HOW TO IMPROVE HUMAN AND MACHINE TRANSCRIPTIONS OF SPONTANEOUS SPEECH

Diana Binnenpoorte, Simo Goddijn & Catia Cucchiarini

A^{2}RT, Department of Language and Speech, University of Nijmegen, The Netherlands

{d.binnenpoorte, s.goddijn, c.cucchiarini}@let.kun.nl

ABSTRACT

This paper reports on an experiment aimed at measuring the quality of automatic transcriptions and human transcriptions of different speech styles that were produced within the framework of a large speech corpus project, the CGN. The results indicate that the procedure adopted in the CGN to improve the quality of phonetic transcriptions does indeed contribute to achieving this aim. However, better transcriptions of spontaneous speech could probably be obtained by resorting to ASR techniques for pronunciation variation modeling. Our research indicates how this could be achieved.

1. INTRODUCTION

In spite of the considerable progress that has been made in the field of speech technology in the last decades, state-of-the-art speech technology still performs significantly better for carefully pronounced speech like read speech than for more natural types of speech like extemporaneous, spontaneous and/or conversational speech. Various factors have contributed to this state of affairs. First, the fact that the majority of our knowledge of speech processes stems from research based on so-called, "laboratory speech", which is often read speech, while spontaneous speech processes have been much less investigated [1]. Second, which is of course very important for speech technology, most of the databases that have been used for training purposes in speech technology also contained rather artificial speech, while databases containing spontaneous, conversational speech have become available only in recent years.

A project aimed at compiling a huge corpus of spoken Dutch (the Spoken Dutch Corpus, CGN) [2] is now being carried out in the Netherlands and Flanders. The experience gained in this project has shown that although the availability of spontaneous speech databases certainly constitutes a boon to speech research, one should not underestimate the difficulties that are encountered in making these databases really accessible for research purposes. These difficulties concern, among other things, the annotations that have to be made in order to be able to really employ the speech data. One type of annotation that poses considerable problems in terms of time, and therefore money, is phonetic annotation. Still, this type of annotation is essential, especially when it comes to understanding which processes are really going on in spontaneous speech.

For this reason researchers have been looking for efficient methods of obtaining phonetic transcriptions of good quality at reasonable costs. This holds in particular for spontaneous speech, because transcriptions of this type of speech are known to be even more time-consuming and therefore expensive.

Within the framework of the CGN project such attempts were also made. On the basis of a small-scale experiment it was decided that it would be more efficient to have transcribers correct an automatically derived phonetic transcription rather than having them transcribe all speech fragments from scratch. Such an automatic transcription (AT) can be easily obtained for the CGN material because in this project a list is compiled of all words contained in the corpus, the CGN lexicon, which provides the orthographic and the corresponding phonetic representation of each word. As for each utterance in the corpus an orthographic transcription is available, a corresponding AT can be obtained by a simple lexicon-lookup procedure. Although the experiment revealed that editing such an AT took less time than transcribing from scratch, it did not provide any information as to the quality of the transcriptions obtained with this procedure. However, for the usability of the corpus for research and applications it is important to get insight into the quality of the transcriptions too.

For this reason we first studied the quality of the AT to be given as the starting point. The idea was that if this AT turned out to be good enough, no human correction would be required. One can imagine that at some points the AT deviates from what was actually realized, and since the AT is based on canonical representations, the deviations are likely to be more substantial for spontaneous speech than for read speech. For this reason in this study various types of speech, ranging from read speech to spontaneous speech, were taken into consideration. Our results indeed showed that for some types of speech, such an AT simply obtained by concatenation already achieved reasonable quality levels. In this study we also found that this AT could be further improved through an equally simple

intervention: static modeling of cross-word voice assimilation (both regressive and progressive). As a matter of fact, this intervention reduced the number of substitutions dramatically, and for some speech styles, i.e. read speech, the agreement levels observed between AT and RT after applying cross-word voice assimilation were comparable to the levels of agreement normally observed between human transcriptions. For the other speech styles it was clear that even this improved AT would not be good enough and that some other measures should be taken, such as, for instance, further improvement through human correction. Unfortunately, with respect to phonetic transcription, it cannot be taken for granted that correction by human transcribers always implies quality improvement. As is well known in phonetic research, phonetic transcriptions are likely to be subjective and inconsistent [3]. It is for this reason that we decided to carry out an experiment to determine whether correction by human transcribers of the CGN ATs does indeed lead to transcriptions of better quality. This is particularly important for speech styles other than read speech, since for read speech ATs already turned out to be of sufficient quality. In the rest of this paper we report on the results of this experiment.

2. EXPERIMENT

In the present experiment the original automatic transcription (AT) the transcribers had received as the starting point was compared to a reference transcription (RT). Then the revised ATs by four different transcribers (HT) were also compared to the reference transcription (RT). The RT is used to evaluate the quality of both the transcriptions made by the four transcribers and the AT. In phonetic research the difficulties in obtaining such a transcription are well known, and it is generally acknowledged that there is no absolute truth of the matter as to what phones a speaker produced in an utterance [3]. Hence there is no reference transcription that can be considered correct. To try and circumvent this problem as much as possible, phoneticians have been looking for procedures that can approach a reference transcription, such as a transcription made by two or more transcribers after they have agreed on each individual symbol: a consensus transcription [4]. This is the procedure that was adopted in the present experiment to obtain an RT that can be used to evaluate both automatic transcriptions and human transcriptions.

To measure transcription quality we resorted to some sort of alignment between the RT and the transcriptions to be evaluated, with a view to determining a distance measure which also provides a measure of transcription quality. For this purpose a dynamic programming algorithm was used in which the distance between corresponding phonetic symbols is calculated on the basis of articulatory features defining the speech sounds the symbols stand for [3]. In addition to aligning two transcriptions, this algorithm compares the two transcriptions and returns various data such as an overall distance measure, the number of insertions, deletions and substitutions of phonemes, and data indicating to which articulatory features substitutions are related. In the present experiment this kind of information is extremely valuable to establish how the revised transcriptions differ from the RT and from the AT.

2.1 Method

2.1.1 Speech material

The present experiment was limited to the Dutch language varieties spoken in the Netherlands. The speech material selected varies with respect to speech style and speaker, thus constituting a plausible sample of the Northern Dutch part of the CGN, and consists of 16 different fragments representing four speech styles in increasing order of spontaneity: read speech (RS), lectures (LC), interviews (IN), and spontaneous conversations (SC). A total of about 16 minutes of speech, containing 2712 words, was transcribed through consensus, by the machine (ATs), and the ATs were then corrected by four human transcribers.

2.1.2 Reference transcription (RT)

The RT of this speech material was made by two phonetically trained listeners who had experience in transcribing speech. They transcribed together from scratch and had to agree on each symbol included in the transcript (consensus transcription). They used the CGN symbol set, which is an adaptation of the SAMPA set for Dutch. The original orthographic transcription was available and could be consulted in case of doubt.

2.1.3 Automatic transcription (AT)

The AT was obtained by concatenating phonetic representations of the orthographic words through simple lookup in the CGN lexicon. All so-called obligatory word internal processes [5] are applied, whereas optional word internal processes are not applied. Optional word boundary processes like progressive and regressive voice assimilation and degemination are applied using statically modeled phonological rules.

2.1.4 *Correction by human transcribers (HT)*

Four human transcribers (HT1, HT2, HT3 and HT4) who were employed in the CGN project are asked to check and, where necessary, modify the optimized AT of the selected speech material. These transcribers first received some training for this specific task. Moreover, an extensive protocol containing rules and instructions for what to do in case of doubt was made available to them. This procedure was followed for the RS, the LC and the IN fragments, while for the SC fragments a double check was applied, as is the case in the CGN project: a transcriber first checks and modifies the AT and then a second transcriber again

checks and modifies the output of the first check. This decision was based on the assumption that spontaneous conversations are particularly difficult to transcribe.

2.1.5 Alignment

All transcriptions revised by the four transcribers (HTs) and the AT were aligned with the RT by using the Align program [3], in which the distance between corresponding phonemes is calculated on the basis of articulatory features like place and manner of articulation, voice, lip rounding, length, etc. For example, substituting a /t/ for a /d/ has a lower cost than substituting a /t/ for a /x/.

2.2 **Results**

2.2.1 Quantitative results

In analyzing the transcriptions we first performed an alignment between the RT and the AT to find out to what extent the two differ from each other. The percentages of substitutions, deletions and insertions in this alignment are displayed in the top panel of Table 1.

%		substitution	deletion	insertion	**Total**
AT	**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**
RS	HT1	4.3	1.0	1.0	**6.3**
	HT2	4.1	0.9	1.1	**6.1**
	HT3	4.8	0.5	1.4	**6.7**
	HT4	4.5	0.5	1.2	**6.2**
LC	HT1	5.7	2.9	2.7	**11.3**
	HT2	5.8	1.4	2.8	**10.0**
	HT3	6.1	1.0	4.0	**11.1**
	HT4	5.5	1.3	3.8	**10.6**
IN	HT1	5.1	3.7	2.2	**11.0**
	HT2	5.6	1.4	3.6	**10.6**
	HT3	5.6	0.7	3.8	**10.1**
	HT4	4.7	1.6	3.9	**10.2**
SC	HT1	7.3	3.1	3.5	**13.9**
	HT2	6.5	2.1	4.8	**13.4**
	HT3	7.8	1.5	5.4	**14.7**
	HT4	7.9	1.6	6.2	**15.7**

Table 1 Deviations per speech style between AT as opposed to RT and the HTs as opposed to RT

The total percentage of deviations (last column) shows that for read speech (RS) the distance between the AT and the RT is around 10%, a percentage very similar to that observed between human transcriptions of the same material made by different transcribers [6]. For the other fragments the percentage of deviations increases as the degree of planning in speech decreases. Here the percentages of deviations are so high that human correction of the AT seems absolutely necessary.

The lower panels of Table 1 present the results of the alignment between the RT and the human-corrected transcriptions (HTs) for all four styles. The percentages of deviations appear to be much lower for all speech types, but are still clearly related to degree of planning in speech: as planning decreases, the percentages of deviations increase. In spite of these considerable improvements, there remain many discrepancies between the HTs and the RT, especially for spontaneous speech. Moreover, as appears from Table 1, the majority of these deviations derive from substitutions. In order to get a better understanding of how these HTs differ from the RT, in the following section we proceed to a more detailed, qualitative analysis of these discrepancies, which can be easily performed with the Align program.

2.2.2 Qualitative results

Owing to space limitations, we now confine ourselves to a qualitative analysis of substitutions alone, because these appear to be the most frequent discrepancies. Closer examination of the data reveals that for all speech types the most frequent substitutions are related to the feature voice, see Table 2. While for RS this type of substitution is responsible for all frequent deviations, in the other speech fragments other sorts of substitutions are also observed, such as confusions between long and short vowels or between short vowels and schwa. Moreover, the frequency of the latter types of vowel substitutions is considerably higher in spontaneous conversations.

	HT1		HT2		HT3		HT4	
RS	v,f	15	x,G	26	x,G	23	x,G	27
	s,z	12	t,d	16	t,d	14	t,d	17
	k,g	10	f,v	13	v,f	12	f,v	12
	t,d	8	s,z	10	z,s	6	k,g	8
	G,x	7	k,g	9	k,g	5	v,f	7
LC	s,z	19	x,G	26	x,G	24	x,G	23
	t,d	16	t,d	22	t,d	19	k,g	17
	k,g	13	k,g	18	k,g	16	t,d	15
	O,o	12	s,z	17	@,E	10	@,E	9
	A,a	10	@,E	12	O,o	9	o,O	8
IN	k,g	8	x,G	11	t,d	10	t,d	10
	t,d	7	t,d	11	x,G	9	x,G	9
	O,o	6	k,g	10	@,A	8	k,g	8
	@,A	4	s,z	7	k,g	6	@,A	5
	I,@	4	@,A	5	I,@	4	I,@	5
SC	t,d	16	x,G	17	@,E	12	k,g	16
	@,E	12	k,g	15	k,g	12	x,G	15
	k,g	12	t,d	15	t,d	10	t,d	13
	A,a	8	@,E	8	@,=	8	@,E	8
	v,f	7	n,=	8	j,=	5	@,A	7

Table 2 Top five substitutions

3. DISCUSSION

In this paper we have described an experiment aimed at measuring the quality of automatic transcriptions and human transcriptions of different types of speech that were produced within the framework of a large speech corpus project, the CGN. Inspection of the quality of ATs revealed that while these can achieve reasonable quality levels for read speech, their quality is definitely insufficient for less planned speech such as lectures, interviews and spontaneous conversations. This suggests that some form of correction is required to be able to achieve reasonable quality levels for these styles too. In the CGN project the choice was to have human transcribers check and modify the ATs. However, since in the case of phonetic transcription correction by human beings does not by definition imply improvement, in our experiment we checked whether the human transcribers indeed improved the ATs. The results showed that correction by transcribers led to an improvement in transcription quality across the board. For spontaneous speech the percentages of deviations after correction by human transcribers appeared to be in the same order of magnitude as those observed in similar investigations with spontaneous speech [7]. However, these percentages are clearly higher than those obtained for the other speech types, in spite of the fact that for spontaneous conversations a double-check procedure was applied.

One might of course wonder whether this should be accepted as a fact or whether one should try, somehow, to improve this result. One thing to be noted in this respect is, for example, that our results on spontaneous speech were obtained in a situation in which the transcribers did not transcribe from scratch, but received an AT to edit. One can imagine that if this AT is very different from the speech signal, which in this case is plausible given that the AT was derived from canonical forms, the transcriber will have to modify many symbols to come as close as possible to the speech signal. In any case, it is clear that in SC the transcribers will have to change more symbols than in RS. Under such conditions it is reasonable to assume that some ceiling effect takes place. In other words, it sounds plausible to assume that there is a maximum number of characteristics a transcriber can attend to when editing a transcription. If the number of symbols to be changed exceeds this maximum, the human-corrected transcription will still be considerably different from the RT. This explanation is corroborated by the analysis of the most frequent substitutions observed in the four speech styles. In RS these are predominantly voice substitutions, whereas in LS, IN and SC all sorts of vowel substitutions are observed that correspond to processes typically observed in spontaneous speech, which, as can be expected, are not modeled in the canonical representations and thus not in the optimized AT. What these observations seem to suggest is that a considerable improvement in transcription quality could be obtained by adopting different ATs for different speech styles. In other words, further optimization and, possibly, speech style adaptation of ATs could be the key to obtaining higher indices of transcription quality for spontaneous speech.

Further optimization could be achieved, for instance, by obtaining the AT not through lexicon lookup, but through a CSR that uses pronunciation variation modeling [5]. This would give the possibility of modeling not only the processes that cause substitutions in spontaneous speech, but also those responsible for the numerous insertions in spontaneous speech (see top panel Table 1). Previous research [5] has indeed shown that, for instance, by varying the length of the HMMs a CSR can be tuned to producing more or fewer insertions, thus improving the quality of ATs.

4. CONCLUSION

The research presented in this paper allows us to draw the following conclusions: Human correction of ATs does indeed lead to quality improvement, albeit to a limited extent, at least for more spontaneous speech styles. There are indications that further quality improvement could be achieved through previous optimization and speech style adaptation of ATs. Our results also indicate in which direction this optimization and adaptation of ATs of more spontaneous speech styles should be sought.

5. REFERENCES

[1] Mehta, G & Cutler, A. (1998) Detection of target phonemes in spontaneous and read speech. *Language and Speech*, 31, 135-156.

[2] Oostdijk, N. & Boves, L. Spontaneous speech in the Spoken Dutch Corpus. *This workshop*.

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

[4] Shriberg, L.D., Kwiatkowski, J., and Hoffman, K. "A Procedure for Phonetic Transcription by Consensus". *Journal of Speech and Hearing Researc*h, 27, 456-465, 1984.

[5] Wester, M. Kessens, J.M. Cucchiarini, C. and Strik H. Obtaining phonetic transcriptions: a comparison between expert listeners and a continuous speech recognizer. *Language and Speech*, 2001, 44(3), 377-403.

[6] Kipp, A., Wesenick, B., and Schiel F. Automatic detection and segmentation of pronunciation variants in German speech corpora. *Proceedings ICSLP '96*, 106-109, 1996.

[7] Kipp, A., Wesenick, B., and Schiel F. Pronunciation modeling applied to automatic segmentation of spontaneous speech. *Proceedings EUROSPEECH '*97, 1023-1026, 1997.

6. ACKNOWLEDGEMENT

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.