TWO AUTOMATIC APPROACHES FOR ANALYZING CONNECTED SPEECH PROCESSES IN DUTCH

Mirjam Wester, Judith M. Kessens & Helmer Strik

A²RT, Dept. Of Language and Speech, University of Nijmegen, The Netherlands
{wester,kessens,strik}@let.kun.nl

ABSTRACT

This paper describes two automatic approaches used to study connected speech processes (CSPs) in Dutch. The first approach was from a linguistic point of view - the top-down method. This method can be used for verification of hypotheses about CSPs. The second approach - the bottom-up method - uses a constrained phone recognizer to generate phone transcriptions. An alignment was carried out between the two transcriptions and a reference transcription. A comparison between the two methods showed that 68% agreement was achieved on the CSPs. Although phone accuracy is only 63%, the bottom-up approach is useful for studying CSPs. From the data generated using the bottom-up method, indications of which CSPs are present in the material can be found. These indications can be used to generate hypotheses which can then be tested using the top-down method.

1. INTRODUCTION

Connected speech processes (CSPs) are responsible for a large amount of variation in spontaneous speech. A precise understanding of these phenomena is therefore useful when modeling pronunciation variation for the purpose of improving automatic speech recognition (ASR). In [1] we described an attempt to modeling CSPs so as to improve the performance of our Continuous Speech Recognizer (CSR), by using a top-down approach. It appears that modeling CSPs does indeed improve recognition performance. Moreover, experiments that were carried out to determine whether our top-down approach works satisfactorily revealed that the performance of the CSR is comparable to that of expert listeners [3].

However, in many cases using a top-down approach alone will not be sufficient for modeling CSPs for ASR. First, because a top-down approach only works for hypothesis testing and cannot be used for explorative purposes. Second, because a top-down approach requires information on the application of CSPs, but this information is limited, for various reasons. The fact that CSPs are influenced by many complex factors such as speech style, speech rate, word frequency, information load, dialectal variation and, last but not least, individual variation makes them less amenable to observation according to traditional methods. Furthermore, for pronunciation modeling in ASR, statistical data are needed on the relative frequency of occurrence of the various processes and this type of information is not available in the literature. Another problem with modeling CSPs for ASR is that in ASR we often have to do with man-machine interactions and it is it not known whether the CSPs that may apply in conversations between humans will also apply when the interlocutor is a machine. This all suggests that it might be worthwhile to test whether a different approach could provide additional information on CSPs and thus contribute to better modeling of pronunciation variation in ASR.

With this in mind, we set out to determine whether a bottom-up approach [2] could be used for this purpose. In this approach, a phone recognition is performed to obtain information on CSPs. The problem with a bottom-up method might be that phone recognition is not good enough to produce reliable information on CSPs. However, an approach of this kind could be sufficient to give indications of the type of processes that might occur in the speech material and that can further be tested by means of a top-down approach. In this paper, we report on an experiment that was aimed at determining whether top-down and bottom-up approaches can be used to obtain information on CSPs.

2. METHOD

In the present research, two procedures for obtaining information about CSPs are studied and compared. To this end, a corpus with connected speech is needed. The speech material selected for this purpose is described in section 2.1. Next, the general characteristics of the CSR are provided in section 2.2. Different versions of this CSR were employed in the two approaches. In the first procedure, forced recognition was applied to select between pronunciation variants (section 2.3). In this way, a top-down transcription (T_{td}) was obtained. In the second approach, a phone recognizer produced a string of phones, the bottom-up transcription (T_{bu}) (section 2.4). The transcriptions were aligned by means of a dynamic programming (DP) algorithm (section 2.5). Top-down and bottom-up transcriptions were compared with each other, but also with a third transcription: the reference

transcription (T_{ref}). The latter was obtained by transcribing each word in the utterances using its corresponding canonical transcription in the lexicon.

2.1. Speech Material

The speech material used in this experiment was selected from a data base named VIOS, which contains a large number of telephone calls recorded with the on-line version of OVIS [4]. OVIS is a spoken dialogue system which is employed to automate part of an existing Dutch public transport information service. Currently OVIS can be used to obtain information about the Dutch train times. The database VIOS thus contains speech from man-machine interactions. For training, 25,104 VIOS utterances (83,890 words) were used. From the VIOS database 50,000 short utterances, containing 82,101 words, were selected for the present research. The selection criteria will be explained in section 2.5.

2.2. CSR

For our research, we used the CSR which is part of OVIS [4]. The most important characteristics of the CSR are as follows. Feature extraction is done every 10 ms for frames with a width of 16 ms. The first step in feature analysis is an FFT analysis to calculate the spectrum. Next, the energy in 14 Mel-scaled filter bands between 350 and 3400 Hz is calculated. Finally, a discrete cosine transformation on the log coefficients and cepstral mean subtraction are applied. Besides 14 cepstral coefficients (c_0-c_{13}), 14 delta coefficients are also used. This makes a total of 28 feature coefficients. The CSR uses acoustic models (HMMs), language models (unigram and bigram), and a lexicon. The continuous density HMMs consist of three segments of two identical states, one of which can be skipped. In total 38 HMMs were trained: for non-speech sounds 1 model, for each of the phonemes /l/ and /r/ 2 models, and for each of the other 33 phonemes 1 model. For /l/ and /r/ a difference was made between prevocalic (/l/ and /r/) and postvocalic position (/L/ and /R/). Beside these special symbols for the allophones of /l/ and /r/, we will use standard SAMPA notation for all other phonemes in this article.

2.3. Top-down Approach

The top-down approach can be used to do hypothesis verification. In our case, we used the CSR to determine whether rules were applied or not. For this purpose the following five optional phonological rules of Dutch were selected:

1. /n/-deletion: syllable final: @ + n → @
   Ex: /rEiz@n/ → /rEiz@/
2. /r/-deletion: @ + r + cons → @ + cons
   unstressed short vowel + r + cons → unstressed short vowel + cons
   long vowel + r + cons → long vowel + cons
   Ex: /Amst@RdAm/ →/Amst@dAm/
3. /t/-deletion: obs + t + cons → obs + cons
   son + t + obs → son + obs
   word final: obs + t → obs
   Ex: /ytr@xt/→ /ytr@x/
4. /@/-deletion: obs + @ + liq + @ → obs + liq + @
   Ex: /And@r@/ → /Andr@/
5. /@/-insertion: in nonhomorganic clusters in coda position
   Ex: /dELft/ → /dEl@ft/

The five phonological rules describe insertion or deletion processes within a word. They were selected mainly because they are frequently applied in Dutch and well described in the literature [5, 6]. The rules were automatically applied to all the words in the lexicon, whenever the condition for its application was met.

In the top-down approach, forced recognition mode is used. In this mode, the recognizer does not choose between all the words in the lexicon but only between different pronunciation variants of the word. Therefore, if pronunciation variants are present for a word, the CSR will select the one that best matches the acoustic signal. In this way a top-down transcription (T_{td}) is automatically acquired. It is important to note here that we checked carefully that none of the rules were applied in the canonical transcriptions of the words (and thus in T_{ref}).

2.4. Bottom-up Approach

In the bottom-up approach a constrained phone recognizer is employed to obtain a phone transcription for each utterance: the bottom-up transcription (T_{bu}). In the bottom-up approach the same 38

monophone models are used as in the top-down approach (see section 2.2.). The recognition process is constrained by using phone language models (unigram and bigram), which were trained on the reference transcriptions of the 25,104 utterances in the training corpus. These constraints are thus general phonotactic constraints.

2.5. DP-Alignment

In order to time-align the various transcriptions a DP algorithm was used. We first implemented a simple DP algorithm in which the penalty for an insertion, deletion or substitution is 1. However, when using DP algorithm 1 we often found sub-optimal alignments, like the following example:

T_{ref} = / A m s t @ R d A m /
T_{bu} = / A m s # @ t a: n # / (# = insertion)

For this reason, we decided to make use of a more sophisticated DP alignment procedure [7]. In this second DP algorithm, the distance between two phones is not just 0 (when they are identical) or 1 (when they are not identical) but more gradual. The distance between two phones is calculated on the basis of the features of the phones which are compared. More details about this DP algorithm can be found in [7]. Using this second DP algorithm the following alignment was found for the example mentioned above:

T_{ref} = / A m s t @ R d A m /
T_{bu} = / A m s # @ # t a: n /

It is obvious that the latter alignment obtained with DP algorithm 2 is better than the alignment calculated with DP algorithm 1. Since in general the alignments obtained with DP algorithm 2 were more plausible than those obtained with DP algorithm 1, DP algorithm 2 was used to determine the alignments. In calculating alignments with DP algorithm 2, we found that the CPU time of the program increases enormously as the compared transcriptions become longer and when the differences between the compared transcriptions becomes larger. This turned out to be a problem for bottom-up transcriptions because they often deviate a great deal from the reference transcriptions. In order to keep the CPU time within reasonable bounds we used the following two criteria to select our speech material:

1. The number of symbols in T_{ref} is smaller than 50: $|T_{ref}| \leq 50$
2. The difference between the compared transcriptions is restricted: $|T_{ref}| - |T_{bu}| / \sqrt{|T_{ref}|} \leq 2$

In this way, 50,000 short utterances were selected, for which DP alignments could be obtained within a reasonable time limit.

3. RESULTS

In this section, we will show that the DP alignments of the 50,000 utterances can be used to extract information about CSPs. First, we will present results obtained with the top-down approach (section 3.1) and the bottom-up approach (section 3.2). Next, the two approaches will be compared in section 3.3.

3.1. Top-down approach

In section 2.3, the five phonological rules and their conditions were specified. First, we examined how often each of these conditions were met in our speech material, i.e. how often a rule could have been applied. The total number of possible applications of each rule are given in row 2 of Table 1 (# possible). Next, we used the top-down approach to determine the number of times a rule was applied in the 50,000 utterances. This was simply done by counting frequency of occurrence in the DP alignments of T_{td} and T_{ref}. The number of times the five rules were applied are shown in the third row of Table 1. The absolute numbers (# applied) and the relative frequency (# applied / # possible) are given.

	n-del	r-del	t-del	@-del	@-ins
# possible	4,832	4,503	4,055	142	2,967
# applied	1,827 (38%)	1,344 (29%)	771 (19%)	61 (43%)	528 (18%)

Table 1: Application of five phonological rules

3.2 Bottom-up Approach

In order to obtain the results of the bottom-up procedure, the DP alignments of T_{bu} and T_{ref} were analyzed. In row 2 of Table 2, the total number of phones are given, followed by the number of identical phones, substitutions, deletions and insertions. On the basis of these numbers phone accuracy was calculated. Phone accuracy is the number of insertions subtracted from the number of identical phones and divided by the total number of phones. Thus, the phone accuracy is 63% for all phones, 58% for the consonants, and 71% for the vowels. It can be seen in Table 2 that the number of deletions is much higher than the number of insertions, both for the absolute numbers and the percentages. Since consonants occur much more often than vowels, it is better to use the frequency data to compare the two. They reveal that vowels remain identical more often, mainly because in comparison to the consonants, they are deleted less often. The frequencies for the substitutions and the insertions do not differ much.

	all phones	%	consonants	%	vowels	%
TOTAL	264,556	100	159,789	100	104,767	100
identicals	177,804	67	100,673	63	77,131	74
substitutions	40,216	15	24,417	15	15,799	15
deletions	46,536	17	34,699	22	11,837	11
insertions	11,072	4	7,883	5	3,189	3

Table 2: Number of identicals, substitutions, deletions and insertions in T_{bu}

In the introduction we already noted that the top-down approach can be used to do hypothesis verification. However, it cannot be used to obtain new hypotheses. The bottom-up approach, on the other hand, could be useful for this purpose. Therefore, we decided to study whether the bottom-up approach can give indications of CSPs. This was done in the following way. First, we counted the number of context specific changes (substitutions, deletions and insertions) in the DP alignments of T_{bu} and T_{ref}. During counting we also used information about utterance and word boundaries. The latter are denoted by the symbol "|". We only used the left and right symbol in the reference transcription as context information (the term symbols is used here instead of phonemes because the context can also be an utterance or word boundary symbol). Then, we selected the CSPs that occurred more than 500 times. Some examples of frequently observed CSPs are given in Table 3.

CSPs	left context	deleted phone	right context	count
1	@	n	|	1144
2	A	t	|	1063
3	x	@	n	909
4	x	t	|	642
5	i:	t	|	558
6	t	@	R	519

Table 3: Examples of frequent CSPs, with left and right contexts and number of occurrences.

It can be seen that all examples in Table 3 concern deletions, which is not surprising as the majority of the changes observed are deletions. The reason these examples have been chosen is because they all are examples of plausible CSPs. The results were compared with phonological rules described in the literature, e.g. the five rules specified in section 2.3. Since the context in Table 3 is limited to the immediate left and right symbol, it was not always possible to inspect whether these CSPs completely meet the conditions of a rule. However, we could analyze whether the CSPs did not violate the conditions. Of course, in our transcriptions the complete context is available, so we could perform analyses for different kinds of contexts. However, it is clear that as the context is increased it becomes more specific, and therefore the counts of the observed CSPs will rapidly decline.

The CSPs 1, 4 and 6 do not violate the conditions of the /n/-deletion rule, /t/-deletion rule and /@/-deletion rule, respectively. Therefore, they could be occurrences of these rules. CSP 2 shows that a word-final /t/ following an /A/ is frequently deleted. Closer inspection of our data showed that the majority of these cases concerns the word "dat". Booij [5] describes that certain sequences of function words can be contracted: for example "dat" + personal pronouns. Furthermore, in some cases "is" following a function word. can reduce to /s/. Thus, "dat is" /dAt|Is/, contracts to /dAs/. CSP 2 is an example of such a contraction process. CSP 3 is an example of the deletion of the /@/ in the context /x/ _ /n/ which occurs in some regional Dutch dialects. The deletion of word-final /t/ after /i:/ also occurs

frequently (CSP 5). In our data, this occurs most often for the word “niet”. Booij [5] mentions that in standard Dutch /t/ deletion is possible in word-final position after a vowel in function words like “niet” /ni:t/, which can be pronounced as /ni:/ in informal language use, which is probably the case for CSP 5. To summarize, these examples show that bottom-up analysis can be used to reveal frequent CSPs. These examples are all related to well-known CSPs. However, it is likely that with this procedure it is also possible to find indications of CSPs which are less well or not known.

3.3. Comparing Top-down and Bottom-up Approaches

Both approaches were compared to calculate the agreement between the approaches. For comparison only a subset of the data could be used, i.e. the data relating to the five rules used in the top-down approach. For all of the cases in which a rule could have been applied, we compared T_{td} with T_{bu}. In order to illustrate the analysis procedure the numbers for the /n/-deletion rule are given in Table 4.

		bottom-up		
	/n/-deletion	not applied	applied	other
top- down	not applied	2,001	481	510
	applied	314	1,019	507

Table 4: Comparison between top-down and bottom-up approach for the /n/-deletion rule.

The top-down approach can only choose between variants of a word, i.e. it can only determine whether a rule is applied or not. Only these two outcomes are possible. However, in the case of the bottom-up procedure there are three possibilities: 1. the /n/ is still present in T_{bu} and thus the rule is not applied, 2. the /n/ is deleted and thus the rule is applied, and 3. another phoneme is present in T_{bu} instead of the /n/ (i.e. a substitution). In the last case, it is impossible to determine whether the /n/-deletion rule was applied or not. One could say that the /n/ is deleted because it is not present in T_{bu}. But on the other hand one could also argue that the other phoneme in the position of the /n/ in T_{bu} indicates that there is something there. For instance, this other phoneme could be one that is reasonably similar to /n/, like another nasal, or it could be a completely different phoneme. In any case, we decided not to use these ‘other’ data for our analysis of agreement. Therefore, agreement is calculated on the basis of the numbers in the 2 by 2 matrix on the left. The numbers on the diagonal are the cases in which both approaches agree (2,001 + 1,019 = 3,020), and the off-diagonal numbers show they disagree in 314 + 481 = 795 cases. The total number of cases (3,815 = 3,020 + 795) can then be used to calculate that the percentage agreement is 79% (3,020 / 3,815). In the same, way percentage agreement was calculated for the other rules. In the second row of Table 5, the number of cases of agreement and the percentage agreement are given for the individual phonological rules, and for all rules together. In the last row of table 5, the total number of cases is given for each rule.

	n-del	r-del	t-del	@-del	@-ins	all rules
agreement	3,020 (79%)	2,263 (65%)	1,618 (49%)	46 (46%)	1,615 (85%)	8,562 (68%)
total	3,815	3,502	3,298	99	1,905	12,619

Table 5: Agreement between T_{bu} and T_{td} for the 5 phonological rules and all rules together

4. DISCUSSION AND CONCLUSIONS

In ASR research the focus has gradually shifted from isolated words to connected speech [2]. It is clear, that in going from isolated words to connected speech, the amount of pronunciation variation increases. A precise understanding of the processes which occur in connected speech is crucial when modeling pronunciation variation for ASR. Moreover, statistical data are needed on the relative frequency of occurrence of the various processes. In this paper, we have presented two approaches which can be used to obtain this information: a top-down and bottom-up approach. Both approaches have been applied in studies about pronunciation modeling for ASR [2].

In [3] we conducted an experiment in which we tested the reliability of the top-down approach by comparing its performance to that of nine listeners. We found that the agreement between the listeners and the CSR was 78%, which was only slightly lower than the average agreement of 82% between listeners. Although the CSR performs somewhat worse than the listeners, its behavior was fairly similar to that of the humans [3]. Therefore, it can be concluded that the top-down method is reliable. In

section 3.1 we showed that by using this method information on the frequency of CSPs can be obtained.

Information about CSPs can also be obtained with the bottom-up method. Some examples of plausible frequent CSPs that were detected by this method were presented in section 3.2. Given that the phone accuracy is 63% and that the average agreement between bottom-up and top-down methods is 68%, it can be concluded that the reliability of the bottom-up method is probably lower than that of the top-down method. Therefore, information obtained with the bottom-up procedure should be handled carefully.

The two approaches not only give qualitative information about CSPs (which process is applied under which conditions), but also quantitative information (how often does a CSP occur). For ASR research the quantitative information is important because frequent CSPs are expected to have a larger influence on the performance of the recognizer than less frequent ones. Furthermore, the quantitative information can be used to calculate probabilities of the CSPs, which is needed for most speech recognizer.

Although the top-down approach is probably more reliable than the bottom-up method, it has the important drawback that it can only be used to verify hypotheses. Many hypotheses on CSPs can be found in the literature. However, probably the information in the literature is not complete. Therefore, it is necessary to obtain new hypotheses about CSPs. These new hypotheses could be obtained by using the bottom-up procedure. In future work, we plan to generate hypotheses by analyzing the bottom-up data carefully, subsequently we will test them in a top-down manner, and additionally, we will use this information to improve the performance of the ASR.

It should be kept in mind that the results of both approaches depend for a great deal on the properties of the CSRs used for recognition. However, the aim of the present research was not to investigate how the performance of both approaches can be improved, e.g. by using more appropriate phone models and tuning recognition parameters. Therefore, we simply used CSRs that were available at the start of this research. In the future we will be looking at ways of optimizing the approaches in such a way that the reliability of the obtained information is increased.

5. ACKNOWLEDGMENTS

The research by J.M. Kessens was carried out within the framework of the Priority Programme Language and Speech Technology, sponsored by NWO (Dutch Organisation for Scientific Research). The research by Dr. H. Strik has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences. We would like to thank Dirk Janssen, for his useful adaptations of the DP-script “align”, and for his help, when using the program.

6. REFERENCES

1. Wester, M., Kessens, J.M., and Strik, H. “Improving the Performance of a Dutch CSR by Modeling Pronunciation Variation,” *Proc. of the Workshop Modeling Pronunciation Variation for Automatic Speech Recognition*, *Kerkrade, 145-150,* 1998.

2. Strik, H. and Cucchiarini, C. “Modeling Pronunciation Variation for ASR: Overview and Comparison of Methods,” *Proc. of the Workshop Modeling Pronunciation Variation for Automatic Speech Recognition*, *Kerkrade, 137-144,* 1998.

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, Sydney, Australia*, 1998.

4. Strik H., Russel A., Van den Heuvel, H., Cucchiarini, C., Boves, L. ”A Spoken Dialogue System for the Dutch Public Transport Information Service”, *Int. Journal of Speech Technology, Vol. 2, No. 2, pp. 119-129*, 1997.

5. Booij G., *The Phonology of Dutch*: Clarendon press, Oxford, 1995.

6. Cucchiarini, C., van den Heuvel, H. “/r/ Deletion in Standard Dutch”, *Proc. of the Dept. of Language & Speech, University of Nijmegen, Vol. 19, pp. 59-65,* 1995.

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