Application-oriented validation of phonetic transcriptions: preliminary results

Christophe Van Bael, **Wilhelmus Strik** and **Henk van den Heuvel**

A^{2}RT, Department of Language and Speech, University of Nijmegen, The Netherlands
{C.v.Bael, W.Strik, H.v.d.Heuvel}@let.kun.nl

ABSTRACT

There is an increasing need for automatic procedures to generate and validate phonetic transcriptions. As the production of manual phonetic transcriptions tends to be time-consuming, error-prone and costly, procedures have been developed to derive phonetic transcriptions automatically by means of automatic speech recognition technology. Such automatic phonetic transcriptions are usually validated by comparing them with manual phonetic transcriptions. Even though this appears to be a plausible procedure at first sight, it might be troublesome. We believe that phonetic transcriptions should ideally be validated with potential applications in mind. The application focused on in this paper is the reduction of word error rates in automatic speech recognition.

1 INTRODUCTION

Phonetic Transcriptions (PTs) are required for various kinds of research. Linguistic research for example may require them to study pronunciation variation [1], whereas speech synthesis research may require PTs for unit selection and the training of duration models [2], and speech recognition research for the training of phone models.

Over the years, researchers became aware of the fact that the production of Manual Phonetic Transcriptions (MPTs) is time-consuming, costly and error-prone due to fatigue and subjective judgements of human transcribers.

Therefore research has shifted to investigating the usability of Automatic Phonetic Transcriptions (APTs). Resulting procedures can ideally be used to automatically provide PTs of large speech corpora, to serve as a reference with which human transcribers can compare their transcriptions [3] or which human transcribers can use as a starting point, as is done in the context of the Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN)[4, 5].

In previous research, among which recently [5] and [6], the quality of APTs was often estimated as a direct function of its similarity to a reference transcription. This may be problematic, as reference transcriptions (as all MPTs) are subject to errors [7]. Moreover, recent research [8] has proven that there is no direct relation between the performance of a recogniser and the resemblance between PTs generated with that recogniser and a manually generated reference transcription. Kessens and Strik [8] report that *"Lower WERs [Word Error Rates] do not guarantee better transcriptions"*. A *better* trancription for them meant a transcription resembling a reference MPT more.

We think the point made in [8] may also hold the other way around: transcriptions that resemble a reference transcription more might not guarantee lower WERs. We also think this can be generalised: a PT suitable for one application may not be the most optimal transcription for another. For some applications certain deviations might be less important than for others.

In this paper first a new PhD project is introduced that will serve as a general framework in which new application-oriented validation procedures will be developed. Then preliminary results are presented that are focused on the reduction of word error rates in automatic speech recognition. The results support the development and use of application-oriented validation procedures.

2 THE FRAMEWORK OF THIS RESEARCH

2.1 Goals

The experiment reported in this paper is embedded in a larger PhD project which started on October 1st, 2002 for a period of four years. This project has two main goals. On the one hand we will investigate a diverse set of procedures to automatically generate PTs in order to facilitate the generation of APTs for large speech corpora. On the other hand, as there are still no fixed rules or procedures to validate PTs [9], and as common procedures might be troublesome, we will test new validation procedures for PTs.

We consider PTs as a tool, the value of which can only be assessed indirectly, viz. through the quality of the product that is produced with the tool. Accordingly, we propose to validate PTs on the basis of their contribution to the development of a number of applications. By doing so, reference transcriptions are no longer considered as the ideal to be approximated. Rather, it may appear that transcriptions that deviate from the reference transcription perform better in the applications. In our research we will focus on at least three applications where PTs are commonly used.

2.2 Application-oriented validation of phonetic transcriptions

Firstly we will investigate the influence of APTs on the accuracy of recognisers. In Automatic Speech Recognition (ASR), APTs are required to train acoustic models, and it is to be expected that using different PTs will result in different recognition performances. Therefore a recogniser will be trained on several APTs and an MPT comprising different speech styles. The performances (in terms of WER) will be interpreted (validated) with regard to the distance between the PTs and a reference transcription.

Secondly we will investigate the effect of different speech styles on APTs generated by a recogniser. In spontaneous speech for example more phoneme reductions and deletions can be expected than in read speech. We'll investigate whether APTs generated through forced recognition will show similar differences when generated for different speech styles. The resulting APTs and the speech-specific pronunciation rules will again be interpreted with regard to the distance between the PTs and a reference transcription.

Finally the influence of APTs on segment duration statistics will be analysed. The APTs will again be generated through forced recognition. This time also the segment durations will be investigated. We expect that the quality of the segment durations is directly related to the quality of the APTs itself. The segment durations of the APTs will be compared with similar durations of an MPT of the same material (the IFA corpus [10]). We will investigate whether there are significant differences between the APTs and the MPT, and we will pay attention to the relevance of the differences in terms of Just Noticeable Differences (JNDs).

To conclude, we will always validate PTs using data comprising different speech styles. We expect that PTs resembling a reference transcription more may not guarantee the best performance in all applications investigated. If this is the case, an application-oriented validation of PTs should be preferred when the actual applications are known.

At present we are investigating the influence of different PTs on a recogniser's performance. Preliminary results are reported and discussed below.

3 MATERIAL AND METHOD

3.1 Material

3.1.1 Corpora: The corpus used to compute the distance between the APT, the MPT and the reference transcription, Tref, contains 16 minutes of speech (2712 words), divided over 4 speech styles: read speech (RS), lectures (LC), interviews (IN) and spontaneous conversations (SC). Tref is a consensus transcription, generated from scratch by two expert listeners. The data covered by Tref were not used for training, tuning or testing of the recogniser.

The training, development and test corpus for the actual recognition task were extracted from the core corpus of the CGN (release 6)[4, 5]. This core corpus provided an MPT for all data used in the recognition experiment. The recogniser's language model scaling factor (to scale the influence of the language model and that of the phone models with regard to each other), the word insertion log probability (to control insertions and deletions) and the pruning factor were optimised on a separate representative development set. Table 1 provides the details of these data sets.

speech style	ref	train	dev	test
RS	682	40934	425	13639
LC	892	9765	102	3263
IN	523	14097	173	4715
SC	615	14679	132	4903
total	2712	79475	902	26520

Table 1: Number of words in the data sets. (sentence boundaries included as words)

3.1.2 Transcriptions: The APT was generated by concatenating phonetic representations from the canonical CGN lexicon. The transcriptions for the out of vocabulary words were inserted from the Celex English database, Onomastica and a grapheme-to-phoneme converter. The MPT used was provided in the CGN. One MPT was available per sound file.

3.1.3 The alignment program and the recogniser: To compare the MPT and the APT with Tref, Align [7] was used. In this program the distance between corresponding phoneme strings is calculated on the basis of articulatory features defined by the user.

The recogniser was built with the Hidden Markov Modelling toolkit HTK [11]. The system used 2 series of 38 left-right context-independent phone models (continuous density Hidden Markov Models (HMMs) with 32 Gaussian mixture components per state: 35 3-state phone models, one 3-state silence model, one 1-state silence model to capture the optional short pauses after words and one model to capture sounds that couldn't be transcribed). The data were parameterised as Mel Frequency Cepstral Coefficients (MFCCs) with 39 coefficients per frame.

The training lexicon used is a 15K enriched version of the canonical CGN lexicon (see 3.1.2). Two test lexica were used: one canonical lexicon (a 5K subset of the 15K training lexicon) and one 7K multiple pronunciation lexicon comprising all pronunciations in the MPT. As the CGN data are provided as chunks of wave files, orthographic and PTs, the *language model* trained was a backed-off bigram *chunk model* . Because of the difficulty of the task (the recognisers were trained on a mixed data set comprising 4 different speech styles, no typical language model but only a chunk model could be trained) this language model was trained on the test set, thus facilitating the recognition task. All speech styles were represented in all lexica.

3.2 Method

In this paper we investigate whether there is a relation between the distance between PTs and a reference transcription and their influence on a recogniser's accuracy (presented in terms of WER).

First an APT was generated using data from the CGN (see 3.1.1). An MPT of this material was already available. In order to estimate the mutual distance between the APT and the MPT, the distances were computed between Tref and the APT on the one hand, and Tref and the MPT on the other hand. As the data set covered by Tref did not overlap with the training, tuning or corpus used in the recognition experiment, the differences between Tref and the MPT on the one hand and Tref and the APT on the other hand, as measured on the reference corpus, are only estimates of the *quality* of the MPT and the APT for the much bigger training and test corpora.

Next, the influence of the transcriptions on the recognition accuracy was investigated. In order to perform ASR, phone models have to be trained by providing a phonetic transcription (whether automatically generated or human-made) of the training data to the system. Using different transcriptions to train acoustic models may affect the recogniser's accuracy.

Both the APT and the MPT were used to train phone models. At recognition time, the two series of models were inserted separately in a system using the same language model (trained on the test set) and lexicon (only covering the data in the test set). The resulting two recognisers were tuned separately. The performances were then interpreted against the background of the distances between the PTs.

4 EXPERIMENTS

4.1 Validation of the phonetic transcriptions by means of their distance to Tref

First the MPT and the APT were validated using the common procedure of computing their distance to the reference transcription Tref. As was done in [9], the Align program [7] was used to get a detailed report of the distance between the transcriptions in terms of substitutions (Sub), deletions (Del) and insertions (Ins). The results per speech style and transcription type are presented in table 2.

PT	Style	Sub (%)	Del (%)	Ins (%)	Tot (%)
MPT	RS	3.1	0.5	1.4	5.0
	LC	4.6	1.5	3.3	9.4
	IN	4.4	0.9	4.3	9.6
	SC	6.8	1.6	7.7	16.1
APT	RS	7.0	2.4	2.9	12.3
	LC	6.7	1.7	6.1	14.5
	IN	7.1	1.6	8.1	16.8
	SC	8.5	1.6	11.1	21.2

Table 2: Distance between the transcriptions and Tref.

4.2 Validation of the phonetic transcriptions by means of their influence on the WER

Two experiments were performed. In the first experiment the APT was used to train the acoustic models. At recognition time, the canonical CGN lexicon was used. In the second experiment the MPT was used to train the acoustic models, and a multiple pronunciation lexicon comprising all pronunciation variants of the phone transcriptions in the MPT was used at recognition time. The results of these experiments are presented in terms of WER in table 3. As the performance on the SC and IN data were too low to draw valid conclusions, they are excluded from table 3. It is important to stress that in both experiments each time one recogniser was trained on a data set comprising all four speech styles; no *speech style-specific* recognisers were used.

PT	lexicon	speech style	WER(%)
MPT	mult.	RS	13.3
		LC	41.7
APT	canon.	RS	11.7
		LC	42.6

Table 3: Recognition results with different transcriptions.

5 DISCUSSION

As far as the validation of the PTs with regard to the reference transcription is concerned, table 2 shows the same tendencies as reported in [6]. Both the MPT and the APT show similar differences in distance from the reference transcription. The MPT and the APT of the Read Speech data seem to match the reference transcription best, whereas the PTs of the Sponteneous Conversations seem to differ most from Tref. Moreover, it is clear that in all cases the MPT resembles the reference transcription more than the APT, which is not surprising as the APT is only a concatenation of canonical transcriptions, whereas the MPT and the ref-

erence transcription are both hand-made: the former by one person, the latter as a consensus transcription.

However, when the results from table 3 are interpreted in terms of table 2, it is clear that the transcription resembling the reference transcription most may not always yield the best recognition performance.

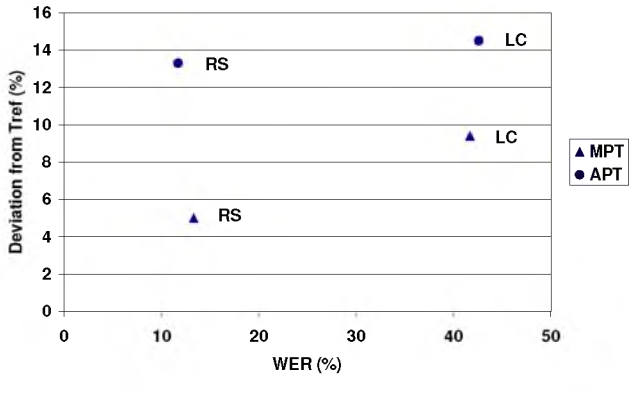

Figure 1: Recognition results with MPT and APT

Figure 1 shows that the Read Speech data are better recognised using the APT, whereas the Lectures data are better recognised using the MPT. This supports our hypothesis that validating PTs should preferably not be conducted by making them more similar to a reference transcription, but rather as an application-oriented proces when the application is known.

6 CONCLUSION

This paper introduced a new PhD project in which phonetic transcriptions will be generated on a large scale using ASR technology. The main aim of the research though, is to come up with new validation procedures for (automatic) phonetic transcriptions. Until now, validation of phonetic transcriptions was typically performed by lining up each symbol of the phonetic transcription with a reference transcription. We believe that transcriptions that are optimal for one application won't necessarily be the best for another one. Therefore we believe that the adequateness of phonetic transcription depends on the possible applications they are used for.

A pilot study supported our hypothesis. We showed that a recogniser trained with a basic automatic phonetic transcription (a concatenation of canonical transcriptions) can already outperform a recogniser trained with a manual transcription on the same recognition task. In the near future speech style-specific recognisers will be built to perform new large scale recognition experiments involving other speech styles and different applications.

ACKNOWLEDGEMENTS

This research was funded by the "Stichting Spraaktechnologie" (Foundation for Speech Technology).

REFERENCES

[1] S. Greenberg, "Speaking in shorthand - a syllable-centric perspective for understanding pronunciation variation," *Speech Communication*, vol. 29, pp. 159–176, 1999.

[2] A. Vorstermans, J.-P. Martens, and B. Van Coile, "Automatic segmentation and labelling of multilingual speech data," *Speech Communication*, vol. 19, pp. 271–293, 1996.

[3] M. Wester, J.M. Kessens, C. Cucchiarini, and H. Strik, "Obtaining phonetic transcriptions: A comparison between expert listeners and a continuous speech recogniser," *Language and Speech*, vol. 44, pp. 377–403, 2001.

[4] N. Oostdijk, "The Spoken Dutch Corpus: Overview and first evaluation," in *Proceedings LREC '00*, 2000, pp. 887–893.

[5] C. Cucchiarini, D. Binnenpoorte, and S. Goddijn, "Phonetic transcriptions in the Spoken Dutch Corpus: how to combine efficiency and good transcription quality," in *Proceedings Eurospeech '01*, 2001, pp. 1679–1682.

[6] D. Binnenpoorte, S. Goddijn, and C. Cucchiarini, "How to improve human and machine transcriptions of spontaneous speech," in *Proceedings ISCAA and IEEE Workshop on Spontaneous Speech Processing and Recognition (to appear)*, 2003.

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

[8] J.M. Kessens and H. Strik, "Lower WERs do not guarantee better transcriptions," in *Proceedings of Eurospeech '01*, 2001, pp. 1721–1724.

[9] C. Cucchiarini and D. Binnenpoorte, "Validation and improvement of automatic phonetic transcriptions," in *Proceedings ICSLP '02*, 2002, pp. 313–316.

[10] R.J.J.H. van Son, D.M. Binnenpoorte, H. van den Heuvel, and L.C.W. Pols, "The IFA Corpus: a phonemically segmented Dutch "open source" speech database," in *Proceedings of Eurospeech '01*, 2001, pp. 2051–2054.

[11] S. Young et al., "The HTK book (for HTK version 3.2)," Tech. Rep., Cambridge University Engineering Department, 2003.