Automatic Phonetic Transcription: An overview

Catia Cucchiarini and Helmer Strik
A^{2}RT, Department of Language and Speech,
University of Nijmegen, The Netherlands
(C.Cucchiarini, H.Strik)@let.kun.nl

ABSTRACT

Both in linguistics and in speech technology phonetic transcriptions (PTs) are often needed. Given the many drawbacks in making manual PTs, researchers have been looking for ways to obtain PTs automatically. In this paper an overview is presented of automatic phonetic transcription (APT). Several aspects of APT are discussed: evaluation, generation and usability. Evaluation is needed to determine the quality of APTs. Usually this is done by comparing the APTs with human reference transcriptions. Generating APTs can be done in several ways, e.g. by means of phone recognition or forced recognition. The quality of the generated APTs can be enhanced by optimizing the automatic speech recognition systems used to make the APTs. In spite of the current limitations of ASR technology, APTs already offer some important advantages for phonetic research. In this paper we explain how.

1. INTRODUCTION

The use of computers in phonetic science is now pervasive: time-consuming tasks that were previously done by hand can now be done automatically with considerable gains in time, costs, and often also in accuracy and/or quality.

One of the tasks that appear to be particularly time-consuming, costly and error-prone is phonetic transcription [16, 3]. Transcriptions have long been used in linguistic research, both for explorative and hypothesis testing purposes. More recently, phonetic transcription has proved to be very useful for speech technology too, e.g. for automatic speech recognition and for speech synthesis. Because of the problems involved in obtaining phonetic transcriptions - the time required, the high costs incurred, the often limited accuracy obtained, and, especially for speech technology applications, the need to transcribe large amounts of data - researchers have been looking for ways of automating this process, for example by employing speech recognition algorithms.

In this paper we provide an overview of the state of the art in automatic phonetic transcription (APT), paying special attention to the most relevant methodological issues and the ways they have been approached. In section 2, evaluation of APTs is discussed, while in section 3 generating the APTs is addressed. Finally, in section 4 the usability of APTs is discussed.

2. APT EVALUATION

Before APTs can be used, it is important to know how accurate they are. The problem of transcription quality assessment is not new, since human phonetic transcriptions pose exactly the same problem: before using them for research one needs to know to what extent they are accurate [16, 3, 4, 18]. Phonetic transcriptions, whether they are obtained automatically or are produced by human transcribers, are generally used as a basis for further processing (research, ASR training, etc.). They can be viewed as representations or measurements of the speech signal and it is therefore legitimate to ask to what extent they live up to the quality standards of reliability and validity that are required of any form of measurement. With respect to automatic transcriptions, the problem of quality assessment is complex because comparison with human performance, which is customary in many fields, is not straightforward, owing to the subjectivity of human transcriptions and to a series of methodologically complex issues that will be explained below.

2.1 Reliability and validity

In general terms, the reliability of a measuring instrument represents the degree of consistency observed between repeated measurements of the same object made with that instrument. It is an indication of the degree of accuracy of a measuring device. Validity, on the other hand, is concerned with whether the instrument measures what it purports to measure. In fact, the definitions of reliability and validity used in test theory are much more complex and will not be treated in this paper. The description provided above indicates an important difference between the reliability of human-made as opposed to automatic transcriptions and is related to the fact that human transcriptions suffer from intra-subject and inter-subject variation and repeated measurements of the same object will differ from each other. With automatic transcriptions this can be prevented because a machine can be programmed in such a way that repeated measurements of the same object always give the same result, thus yielding a reliability coefficient of 1, the highest possible. It follows that with respect to the quality

ISBN 1-876346-48-5 © 2003 UAB

of automatic transcription only one, albeit not trivial, question needs to be answered, viz. that concerning validity.

2.2 Defining a reference

The description of validity given above suggests that any validation activity implies the existence of a correct representation of what is to be measured, a so-called benchmark or 'true' criterion score (as in test theory). The difficulties in obtaining such a benchmark 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, 4, 19]. For instance, in an experiment we asked nine experienced listeners to judge whether a phone was present or not for 467 cases [19]. The results showed that all nine listeners agreed in only 246 of the 467 cases, which is less than 53%. Furthermore, a substantial amount of variation was observed between the nine listeners. The values of Cohen’s kappa varied from 0.49 to 0.73 for the various listener pairs. It follows that one cannot establish the validity of an automatic transcription simply by comparing it with an arbitrarily chosen human transcription, because the latter would inevitably contain errors. Unfortunately, this seems to be the practice in many studies on automatic transcription.

To try and circumvent the problems due to the lack of a reference point as much as possible, different procedures have been devised to obtain reference transcriptions. One possibility consists in using a consensus transcription, which is a transcription made by at least two experienced phoneticians after having reached a consensus on each symbol contained in the transcript [15]. The fact that different transcribers are involved and that they have to reach a consensus before writing down the symbols may be seen as an attempt to minimize errors of measurement, thus approaching 'true' criterion scores. Another option is to have more than one transcriber transcribe the material and to use only that part of the material for which all transcribers agree or at least the majority of them [12, 10].

2.3 Comparing automatic and reference transcription

Another issue that has to be defined in APT is how to determine whether the quality of a given APT is satisfactory. Once a reference transcription (RT) has been defined, the obvious choice would be to carry out some sort of alignment between the RT and the APT, with a view to determining a distance measure which will also provide a measure of transcription quality. For this purpose, dynamic programming algorithms with different weightings have been used by various authors [13, 8]. In our research we have used a dynamic programming algorithm 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, 4]. 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 more detailed data indicating to which features substitutions are related. This kind of information can be extremely valuable if one is interested to know how the APT differs from the RT, and, how the APT could be improved [see e.g. 6, 5, 1].

2.4 Determining when an APT is of satisfactory quality

After having established how much an APT differs from an RT, one would probably need some reference data to determine whether the degree of distance observed is acceptable or not. In other words, how can we determine whether the quality of a given APT is satisfactory? Again human transcriptions could be used as a point of reference. For instance, one could compare the degree of agreement observed between the APT and the RT with the degree of agreement observed between human transcriptions of the same utterances that are of the same level of detail and that are made under similar conditions, because this agreement level constitutes the upper bound, as in the study reported in [18]. If the degree of agreement between the APT and the RT is comparable to what is usually observed between human transcriptions, one could accept the APT as is (see also section 3.2); alternatively, if the degree of agreement between the APT and the RT is lower than what is usually observed between human transcriptions, the APT should first be improved. However, the problem with this approach is that it is difficult to find data on human transcriptions to be used as reference (see for more information on this point [5]).

Whether a transcription is of satisfactory quality will also depend on the purpose one needs the transcription for. Some differences in transcriptions can be important for one application, but less important for another application. Therefore, for meaningful evaluation of transcriptions a context is needed. Such an evaluation within the context of various applications is the topic of a PhD project that has recently started at our department [see 17].

3. HOW TO OBTAIN APTs?

In speech technology, various tools have been developed that go some way toward obtaining phonetic representations of speech in an automatic manner. Some are described in this section.

3.1 Phone recognition

One way to obtain APTs is by using phone recognition. For this purpose, completely unrestricted phone recognition can be used, but usually some phonotactic constraints are employed in the form of a phone language model. Phone accuracy turns out to vary between roughly 50% and 70%. For instance, for one of our ASR systems we measured a phone accuracy level of 63% for extemporaneous speech [20]. In general, such levels of phone accuracy are too low, and thus the resulting APTs cannot be used directly for most applications. Still, phone recognition can be useful. For our ASR system with a phone accuracy of 63% we examined the resulting phone strings by comparing them to canonical transcriptions [20]. The results showed that the number of insertions (4%) was much smaller than the number of deletions (17%) and substitutions (15%). Furthermore, the vowels remain identical more often than the consonants, mainly because in comparison to the consonants they are deleted less often. Finally, we studied the most frequently observed processes, which were all deletions. It turned out that these frequent processes are plausible connected speech processes (see [20]), some of which are related to Dutch phonological processes that have been described in the literature (e.g. /n/-deletion, /t/-deletion and /@/-deletion are described in [2]). Phone recognition can thus be used for hypothesis generation. However, owing to the considerable number of inaccuracies in unsupervised phone recognition, it is often necessary to check or filter the output of phone recognition. The latter can be done by applying decision trees [7] or forced recognition [19].

3.2 Forced recognition

In forced recognition the goal is not to recognize the string of words that was spoken, as in standard ASR. On the contrary, in forced recognition this string of words (the orthographic transcription) has to be known. Given the orthographic transcription, forced recognition can determine for each word which pronunciation most closely matches the signal. In order to do so, multiple pronunciations of words have to be present. Forced recognition, which is also referred to as forced (Viterbi) alignment, can thus be used for hypothesis verification.

In section 2.4 we mentioned that one of the ways to determine whether an APT is of satisfactory quality is by comparing the degree of agreement between the APT and the RT with the degrees of agreement usually observed between comparable transcriptions made by human transcribers. This has been done for forced recognition. For spontaneous speech the following interlistener agreement levels were reported: 79%-83% [11, 14], and 75%-85% [19]. The agreement levels observed between manual and machine transcriptions are 72%-80% [11, 14], and 76%-80% [19]. These data indicate that forced recognition produces satisfactory results, because the choices made by the ASR are comparable to those made by human listeners [19].

3.3 How to optimize APTs?

Given that APTs are used more and more, it is remarkable that relatively little research has been conducted on optimizing APTs. Much more research has been done on lowering the word error rates (WERs) of ASR systems. Whenever APTs have to be generated by means of ASR systems, the available ASR system with the lowest WER is likely to be chosen to generate the APTs. However, in [9] we showed that lower WERs do not always guarantee better transcriptions. For instance, using context-dependent HMMs usually leads to lower WERs, but not always to better quality transcriptions. Therefore, in order to increase the quality of APTs, one should not simply take the ASR system with the lowest WER. Instead, specific ASR systems have to be optimized for this task (i.e. to generate optimal APTs). Our research made clear that by combining the right properties of an ASR, the resulting APTs can be improved. In [9] this was achieved by training the HMMs on read speech (instead of spontaneous speech), by shortening the topology of the HMMs, and by means of pronunciation variation modeling.

4. USABILITY OF APTs

The question that arises at this point is what advantages APT can offer. It is obvious that APT cannot be used for explorative purposes to obtain phonetic transcriptions of complete utterances from scratch. However, by means of phonetic recognition information can be obtained about frequently occurring processes (see section 3.1), e.g. in the form of rewrite rules. Furthermore, forced recognition can be employed for hypothesis verification (see section 3.2), which is probably the most common way of using phonetic transcriptions in phonetics, phonology, sociolinguistics, and dialectology.

In practice, APT can be used in all research situations in which the phonetic transcriptions have to be made by one person. Given that an ASR does not suffer from tiredness and loss of concentration, it could assist the transcriber who is likely to make mistakes owing to concentration loss. By comparing his/her own transcriptions with those produced by the ASR a transcriber could spot possible errors that are due to absent-mindedness.

Furthermore, this kind of comparison could be useful for other reasons. For instance, a transcriber may be biased by his/her own hypotheses and expectations with obvious consequences for the transcriptions, while the biases in APT can be controlled. Checking the automatic transcriptions may help discover possible biases in the listener’s data. In addition, APT can be employed in those situations in which more than one transcriber is involved, in order to solve possible doubts about what was actually realized. It should be noted that using APT will be less expensive than having an extra transcriber carry out the same task.

The advantages of APT become really evident when it comes to exploring large speech databases. First, because, as mentioned above, APT makes it possible to achieve uniformity in phonetic transcription. With human phonetic transcription this would be utopian: large amounts of speech data cannot possibly be transcribed by one person, and the more transcribers are involved, the less uniform the transcriptions are going to be. Eliminating part of this subjectivity in transcriptions can be very advantageous, especially when analyzing large amounts of data. Second, because with APT it is possible to generate phonetic transcriptions of huge amounts of data that would otherwise remain unexplored. The fact that these large amounts of material can be analyzed in a relatively short time, and with relatively low costs makes APT even more important. The importance of this aspect for the generalizability of the results cannot be overestimated. And although the automatic procedures used to generate APTs are not infallible, the advantages of a very large dataset might very well outweigh the errors introduced by the mistakes the automatic procedures make.

ACKNOWLEDGEMENTS

We would like to thank many colleagues of our department for their comments on previous versions of this paper.

REFERENCES

The references below are listed in alphabetical order.

[1] D. Binnenpoorte and C. Cucchiarini. “Phonetic Transcriptions of Large Speech Corpora: How to Boost Efficiency without Affecting Quality”, This conference 2003.

[2] G. Booij. The phonology of Dutch. Clarendon press, Oxford, 1995.

[3] C. Cucchiarini. Phonetic transcription: a methodological and empirical study. PhD Thesis, University of Nijmegen, 1993.

[4] C. Cucchiarini. “Assessing transcription agreement: methodological aspects”, *Clinical Linguistics & Phonetics*, 2, 131-155, 1996.

[5] C. Cucchiarini and D. Binnenpoorte. Validation and improvement of automatic phonetic transcriptions, Proceedings ICSLP2002, Denver, 2002.

[6] C. Cucchiarini, D. Binnenpoorte and S. Goddijn. “Phonetic Transcriptions in the Spoken Dutch Corpus: how to Combine Efficiency and Good Transcription Quality”, *Proceedings EUROSPEECH ’01*, 1679-1682, 2001.

[7] E. Fosler-Lussier. Dynamic Pronunciation Models for Automatic Speech Recognition. PhD Thesis, Univ. of California at Berkeley, 1999.

[8] P. Hanna, D. Stewart, J. Ming & F.J. Smith. ' A Improved DP Match for Automatic Lexiconn Generation' , Proceedings of International Congress of Phonetic Science (ICPhS 99), pp. 1717-1720, San Francisco, USA, 1999.

[9] J.M. Kessens & H. Strik. Lower WERs do not guarantee better transcriptions. Proceedings of Eurospeech-2001, Aalborg, Denmark, 1721-1724, 2001.

[10] J.M. Kessens, M. Wester, C. Cucchiarini & H. Strik. The Selection of Pronunciation Variants: Comparing the Performance of Man and Machine. Proc. ICSLP-1998, Sydney, Vol. 6, pp. 2715-2718, 1998.

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

[12] C. Kuipers, and W. Van Donselaar. The influence of rhythmic context on schwa epenthesis and schwa deletion in Dutch. *Language and Speech,* 41 (1), 87-108, 1997.

[13] J. Picone, K.M. Goudie Marshall, G.R. Doddington and W. Fisher. Automatic text alignment for speech system evaluation. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 34, 780-784, 1986.

[14] F. Schiel, A. Kipp, H.G. Tillmann. Statistical modelling of pronunciation: it' s not the model, it' s the data. Proc. of the ESCA workshop ' Modeling Pronunciation Variation for Automatic Speech Recognition' , Rolduc, Kerkrade, pp. 131-136, 1998.

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

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

[17] C. Van Bael, W. Strik & H. van den Heuvel. “Application-oriented validation of phonetic transcriptions: preliminary results”. *Proceedings of ICPhS 2003*, Barcelona, Spain, 2003.

[18] M-B. Wesenick and A. Kipp. “Estimating the Quality of Phonetic Transcriptions and Segmentations of Speech Signals”, *Proceedings of ICSLP 1996*, Philadelphia, USA, 129-132, 1996.

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

[20] M. Wester, J.M. Kessens, and H. Strik. Two automatic approaches for analyzing the frequency of connected speech processes in Dutch. *Proceedings of the International Conference on Spoken Language Processing, Student Day*, **7**, 3351-3356, 1998.