The RS was recorded at 16kHz with high-quality table-top microphones for the compilation of a library for the blind. The TD, comprising much more spontaneous speech, were recorded at 8kHz through a telephone platform. As part of the orthographic transcription process all speech material was manually segmented into chunks of approximately 3 seconds. The transcribers were instructed to put chunk boundaries in naturally occurring pauses; only if speech stretched for substantially longer than 3 seconds they had to put chunk boundaries between two words with minimal cross-word co-articulation. The experiments in this study have taken chunks as basic fragments. In order to be able to focus on phonetic transcription proper, we excluded speech chunks that, according to the orthographic transcription, contained salient non-speech sounds, broken words, unintelligible speech, overlapping and foreign speech.

The statistics of the data are presented in Table 1. The data from each speech style were divided into a training set, a development set, and an evaluation set. All data sets were mutually exclusive but they comprised similar material.

		Transcription sets		
Speech style		Training	Development	Evaluation
RS	# words	532,451	7,940	7,940
	hh:mm:ss	44:55:59	0:40:10	0:41:39
TD	# words	263,501	6,953	6,955
	hh:mm:ss	18:20:05	0:30:02	0:29:50

Table 1: Statistics of the phonetic transcriptions.

2.2. Canonical Lexicon

We used a comprehensive multi-purpose in-house lexicon that was compiled by merging various existing electronic lexical resources. The pronunciation forms in this lexicon reflected the pronunciation of words as carefully pronounced in isolation according to the obligatory word-internal phonological processes of Dutch (Booij, 1999). Each lexical entry was represented by just one standard broad phonetic transcription. Information about syllabification and syllabic stress was ignored in order to ensure the applicability of the transcription procedures to languages lacking a lexicon with such specific linguistic information.

2.3. Reference Transcription (RT)

Since we aimed at approximating the manually verified phonetic transcriptions of the Spoken Dutch Corpus, we used these transcriptions as Reference Transcriptions (RT) to tune (development set) and evaluate (evaluation set) our transcription procedures. The RTs were generated in three steps. First, a canonical transcription was generated through a lexicon-lookup procedure in a canonical lexicon. Subsequently, two phonological processes of Dutch, voice assimilation and degemination, were applied to the phones at word boundaries. This was justified by previous research indicating that these processes apply on more than 87% of the word boundaries where they can actually apply (Binnenpoorte & Cucchiarini, 2003). The enhanced transcriptions were verified and corrected by trained students. The transcribers acted according to a strict protocol instructing them to change the canonical example transcription only if they were certain that the example transcription did not correspond to the speech signal. The use of an example transcription resulted in reasonably consistent phonetic transcriptions, but the constraints imposed on the human transcribers also implied the risk of biasing the resulting transcriptions towards the canonical example transcription (Binnenpoorte, 2006).

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) based on Hidden Markov Models and 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, 1 model for silences of 30 ms or more and 1 model for the optional silence between words).

The acoustic models were trained in three stages using the CAN-PTs (cf. 3.1.1.1) of the training data. First, flat start acoustic models with 32 Gaussian mixture components were trained through 41 iterative alignments. Subsequently, these models were used to obtain more realistic segmentations of the speech material. These segmentations were then used to bootstrap a new set of acoustic models, which were retrained (through 55 iterations) to acoustic models with 128 Gaussian mixture components per state.

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

ADAPT (Elffers et al., 2005) is a dynamic programming algorithm designed to align strings of phonetic symbols according to the articulatory distance between the individual symbols. In this study, ADAPT was used to align phonetic transcriptions for the generation of lexical pronunciation variants, and to assess the quality of the automatic phonetic transcriptions through their alignment with a reference transcription.

3. Methodology

In Section 3.1, we introduce ten automatic transcription procedures to generate low-cost APTs. Section 3.2 describes the evaluation procedure with which the APTs and, consequently, the procedures were assessed.

3.1. Generation of phonetic transcriptions with different transcription procedures

Figure 1 shows ten APTs. The procedures from which they result can be divided into two categories: two procedures that did not rely on the use of a lexicon with multiple pronunciation variants per word, and eight procedures that did rely on the use of a multiple pronunciation lexicon in combination with a CSR. The latter procedures can be further categorised according to the way the pronunciation variants were generated. These variants were either based on knowledge from the literature, they were obtained by combining canonical, data-driven and knowledge-based transcriptions, or they were generated with decision trees trained on the alignment of the APTs and the RT of the development data. Most of the procedures required several parameters to be tuned to better approximate the RT of the development data. The optimal parameter settings were subsequently applied for the transcription of the data in the evaluation set.

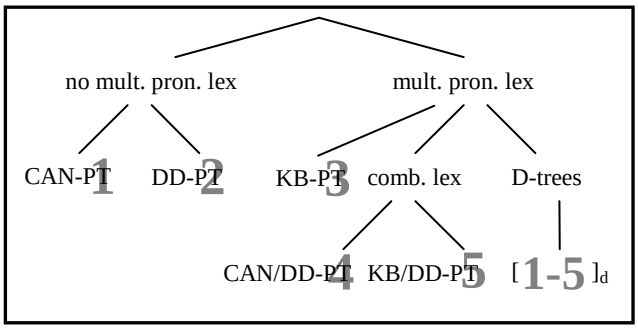

Figure 1: 10 different automatic phonetic transcriptions.

3.1.1. Transcription procedures without a multiple pronunciation lexicon

3.1.1.1. Canonical transcription (CAN-PT)

The canonical transcriptions (CAN-PTs) were generated through a lexicon look-up procedure. Cross-word assimilation and degemination were not modelled. Canonical transcriptions are easy to obtain, since many corpora feature an orthographic transcription and a canonical lexicon of the words in the corpus.

3.1.1.2. Data-driven transcription (DD-PT)

The data-driven 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 using its acoustic models and a 4-gram phonotactic model trained with the reference transcriptions of the development data in order to approximate human transcription behaviour. Transcription experiments with the data in the development set indicated that for both speech styles 4-gram models outperformed 2-gram, 3-gram, 5-gram and 6-gram models.

3.1.2. Transcription procedures with a multiple pronunciation lexicon

The transcription procedures described in this section differ in the way pronunciation variants were generated. The variants were always listed in speech style-specific multiple pronunciation lexicons. For every word, the best matching variant was selected through the use of a CSR that chose the best matching pronunciation variant from the lexicon given the orthography, the acoustic signal and a set of acoustic models. The development set was used to optimise various parameters in the individual procedures in order to optimise the selection of the lexical pronunciation variants of the words in the evaluation set.

3.1.2.1. Knowledge-based transcription (KB-PT)

In particular ASR research often draws on the literature for the extraction of linguistic knowledge with which lexical pronunciation variants can be generated (Kessens et al., 1999; Strik, 2001). We generated so-called knowledge-based transcriptions (KB-PTs) in three steps.

First, a list of 20 prominent phonological processes was compiled from the linguistic literature on the phonology of Dutch (Booij, 1999). These processes were implemented as context-dependent rewrite rules modelling both within-word and cross-word contexts in which phones from a CAN-PT can be deleted, inserted or substituted with another phone. Most of the processes identified by Booij (1999) can be described in terms of phonetic symbols or articulatory features. However, some of the processes can only be described with information about the prosodic or syllabic structure of words. Most of these processes were reformulated in terms of phonetic symbols and features, since we wanted to exclude non-segmental information (see Section 2.2). The rules were implemented conservatively to minimise the risk of over-generation. The resulting rule set comprised some rules specific for particular words in Dutch, and general phonological rules describing progressive and regressive voice assimilation, nasal assimilation, syllable-final devoicing of obstruents, t-deletion, n-deletion, r-deletion, schwa deletion, schwa epenthesis, palatalisation and degemination. The reduction and the deletion of full vowels, two prominent processes in Dutch, could not be easily formulated without the explicit use of syllabic and prosodic information.

In the second step, the phonological rewrite rules were ordered and used to generate optional pronunciation variants from the CAN-PTs of the speech chunks. The rules applied to the chunks rather than to the words in isolation to account for cross-word phenomena. The rules only applied once, and their order of application was manually optimised. Informal analysis of the resulting pronunciation variants suggested that few - if 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 the rewrite rules is needed for the transcription of other languages.

In the third step of the procedure, chunk-level pronunciation variants were listed. Since the literature did not provide numeric information on the frequency of phonological processes, the pronunciation variants did not have prior probabilities. The optimal knowledge-based transcription (KB-PT) was identified through forced recognition.

3.1.2.2. Combined transcriptions (CAN/DD-PT, KB/DD-PT)

After having generated the CAN-PTs, DD-PTs and KB-PTs, these transcriptions were combined to obtain new transcriptions. This time lexical pronunciation variants were generated through the alignment of two APTs at a time. Since the KB-PTs were based on the CAN-PTs, we only combined the CAN-PT with the DD-PT (CAN/DD-PT) and the KB-PT with the DD-PT (KB/DD-PT). Figure 2 illustrates how different pronunciation variants were generated through the alignment of the phones in the CAN-PT and the DD-PT.

CAN-PT: d @	A p @ l t a r t
	+
DD-PT: d -	A b @ l t a - t

Multiple pronunciation variants in CAN/DD-PT :

d @	A p @ l t a r t
d	A p @ l t a r t
d @	A b @ l t a r t
d	A b @ l t a r t
d @	A p @ l t a t
d	A p @ l t a t
d @	A b @ l t a t
d	A b @ l t a t

Figure 2: Generation of pronunciation variants through the alignment of two phonetic transcriptions.

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

3.1.2.3. Phonetic transcription with decision trees

The use of DD transcription procedures can result in too many, too few or very unlikely lexical pronunciation variants (Wester, 2003). In ASR research, the use of decision trees defining plausible alternatives for a phone given its context phones has often reduced the number of unlikely pronunciation variants and optimised the number of plausible pronunciation variants in recognition lexicons (Riley, 1999; Wester, 2003). We generated decision trees with the C4.5 algorithm (Quinlan, 1993), provided with the Weka package (Witten & Frank, 2005). The procedure pursued to successively improve the CAN-PTs, DD-PTs, KB-PTs, CAN/DD-PTs and KB/DD-PTs comprised four steps.

First, the APT (each of the aforementioned transcriptions consecutively) and the RT of the development data were aligned. Second, all the phones and their context phones in the APT were enumerated. The size of these "phonetic windows" was limited to three phones: the core phone, one preceding and one succeeding phone. The correspondences of the phones in the APT and the RT and the frequencies of these correspondences were used to estimate:

$$P(RT\_phone|APT\_phone,APT\_context\_phones) \quad (1)$$

i.e. the probability of a phone in the reference transcription given a particular phonetic window in the APT. In the third step of the procedure, the resulting decision trees were used to generate likely pronunciation variants for the APT of the unseen evaluation data. The decision trees were now used to predict:

$$P(pron\_variants|APT\_phone,APT\_context\_phones) \quad (2)$$

i.e. the probability of a phone with optional pronunciation variants given a particular phonetic window in the APT. All pronunciation variants with a probability lower than 0.1 were ignored in order to reduce the number of pronunciation variants and, more importantly, to prune unlikely pronunciation variants originating from idiosyncrasies in the original APT.

In the fourth and final step of the procedure, the pronunciation variants were listed in a multiple pronunciation lexicon. The probabilities of the variants were normalised so that the probabilities of all variants of a word added up to 1. Finally, our CSR selected the most likely pronunciation variant for every word in the orthography. The consecutive application of decision tree expansion to the CAN-PTs, DD-PTs, KB-PTs, CAN/DD-PTs and KB/DD-PTs resulted in five new transcriptions hereafter referred to as $[CAN\text{-}PT]_d$, $[DD\text{-}PT]_d$, $[KB\text{-}PT]_d$, $[CAN/DD\text{-}PT]_d$ and $[KB/DD\text{-}PT]_d$.

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 human RT. The comparison was conducted with ADAPT (Elffers et al., 2005). The disagreement metric was formalised as:

$$\%disagreement = \left(\frac{Sub + Del + Ins}{N} \right) * 100 \quad (3)$$

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*). 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 triggered by the use of the different transcription procedures.

4. Results

The figures in Table 2 describe 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 telephone dialogues and the read speech. In addition, the statistics of the substitutions (sub), deletions (del) and insertions (ins) are presented to provide basic insight in the nature of the disagreements.

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

Table 2: Comparison of APTs and human RTs. Fewer disagreements indicate better APTs.

The proportions of disagreements observed in the CAN-PTs and the KB-PTs were significantly different from each other ($p < .01$). The CAN-PT of the read speech was more similar to the RT than the KB-PT (Δ = 6.3% rel.) while the opposite held for the telephone dialogues (Δ = 5.9% rel.). The proportion of substitutions was about equal for the CAN-PTs and the KB-PTs. Most mismatches in the CAN-PTs were due to substitutions and insertions. There were more deletions than insertions in the KB-PT of the read speech, but there were fewer deletions than insertions in the KB-PT of the telephone dialogues. Detailed analysis of the aligned transcriptions showed that most frequent mismatches in the CAN-PTs and the KB-PTs of the two speech styles were due to voiced/unvoiced classifications of obstruents, and 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 cross-word pronunciation modelling.

The disagreement scores obtained for the DD-PTs were much higher than the scores for the CAN-PTs and the KB-PTs. This holds for both speech styles. Most discrepancies between the DD-PTs and the RTs were substitutions and deletions. When compared to the CAN-PTs and the KB-PTs, in particular the high proportion of deletions and the wide variety of substitutions were striking. Not only did we observe consonant substitutions due to voicing, we also observed various consonant substitutions due to place of articulation, and vowel substitutions with schwa (and vice versa).

The proportion of disagreements in the CAN/DD-PTs and the KB/DD-PTs was lower than in the DD-PTs, but the individual CAN-PTs and KB-PTs resembled the RT better than the CAN/DD-PTs and the KB/DD-PTs. The CAN/DD-PTs and the KB/DD-PTs comprised twice as many substitutions and even more deletions than the CAN-PTs and the KB-PTs. Whereas the increased number of deletions in the CAN/DD-PT of the telephone dialogues 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.

Decision trees were applied to the ten aforementioned APTs (5 procedures x 2 speech styles). In nine out of ten cases, the application of decision trees improved the original transcriptions; only the $[DD\text{-}PT]_d$ of the telephone dialogues comprised more disagreements than the original DD-PT. The magnitude of the improvements differed substantially, though. The differences were negligible for the DD-PTs, somewhat larger for the APTs emerging from the combined procedures, and most outspoken for the CAN-PTs and KB-PTs. For both speech styles, the $[CAN\text{-}PT]_d$ proved most similar to the RT. The $[KB\text{-}PTs]_d$ were slightly worse. The $[CAN\text{-}PTs]_d$ comprised on average 20.5% fewer 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% fewer mismatches in the $[KB\text{-}PTs]_d$ than in the original KB-PTs ($p < .01$).

5. Discussion

5.1. Reflections on the evaluation procedure

In this study, the reference transcriptions 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 experts preferred not to change the example transcription in the presence of contradicting acoustic cues, and cases where human experts approved phones in the example transcription that had no trace in the signal.

This observation is important for our study, since our RTs may have been biased towards the canonical example transcription they were based on. Considering that both the RTs and the KB-PTs were based on the CAN-PTs, the quality assessment of the CAN-PTs and the KB-PTs may have been positively biased. Consequently, the assessment of the DD-PTs may have been negatively biased, since the DD-PTs were based on the signal. Their assessment may have suffered from the human tendency to accept the canonical example transcription irrespective of the information in the acoustic signal (most probably because the human transcribers were instructed to change the example transcription only in case of obvious discrepancies).

In corpus creation projects, however, manually verified phonetic transcriptions are often preferred over automatic phonetic transcriptions. Therefore, in the light of the phonetic transcription of large speech corpora, our automatic procedures were tuned towards and evaluated in terms of this type of transcription.

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

5.2.1. Canonical transcription

The quality of the CAN-PT of the telephone dialogues (18% disagreement) already compared favourably to human inter-labeller disagreement scores reported in the literature. Greenberg et al. (1996), for example, reported 25 to 20% disagreements between manual transcriptions of American English telephone conversations, and Kipp et al. (1997) reported 21.2 to 17.4% inter-labeller disagreements between manual transcriptions of German spontaneous speech. Binnenpoorte (2006), however, reported better results: from 14 to 11.4% disagreements between manual transcriptions of Dutch spontaneous speech. The proportion disagreement between the CAN-PT and the human RT (10.1% disagreement) of the read speech was not yet at the same level as 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% disagreements between human transcriptions of Dutch read speech.

The apparent contradiction that the quality of the CAN-PT of the telephone dialogues already compared well to published human inter-labeller disagreement scores, whereas the CAN-PT of the read speech did not, may be explained by the different degrees of spontaneity in the speech samples. There is a higher chance for human inter-labeller disagreement in transcriptions of spontaneous than of well-prepared speech, since human transcribers have to transcribe or verify more phonological processes as speech becomes more spontaneous (Binnenpoorte et al. 2003). Nevertheless, considering the trade-off between overall transcription quality and the time and expenses involved in the human transcription and verification process, and considering the similarities with previously published human inter-labeller disagreement scores, we can conclude that the CAN-PTs were of a satisfactory quality. However, the high proportion of substitutions and insertions at word boundaries still implied the necessity of pronunciation variation modelling to better resemble the RT.

5.2.2. Data-driven transcription

Constrained phone recognition proved suboptimal for the generation of the targeted type of transcriptions. The high number and the wide variety of substitutions suggest that the use of a phonotactic model did not sufficiently tune our CSR towards the RT. 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. An informal inspection of the DD-PTs revealed that many deletions were unlikely, thus ruling out the possibility that the CSR analysed the signal more accurately than the human experts did. Kessens & Strik (2004) observed that the use of shorter acoustic models (e.g. using 20 ms models instead of 30 ms models) may reduce this tendency for deletions, but the diverse nature of the deletions in our study makes a substantial reduction of deletions through the mere use of different acoustic models rather unlikely.

5.2.3. Knowledge-based transcription

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. This was probably due to the different degree of spontaneity in the two speech styles; the availability of pronunciation variants is probably more beneficial for the transcription of spontaneous speech, since more spontaneous speech comprises more pronunciation variation than well-prepared speech (Goddijn & Binnenpoorte, 2003). Most probably, the CSR preferred non-canonical variants in the read speech where the human transcribers adhered to the canonical example.

The knowledge-based recognition lexicon of the telephone dialogues comprised on average 1.39 pronunciation variants per lexeme, the lexicon of the read speech 1.47 variants per lexeme. The higher average number of pronunciation variants in the read speech lexicon is not contradictory, since the pronunciation variants of both speech styles were based on the canonical transcription, and not on the actual speech signal (which would, most probably, have highlighted more pronunciation variation in the telephone dialogues than in the read speech). Moreover, 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, resp.), the canonical transcription of the telephone dialogues was less susceptible to the application of rewrite rules than the CAN-PT of the read speech.

In order to estimate the possible impact of the application of KB rewrite rules on the CAN-PTs, we computed the maximum and minimum accuracy that could be obtained with the two KB recognition lexicons. For every chunk, every combination of the pronunciations of the words was consecutively aligned with the RT, and the highest and the lowest disagreement measures were retained. We found that the KB recognition lexicon of the telephone dialogues was able to provide KB-PTs of which 22.6 to 13.2% phones differed from the RT. The KB lexicon of the read speech was able to provide KB-PTs of which 16.3 to 7.4% 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, but that the acoustic models of our CSR often opted for suboptimal transcriptions. In this respect, the use of acoustic models trained on a KB-PT instead of a CAN-PT might have improved the selection of pronunciation variants.

5.2.4. Combined transcriptions

The blend of DD pronunciation variants with canonical or KB 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 and the KB transcription procedure. The DD-PT benefited from the blend with the canonical and the KB pronunciation variants, while the influence of DD pronunciation variants increased the number of discrepancies between the resulting transcriptions and the RTs (as compared to the original CAN-PTs and KB-PTs).

5.2.5. Phonetic transcription with decision trees

Contrary to our expectations, the $[DD\text{-}PT]_d$ of the telephone dialogues comprised more (though not significantly more, $p > .1$) mismatches than the original DD-PT. The $[DD\text{-}PT]_d$ of the read speech was only slightly (again, not significantly, $p > .1$) better than the original DD-PT. This was probably due to the increased confusability in the recognition lexicons. The size of the lexicons had grown to an average of 9.5 variants per word in the recognition lexicon for the telephone dialogues, and an average number of 3.5 variants per word in the lexicon for the read speech. Note that, contrary to the pronunciation variants in the KB recognition lexicons, the pronunciation variants in the $[DD\text{-}PT]_d$ 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 $[DD\text{-}PTs]_d$ 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 pollute the recognition lexicon.

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 through the same line of reasoning. The numerous discrepancies between the CAN/DD-PTs and the KB/DD-PTs and the RTs yielded numerous pronunciation variants in the resulting recognition lexicons (though less than in the DD-PT lexicons). The higher similarity between the original $[CAN\text{-}PT]_d$, the $[KB\text{-}PTs]_d$ and the RTs, led to fewer branches in the decision trees and fewer pronunciation variants in the resulting recognition lexicons. Moreover, the corresponding lexical probabilities were intrinsically more robust than the probabilities in the DD lexicons comprising more pronunciation variants per lexeme. Since the $[CAN\text{-}PTs]_d$ were better than the $[KB\text{-}PTs]_d$ of both speech styles, and since informal inspection of the rules seems to suggest that the KB-PTs and the $[KB\text{-}PTs]_d$ could not be drastically improved through the modelling of vowel reduction and vowel deletion, we conclude that prior knowledge about the phonological processes of a language, and the subsequent implementation of knowledge-based phonological rules are not necessary to approximate the quality of manually verified phonetic transcriptions of large speech corpora. Instead, the use of decision trees and a small sample of manually verified phonetic transcriptions suffice to make canonical transcriptions approximate human transcription behaviour.

5.3. What about the remaining discrepancies?

The number of remaining discrepancies in the $[CAN\text{-}PTs]_d$ 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 speech, and between 6.2 and 3.7% disagreements on transcriptions of Dutch read speech. A closer look at the 20 most frequent dissimilarities distinguishing the $[CAN\text{-}PTs]_d$ 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. Similar differences were observed between manual transcriptions that were based on the same example transcription (Binnenpoorte et al., 2003). The remaining mismatches can be largely attributed to the very nature of human transcription behaviour. Varying disagreement scores like the ones reported in Binnenpoorte et al. (2003) seem to suggest that it is intrinsically very hard, if not impossible, to model the often whimsical human transcription behaviour with one automatic transcription procedure. Therefore, we are inclined to believe that we should not try to further model the inconsistencies in manual transcriptions of speech, and we conclude that we found a very quick, simple and cheap transcription procedure approximating human transcription behaviour for the transcription of large speech samples. Our procedure uniformly applies to well-prepared and spontaneous speech.

6. Conclusions

The aim of our study was to find an automatic transcription procedure to substitute human efforts in the phonetic transcription of large speech corpora whilst ensuring high transcription quality. To this end, ten automatic transcription procedures were used to generate a phonetic transcription of spontaneous speech (telephone dialogues) and well-prepared speech (read-aloud texts). The resulting transcriptions were compared to a manually verified phonetic transcription, since this kind of transcription is often preferred in corpus design projects.

An analysis of the discrepancies between the different transcriptions and the reference transcription showed that purely data-driven transcription procedures or procedures partially relying on data-driven input could not approximate the human reference transcription. Much better results were obtained by implementing phonological knowledge from the linguistic literature. The best results, however, were obtained by expanding canonical transcriptions with decision trees trained on the alignment of canonical transcriptions and manually verified phonetic transcriptions. In fact, our results show that 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 are sufficient to approximate human transcription quality in projects aimed at generating broad phonetic transcriptions of large speech corpora.

Our procedures uniformly applied to well-prepared and spontaneous speech. Hence, we believe that the performance of our procedures will generalise to other speech corpora, provided that the emerging automatic phonetic transcriptions are evaluated in terms of a similar reference transcription, viz. a manually verified automatic phonetic transcription of speech.

Acknowledgement

The work of Christophe Van Bael was funded by the Speech Technology Foundation (Stichting Spraaktechnologie, Utrecht, The Netherlands).

References

Bellegarda, J.R. (2005). Unsupervised, language-independent grapheme-to-phoneme conversion by latent analogy. In: *Speech Communication*, vol. 46/2, pp. 140-152.

Binnenpoorte, C., Goddijn, S.M.A., 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.

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

Binnenpoorte, D., (2006). *Phonetic transcription of large speech corpora*. Ph.D. thesis, Radboud University Nijmegen, the Netherlands.

Booij, G. (1999). *The phonology of Dutch*. Oxford University Press, New York.

Cucchiarini C. (1993). *Phonetic transcription: a methodological and empirical study*. Ph.D. thesis, University of Nijmegen.

Demuynck, K., Laureys, T., Gillis, S. (2002). Automatic generation of phonetic transcriptions for large speech corpora. In: *Proceedings of International Conference on Spoken Language Processing (ICSLP)*, Denver, USA, pp. 333-336.

Demuynck. K., Laureys, T., Wambacq, P., Van Compernolle, D. (2004). Automatic phonemic labeling and segmentation of spoken Dutch. In: *Proceedings of LREC*, Lisbon, Portugal, pp. 61-64.

Elffers, B, Van Bael, C., Strik, H. (2005). *ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcriptions*. Internal report, CLST, Radboud University Nijmegen. http://lands.let.ru.nl/literature/elffers.2005.1.pdf.

Godfrey, J., Holliman, E. and McDaniel, J. (1992) SWITCHBOARD: Telephone speech corpus for research and development. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, San Francisco, USA, pp. 517-520.

Goddijn, S.M.A. & Binnenpoorte, D. (2003). Assessing Manually Corrected Broad Phonetic Transcriptions in the Spoken Dutch Corpus. In: *Proceedings of ICPhS*, Barcelona, Spain, pp. 1361-1364.

Greenberg, S., Hollenback, J. and 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.

Hess, W., Kohler, K.J., Tillman, H.-G. (1995) The Phondat-Verbmobil speech corpus. In: *Proceedings of Eurospeech*, Madrid, Spain, pp. 863-866.

Jande, P.A. (2005). Inducing Decision Tree Pronunciation Variation Models from Annotated Speech Data. In: *Proceedings of Interspeech*, Lisbon, Portugal, pp. 1945-1948.

Kessens, J.M., Wester, M., Strik, H. (1999). Improving the performance of a Dutch CSR by modelling within-word and cross-word pronunciation variation. In: *Speech Communication*, vol. 29, pp. 193-207.

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(2), pp. 123-141.

Kipp, A., Wesenick, M.-B., Schiel F. (1996) Automatic detection and segmentation of pronunciation variants in German speech corpora. In: *Proceedings of ICSLP*, Philadelphia, USA, pp. 106-109.

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.

Koskenniemi, K. (1983) *Two-level morphology: A general computational model of word-form recognition and production*. Tech. Rep. Publication No. 11, Dept. of General Linguistics, University of Helsinki.

Maekawa, K. (2003). Corpus of Spontaneous Japanese: Its design and evaluation. In: *Proceedings of ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR)*, Tokyo, Japan.

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.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

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

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.

Strik, H. (2001). Pronunciation adaptation at the lexical level. In: *Proceedings of the ISCA Tutorial & Research Workshop (ITRW) 'Adaptation Methods for Speech Recognition'*, Sophia-Antipolis, France, pp. 123-131.

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

Tjalve, M., Huckvale, M., (2005). Pronunciation variation modelling using accent features. In: *Proceedings of Interspeech*, Lisbon, Portugal, pp.1341-1344.

Van Bael, C., Van den Heuvel, H., Strik, H. (2006). Validation of phonetic transcriptions in the context of automatic speech recognition. Submitted to: *Language Resources and Evaluation.*

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.

Wesenick, M.-B., Kipp, A. (1996) Estimating the quality of phonetic transcriptions and segmentations of speech signals. In: *Proceedings of ICSLP*, Philadelphia, USA, pp. 129-132.

Wester, M. (2003). Pronunciation modeling for ASR - knowledge-based and data-derived methods. In: *Computer Speech & Language*, vol. 17/1, pp. 69-85.

Witten, I.H., Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, USA.

Yang, Q., Martens, J.-P., (2000). Data-driven lexical modelling of pronunciation variations for ASR. In: *Proceedings of ICSLP*, Beijing, China, pp. 417-420.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., Woodland, P. (2001). *The HTK book (for HTK version 3.1)*, Cambridge University Engineering Department.