Validation of Phonetic Transcriptions Based on Recognition Performance

Christophe Van Bael, Diana Binnenpoorte, Helmer Strik, Henk van den Heuvel

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

Abstract

In fundamental linguistic as well as in speech technology research there is an increasing need for procedures to automatically generate and validate phonetic transcriptions. Whereas much research has already focussed on the automatic generation of phonetic transcriptions, far less attention has been paid to the validation of such transcriptions. In the little research performed in this area, the estimation of the quality of (automatically generated) phonetic transcriptions is typically based on the comparison between these transcriptions and a human-made reference transcription. We believe, however, that the quality of phonetic transcriptions should ideally be estimated with the application in which the transcriptions will be used in mind, provided that the application is known at validation time. The application focussed on in this paper is automatic speech recognition, the validation criterion is the word error rate. We achieved a higher accuracy with a recogniser trained on an automatically generated transcription than with a similar recogniser trained on a human-made transcription resembling a human-made reference transcription more. This indicates that the traditional validation approach may not always be the most optimal one.

1. Introduction

In the last decade, many large speech corpora have become available for fundamental and application-oriented research. Whereas almost all corpora provide orthographic transcriptions, they often lack Phonetic Transriptions (PTs). This is troublesome, as PTs are often required for phonetic, phonological and pathological research, as well as for speech synthesis and speech recognition applications.

The first attempts to fulfill the need for PTs focussed on the generation of Manual Phonetic Transcriptions (MPTs). However, the production of MPTs proved to be time-consuming and expensive. Moreover, MPTs tend to be error-prone due to fatigue and subjective judgements of the transcribers [1]. Therefore research has shifted to investigating the usability of Automatically generated Phonetic Transcriptions (APTs).

A wide range of procedures to automatically generate phonetic transcriptons has already been developed. The resulting APTs can be used as an alternative to MPTs, as a reference with which human transcribers can compare their transcriptions, or as a starting point human transcribers can modify. The latter approach is implemented in the context of the Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN) [2], a joint Dutch-Flemish project compiling a 10 million words corpus of which 1 million words will receive an MPT (i.e. an APT modified by human transcribers) [3], and 9 million words an APT (generated without the intervention of human transcribers).

The general goal of our research is to acquire knowledge about how to automatically generate and validate PTs in the best possible way. In this paper we focus on the validation of PTs. Until now, (automatically generated) PTs have been typically validated by comparing them to a human-made reference transcription, because at validation time often no specific applications are known in which the PTs will be used. However, if such applications are known, we believe that these applications should be taken in consideration when estimating the quality of the PTs, as the importance of differences between a PT and a reference transcription may vary per application. In this paper we focus on Automatic Speech Recognition (ASR) as an application in which PTs are commonly used, and we use the Word Error Rate (WER) as the validation criterion.

Recent research [4] has shown that there is no direct relation between the performance of a recogniser and the similarity between an APT generated by that recogniser with a consensus transcription. [4] proved that *lower WERs do not guarantee better transcriptions*, where a *better* transcription meant a transcription resembling a consensus transcription more. [5] showed this can also hold the other way around: transcriptions more similar to a human-made reference transcription and used to train recognisers do not guarantee lower WERs. It was shown that read speech was better recognised by a recogniser trained on a simple APT than by a similar recogniser trained on an MPT more similar to a consensus transcription.

Whereas in [5] PTs were validated in terms of the accuracy obtained with recognisers trained on material comprising four different speech styles, in this paper PTs are evaluated by means of their contribution to the accuracy of speech-style specific recognisers. The rationale was that if a recogniser trained on an APT would again show a higher recognition accuracy than a similar recogniser trained on a PT resembling a human-made reference transcription more, this would again support our belief that PTs should ideally be validated with the applications in which these transcriptions will be used in mind, rather than by simply comparing the PTs with a consensus transcription.

We trained three recognisers, each one on a different type of PT. The first recogniser was trained on an MPT, the second one on an APT, and the third one on an APT in which several optional phonological rules had been applied. The application of the rules is based on the work of [6]. The three PTs were validated with respect to the distance between the transcriptions and a human-made reference transcription on the one hand and with respect to the performance yielded by the recognisers trained on those PTs on the other hand. The outcomes of these two validation approaches were then compared to each other.

In what follows, first the material and the general idea behind the experiments are introduced. Then the results are presented and dicussed, followed by a conclusion and our ideas for future research in this area.

2. Material and method

2.1. Material

2.1.1. Corpora

Phonetic transcriptions comprising data from two speech-styles were used: read speech (RS) and lectures (LC). Two corpora were used for the experiments with the RS data, and two corpora for the experiments with the LC data (see table 1). Each time one corpus (*RefCorp*) was used to compute the distance between the PTs (one MPT and two APTs) and the reference transcription, and the other corpus (*RecCorp*) was used to perform the recognition experiments. The latter corpus was always divided in three separate data sets comprising data to train, tune and test the recognisers. A separate data set for tuning was needed in order to scale the weight of the recognisers' language models with regard to the acoustic models and to determine the optimal word insertion penalties to control the number of insertions and deletions. There was no overlap between the corpora.

All data sets were extracted from the so-called *core corpus* of the CGN (release 6)[2]. They all comprised similar data per speech style, thus the recognisers were trained on data representative of the test data. Table 1 provides the details of the data sets.

corpus	RefCorp	RecCorp		
data set / speech style	reference set	train set	tune set	test set
RS	682	49898	998	16610
LC	892	10800	999	3579

Table 1: *Number of words in the data sets.*

2.1.2. Transcriptions

In all, 13 PTs were used (see table 2). Per speech style (RS and LC), three types of transcriptions (MPTs, APTs and *enhanced* APTs) were used to train the recognisers, and three similar transcriptions were used to compute the string edit distance between these transcriptions and the reference transcription.

The MPTs were already provided in the core corpus of the CGN. One MPT was available per sound file. The first APT (*APT1* hereafter) was generated by concatenating PTs 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 [7]. All obligatory word-internal phonological processes [8] were applied on all PTs in this lexicon, according to previous research, among which [7]. The second APT (*APT2* hereafter), was an enhanced version of APT1. Progressive and regressive cross-word assimilation, as well as cross-word degemination rules were applied on APT1, thus resulting in APT2. This procedure is based on [7] and [6], who applied the same rules on their APTs to closer resemble a human-made consensus transcription.

The reference transcription (*Tref* hereafter) of RefCorp was a consensus transcription, generated from scratch by two expert listeners [9]. It was used to compute the distances between the MPT, APT1 and APT2 of the RS and LC data in RefCorp and the reference transcription. The transcriptions of RefCorp were generated in a similar way as and they were thus representative of the transcriptions of the training data in RecCorp.

task / style	training acoustic models (RecCorp)	computing distance with Tref (RefCorp)	
RS	MPT	MPT	Tref
	APT1	APT1	
	APT2	APT2	
LC	MPT	MPT	
	APT1	APT1	
	APT2	APT2	

Table 2: *13 Different phonetic transcriptions.*

2.1.3. Lexica

For both speech styles, three sets of lexica were used, one set for each recogniser (see table 3). Those sets comprised a training lexicon to derive PTs from (except for the MPTs, as those transcriptions were already available), and one tune-test lexicon comprising only the pronunciation variants occuring in the tune and test sets. The tune-test lexica were compiled from the transcriptions of the tune and test sets. The transcriptions of these data were only used for the purpose of compiling those lexica.

As mentioned, no lexica were used to derive MPTs from. The lexicon covering the RS data used to tune and test the recognisers trained with the MPTs had a pronunciaton/lexeme ratio of 1.25, the lexicon covering the LC data had a ratio of 1.33.

For the recognisers trained on the APTs, lexica were also used to derive these APTs for the training data in order to train the acoustic models. The tune-test lexica used for the tuning and testing of the recognisers built with the APT1s were canonical lexica. The lexica used for the tuning and testing of the recognisers trained with the APT2s were multiple pronunciation lexica similar to the lexica used for training. They were generated by applying the phonological rules to the APT1s of the tune and test sets of RecCorp (so for practical reasons the lexica were built from the PTs here).

The training lexicon comprising the RS training data had a pronunciaton/lexeme ratio of 1.08, the lexicon covering the RS tune and test data a ratio of 1.07. The training lexicon covering the LC training data had a ratio of 1.1 and the lexicon used for tuning and testing that recogniser had a ratio of 1.07. Whereas [10] found best recognition results with a ratio of 1.4 and good results up to a ratio of 2.5, for now we chose to stay as close as possbile to the phonological rules applied in and the resulting pronunciation variants generated in [7]. One important drawback in this procedure is that only 38 phone models were trained, whereas the CGN-phoneset used by [7], comprised 46 phones. Therefore undoubtly phonetic detail was lost in our transcription with respect to the one used in [7]. Moreover, some phonological rules (in particular the ones involving the voiced velar stop and the voiced velar fricative) could not be applied, as those phones were not present in our phoneset. Expanding the phone set and increasing the lexical variability may be a topic for further research. Table 3 presents the lexica (*mult.* representing *multiple pronunciation lexicon* and *can.* respresenting *canonical lexicon*) used for the training, tuning and testing of the recognisers, as well as their average number of pronunciations per lexeme (in brackets).

2.1.4. The alignment program and the architecture of the recognisers

To compare the MPT and the APTs with Tref, the Align program [1] was used. This program computes the string edit dis-

task / speech style	phonetic transcription	training	tuning and testing
RS	MPT	no lex. used	mult. (1.25)
	APT1	can. (1)	can. (1)
	APT2	mult. (1.08)	mult. (1.07)
LC	MPT	no lex. used	mult. (1.33)
	APT1	can. (1)	can. (1)
	APT2	mult. (1.10)	mult. (1.07)

Table 3: *Different lexica and the average number of pronunciations per lexeme.*

tance (the sum of all substitutions, insertions and deletions divided by the total amount of characters in Tref) between corresponding phoneme strings as well as a weighted distance based on articulatory features. Only the string edit distance was taken in account here.

The recognisers were built with the Hidden Markov Modelling toolkit HTK [11]. The systems used 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. All data were parameterised as Mel Frequency Cepstral Coefficients (MFCCs) with 39 coefficients per frame. The language models were backed-off bigram models trained per recogniser on the tune and test set data.

2.2. Method

The PTs were validated in two ways. First the traditional approach was followed by estimating the quality of the PTs by means of their string edit distance to Tref. In this approach the transcription that best matches the manually created reference transcription is considered to be the most optimal one.

Next the PTs were validated by means of their influence on the accuracy of the recognisers that used the transcriptions to train their acoustic models. By using different test lexica, one might argue that an extra variable was introduced possibly masking the effect of the PTs on the recognition accuracy. This procedure was preferred, though, because no other PTs and lexica than the ones involved in the experiments are likely to be available in reality. In all, 6 recognisers were trained and tested: 2 series of 3 recognisers, one series per speech style. Per speech style, one recogniser was trained on an MPT, one on an APT1 and one on an APT2. The six recognisers will be called *RS/MPT*, *RS/APT1*, *RS/APT2*, *LC/MPT*, *LC/APT1* and *LC/APT2* hereafter. In this approach the transcription leading to the lowest WER is considered to be the most optimal one.

The outcomes of these two validation techniques were then compared to each other.

3. Results and discussion

Our initial belief was that PTs should ideally be validated with their potential applications in mind. We believe a transcription better resembling a human-made reference transcription does not always yield the best results in all applications, and that therefore the traditional approach to the validation of phonetic transcriptions may not always be the most optimal one. The results obtained in the experiments support our belief.

3.1. Validation of the PTs by means of their distance to Tref

In this experiment the PTs were validated according to the traditional approach by comparing them to a human-made reference transcription. Table 4 presents the results in terms of substitutions (sub), deletions (del) and insertions (ins).

The MPTs of both the RS and the LC data resemble more to Tref than the two APTs. For both data sets, APT2 slightly resembles Tref more than APT1 does, but two times it's a close call. The results generally resemble the results reported in [6], but the differences in distance between APT1 and Tref on the one hand and APT2 and Tref on the other hand are much more outspoken in [6]. The differences with [6] are mainly due to the fact that we used a smaller phone set. Hence several rules could not be applied to APT1 in order to generate an APT2 that closer resembled the consensus transcription (see 2.1.3). Also, whereas all PTs of all RS data in RefCorp could be aligned with Tref, we found that 1.4% of the phones in the MPT of the LC data could not be aligned to the reference transcription due to practical reasons. In the alignment between APT1 of the LC data and Tref 9.1% of the phones could not be aligned and in the alignment between APT2 of the LC data and Tref 5.5% of the phones could not be aligned. The results in 4 are solely based on the successful alignments, thus neglecting the cases where no alignment could be conducted.

Still we can conclude that according to the traditional approach to validating PTs (estimating their quality with regard to their overall distance to a reference transcription), for both data sets, the MPTs *proved* to be the *best* transcriptions, followed by the APT2s and the APT1s.

style	PT	sub (%)	del (%)	ins (%)	tot (%)
RS	MPT	3.1	0.5	1.4	5.0
	APT1	7.0	2.4	2.9	12.3
	APT2	6.1	2.7	2.5	11.3
LC	MPT	4.7	1.5	3.4	9.6
	APT1	7.3	1.8	6.6	15.7
	APT2	6.7	2.2	6.4	15.3

Table 4: *Distances between the transcriptions and Tref.*

3.2. Validation of the PTs by means of their influence on the WER

In this experiment the transcriptions were evaluated with a particular application (ASR) in mind. Therefore our evaluation criterion was the WER (the lower, the better). The recognisers' performances (in terms of WER) are presented in table 5. The performances are plotted against the distances of the PTs to Tref in figure 1. Whereas the LC data were significantly better recognised with recogniser LC/MPT than with recognisers LC/APT1 and LC/APT2 (this indicates that the transcription resembling the reference transcription most was the most optimal transcription *in this particular case, for these specific data*), the RS data were better recognised with recogniser RS/APT1 than with recogniser RS/MPT. This resembles the results obtained in [5]. Recogniser RS/APT1 also outperformed recogniser RS/APT2 trained on the *enhanced* APT and using a multiple pronunciation lexicon. This is probably due to the fact that the RS data were more carefully pronounced than the LC data (thus leaning more towards a canonical transcription), so that the RS recognisers suffered more from having multiple pronunciations in the test lexica than gaining from it. The pronunciation variants in

the more extensive lexicon covering the MPT of the tune and test RS data seem to have fit the data better than the transcriptions in the lexicon covering the APT2 of these data.

speech style	phonetic transcription	lexicon	WER(%)
RS	MPT	mult. (1.25)	9.6 ($\pm$0.5)
	APT1	can. (1)	8.3 ($\pm$0.5)
	APT2	mult. (1.07)	10.2 ($\pm$0.5)
LC	MPT	mult. (1.33)	21.4 ($\pm$1.4)
	APT1	can. (1)	25.5 ($\pm$1.4)
	APT2	mult. (1.07)	23.4 ($\pm$1.4)

Table 5: *Recognition results with different transcriptions. Between brackets 95% confidence interval.*

So, the recognition results from the recognisers trained, tuned and tested on read speech seem to support our belief that a PT resembling a human-made reference transcription more may not be the most optimal transcription for all applications. Here APT1 proved to be a better choice than APT2 and MPT (both resembling Tref more than APT1 did) to obtain a better recognition performance on the RS data.

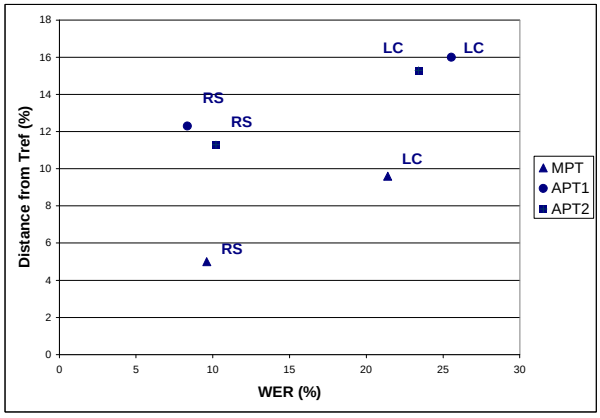

Figure 1: *Recognition results with MPT, APT1 and APT2.*

4. Conclusions

Vast amounts of phonetic transcriptions are required both for fundamental and for application-oriented research. Whereas many procedures have already been developed to automatically generate phonetic transcriptions, far less procedures or tests have been defined to validate such transcriptions.

We believe that phonetic transcriptions should ideally be validated on the basis of their contribution to the development of applications, rather than by a comparison with a human-made reference transcription (as is usually done). In this paper we have focussed on automatic speech recognition as an application for which phonetic transcriptions are commonly used. We used the word error rate as a validation criterion for our phonetic transcriptions. Our results support our belief that a phonetic transcription closer resembling a human-made reference transcription does not always guarantee best recognition performance. This indicates that the traditional approach to the validation of phonetic transcriptions may not always be the most optimal one.

5. Future research

In future research we will further investigate the relation between phonetic transcriptions and recognition accuracy. We will also study the effect of different speech styles on transcriptions generated by a recogniser. We will investigate whether the transcriptions and the pronunciation rules generated through forced recognition will show similar differences when generated for different speech styles. Finally, also the influence of APTs on segment duration statistics will be analysed. We expect that the quality of the estimation of the segment durations is directly related to the quality of the APTs itself.

6. Acknowledgements

This research was funded by the "Stichting Spraaktechnologie" (Foundation for Speech Technology), Utrecht, The Netherlands. The authors would like to thank Johan de Veth at A^{2}RT for useful suggestions concerning and practical help with the research.

7. References

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

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

[3] S. Goddijn and D. Binnenpoorte, "Assessing manually corrected broad phonetic transcriptions in the Spoken Dutch Corpus," in *Proceedings of ICPhS '03*, 2003, (to appear).

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

[5] C. Van Bael, W. Strik, and H. van den Heuvel, "Application-oriented validation of phonetic transcriptions: preliminary results," in *Proceedings of ICPhS '03*, 2003, (to appear).

[6] D. Binnenpoorte and C. Cucchiarini, "Phonetic transcription of large speech corpora: How to boost efficiency without affecting quality," in *Proceedings of ICPhS '03*, 2003, (to appear).

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

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

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

[10] J.M. Kessens, *Making a difference. On automatic transcription and modeling of Dutch pronunciation variation for automatic speech recognition*, Ph.D. thesis, University of Nijmegen, The Netherlands, 2002.

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