Phonetic Transcription of Large Speech Corpora: How to boost efficiency without affecting quality

Diana Binnenpoorte and **Catia Cucchiarini**
A^2RT, Department of Language and Speech,
University of Nijmegen, The Netherlands
(d.binnenpoorte, c.cucchiarini)@let.kun.nl

ABSTRACT

This paper reports on an experiment aimed at improving an automatically generated phonetic transcription of the Spoken Dutch Corpus (CGN). Different techniques are explored to improve an automatically generated phonetic transcription (AGT). The different AGTs are compared to a reference transcription to determine their quality. The results indicate that implementing phonological rules does improve the AGT for all speech styles considered in the experiment. Applying ASR techniques to model phonological rules that are less frequent in continuous speech results in a decrease of substitution errors.

1. INTRODUCTION

In the last decennia large speech corpora have been collected for many languages for the purpose of developing applications and conducting research. A project aimed at compiling a 10 million word corpus of spoken Dutch, the Spoken Dutch Corpus (*Corpus Gesproken Nederlands*, CGN), is being carried out in the Netherlands and Flanders. The main objective of the project is to create a language resource for research in various linguistic disciplines and for the development of applications in language and speech technology. The speech material in the corpus will be orthographically transcribed, lemmatized and enriched with part-of-speech tagging. Furthermore, automatic word segmentations will be created for all the speech material. For about 1 million words more detailed information will be provided, such as a syntactic annotation, a manually verified broad phonetic transcription and a hand-checked word segmentation. The latter will be based on the manually created broad phonetic transcriptions, whereas the automatic word segmentation for the bulk of the material will be based on an automatically generated phonetic transcription (AGT). Manual phonetic transcriptions are known to be very time-consuming and thus costly. Furthermore, manual phonetic transcriptions are the result of judgments of the human transcriber and thus contain an element of subjectivity. The transcriptions differ when made by different transcribers [1], thus lacking in reliability. Automatic transcriptions, on the other hand, do not suffer from subjectivity; moreover, repeating the transcription with the same machine will guarantee that the results will be identical. This is important when creating a large speech corpus as the CGN. Moreover, automatic transcriptions can be created for a fraction of the costs of manually generated transcriptions.

In the CGN project it was decided that the manually verified phonetic transcriptions for the 1 million words would be made in a two-pass process. First, an optimized automatic transcription is created, and second human transcribers auditorily check and then correct these transcriptions according to an extended transcription protocol [2]. The better the AGT presented to the human transcribers, the more efficient the transcription procedure is, and the better the ultimate transcription quality will be. Furthermore, the AGT is also used for an automatic word segmentation that will be provided for the remaining words in the CGN (see for procedure of word segmentation [3]). For this purpose, it is especially important to investigate the phonological phenomena at word boundaries, as that is where the marks must be set in the acoustic signal.

In the research reported here three AGTs were created and validated. The first AGT, AGT*basic*, was a simple concatenation of phonetic transcriptions as found in the lexicon. The second one, AGT*static*, was the result of implementing the phonological processes of voice assimilation and degemination at word boundaries on the previous AGT*basic*. This AGT is called static because the rules are always applied whenever the context is met in which the rule could be applied. In the last AGT, AGT*dynamic*, other less frequently applied processes, were modeled in a dynamic way. This means that a continuous speech recognizer had to choose the best matching transcription for the acoustic signal.

In the rest of this paper we report on the creation of the three AGTs and the result of the validations.

2. EXPERIMENT

In previous experiments we investigated to what extent the AGT*basic* deviates from a reference transcription (RT). Not only quantitative results, such as the number of deviations, but also qualitative results, such as the nature of the discrepancies, were presented. We learned that for some phonological processes static modeling of those processes, AGT*static*, gave a substantial improvement.

In the following paragraph we describe the speech material that was used, how the reference transcription was created,

and how it is used as a benchmark for validating three AGTs. Finally the new AGT, AGT*dynamic*, is explained.

2.1 Speech material

The speech material used in this experiment was taken from the CGN. The subcorpus on which the experiment was carried out consists of 16 minutes of speech, containing 2712 words. The subcorpus contains fragments of four different speech styles: read speech (RS), lectures (LC), interviews (IN) and spontaneous conversations (SC). The material also varies with respect to the speakers. It was produced by twenty different speakers, eleven males and nine females, whose ages vary between 20 and 73, and who were born in different regions in the Netherlands. In this way a plausible sample of Northern Dutch was collected.

2.2 Reference transcription, RT

In [4] we described a method to validate phonetic transcriptions. It was explained why one should use a reference transcription (RT) in order to be able to measure transcription quality. A reference transcription can serve as a benchmark or a 'true' transcription against which other transcriptions can be validated. A consensus transcription is probably the best possible operationalisation to approach a 'true' transcription [5].

Two phonetically trained and experienced listeners were asked to make the consensus transcription of the speech material. They transcribed from scratch and had to agree on each symbol included in the transcript. They used the CGN symbol set, which is derived from the SAMPA set for Dutch. This resulted in a broad phonetic consensus transcription, which will serve as the reference transcription throughout the experiments.

2.3 Alignment

In order to determine the quality of the AGT, the transcription is compared to the reference transcription. A dynamic programming algorithm was used to make an alignment between the two transcriptions. This alignment provides, among other things, the number of substitutions, deletions and insertions. Each of these errors is assigned a weighting, which is used as a distance measure during the alignment procedure. The weightings are calculated in terms of articulatory features, such as place and manner of articulation, voice, lip rounding, length, etc. As such, substituting a /t/ for a /d/ causes a difference in the feature 'voice' and has lower costs than a substitution between a /t/ and a /b/, which causes not only a differences in the feature 'voice' but also a difference in the feature 'place'. In this way it is clear in what respect the AGT differs from the reference transcription.

2.4 Automatically generated transcription, AGT

An orthographic transcription is available for all the speech material in the CGN. In the CGN lexicon a canonical phonetic transcription of the words in the orthography is available. The transcriptions in the CGN lexicon were obtained by means of TREETALK [6], which is a grapheme-to-phoneme converter trained on CELEX. For orthographic words that were not (yet) included in the CGN lexicon the phonetic transcriptions were obtained from other sources, such as the CELEX English database, Onomastica, and a rule-based grapheme-to-phoneme converter. In the resulting phonetic representations all so-called obligatory word-internal processes [7] were applied, but optional word-internal processes were not. Three different AGTs with increasing degree of optimization were created.

AGT*basic*
AGT*basic* was the product of the simple concatenation of the phonetic transcriptions from the lexicon. No further adaptations were made for this AGT.

AGT*static*
AGT*static* was created by applying static phonological rules to AGT*basic*. The phonological rules concerned are progressive and regressive crossword voice assimilation and degemination. Previous experiments [8] showed that these phonological rules could indeed be statically applied, because it was found that the process was applied in more than 87% of the possible contexts where it could have been applied in the reference transcription.

AGT*dynamic*
Other processes, which were the source of many discrepancies between the AGT*basic* and the reference transcription, were word-final deletions of /n/, /m/ and /@/ and word-final insertions of /n/, /r/ and /t/. The relative frequency of these processes, which is the number of times a process is applied divided by the number of times the process could have been applied because the conditions for application were met, is about 50%. This means that static modeling results in as many improvements as deteriorations. To circumvent this problem multiple pronunciation variants were allowed in the recognition lexicon. For example in the word 'heeft' in 'Hij *heeft* gelijk' (he is right) the word-final /t/ can be deleted according to the rules that were found in the comparison of AGT*basic* and RT. This results in:

heeft => /heft/ or /hef/

Both realizations are possible and both are included in the lexicon. A continuous speech recognizer (CSR) had to decide which one of the two best fits the acoustic signal. To create this third AGT, AGT*dynamic*, the CSR decided which of the multiple variants was the most likely.

2.5 Forced recognition

The canonical phonemic representations obtained from the different lexical sources (CGN lexicon, English CELEX, rule-based grapheme-phoneme converter) appear to vary with respect to the application of the process of /n/-deletion after schwa. In order to generate variants in the lexicon, all canonical representations were rewritten to forms containing an /n/ after schwa. Another rule that was found through the alignment between RT and AGT*basic* is schwa-deletion in word-final position. Because the

canonical forms in the lexicon vary with respect to word-final /n/, the /@/ often happens to be in word-final position in the canonical forms. These forms are also partly responsible for the word-final schwa deletion rule that was found in the above alignment between AGT*basic* and RT. It is because of this fact that all canonical forms ending with /..@n/ were rewritten to a form in which first the /n/ and subsequently the /@/ were deleted. This resulted in three variants in the lexicon.

The CSR that had to choose the most likely variant from the lexicon is described in [9]. The CSR uses acoustic models, word-based language models and a multiple pronunciation lexicon containing the added pronunciation variants according to [10]. The acoustic models are continuous density hidden Markov models (HMMs) with 32 Gaussians per state. Each HMM consists of six states, three parts of two identical states, one of which can be skipped [11]. In total, 39 HMMs were trained. In addition, one model was trained for non-speech sounds and a model consisting of only one state was employed to model silence. In order to make sure the CSR was not trying to recognize words which were not uttered at all, each utterance had its own language model and a lexicon that contained all pronunciation variants.

3. RESULTS

3.1 Substitutions, deletions and insertions

As described above, all three AGTs were aligned to the RT for all four speech styles. The percentages of substitutions, deletions and insertions are presented in Table 1. The last column gives the total percentage of deviations between the different AGTs and the RT.

AGT	substitutions	deletions	insertions	total
basic				
RS	7.2	2.1	4.7	14.0
LC	9.7	3.8	9.3	22.8
IN	8.4	4.2	11.4	24.0
SC	11.6	3.3	17.4	32.3
static				
RS	6.9	2.3	1.3	10.5
LC	7.9	1.3	7.5	16.7
IN	7.6	1.7	10.1	19.4
SC	10.8	2.1	13.9	26.8
dynamic				
RS	6.5	3.9	1.9	12.3
LC	7.6	5.7	5.1	18.4
IN	7.3	5.8	7.3	20.4
SC	9.6	5.3	12.1	27.6

Table 1: Percentage of deviations between three AGTs and RT.

For all AGTs the number of substitutions, deletions and insertions increases as the spontaneity of the speech increases. Table 1 also shows that the best overall results are obtained with AGT*static*. Nevertheless, a closer inspection of the data reveals that the deterioration in AGT*dynamic* relative to AGT*static* is due to the increasing number of deletions. When comparing the deletions of AGT*dynamic* to AGT*static* the same tendency can be found. The number of substitutions and insertions, on the other hand, diminishes when static phonological rules are applied (AGT*static*) compared to AGT*basic*. This is even more the case when besides the static rules also dynamic rules (AGT*dynamic*) are applied.

3.2 Errors on word boundaries

In [8] we showed that roughly half of the substitutions and insertions in AGT*basic* occur at word boundaries, and for deletions this percentage is even higher. This makes it worthwhile to investigate the nature of the deviations that still remain in AGT*dynamic*, in which processes at word boundaries are modeled.

dynamic	substitutions	deletions	insertions
RS	19.8	84.4	32.7
LC	22.5	85.1	35.2
IN	14.4	79.0	46.8
SC	19.2	80.0	50.0

Table 2: Substitutions, deletions and insertions at word boundaries in AGT*dynamic*.

Table 2 presents the percentages of word-boundary deviations compared to the total number of each specific category. So in AGT*dynamic* about 20% of all substitutions take place at word boundary, as opposed to the 50% presented in [8]. The number of insertions at word boundaries is not diminished with the same proportion, but shows nevertheless a slight improvement. It varies between 32% and 50% in AGT*dynamic* as opposed to 42% to 63% in AGT*basic*. The percentage of word-boundary deletions is rather substantial. Closer examination of the data reveals that the number of insertions and deletions is higher in word-final than in word initial position. Furthermore, the absolute number of /@/, /n/ and /m/ deletions for all four speech styles is higher in AGT*dynamic* than in AGT*basic*, except for RS, where fewer /n/ deletions on word-final position occur. However, the number of insertions of /n/, /r/ and /t/ found on word-final positions in AGT*dynamic* is reduced enormously.

4. DISCUSSION

In this paper an experiment has been described in which the quality of three automatically generated phonetic transcriptions was measured. The automatic transcriptions were produced within the framework of the CGN project and their aim is twofold. One purpose is to serve as an example transcription for the human transcription of 10 percent of all data in the CGN. Another purpose of the AGT is to serve as the starting point for an automatic word segmentation in which the AGT is matched to the acoustic signal to determine word boundaries. For both purposes it holds that the better the quality of the AGT the higher its

usefulness. The three automatic transcriptions, AGTs, differed in degree of implementation of phonological processes.

The degree of agreement between the AGT that was explicitly adapted to the speech signals, AGT*dynamic*, and the RT is somewhat lower than expected. Even though the number of substitutions decreases, especially those at word boundaries, the overall number of deviations does not decrease. This is attributed to the number of deletion errors that seem to be introduced in AGT*dynamic*. In this AGT a CSR had to choose the pronunciation variant that best matches the acoustic signal from among those included in the lexicon. Inspection of the data reveals that the CSR had the tendency to choose the variant in which the word-final phoneme is deleted. Moreover, when a variant existed in which two phonemes on word-final position were deleted, in case of a /…@n/-end, the CSR quite often chose that particular variant. In [12] a similar tendency was reported. It seems that the CSR and the two expert listeners who made the RT had different durational thresholds for phoneme detection, especially for /@/ and /r/ detection. The topology of the acoustic models used in our CSR requires that phonemes should at least be 30 ms to be detected at all. This is also reflected in the increase in deletion errors when the spontaneity of the speech increases. It is known that in spontaneous speech speakers tend to reduce phonemes, which is in line with the results we found. Future research plans envisage measuring the durations of the word-final phonemes in the four speech styles in order to support the above. Moreover, it will be investigated whether the segmental transcription of heavily reduced syllables should be replaced by symbols that represent complete syllables. In that manner an automatic transcription machine might be able to detect the 'presence' of segments that were reduced so much that they only leave traces in the form of the features of surrounding sounds.

To summarize, AGT*static* performs best and comes close to human transcription quality for RS. For other speech styles more adaptation is needed. In an attempt to do so, we showed that substitution and insertion errors decrease, whereas the number of deletions increases to a great extent, making the total percentage of deviations higher.

5. CONCLUSION

In this paper we have reported an attempt to further optimize an AGT which was previously adapted through modeling static phonological processes that are frequently found at word boundaries. The results indicate that the proposed improvement in which a CSR was used to decide on less frequent phonological processes at word boundaries, is not sufficient to resemble human transcription performance. We suggest that using a different acoustic model topology in the CSR could lead to better detection of reduced phonemes and thus to an AGT of better quality.

AKNOWLEDGEMENTS

This research was supported by the project "Spoken Dutch Corpus (CGN)", which is funded by the Netherlands Organization for Scientific Research (NWO) and the Flemish Government.

REFERENCES

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

[2] Goddijn, S. and Binnenpoorte, D., "Assessing Manually Corrected Broad Phonetic Transcriptions in the Spoken Dutch Corpus" in *Proceedings of ICPhS 2003.*

[3] Martens, J.P., D. Binnenpoorte, K. Demuynck, R. Van Parys, T. Laureys, W. Goedertier and J. Duchateau, "Word Segmentation in the Spoken Dutch Corpus" in *Proceedings of LREC 2002, 1432-1437.*

[4] Cucchiarini, C. and Binnenpoorte, D., "Validation and Improvement of Automatic Phonetic Transcriptions" in *Proceedings of ICSLP 2002, 313-316.*

[5] Shriberg, L. D., and Lof, L., "Reliability studies in broad and narrow phonetic transcription" in *Clinical Linguistics and Phonetics, 5, 225-279, 1991.*

[6] Hoste, V., Daelemans, W., Tjong Kim Sang, E. and Gillis, S., "Meta-learning for phonemic annotation of corpora" in *Proceedings of ICML-2000, P Langley (ed), 375-382. Stanford University, 2000.*

[7] Booij, G., *The phonology of Dutch*, Clarendon Press, Oxford, 1995.

[8] Cucchiarini, C., Binnenpoorte, D. and Goddijn, S., "Phonetic Transcriptions in the Spoken Dutch Corpus: how to Combine Efficiency and Good Transcription Quality" in *Proceedings EUROSPEECH'01, 1679-1682.*

[9] Strik, H., Russel, A., van den Heuvel, H., Cucchiarini, C. and Boves, L., "A spoken dialogue system for the Dutch public transport information service" in *International Journal Speech Technology 2 (2), 119-129.*

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

[11] Steinbiss, V. et al, "The Philips research system for large-vocabulary continuous-speech recognition" in *Proceedings of EUROSPEECH'93, 2125-2128.*

[12] Wester, M., Kessens, J.M., Cucchiarini, C., and Strik, H., "Obtaining Phonetic Transcriptions: A Comparison between Expert Listeners and a Continuous Speech Recognizer" in *Language and Speech, 2001, 44 (3), 377-403.*