A BOTTOM-UP METHOD FOR OBTAINING INORMATION ABOUT PRONUNCIATION VARIATION

Judith M. Kessens, Helmer Strik & Catia Cucchiarini
A²RT, Dept. of Language and Speech, University of Nijmegen, the Netherlands
{ J.Kessens, W.Strik, C.Cucchiarini}@let.kun.nl; http://lands.let.kun.nl/

ABSTRACT

In this paper a Bottom-Up (BU) method of obtaining information about pronunciation variation is proposed. BU transcriptions (T_{bu}) were obtained by letting a CSR decide for each phone whether it was deleted or not. The T_{bu} were compared to transcriptions obtained automatically with a Top-Down method, and the agreement appeared to be very high. Subsequently, the T_{bu} were aligned with canonical reference transcriptions (T_{ref}) and on the basis of this alignment, deletion rules were derived. The BU rules were employed to generate variants which were used in recognition experiments. The results of these recognition experiments show that the information about pronunciation variation obtained using the BU method can be used to improve recognition performance.

1. INTRODUCTION

In previous work [1], we showed that modeling pronunciation variation on the basis of phonological knowledge significantly improves the recognition performance of our Dutch CSR. However, the information on pronunciation variation that can be found in the literature is limited and probably processes exist that have not been described yet, especially if we consider that we are dealing with extemporaneous speech, which is a still very under-researched type of speech at the moment [2]. Since we would like to test whether recognition performance could be improved even further if we were able to model more of the variation in pronunciation that is present in our speech material, we have been looking for alternative ways of obtaining information on pronunciation variation.

The goal of the research presented in this paper is to investigate whether it is possible to obtain this information „Bottom-Up“ (BU), or in other words: directly from the speech signal. In this study we do essentially three things. First, we illustrate how information on pronunciation variation can be derived from the speech signal. Second, we check whether the information thus obtained is reliable. Finally, since our ultimate goal in modeling pronunciation variation is to improve the recognition performance of our CSR, we check whether this is indeed the case.

These three steps will be described in detail in the rest of this paper. In particular, in section 2, we give more details on the CSR and the speech material that was used. In section 3 we explain the BU transcription method and present the results of the comparison between the BU method and another automatic transcription method. In section 4, we explain how the BU information was formalized in rules and present the selected rules together with the rule statistics. Finally, in section 5, we describe the recognition experiments that were carried out to test whether the information on pronunciation variation obtained with the BU method does indeed improve recognition performance. At the end of the paper we discuss the results and draw some conclusions.

2. CSR AND SPEECH MATERIAL

The CSR uses phone models (continuous density hidden Markov models (HMMs)), language models (unigram and bigram), and a lexicon. The HMMs consist of three segments of two identical states, of which one can be skipped. In total 38 HMMs were trained. For more details on the CSR, see [6].

The speech material used in these experiments was selected from the VIOS database, which contains a large number of telephone calls recorded with the on-line version of a spoken dialogue system called OVIS [6]. OVIS is employed to automate part of an existing Dutch public transport information service. The speech material consists of interactions between man and machine. From the VIOS corpus, 11,247 utterances (83,447 words) were selected for making the BU transcriptions.

The baseline lexicon contains one transcription for each word. These transcriptions are automatically obtained using a Text-to-Speech system (TTS) for Dutch [5]. The baseline phone models (PMs) are trained using the training corpus (25,104 VIOS utterances) and the corresponding transcriptions in the baseline lexicon. Forced recognition is performed using the baseline PMs. Recognition performance was measured on an independent test set consisting of 6,267 VIOS utterances (21,106 words).

3. BU TRANSCRIPTIONS

In this study we started off by only looking at the deletions of phones because we expect this type of variation to be frequent in our speech material. Furthermore, we expect deletions (and insertions) to be more important than substitutions, since substitutions can be implicitly modeled within the phone models.

3.1. Obtaining the BU transcriptions

We obtained T_{bu} by performing the following steps:

1. A reference transcription (T_{ref}) is used as a starting point for the generation of the transcription variants. T_{ref} is obtained by looking up the word in the baseline lexicon and by canceling all deletion rules that were already applied. For example, the word final /n/ after a schwa is deleted in the baseline lexicon, and this rule is canceled by the restitution of /n/ in T_{ref}.
2. The transcription variants were automatically generated by making each phone in T_{ref} optional, leaving at least one phone per syllable. For example: Suppose T_{ref} is “/wIL/” (to want), then the following variants are possible: /wIL/, /wI/, /wL/, /IL/, /w/, /I/ and /L/.

3. Forced recognition was performed to determine which of all possible variants best matched the acoustic signal. In this way, we obtained the T_{bu} of the speech material.

3.2. Evaluation of the BU transcriptions

To establish whether the transcriptions obtained with the BU method were at all reliable, we decided to compare them to those obtained with another automatic transcription method, which we will call the Top-Down (TD) transcription method, because it uses knowledge obtained from the literature as the starting point [2]. For the TD method, variants were generated for four frequently occurring Dutch deletion processes: /n/-deletion, /r/-deletion, /t/-deletion, /@/-deletion. Next, variants were selected by means of forced recognition. The TD method had been previously evaluated by comparing the TD transcriptions (T_{td}) with human transcriptions. In these experiments the same material was transcribed by nine experienced listeners and by a CSR according to the TD method. A reference transcription was composed on the basis of the nine human transcriptions. It appeared that the degree of agreement between T_{td} and the reference transcription was comparable to the degree of agreement between the listeners and the reference transcription [3].

The essential difference between the BU transcription method and the TD method is that for the TD method the CSR could choose among a small number of transcription variants per word (i.e. 2-16 variants), whereas the number of variants among which the CSR could choose was much larger for our BU method. The number of BU variants per word is given by the following formula:

$$\prod_{i=0}^{M}(2^{n_i}-1) \qquad \begin{array}{l} M = \text{number of syllables} \\ n_i = \text{number of phones for syllable i} \end{array}$$

The consequence is that the possibility for the deletion of phones is more limited for the TD method compared to the BU method. Therefore, the question we would like to answer is whether this limitation influences the choices the CSR makes in deciding whether a phone is present or absent.

First, we used the TD method to make a transcription for the same material as for T_{bu} were made (11,247 utterances). Next, for the phones to which one of the four deletions rules could apply, we analyzed whether the same phone was deleted/present in T_{bu}. Only the phones were taken into account for which the preceding and following phone was not deleted in T_{bu}, since we used this as a criterion for the selection of the rules to be used in the recognition experiments. As a measure of agreement we used kappa (κ), which corrects for chance agreement (P_c):

$\kappa = (P_o - P_c) / (1 - P_c)$; $-1 < \kappa < 1$
P_o = observed proportion of agreement
P_c = proportion of agreement on the basis of chance

In Table 1, the results are given for the comparison between T_{bu} and T_{td}. In this Table, "yes" means that the phone is deleted, whereas "no" means that the phone is not deleted. Finally, in the last two columns, percentage agreement (%) and kappa (κ) are shown.

From Table 1 it can be concluded that agreement is very high. This means that for the specific phones under investigation (/n/, /r/, /t/ and /@/), the BU method makes essentially the same choices as to the presence or absence of these four phones as the TD method, regardless of whether other phones can be deleted in the same word.

BU	no	yes	no	yes	agreement	
TD	no	yes	yes	no	%	κ
/n/-del	2996	2984	21	160	97.1	0.99
/r/-del	3267	1520	9	72	98.5	0.94
/t/-del	1688	458	6	51	97.4	0.97
/@/-del	64	73	0	1	99.3	0.92

Table 1: Agreement for each rule for a comparison between the BU method (BU) and the TD method (TD), expressed in percentage agreement (%) and kappa (κ).

Since earlier research showed that agreement between T_{td} and a human reference transcription is high, we have evidence that this might also hold for the proposed T_{bu}. However, one should bear in mind that these results only hold for the deletion rules for which the preceding and following phone are unaffected.

4. BU RULES

4.1. Obtaining the BU rules

The information contained in the BU transcriptions was formalized in the form of a set of rules by performing the following steps:

1. T_{bu} was time-aligned with T_{ref} using a dynamic programming algorithm, for which the distance between two phones was calculated on the basis of the features defining the two phones in question [4].
2. After time-alignment, for each target phone we formulated deletion rules as follows:

   $\{L\ X\ R\}_{ref} \Rightarrow \{L\ -\ R\}_{bu}$.

   This means that the target phone "X" in T_{ref} following the phone "L" (left context) and preceding the phone "R" (right context) is deleted in T_{bu} ("-" = deletion). "L" or "R" can be a phone or a word boundary.
3. We then calculated the absolute and relative rule application. The relative rule application is defined as the absolute number of times the rule applies divided by the number of times the rule could have applied.
4. In this paper, we only present the rules that were used in the recognition experiments described in section 5. These rules were selected on the basis of the following criteria:
   - The left and right adjacent phones had to be identical in T_{ref} and T_{bu}: $L_{ref}=L_{bu}$ and $R_{ref}=R_{bu}$. We applied this criterion because we expect that T_{bu} will contain more errors if besides the target phone, also the context in which it occurs is altered.
   - L_{ref} and R_{ref} should not be the beginning or end of an utterance. This criterion was applied because it can be expected that the beginning and end of an utterance contain more acoustic artifacts like noise or truncation of the speech signal.
   - For the recognition experiments we only used the rules for which the absolute rule application is

higher than 100, and for which the relative rule application is higher than 0.2. We adopted this criterion because we assume that rules that are (absolutely and relatively) frequently applied are most important for our goals.

4.2. Statistics of the BU rules

In Table 2, the BU rules used in the recognition experiments are shown. In the second column, the deleted phone is shown, whereas in the third column the specific deletion rule is given. The symbol „|“ denotes a word boundary. In the last two columns, the relative (%appl.) and absolute (#appl.) rule application is given.

	phone	BU deletion rule	%appl.	#appl.
1	R	{@ R m} → {@ - m}	63	125
2	d	{n d I} → {n - I}	60	107
3	R	{@ R d} → {@ - d}	51	1095
4	n	{@ n |} → {@ - |}	40	1653
5	R	{@ R t} → {@ - t}	38	346
6	d	{| d @} → {| - @}	32	176
7	t	{s t @} → {s - @}	30	499
8	@	{v @ R} → {v - R}	30	444
9	t	{n t s} → {n - s}	27	131
10	@	{d @ R} → {d - R}	27	126
11	d	{n d @} → {n - @}	27	118
12	h	{| h E} → {| - E}	26	127
13	R	{@ R |} → {@ - |}	24	105
14	t	{i t |} → {i - |}	23	283
15	n	{@ n t} → {@ - t}	23	175

Table 2: Set of selected BU rules with relative (%appl.) and absolute (#appl.) rule application.

5. RECOGNITION EXPERIMENTS

5.1. Testing Conditions

The baseline performance was measured by performing a recognition test in which no pronunciation variation was modeled.

Since we want to compare the results of the modeling of the BU pronunciation variants to the results obtained with the TD method, the same three testing conditions were analyzed as in [1]. In short, these testing conditions consist of incorporating pronunciation variants at the three levels the CSR consists of:

1. Pronunciation variants are added to the *lexicon*.
2. Additionally, pronunciation variation is included in the training of the *PMs* by retraining the PMs on the basis of a training corpus in which pronunciation variants are transcribed by performing forced recognition.
3. Additionally, pronunciation variants are used in the *language model (LM)*, meaning that in general different variants have different probabilities in the LM.

For the TD method, variants were generated using rules that were based on the following five frequently occurring Dutch phonological processes: /n/-deletion, /r/-deletion, /t/-deletion, /@/-deletion and /@/-insertion [1]. Since for our BU method only deletion rules can be obtained, we repeated the tests for the TD method without including the @-insertion rule. Next, we repeated the tests for the set of 15 selected BU rules. Finally, we did the same for a set of four of the 15 BU rules which were selected because they are very frequent, either in absolute terms (>1000) or in relative terms (>0.5) (rules 1 to 4 in Table 2). The motivation for this choice is to get a first indication of what the optimal number of modeled BU variants is. If this amount is too high when using 15 BU rules, we expect to find lower WERs for the four frequent BU rules.

5.2 Results of recognition experiments

When comparing the two methods of modeling pronunciation variation with each other, one should bear in mind that the rules of the TD method are formulated differently from the rules of the BU method. The only conditions for our BU rules are a specific phone (or word boundary) on the left and right. The conditions for a TD rule can be broader in two ways: (1) classes of phones (e.g. vowels, obstruents) are often used, and (2) the context on the left and right can be larger than just one phone.

In general, more variants were generated with a specific TD rule compared to a specific BU rule. To get an idea of the number of variants that play a role during recognition, the number of variants that were added to the test lexicon are given in Table 3 (column 3). Since the number of variants varies for the different lexica, the number of variants that were transcribed in the training corpus also varies. By transcription of a variant we mean that during forced recognition a different transcription is selected than the baseline transcription. Column 5 shows the number of variants that were transcribed in the training corpus. Between brackets, the percentage of words is given that is transcribed as a variant. Furthermore, column 4 shows the percentage of the BU variants that were also present in the TD lexica (%TD).

	# rules	lexicon	%TD	training corpus
Top Down	5	1125	-	8541 (10.5%)
	4	971	-	8123 (10.0%)
Bottom Up	15	627	48	7249 (8.9%)
	4	229	96	5051 (6.2%)

Table 3: Number of variants added to baseline test lexicon, percentage overlap between BU and TD variants, and number of variants transcribed in training corpus

From Table 3 it can be concluded that despite the fact that the BU rules and TD rules are obtained differently and that the way the rules are defined are very differently, there is a lot of overlap between the variants that are generated with the BU rules and the TD rules. This overlap appears to be extremely large for the four most frequent BU rules (96%).

The Word Error Rate (WER=(S+D+I)/N)) for our baseline system was 12.75%. In Table 4, the WERs are given for the three different testing conditions. In the first row, the level at which pronunciation variants are incorporated is indicated. In the second row, the results are given for the five TD rules as

reported in [1]. In the next row, the results are given for the four TD rules (/@/-insertion excluded). In the fourth row, the results are given for the BU method in which variants were generated by using the 15 deletion rules. Finally, in the last row, the results are given for the four selected BU rules.

	# rules	lexicon	PMs	LM
Top Down	5	12.44	12.22	12.07
	4	12.52	12.42	12.41
Bottom Up	15	12.48	12.49	12.18
	4	12.74	12.48	12.36

Table 4: WERs for the Top-Down and Bottom-Up method for the 3 different testing conditions and different number of rules.

From Table 4 it can be concluded that modeling of the BU variants leads to a reduction in WER compared to the baseline. With the set of four most frequent BU rules the same order of improvement is obtained as with the four TD rules, whereas the number of variants that are added to the lexicon is only one quarter of the number used in the TD method.

6. DISCUSSION

In the present study we have investigated the adequacy of a BU method for obtaining information about pronunciation variation directly from the speech signal. The results of this study provide various indications that the BU method proposed here is suitable for this purpose. First of all, the BU transcriptions contain information about numerous phonological processes. Some of these processes were already known from the literature and either had already been used in the TD method (the rules 1, 3, 4, 5, 9 and 15 in Table 2 overlap to a great extent with the TD deletion rules), or had not been modeled yet (like rules 12 and 14 in Table 2). Some others had not been described yet, but appear to be plausible connected speech processes (e.g. rules 2, 7 and 11 in Table 2). Second, for the processes investigated the information derived from the BU transcriptions appears to be reliable since the BU transcriptions are extremely similar to the transcriptions obtained by means of a TD method whose accuracy was checked against transcriptions made by human listeners. Third, when the most frequent rules derived by means of the BU method were used to model pronunciation variation, recognition performance turned out to improve. In addition, the improvement obtained with the four most frequent BU rules was larger than the one obtained by using the four frequently applied TD rules in [1], whereas only one quarter of the number of variants was used.

It can be expected that if more variation is modeled, more errors can be solved. On the other hand, this would increase the complexity of the recognition task, thus leading to more errors. Since the BU rules are selected on the basis of frequency, we expect that the balance between errors that are solved and errors that are introduced will be positive when few frequent rules are used, whereas when more rules are used, the balance will shift to the other side. This might indicate that the amount of pronunciation variation that is modeled is not yet optimal. Recognition experiments in which the number of variants that are included in the lexicon is systematically varied in relative and/or absolute frequency will have to reveal whether this hypothesis is true. Since recognition performance is improved when 15 BU rules are used compared to using the four most frequent BU rules, we have an indication that the optimal amount of BU pronunciation variation to model has not been reached yet.

Our ultimate goal is to find the optimal rule set, in the sense that recognition performance is optimally improved. Clearly, we expect this optimal rule set to be a combination of TD and BU rules. However, earlier research [1] showed that finding the optimal rule set is not straightforward, since no adequate measure exists to decide whether a rule (or variant) should be included or not. In another paper presented at this conference [7], we conclude that error rates alone are not a good measure for expressing the effect of modeling pronunciation variation and that the results are corpus dependent. Therefore, in future we intend to develop a more appropriate test corpus and to perform more detailed error analyses on this test corpus in order to gain more insight into the question concerning the type of variation that should be modeled.

7. ACKNOWLEDGMENTS

The research by Judith M. Kessens was carried out within the framework of the Priority Programme Language and Speech Technology, sponsored by NWO (Dutch Organization for Scientific Research). The research by Helmer Strik has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

8. REFERENCES

1. Kessens, J.M., Wester, M., and Strik, H. ”Improving the Performance of a Dutch CSR by Modeling Within-word and Cross-word Pronunciation”, *Speech Comm., Vol.* 29, *1999, p 193-207.*

2. Strik, H. & Cucchiarini, C. “Modeling pronunciation variation for ASR: a survey of the literature”, *Speech Comm., Vol. 29, 1999, p 225-246.*

3. Kessens, J.M., Wester, M., Cucchiarini, C. and Strik, H. “The selection of pronunciation variants: comparing the performance of man and machine”, *Proc. ICSLP-98:* 2715-2718, 1998.

4. Cucchiarini, C. “Assessing Transcription Agreement: Methodological Aspects”, *Clinical Linguistics & Phonetics, Vol. 10, no.2, 1996, p 131-155.*

5. Kerkhoff, J., Rietveld, T. “Prosody in Niros with Fonpars and Alfeios”, *Proc. of the Department of Language and Speech, University of Nijmegen, Vol. 18:* 107-119, 1994.

6. Strik, H., Russel, A.J.M., van den Heuvel, H. Cucchiarini, C. & Boves, L. “A spoken dialog system for the Dutch public transport information service”, *Int. Journal of Speech Technology, Vol. 2, No. 2, 1997, p 119-129.*

7. Wester, M., Kessens, J.M., Strik, H. “Pronunciation Variation in ASR: Which Variation to model?”, *Proc. ICSLP-00, Beijing, China*, 2000.