Automatic Phonetic Transcription of Large Speech Corpora: A Comparative Study

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 investigates whether automatic transcription procedures can approximate manual phonetic transcriptions typically delivered with contemporary large speech corpora. We used ten automatic procedures to generate a broad phonetic transcription of well-prepared speech (read-aloud texts) and spontaneous speech (telephone dialogues). The resulting transcriptions were compared to manually verified phonetic transcriptions. We found that the quality of this type of transcription can be approximated by a fairly simple and cost-effective procedure.

Index Terms: automatic phonetic transcription, speech corpora

1. INTRODUCTION

The usability of large speech corpora depends on the availability of appropriate annotations of the data. In particular a good phonetic transcription increases the value of a corpus for scientific research and for the development of applications such as automatic speech recognition (ASR). Since manual transcriptions have proven to be time-consuming, expensive, and, moreover, prone to inconsistencies, automatic procedures may offer a quicker, cheaper and more consistent alternative, especially when large amounts of speech are to be transcribed.

Several studies already reported the benefits of automatic phonetic transcriptions (APTs) for the development of ASR systems [1], and for the improvement of speech synthesis systems [2]. In these studies, the transcriptions were used as tools to improve the performance of a specific system. Hence, they were not evaluated in terms of their similarity with a human reference transcription (RT). Studies reporting such evaluations, however, typically described the use and evaluation of one or a limited number of similar procedures at a time. To our knowledge, no previous study has systematically compared the performance of different transcription procedures in terms of their ability to approximate a human RT. Neither do we know of any attempts to study the potential synergy of the combined use of existing transcription procedures.

The aim of this study is two-fold: we will compare the applicability of established automatic procedures in large-scale transcription projects, and we will investigate whether combinations of these procedures yield a better performance. In order to ensure the applicability of the procedures in projects with limited funding, the procedures were optimised with limited resources and minimal human effort.

This paper is organised as follows. In Section 2, we introduce our material and tools. Section 3 sketches the transcription procedures. In Section 4 and 5, we evaluate and discuss their performance, and in Section 6, general conclusions are formulated.

2. MATERIAL AND TOOLS

We extracted Dutch speech from the Spoken Dutch Corpus [3]. The material was manually segmented into speech chunks of approximately 3 seconds. We adhered to this chunk-level annotation. In order to focus on phonetic transcription proper, we excluded chunks with non-speech, broken words, unintelligible, non-native and overlapping speech.

In order not to restrict our study to one speech style, we selected read speech (RS) and spontaneous telephone dialogues (TD). Per speech style, the data were divided into a training set, a development set and an evaluation set (Table 1). These data sets were formed by listing all chunks of all speakers, randomising their ordering, and extracting the subsets. This guaranteed mutually exclusive data sets (though speakers could occur in several sets) with similar material.

Table 1: Statistics of the speech material (#words).

Speech Style	Training	Development	Evaluation
RS	532,451	7,940	7,940
TD	263,501	6,953	6,955

All words were comprised in a canonical pronunciation lexicon, with only one standard broad phonetic transcription for every entry. The transcriptions reflected the obligatory word-internal phonological processes described in the literature [4]. Information about syllabification and syllabic stress was ignored to ensure the applicability of the procedures in projects lacking such specific linguistic information.

The human reference transcriptions (RTs) were extracted from the Spoken Dutch Corpus. The RTs were based on a canonical transcription enhanced to model two frequent cross-word processes in Dutch. The example transcription was manually verified by trained linguistics students who were instructed to change the example transcription only if they were absolutely sure that it did not match the acoustic data [5].

Except for the canonical transcriptions, all APTs were generated with an HMM-based continuous speech recogniser (CSR) that was implemented with HTK [6]. Our CSR used 39 gender- and context-independent, but speech style-specific acoustic models (37 phone models, 1 long silence and 1 short pause model) that were trained through a bootstrap procedure with the canonical transcriptions of the training data.

ADAPT [7] is a dynamic programming algorithm designed to align strings of phonetic symbols according to their articulatory distance. We used ADAPT to align APTs for the generation of lexical pronunciation variants, and to evaluate APTs by comparing them to the human RTs.

3. TRANSCRIPTION PROCEDURES

Figure 1 shows ten APTs. The procedures from which they result can be divided into a set of procedures that did not rely on the use of a multiple pronunciation lexicon, and a set of procedures that did rely on the use of such a lexicon. The latter set can be further categorised according to the way lexical pronunciation variants were generated. The variants were either based on knowledge from the literature, they were obtained by combining APTs, or they were generated with decision trees trained on the alignment of the aforementioned APTs and the corresponding RTs of the development data. The variants were listed in speech style-specific lexicons which were used for forced recognition.

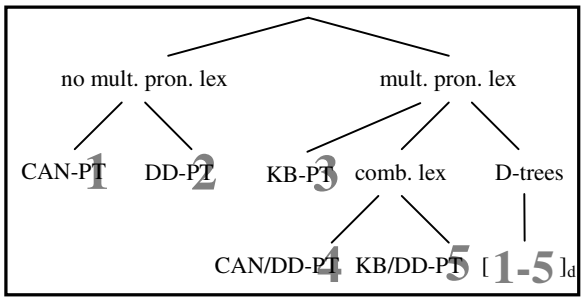

Figure 1: Ten automatic phonetic transcriptions.

The *canonical transcriptions* (CAN-PTs) were generated by substituting each orthographic word with its canonical pronunciation. Cross-word processes were not modelled.

The *data-driven transcriptions* (DD-PTs) were generated through constrained phone recognition; our CSR labelled the signal with its acoustic models and a 4-gram phonotactic model trained with the RTs of the development data.

For our *knowledge-based transcriptions* (KB-PTs), we compiled a list of 20 prominent phonological processes from the literature on Dutch [4]. These processes were formulated as context-dependent rewrite rules modelling within-word and cross-word contexts in which phones from the CAN-PT could be deleted, inserted or substituted. The resulting rule set comprised rules specific for particular words in Dutch, and general rules describing progressive and regressive voice assimilation, nasal assimilation, syllable-final devoicing of obstruents, t-deletion, n-deletion, r-deletion, schwa deletion and insertion, palatalisation and degemination. The rules were ordered and used to generate pronunciation variants from the CAN-PTs of the speech chunks. The rules applied only once, and their order of application was manually optimised. They applied to chunks rather than to words in isolation to account for cross-word phenomena. Analysis of the resulting pronunciation variants proved that few -if any- implausible variants were generated, and that no obvious variants were missing. The chunk-level pronunciation variants (among which the original CAN-PT) were listed in multiple pronunciation lexicons. Since the literature did not provide numeric information on the frequency of phonological processes, the pronunciation variants did not have prior probabilities. The optimal KB-PT was identified through forced recognition.

The *combination* of the CAN-PTs and the DD-PTs on the one hand, and the CAN-PTs and the KB-PTs on the other hand, was aimed at providing our CSR with additional linguistically plausible pronunciation variants. After all, CAN-PTs do not model pronunciation variation, and our KB-PTs only modelled the variation that was implemented in the form of phonological rewrite rules. The DD-PTs, however, were based on the speech signal. Therefore, they had the potential of better representing the actual speech, at the risk of being linguistically less plausible than the CAN-PTs and the KB-PTs. Since the KB-PTs were based on the CAN-PTs, we only combined the CAN-PTs with the DD-PTs (CAN/DD-PTs) and the KB-PTs with the DD-PTs (KB/DD-PTs). Figure 2 illustrates how chunk-level variants (Figure 2 only presents the variants of two successive words) were generated through the alignment of the phones in a CAN-PT and a DD-PT.

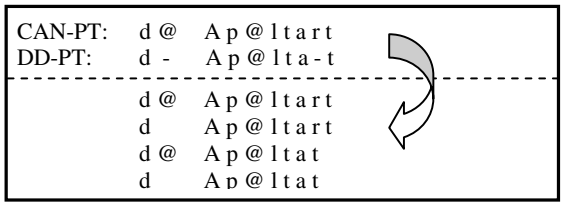

Figure 2: Generation of pronunciation variants through the alignment of two phonetic transcriptions.

Our *decision trees* were generated with the C4.5 algorithm, provided with the Weka package [8]. First, we aligned the APT and the RT of the development data. Subsequently, we made a list of all 'phonetic contexts' in the APT (phones and their two immediate context phones). The correspondences of phones in the APT and the RT, as well as the frequencies of these concurrences, were formalised as decision trees estimating the probability of a phone in the RT given a phonetic context in the APT. Next, the decision trees were used to generate pronunciation variants for the APT of the evaluation data. They now predicted the probability of a phone with optional variants given a particular phonetic window in the APT. All variants (phones) with a probability lower than 0.1 were ignored in order to reduce the number of pronunciation variants and to prune unlikely variants originating from idiosyncrasies in the original APT.

The remaining phone-level variants were combined to word-level variants, which 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. The optimal transcription was identified through forced recognition. The consecutive application of decision tree filtering to the CAN-PTs, DD-PTs, KB-PTs, CAN/DD-PTs and KB/DD-PTs resulted in five new transcriptions hereafter referred to as $[\text{CAN-PT}]_d$, $[\text{DD-PT}]_d$, $[\text{KB-PT}]_d$, $[\text{CAN/DD-PT}]_d$ and $[\text{KB/DD-PT}]_d$.

The APTs of the data in the evaluation sets were evaluated in terms of their deviations from the human RTs. The disagreement metric was formalised as the sum of all phone substitutions (*Sub*), 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.

4. RESULTS

Table 2 presents the disagreement scores (%dis) and the statistics of the substitutions (sub), deletions (del) and insertions (ins) between the APTs and the RTs.

Table 2: Assessment of phonetic transcriptions.

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-PT]$_D$	7.1	3.3	4.2	14.6	4.8	1.6	1.7	8.1
[DD-PT]$_D$	26.0	18.6	3.8	48.3	15.7	7.4	3.5	26.7
[KB-PT]$_D$	7.1	3.5	4.2	14.8	5.0	3.2	1.2	9.4
[CAN/DD-PT]$_D$	20.1	7.2	5.5	32.8	12.0	2.3	4.3	18.5
[KB/DD-PT]$_D$	19.3	9.4	4.5	33.1	11.6	3.1	3.1	17.8

The CAN-PTs and the KB-PTs resembled the RTs much better than the DD-PTs. The most frequent discrepancies in the CAN-PTs and the KB-PTs regarded voiced/unvoiced classifications of obstruents, insertions of schwa and insertions of /r/, /t/ and /n/. About 62-75% of the substitutions and deletions occurred at word boundaries, but the absolute numbers in the KB-PTs were lower due to cross-word pronunciation modelling.

Most discrepancies between the DD-PTs and the RTs were substitutions and deletions. In particular the high proportion of deletions and the wide variety of substitutions were striking. In addition to consonant substitutions due to voicing, we also observed substitutions due to place of articulation, and vowel reductions to schwa.

The proportion of disagreements in the CAN/DD-PTs and the KB/DD-PTs was lower than in the DD-PTs, but much higher than in the original CAN-PTs and the KB-PTs.

The application of decision trees improved the original APTs; only the [DD-PT]$_d$ of the TD 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 CAN/DD-PTs and the KB/DD-PTs, and most outspoken for the CAN-PTs (Δ=20.5% rel., $p<.01$) and KB-PTs (Δ=14.1% rel., $p<.01$). This is remarkable, because one would expect the largest improvement for the worst baseline. However, our [CAN-PTs]$_d$ proved most similar to the RTs.

5. DISCUSSION

The use of an example transcription and a strict protocol for verification speeds up 'manual' transcription, but it can also tempt human experts into neglecting acoustic cues in favour of other sounds [9]. Since both our RTs and KB-PTs were based on CAN-PTs, the assessments of the CAN-PTs and the KB-PTs may have been positively biased. Consequently, the assessments of the DD-PTs may have been negatively biased, for these transcriptions were based on the signal instead of on CAN-PTs.

[5] Suggested the additional use of a consensus transcription (CT) to minimise the risk of such a biased assessment. A CT is a transcription which is unanimously agreed upon by two or more expert phoneticians, made from scratch to rule out the biasing effect of an example transcription. We did not use a CT because 1) a CT is not necessarily flawless either; transcribers may influence each other, and they may still base their judgements on canonical forms in their mental lexicon 2) the generation of a CT is more expensive than the manual verification of an example transcription and therefore only possible for relatively small speech samples; samples that are likely to be too small to train robust decision trees.

The proportion of disagreements between the CAN-PT and the RT of the TD already compared to human inter-labeller disagreement scores, whereas the CAN-PT of the RS did not [5, p. 26]. This was probably due to the high number of inconsistencies in the RT of the TD. The manual verification of spontaneous speech transcriptions in the CGN yielded almost as many 'errors' as corrections [5]. Nevertheless, the trade-off with the limited costs suggests that CAN-PTs are close to the best one can buy for a tolerable amount of money. The observed proportion of substitutions and insertions at word boundaries, however, does imply the need for pronunciation variation modelling.

The high number and the wide variety of substitutions in the DD-PTs shows that the sole use of a phonotactic model did not sufficiently tune our CSR towards the targeted type of 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 APTs with fewer symbols. Inspection of the DD-PTs proved many deletions unlikely, thus ruling out the possibility that the CSR analysed the signal more accurately than the human experts did.

The availability of knowledge-based pronunciation variants proved most beneficial for the transcription of the more spontaneous TD. The optimal performance that could be obtained with the two knowledge-based recognition lexicons (22.6 to 13.2% disagreement with the TD lexicon, 16.3 to 7.4% disagreement with the RS lexicon) shows that there was still room for improvement, and that the acoustic models of our CSR often preferred suboptimal variants. The use of acoustic models trained on KB-PTs might well further optimise the selection of pronunciation variants.

The combination of DD pronunciation variants with canonical or KB variants allowed our CSR to better approximate the RTs 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 canonical and KB lexical variants clearly caused a bias towards the RTs in the otherwise signal-based recognition lexicons.

The [DD-PTs]$_d$ were not significantly 'better' than the original DD-PTs ($p > .1$). This was probably due to the confusability of the lexical pronunciation variants. The size of the lexicons had grown to an average of 9.5 (TD) and 3.5 (RS) variants per word.

This was due to the high number of mismatches between the original DD-PTs and the RTs. These mismatches were learned by the decision trees, then filtered, and finally reformulated as lexical pronunciation variants. However, the presence of unlikely variants must have polluted the lexicons and weakened the lexical probabilities of the more plausible pronunciation variants. 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-PTs, the KB-PTs 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 lexical probabilities were more robust than the priors in the data-driven lexicons which had more pronunciation variants per entry.

An interesting result of our study is that the CAN-PTs were the best point of departure to approximate the targeted type of RT. We are inclined to believe that this can only be explained by a canonically-oriented bias in our RTs that was so strong that no other point of departure could close the gap. Thus, in order to approximate CGN-like manually verified transcriptions, it is worthwhile learning the most obvious differences between the canonical and the reference transcriptions through the use of decision trees. However, we also believe that the CAN-PTs as point of departure for APTs may be suboptimal to approximate RTs that are not based on a (similar) example transcription. The failure of all our signal-based APTs to approximate the CGN transcriptions raises questions about the degree to which the CGN transcriptions represented the signal in a way similar to a transcription made from scratch.

The number of remaining discrepancies in the $[\text{CAN-PTs}]_d$ of the TD and the RS (14.6% and 8.1% disagreement) was only slightly higher than human inter-labeller disagreement scores reported in the literature. [5] 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. Inspection of the 20 most frequent discrepancies between the $[\text{CAN-PTs}]_d$ and the RTs showed a comparable number of insertions and deletions, and a set of substitutions in which the mismatches between voiced and voiceless phones were dominant. Similar discrepancies were observed between different manual transcriptions that were based on the same example transcription [5]. Since our RTs were generated by different transcribers, and individual transcribers do not always expose consistent transcription behaviour either, we conclude that we should not try to further model the inconsistencies in manual transcriptions of speech, and 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

In this study we compared the applicability of established automatic phonetic transcription procedures in large-scale transcription projects, and we investigated whether the combination of these procedures yields transcriptions closer resembling a manually verified reference transcription.

We found that signal-based procedures could not approximate our reference transcription. A knowledge-based procedure did not give optimal results either. Quite surprisingly, a procedure in which a canonical transcription, 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 discrepancies compared to inter-labeller disagreements reported in the literature. This implies that future corpus designers should consider the use of automatic transcription procedures as a valid and cheap alternative to expensive human experts.

However, we are inclined to believe that the success of the canonical-based procedure and the failure of the signal-based procedures are mainly due to the conservative nature of the manually verified reference transcriptions. If such transcriptions are indeed likely to reflect only very obvious deviations from an example transcription, their costs may no longer be justified in the future.

7. ACKNOWLEDGEMENT

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

8. REFERENCES

1. Wester, M. “Pronunciation modeling for ASR - knowledge-based and data-derived methods”, *Computer Speech & Language*, vol. 17/1, pp. 69-85, 2003.
2. Bellegarda, J.R. “Unsupervised, language-independent grapheme-to-phoneme conversion by latent analogy”, *Speech Communication,* vol. 46/2, pp. 140-152, 2005.
3. Oostdijk N. “The design of the Spoken Dutch Corpus”, Peters P., Collins P., Smith A. (Eds.) *New Frontiers of Corpus Research*. Rodopi, Amsterdam, pp. 105-112, 2002.
4. Booij, G. *“The phonology of Dutch”*, Oxford University Press, New York, 1999.
5. Binnenpoorte, D. *Phonetic transcription of large speech corpora*, Ph.D. thesis, Radboud University Nijmegen, the Netherlands, 2006.
6. Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., Woodland, P. *The HTK book (for HTK version 3.1)*, CUED, 2001.
7. Elffers, B, Van Bael, C., Strik, H. *ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcriptions*. http://lands.let.ru.nl/literature/elffers.2005.1.pdf, 2005.
8. Witten, I.H., Frank, E. *Data Mining: Practical machine learning tools and techniques*, 2nd Edition, Morgan Kaufmann, San Francisco, USA, 2005.
9. Demuynck, K., Laureys, T., Gillis, S. “Automatic generation of phonetic transcriptions for large speech corpora”. *Proc. ICSLP*, Denver, USA, pp. 333-336, 2002.