On the Usefulness of Large Spoken Language Corpora for Linguistic Research

Christophe Van Bael, Helmer Strik, Henk van den Heuvel

Department of Language and Speech, University of Nijmegen, the Netherlands

e-mail: {c.v.bael,w.strik,h.v.d.heuvel}@let.kun.nl

Abstract

In the past, fundamental linguistic research was typically conducted on small data sets that were handcrafted for the specific research at hand. However, from the eighties onwards, many large spoken language corpora have become available. This study investigates the usefulness of large multi-purpose spoken language corpora for fundamental linguistic research. A research task was designed in which we tried to capture the major pronunciation differences between three speech styles in context-sensitive re-write rules at the phone level. These re-write rules were extracted from the alignments of both a manual phonetic transcription and an automatic phonetic transcription with a canonical reference transcription of the same material.

1. Introduction

In the past, fundamental linguistic research was typically conducted on small data sets that were handcrafted for the specific research at hand. However, from the eighties onwards, many large (often multi-purpose) spoken language corpora have become available (Lamel et al. 1986; Godfrey et al., 1992; Oostdijk 2000). Whereas the speech technology community has already made extensive use of such corpora for some years, the use of these spoken language corpora in linguistic research has been quite limited. We believe, however, that also linguistic research might benefit from the use of large spoken language corpora.

The study presented in this paper investigates the possibility of charting the major pronunciation differences of three different speech styles (speech recorded at public lectures, read speech and speech recorded from telephone dialogues) through data-driven research. To this end, speech-style specific context-sensitive re-write rules were retrieved from the alignments of different phonetic transcriptions with a canonical reference transcription of the data. These re-write rules defined in which contexts which phones were substituted, deleted or inserted in the different speech styles.

In a first experiment, a manually verified broad phonetic transcription was aligned with a canonical reference transcription. This experiment investigated the usability of large corpora comprising manual phonetic transcriptions for this type of research. The three resulting rule sets (one rule set per speech style) were statistically compared with each other to chart the major differences between the pronunciation characteristics of the three speech styles.

In a second experiment, an automatic phonetic transcription of the same data was aligned with the same reference transcription in order to investigate the potential of large corpora lacking manual phonetic transcriptions for this type of research. The resulting rule sets were compared to the corresponding rule sets of the first experiment to test whether similar patterns could be found in rule sets obtained from the alignment of an automatic phonetic transcription and a reference transcription on the one hand, and from the alignment of a manually verified phonetic transcription and a reference transcription on the other hand.

This paper is organised as follows. In section 2, the general idea behind the research is introduced, as well as the material used in the experiments. In section 3, the results of the experiments are presented, and in section 4, the results are discussed. Finally, in section 5, general conclusions and plans for future research are presented.

2. Method and Material

2.1. Method

Two experiments were conducted in which pronunciation characteristics of three speech styles were captured in context-sensitive re-write rules at the phone level. The three speech styles represented different degrees of articulatory precision, ranging from well-articulated speech (read speech and, to a lesser extent, public lectures) to conversational speech (telephone dialogues). It was expected that the well-articulated speech styles would be least deviant from each other, and that the speech from the public lectures would differ less from the speech in the telephone dialogues than the read speech.

In the first experiment, re-write rules were obtained from the alignment of a manually verified phonetic transcription (MPT) with a canonical reference transcription (RPT). In the second experiment, re-write rules were obtained from the alignment of an automatically generated phonetic transcription (APT) with the same reference transcription. Per speech style, the RPT, MPT and APT were transcriptions of the same data. Therefore, the rule-sets derived in the two experiments could be compared.

The alignments were obtained with Align (Cucchiarini, 1996). Align is a dynamic programming algorithm that decides on the optimal alignment of two strings of phonetic symbols according to a matrix in which the acoustic distance between different phonetic symbols is defined at the articulatory feature level. Whenever there were mismatches between phones in the MPTs and the APTs with regard to the phones in the RPT, re-write rules were formulated. The left-hand side of these rules consisted of the phone in the RPT, its left context (two phonetic symbols) and its right context (two phonetic symbols). The right-hand side of the re-write rules defined the substituted, deleted or inserted phone found in the

MPT or in the APT. Align was prevented from aligning phones across word boundaries. This means that only the transcriptions of the same words in the MPT and the RPT or in the APT and the RPT could be aligned with each other.

Per speech style, the rules of which the contexts occurred frequently in all RPTs were selected for further research. We opted to select the rules on the basis of the presence of their *contexts* in all RPTs, because we did not want to restrict our study to *rules* that occurred in all speech styles. In our study, also rules that did not occur in *all* speech styles (even though they could have occurred, as their context was frequently present in all RPTs) were investigated. We normalised for the differences in size of the different data sets. In order for a rule to be selected, its context had to occur at least $N_{context}$ times in the RPTs of all data sets. The threshold $N_{context}$ was dependent on the number of phones in the different data sets. As a result, the rule context had to occur at least 4 times in the public lectures, 22 times in the read speech and 9 times in the telephone dialogues in order for the corresponding rules to be selected for further investigation.

After selecting characteristic rules for all data sets, the Rule Application Probability (RAP) of each rule was computed. The RAP of a rule was defined by the number of times the rule was applied in the MPT or the APT (N_{rule}), divided by the number of times the rule could have been applied (i.e. the number of times the context was present in the RPT, $N_{context}$). Thus:

$$RAP = N_{rule} / N_{context}$$

If, for example, a rule was applied 30 times (N_{rule}) in the MPT, whereas the context ($N_{context}$) of the rule occurred 100 times in the RPT, the RAP of the rule in the MPT was 0.3.

By investigating the RAPs of the rules per speech style, and by investigating the RAPs obtained from the data of the other two speech styles, pronunciation differences between the three speech styles could be charted.

2.2. Material

2.2.1. Speech data and orthographic transcriptions

All speech data and the orthographic transcriptions of the data were taken from the sixth release of the Spoken Dutch Corpus (Oostdijk, 2000). The Spoken Dutch Corpus (Corpus Gesproken Nederlands – CGN) is a typical multi-purpose spoken language corpus comprising about 9 million words, of which 1 million words received -among other annotations- a manually verified broad phonetic transcription. Data from three speech styles were selected: speech recorded at public lectures (PL), read speech (RS) and speech recorded from telephone dialogues (TD). The data sets were divided in training and test sets. The speech material and the transcriptions in the training sets were used to train the acoustic models with which the continuous speech recogniser had to generate the APTs (through forced alignment). The transcriptions of the data in the test sets were used to derive the re-write rules. Statistics of the data sets are provided in Table 1.

Speech style	Training sets	Test sets	
Public Lectures (PL)	11,843	3,461	16,977
Read Speech (RS)	51,082	17,011	86,830
Tel. Dialogues (TD)	38,657	8,566	35,065

Table 1: Number of words in the training sets (column 1), number of words (column 2) and number of phones (column 3) in the test sets.

The PL and RS data were recorded at 16kHz with either close-talking or table-mounted microphones. The telephone dialogues were recorded as an 8kHz A-law coded signal through a telephone platform.

2.2.2. Phonetic transcriptions

The RPTs of the data were generated through a lexical lookup procedure with the orthographic transcriptions of the data and CELEX (Baayen et al., 1995), a validated canonical lexicon comprising 381K lexemes and their Dutch pronunciation. The transcriptions of all OOVs were inserted from the Dutch canonical lexicon delivered with the Spoken Dutch Corpus. The transcriptions in this lexicon were generated using a grapheme-to-phoneme converter that was trained on the original CELEX database. All phonetic transcriptions taken from the lexicon of the Spoken Dutch Corpus were manually verified and checked for consistency with the CELEX entries. In the resulting canonical lexicon, all obligatory word-internal phonological processes (Booij, 1999) were applied. No crossword phonological processes (e.g. assimilation of voice, degemination) were applied to the RPTs.

The MPTs were delivered with the Spoken Dutch Corpus. In order to generate these MPTs, expert phoneticians took a standard phonetic transcription as a starting point. This phonetic transcription was an enhanced version of a canonical transcription of the data. The transcribers were specifically asked to change the original transcription only if they felt confident about the changes they were about to make. Therefore, the MPTs should be considered to be manually *verified* phonetic transcriptions, rather than manually *generated* phonetic transcriptions. In addition, a bias towards the canonical transcription is to be expected in the MPTs.

The APTs were generated through forced alignment. Based on the orthography and speech style-specific acoustic models, a continuous speech recogniser was forced to choose the most optimal phonetic transcription for every word from a list of possible transcriptions in a multiple pronunciation lexicon. This lexicon was an extended version of the canonical lexicon used to generate the RPTs. The first two and the last two phones of each transcription in the lexicon could now be deleted or replaced by a small set of alternatives. Phone insertions were excluded at this stage. All pronunciations had to contain at least one phonetic symbol, i.e. no words could be left un-transcribed. The possible deletion of phones at the beginning and end of each word allowed for degemination across word boundaries, and the possible substitution of phones allowed for crossword assimilation of voice.

2.2.3. Continuous Speech Recogniser: HTK

The continuous speech recogniser with which the APTs were generated, was built with the HTK toolkit (Young et al., 2001). Per speech style, 41 left-to-right context- and gender-independent phone models were built with 32 Gaussian mixture components per state: 38 phone models, one garbage model for unintelligible speech and non-speech sounds, one silence model and a short model to capture the optional short pauses after words. All data were parameterised as Mel Frequency Cepstral Coefficients (MFCCs) with 39 coefficients per frame.

3. Results

3.1. Experiment 1

566 rules were selected according to the selection procedure described in section 2.1. Subsequently, the RAPs of these 556 rules were compared in pairs, after which smaller rule sets were retained for further investigation. Of each pair of rule sets, only the rules were selected for which the RAPs differed in the two rule sets. By retaining only those rules, we wanted to investigate whether speech style specific pronunciation differences could be retrieved from the transcriptions of the data.

The figures in Table 2 present the Pearson correlations between the RAPs under investigation (N is the number of RAPs taken into account). In this and in all following tables, the significance levels of the correlations (r) are indicated as follows: one asterisk indicates a significant correlation at the .05 level (2-tailed), two asterisks indicate a significant correlation at the .01 level (2-tailed). The higher the correlation between the RAPs of two speech styles, the more related the RAPs of one speech style are to the RAPs of the other speech style. Also the significance of the difference between the mean RAPs (ΔM) of one speech style and another speech style are presented: one asterisk indicates a significant difference at the .05 level ($p < .05$, 2-tailed paired samples t-test), two asterisks indicate a significant difference at the .01 level ($p < .01$, 2-tailed paired samples t-test). A positive value for the ΔM of the RAPs of two speech styles indicates that the mean RAP of the first speech style was higher than the mean RAP of the second speech style. A negative value for ΔM indicates the opposite.

MPT-RPT	N	r	ΔM
P Lectures – R Speech	191	.833 **	-.018 *
P Lectures – T Dialogues	284	.727 **	-.046 **
R Speech – T Dialogues	350	.777 **	-.027 **

Table 2: Pearson correlation coefficients of the RAPs of the rule sets retrieved from the MPT-RPT alignments.

The rules under investigation were further divided in a set of substitution rules (most of which were the result of an assimilation of place or voice, or a change in vowel length), a set of reduction rules (vowels reducing to schwa), a set of insertion rules and a set of deletion rules. The set of reduction rules was treated as a separate rule set with regard to the set of substitution rules, because vowel reductions to schwa are a typical phenomenon encountered in spontaneous speech (van Bergem, 1995), and because we wanted to check the different speech styles for their behaviour in this respect.

Table 3 presents the separate Pearson correlation coefficients of the RAPs of the substitution rules, the reduction rules, the insertion rules and the deletion rules.

Substitutions	N	r	ΔM
P Lectures – R Speech	76	.862 **	-.040 **
P Lectures – T Dialogues	73	.826 **	-.046 **
R Speech – T Dialogues	104	.928 **	-.003
Reductions			
P Lectures – R Speech	19	.956 **	-.006
P Lectures – T Dialogues	28	.825 **	-.078 **
R Speech – T Dialogues	31	.812 **	-.067 **
Insertions			
P Lectures – R Speech	45	.448 **	-.029
P Lectures – T Dialogues	35	.588 **	.034
R Speech – T Dialogues	59	.259 *	.042 **
Deletions			
P Lectures – R Speech	51	.822 **	.021*
P Lectures – T Dialogues	148	.613 **	-.058 **
R Speech – T Dialogues	156	.634 **	-.062 **

Table 3: Pearson correlation coefficients of the RAPs of the substitution, reduction, insertion and deletion rules retrieved from the MPT-RPT alignments.

3.2. Experiment 2

As in the first experiment, selections of the rule sets derived from the alignments of the APTs and the RPTs of the different speech styles were compared in pairs.

Table 4 presents the Pearson correlations of all RAPs under investigation. Again, the significance levels of the correlation between the RAPs of the different speech styles, and the significance levels of the differences between the mean RAPs of the different speech styles are indicated with asterisks.

APT-RPT	N	r	ΔM
P Lectures – R Speech	263	.572 **	.045 **
P Lectures – T Dialogues	297	.624 **	-.048 **
R Speech – T Dialogues	318	.650 **	-.081 **

Table 4: Pearson correlation coefficients of the RAPs of the rule sets retrieved from the APT-RPT alignments.

4. Discussion

The statistical comparison of the different RAPs provides interesting insights into the nature of the differences between the three speech styles. These insights will guide our future research.

The results in Table 2 show that there are no large differences between the RAPs of the three speech styles investigated. Especially the pronunciation characteristics of the PL and the RS seem to be very similar. This is even more striking because only the rules of which the RAPs differed in the different speech styles were taken into account. The high resemblance between the PL and the RS, and the larger difference with the TD confirm our hypothesis that the pronunciation characteristics of carefully articulated speech (RS and PL) are similar, and that these characteristics are quite different from the pronunciation characteristics of conversational speech (TD). Table 3 gives an explanation for the high resemblance of the two well-articulated speech styles. It appears that PL and RS are quite alike when it comes to the substitution, reduction or the deletions of phones. Whereas more phones tend to be substituted in RS than in PL (ΔM = -.040), phones are more frequently deleted in the PL (ΔM = .021). The overall picture, though, is that apart from the insertion rules, which are quite speech style specific, the pronunciation characteristics of RS and PL resemble each other to a very high degree.

It appears that the largest differences between the carefully pronounced speech styles (PL and RS) and the more sloppy speech style (TD) can be found in the RAPs of the vowel reductions and the deletion rules. Table 3 indicates that for these types of re-write rules, the RAPs of the TD rules are quite deviant from the RAPs of the PL and the RS rules. The differences between the mean RAPs of the reduction and deletion rules also indicate that in telephone dialogues, reductions of vowels to schwa and deletions of phones occur more frequently than in the other speech styles.

Contrary to our expectations, Table 2 also indicates that the RAPs of the TD rules tend to be more correlated to the RAPs of the RS rules than to those of the PL rules. Table 3 reveals that this is most probably due to the high correlation with the RAPs of the substitution rules in the RS.

Table 4 indicates that the alignments of the APTs and the RPTs did not show the same pronunciation differences, as did the alignments of the MPTs and the RPTs. According to our results, the RAPs of the TD rules and the RS rules were correlated most. However, a more detailed study such as the one we conducted on the results of experiment 1 did not give a sufficient explanation for the lower correlation values found in experiment 2. We may therefore conclude that our APT is not yet suited for this kind of linguistic research.

5. Conclusions and Future Research

In this paper, a first attempt was made to investigate the potential of using large spoken language corpora for linguistic research. We tried to capture the major pronunciation differences between three speech styles (representing speech with varying degrees of articulatory precision) in context-sensitive re-write rules at the phone level. In a first experiment, these re-write rules were retrieved from the alignment of a manually verified broad phonetic transcription and a canonical reference transcription. The comparison of the rule application probabilities of the rules of the different speech styles highlighted a high correlation between the phone substitutions, reductions and deletions of read speech and speech recorded at public lectures. The pronunciation differences between these speech styles and more spontaneous speech (telephone dialogues) can be attributed to the high number or vowel reductions and phone deletions in the telephone dialogues. These findings resemble results reported in the linguistic literature. In a second experiment, re-write rules were obtained from the alignment of an automatically generated phonetic transcription and a canonical reference transcription of the same data. The comparison of the rule application probabilities of these rules did not show clear tendencies. Therefore, we may conclude that our automatic phonetic transcription was not yet suited for the task of charting pronunciation characteristics of different speech styles.

The findings in this paper will guide our future research. The alignments of the manually verified broad phonetic transcriptions and the reference transcription already gave a first confirmation of our hypothesis that spoken language corpora may be beneficial for hypothesis generation in linguistic research. Therefore, we will continue our research with a more detailed survey of the actual rules underlying the correlation coefficients presented here. We will also try to improve our automatic phonetic transcription to the degree that similar tendencies can be found as with a manually verified phonetic transcription.

Acknowledgements

The work of Christophe Van Bael was funded by the “Stichting Spraaktechnologie” (Foundation for Speech Technology), Utrecht, the Netherlands. The authors would like to thank Lou Boves and Richard Piepenbrock for their contribution to this research.

References

Baayen, R.H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX Lexical Database (Release 2) [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor].

van Bergem, D. (1995). Acoustic and Lexical Vowel Reduction. Ph.D. thesis. University of Amsterdam, the Netherlands.

Booij, G. (1999). The Phonology of Dutch. Oxford University Press, New York.

Cucchiarini, C. (1996). Assessing transcription agreement: methodological aspects. In: Clinical Linguistics & Phonetics, vol. 10/2 (pp.131-155).

Godfrey, J. J., Holliman, E. C., McDaniel, J. (1992). SWITCHBOARD: Telephone speech corpus for research and development. In: Proceedings of IEEE ICASSP (pp. 517-520).

Lamel, L., Kassel, R., Seneff, S. (1986). Speech database development: Design and analysis of the acoustic-phonetic corpus. In: Proceedings of DARPA Speech Recognition Workshop (pp. 100-109).

Oostdijk, N. (2000). The Spoken Dutch Corpus: Overview and first evaluation. In: Proceedings of LREC 2000 (pp. 887-893).

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Olasson, D., Valtchev, V., Woodland, P. (2001). The HTK book (for HTK version 3.1.), Cambridge University Engineering Department.