

On the Sufficiency of Automatic Phonetic Transcriptions for Pronunciation Variation Research

Christophe Van Bael, Hans van Halteren

Centre for Language and Speech Technology (CLST)
Radboud University Nijmegen, The Netherlands
{c.v.bael, hvh}@let.ru.nl

ABSTRACT

We investigated whether automatic phonetic transcriptions (APTs) can replace manually verified phonetic transcriptions (MPTs) in a large corpus-based study on pronunciation variation. To this end, we compared the performance of both transcription types in a classification experiment aimed at establishing the direct influence of a particular situational setting on pronunciation variation. We trained classifiers on the speech processes extracted from the alignments of an APT and an MPT with a canonical transcription. We tested whether the classifiers were equally good at verifying whether unknown transcriptions represent read speech or telephone dialogues, and whether the same speech processes were identified to distinguish between transcriptions of the two situational settings. Our results not only show that similar distinguishing speech processes were identified; our APT-based classifier yielded better classification accuracy than the MPT-based classifier whilst using fewer classification features.

Index Terms: automatic phonetic transcription, pronunciation variation

1. INTRODUCTION

The increasing availability of large speech corpora offers new opportunities for linguistic research. The release of the Spoken Dutch Corpus (Corpus Gesproken Nederlands; CGN, [1]), a 9M word corpus of contemporary Dutch speech, recently allowed us to start investigating a corpus-based Bayesian model describing the way in which several factors affect pronunciation variation. Since we study pronunciation variation by applying machine learning to phonetic transcriptions, our study depends on the availability of large amounts of annotated corpus material.

In previous experiments, we used ‘manually verified automatic phonetic transcriptions’ (MPTs) from the CGN. Present-day speech corpora are often annotated semi-automatically, for a check-and-correct procedure is attractive in terms of cost reduction. However, the manual verification of the phonetic transcriptions in the CGN still took 15 minutes for one minute of speech in formal lectures and 40 minutes for one minute of spontaneous speech [2]. This explains why the automatic transcription of only a limited amount of speech could be manually verified.

Recently, it was shown that the MPTs of the CGN can be approximated by means of an automatic transcription procedure requiring limited resources and minimal human effort [3]. Since, for our future research, we expect to require more phonetic transcriptions than the MPTs currently available in the CGN, our present study is aimed at testing whether the transcription procedure proposed in [3] can produce automatic

phonetic transcriptions (APTs) that are ‘good enough’ for a large corpus-based study on pronunciation variation.

In this paper, we compare the performance of such an APT and an MPT in a classification experiment aimed at establishing the direct influence of a particular situational setting on pronunciation variation. More specifically, we train classifiers on speech processes extracted from the alignments of an APT and an MPT of read speech and telephone dialogues with a canonical transcription, and we test whether the APT- and MPT-based classifiers are equally good at verifying whether phonetic transcriptions represent read speech or telephone dialogues. In addition, we test whether the APT- and MPT-based classifiers identify the same speech processes to distinguish between transcriptions of the two situational settings.

This paper is organised as follows. Section 2 presents the corpus material and the phonetic transcriptions. In Section 3, we describe our general methodology. Section 4 presents and discusses the results of the classification experiments. In Section 5, we present our conclusions.

2. MATERIAL AND TRANSCRIPTIONS

We experimented with read speech (RS) and telephone dialogues (TD) from the CGN. We excluded speech fragments that could not be reliably transcribed (broken words, overlapping speech, etc.). Table 1 presents the statistics of the data. The development data were used to optimise the automatic transcription procedure, the evaluation data were used to train and test our classification algorithm through standard ten-fold cross-validations. We successively tested on each group using the other nine groups for training.

Table 1: Statistics of the speech material.

Speech style	Development set		Evaluation set	
	Words	Speakers	Words	Speakers
RS	10,399	21	53,359	104
TD	10,175	28	45,469	81

The canonical transcriptions (CanTs) were generated by means of a lexicon-lookup procedure in which every word in the orthography was substituted with its standard pronunciation in a canonical pronunciation lexicon. The CanTs reflected the obligatory word-internal phonological processes of Dutch [4].

The MPTs were extracted from the CGN. They were generated in three steps. First, a canonical transcription was

generated. Second, two prominent phonological processes in Dutch (voice assimilation and degemination) were modelled at the word boundaries. Third, trained linguistics students (one student for the RS, two students (in succession) for the TD) verified the example transcription. They acted according to a strict protocol which allowed them to change the transcription only if it was unmistakably deviant from the acoustic signal [5].

The APTs were based on the CanTs introduced above. A four-step procedure tuned the CanTs towards the MPTs of the CGN. The procedure was individually optimised for the transcription of RS and TD [3]. The procedure first aligned the MPT and the CanT of the development data. Subsequently, it listed all phones in the CanT, along with the left and right neighbour phones, and the corresponding phones in the MPT. These phone mappings between the MPT and the CanT (and their frequencies) were used to estimate the probability of every phone in the MPT given its corresponding phones in the CanT. This knowledge was formalised as a set of decision trees (one tree per phone) which, in a third step, were used to generate pronunciation variants for the CanT of the words in the evaluation set. All phone variants with a probability lower than 0.1 were ignored to minimise the risk of over-generation. In the fourth step of the procedure, the remaining phone-level variants were combined to word-level variants, which were listed in a multiple pronunciation lexicon. Their probabilities were normalised so that the probabilities of all the variants of a word added up to 1. An HMM-based continuous speech recogniser selected the most likely pronunciation variant for each word in the orthography. The recogniser used two sets (RS and TD) of gender- and context-independent acoustic models [3].

3. METHODOLOGY

Our experiments were based on the assumption shown in Figure 1. It states that the influence of one variable (e.g. situational setting) on another variable (e.g. pronunciation variation) can be verified by the ability of a classification algorithm to derive information about the former variable from observations in the latter variable. If, for example, we assume that the situational setting influences pronunciation variation, a classification algorithm should be able to determine which situational setting led to an observed set of pronunciation processes. Furthermore, we assumed that the quality of the classification is indicative of the amount of information about the source variable that can be retrieved from the target variable.

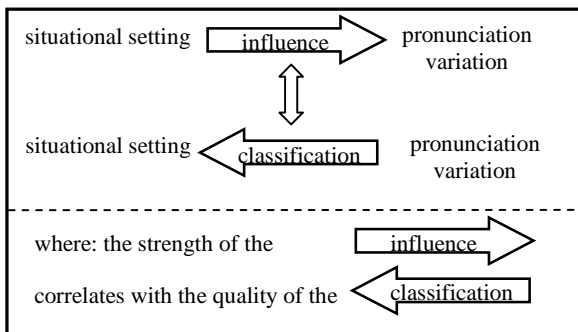


Figure 1: Verification through classification.

This method of “verification-through-classification” requires three components: a classification algorithm (3.2), features that can be used for classification (3.1), and a measure to express the classification quality (3.3).

3.1 Classification Features

We derived classification features (speech processes representing pronunciation variation) from the alignments of MPTs and CanTs on the one hand, and APTs and CanTs on the other hand. Figure 2 illustrates the alignment of an MPT and a CanT, conducted with ADAPT, a dynamic programming algorithm designed to align two strings of phonetic symbols according to their articulatory distance [6]. The top two tiers represent the canonical pronunciation and the observed transcription. Word boundaries are represented as vertical bars, traces of phone insertions in the CanT and traces of phone deletions in the MPT as a dash. The third tier highlights the discrepancies between the CanT and the MPT as substitutions (s), deletions (d) and insertions (i) of phones. Insertions at word boundaries (such as the /w/ in Figure 2) and the remnants of degemination processes (such as the /n/ and the /st/-cluster) were always attributed to the second word. The last two lines in Figure 2 present the Dutch orthography and an English translation.

CanT		w	a	r		z	o		-	e	n		f	o	n		n	a	s	t		s	t	O	n	t		
MPT		w	a	-		s	o	w	@	-		f	o	-		n	a	-		s	t	O	n	t				
processes						d	s				i	s	d			d												
Dutch		w	a	a	r	z	o	é	é	n	f	o	o	n	n	a	a	s	t	s	t	o	n	d				
English		(where one such phone stood next to)																										

Figure 2: Alignment of phonetic transcriptions.

The frequent occurrence of particular words can be highly indicative of a specific socio-situational setting. In particular the transcriptions of the TD were easy to distinguish by means of the frequent occurrence of (transcriptions of) short confirming ‘words’ such as ‘ja’ (yes) and ‘uh-huh’. In order to exclude this lexical influence from our experiments, we measured the frequency of pronunciation processes (retrieved from the alignments) in fixed lexical contexts only.

First of all, we examined the word in which a process occurred. Two processes were only considered the same if they occurred at the same position in the same word. Our classification algorithm had access to classification features of the form:

$$[A] \rightarrow [@] / [[v _ n]]_{\text{CanT}} \quad (1)$$

i.e. a canonical /A/ was reduced to /@/ in the word ‘van’. In addition, the algorithm considered the two adjacent speech processes, which were represented in classification features of the form:

$$[A] \rightarrow [@] / [[(v \rightarrow f) _ (n \rightarrow n)]]_{\text{CanT}} \quad (2)$$

i.e. a canonical /A/ was reduced to /@/ in the word ‘van’ when preceded by a substitution of /v/ with /f/ and when followed by /n/ (null-process).

We also ensured that the speech processes (and their contexts) occurred frequently enough to obtain reasonably reliable probability estimates. We only considered words which occurred at least 100 times (RS and TD combined), and we demanded the canonical context to occur in at least 80 percent of all samples (a sample was defined as the selected utterances from one speaker in one recording). Demanding presence in all samples would have severely limited the number of available classification features. As a result of our selection criteria, the MPT-based classifier ($\text{class}_{\text{MPT}}$) could work with 306 features, and the APT-based classifier ($\text{class}_{\text{APT}}$) with 183 features, all located at word boundaries or in one of 17 frequent words. We will come back to the different number of classification features for the $\text{class}_{\text{MPT}}$ and the $\text{class}_{\text{APT}}$ in the discussion of our results (4.3).

Our final restriction was aimed at eliminating the influence of multi-word expressions, for multi-word expressions can affect the pronunciation of words. We excluded processes in words collocating strongly with their left or right neighbour, i.e. if the words were found adjacent at least 5 times (in the whole set) and if their Mutual Information score was at least 5.

3.2 Classification Algorithm: Linguistic Profiling

The classification experiments were conducted by means of Linguistic Profiling [5]. The training and the use of the algorithm (through ten-fold cross-validations with the APT- and MPT-based classification features, resp.) consisted of four steps.

First, the algorithm determined the norm and the standard deviation for every classification feature in the alignments of the transcriptions of the RS and the TD. The norm of a feature was defined as its mean application probability in all samples. Per sample, the application probability was defined as the count of that feature divided by the number of occurrences of its canonical context. For example, in our MPTs, the application probability of the feature:

$$[k] \rightarrow [g] / [| o _ _ |]_{\text{CanT}} \quad (3)$$

was 0.23, with a standard deviation of 0.30.

Second, for every sample, the algorithm determined how many standard deviations the count of every feature differed from its norm. For example, in sample fn008060/N08082, the application probability of feature (3) was 0.43. This was 0.67 standard deviations above the norm (0.23). The algorithm took the difference (+0.67) as the value of feature (3) for this sample.

Third, the classification algorithm created separate verifiers for RS and TD. The algorithm averaged the values of every feature over all positive training samples, thus producing a verifier aimed at recognising RS (and rejecting TD), and another one aimed at recognising TD (and rejecting RS).

Fourth, in order to classify a held-out sample, the verifiers compared each feature value with the RS and TD model, and determined a distance based on a weighted combination of the values of the individual features in the sample compared to the means in the RS and TD models (see [7] for a more thorough description of Linguistic Profiling and its parameters). The distances for all classification features were combined into a single overall verification score, which was subsequently

compared to the overall scores of the negative training samples. A threshold value determined whether the held-out sample could be accepted as RS or TD.

3.3 Measuring Classification Accuracy

Linguistic Profiling is a verification algorithm. Therefore, in previous studies, the classification scores were mainly evaluated by means of standard verification measures, viz. the False Reject Rate (FRR), False Accept Rate (FAR) and Equal Error Rate (EER). In the current experiments, however, the EER was often zero and could therefore not be used as our main quality measure. Instead, we used a measure which we termed the Cluster Separation Score (CSS). The CSS takes into account both the density of the two clusters and the distance of their centres, and it is formalised as in (4), where S_+ and S_- are the positive and negative test samples:

$$\text{CSS} = \frac{\text{Mean}_{s \in S_+}(\text{Score}(s)) - \text{Mean}_{s \in S_-}(\text{Score}(s))}{\text{Stddev}_{s \in S_+}(\text{Score}(s)) + \text{Stddev}_{s \in S_-}(\text{Score}(s))} \quad (4)$$

Higher CSSs indicate higher classification accuracy and lower EERs. A CSS of 2 or higher indicates near-perfect classification.

4. RESULTS AND DISCUSSION

We tested whether the APT-based classifier and the MPT-based classifier were equally good at verifying whether unknown transcriptions represent read speech or telephone dialogues. In addition, we tested whether the classifiers identified similar distinguishing speech processes for read speech and telephone dialogues.

4.1 Degree of Classification of the Samples

Table 2 shows the average CSSs of the MPT-based classifier and the APT-based classifier. The CSSs of the $\text{class}_{\text{MPT}}$ confirm that classification with the MPT-based classifier was nearly perfect and that the TD verifier outperformed the RS verifier. Similar to the MPT-based classifier, also the APT-based TD verifier outperformed the APT-based RS verifier with a factor of 1.5.

Table 2: Classification accuracy for RS/TD verifiers (in CSS). Higher CSSs reflect higher classification accuracy.

	RS verifier	TD verifier
$\text{class}_{\text{MPT}}$	2.01	3.09
$\text{class}_{\text{APT}}$	3.07	4.55

However, the most interesting result is that the APT-based classifier consistently outperformed the MPT-based classifier. Our classification algorithm must have picked up stronger distinguishing speech processes from the APTs than from the MPTs. This may be due to a smaller number of intra- and inter-transcriber inconsistencies in the APTs than in the MPTs, and due to our automatic transcription procedure which must have applied MPT-based pronunciation variation to the CanT whilst leaving out potential inconsistencies of the MPT from the CGN.

4.2 Classification of the Individual Samples

In order to test whether the individual samples were classified similarly with the APT- and the MPT-based classifiers, we investigated the verification score of the individual samples at a parameter setting giving good results for both the APT- and MPT-based TD verifiers. Figure 3 plots the verification score of the RS and the TD test samples (x's and o's, respectively) according to the TD class_[MPT] (x-axis) and the TD class_[APT] (y-axis).

The graph reflects good classification quality of both TD verifiers. Despite some outliers, the RS and TD clusters are well-separated. Although the spreading of the RS samples is similar in both dimensions, it hardly disturbed the separation of RS and TD samples. Furthermore, the APT-based classifier recognised the TD samples even better than the MPT-based classifier. The vertical spreading of TD samples is small, while there is a large horizontal spreading.

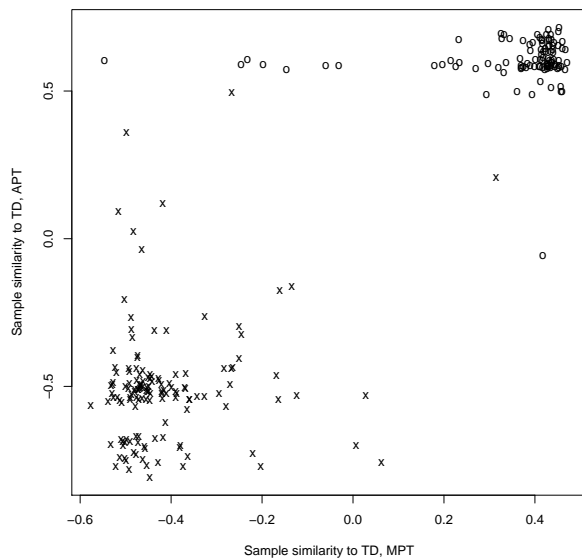


Figure 3: Verification score of RS (x) and TD (o) samples according to TD class_[MPT] and TD class_[APT].

4.3 Similarity of the classification features

Our APT-based classifier which, because of the APTs' closer resemblance with the CanTs, worked with fewer classification features than the MPT-based classifier (183 vs. 306), largely identified the same speech processes as characteristic for either one of the two situational settings. A comparison of the ten most distinguishing classification features of the TD class_[APT] and the TD class_[MPT] investigated in (4.2) showed that the two top-tens had seven features in common. This explains the similar classification behaviour observed in Figure 3.

One of the three features which was absent in the top-ten of the TD class_[APT] only just missed the top-ten of the TD class_[MPT]. The other two features, the reduction of /A/ (to schwa) in the word 'dat' and the deletion of /l/ in the word 'als', can be explained by the conservative nature of our APT. Since the APTs were based on the canonical transcriptions, and since they were tuned towards the MPTs by means of decision

trees generating lexical pronunciation variants, the APTs were bound to be more similar to the canonical transcriptions than the MPTs. We are inclined to believe that only more conservative pronunciation variants made it to the RS and the TD recognition lexicons.

5. CONCLUSIONS

In this study, we investigated whether APTs can replace MPTs in a large corpus-based study on pronunciation variation. More specifically, we compared the influence of APTs and MPTs in a classification experiment aimed at establishing the direct influence of a particular situational setting (read speech or telephone dialogues) on pronunciation variation.

We learned that our APT-based classifier was better at determining which situational setting an unseen phonetic transcription represented. Whereas in general the same speech processes were identified as characteristic for either read speech or telephone dialogues, the overall classification accuracy of our APT-based classifier was higher than the accuracy of our MPT-based classifier. This is encouraging, for it strengthens our belief that automatic phonetic transcriptions may be as suitable for our research on pronunciation variation as manually verified phonetic transcriptions often delivered with contemporary speech corpora. At the same time, our results might even question the justifiability of the expenses involved in the manual verification of phonetic transcriptions in future annotation tasks.

6. ACKNOWLEDGEMENT

The work of Christophe Van Bael was funded by the Speech Technology Foundation, Utrecht, the Netherlands.

7. REFERENCES

1. Oostdijk, N. The design of the Spoken Dutch Corpus. In: Peters P., Collins P., Smith A. (Eds.) *New Frontiers of Corpus Research*. Rodopi, Amsterdam, pp. 105-112, 2002.
2. Demuynck, K., Laureys, T., Gillis, S. Automatic generation of phonetic transcriptions for large speech corpora. In: *Proc. ICSLP*, pp. 333-336, 2002.
3. Van Bael, C., Boves, L., van den Heuvel, H., Strik, H. Automatic Phonetic Transcription of Large Speech Corpora: a Comparative Study. In: *Proc. ICSLP*, 2006.
4. Booij, G. *The Phonology of Dutch*. Oxford University Press, New York, 1999.
5. Goddijn, S., Binnenpoorte, D. Assessing Manually Corrected Broad Phonetic Transcriptions in the Spoken Dutch Corpus In: *Proc. ICPHS*, pp. 1361-1364, 2003.
6. Elffers, B., Van Bael, C., Strik, H. *ADAPT: Algorithm for Dynamic Alignment of Phonetic Transcriptions*. <http://lands.let.ru.nl/literature/elffers.2005.1.pdf>, 2005.
7. Halteren, van, H. Linguistic profiling for author recognition and verification. In: *Proc. ACL*, pp. 200-207, 2004.