AUTOMATIC DETECTION AND VERIFICATION OF DUTCH PHONOLOGICAL RULES

Judith Kessens, Mirjam Wester & Helmer Strik

A2RT, Department Language & Speech, University of Nijmegen, The Netherlands
{J.Kessens, M.Wester, W.Strik)@let.kun.nl, http://lands.let.kun.nl/staff/kessens

Abstract

In this paper, we propose two methods for automatically obtaining hypotheses about pronunciation variation. To this end, we used two different approaches in which we employed a continuous speech recognizer to derive this information from the speech signal. For the first method, the output of phone recognition was compared to a reference transcription in order obtain hypotheses about pronunciation variation. Since phone recognition contains errors, we used forced recognition in order to exclude unreliable hypotheses. For the second method, forced recognition was also used, but the hypotheses about the deletion of phones were not constrained beforehand. This was achieved by allowing each phone to be deleted. After forced recognition, we selected the most frequently applied rules as the set of deletion rules. Since previous research showed that forced recognition is a reliable tool for testing hypotheses about pronunciation variation, we can expect that this will also hold for the hypotheses about pronunciation variation which we found using each of the two methods. Another reason for expecting the rule hypotheses to be reliable is that we found that 37-53% of the rules are related to Dutch phonological processes that have been described in the literature.

1. Introduction

The continuous speech recognizer (CSR) that we used in our research is part of a spoken dialogue system called OVIS, which gives information about public transport (Strik et al., 1997). A large number of telephone calls of the on-line version of OVIS have been recorded and are stored in a database called VIOS. These man-machine

Phonus 5, Institute of Phonetics, University of the Saarland, 2000, 117-128.

interactions clearly show that the manner in which people address OVIS is extremely varied, ranging from using hyper-articulated speech to very sloppy speech. This enormous variation in pronunciation constitutes a serious challenge to our CSR, because pronunciation variation lowers recognition performance if it is not properly accounted for (see e.g. Kessens et al., 1999; Strik & Cucchiarini, 1999).

By using available knowledge about frequent phonological processes we managed to model part of the within-word and cross-word variation, which in turn led to significant improvements in recognition performance (Kessens et al., 1999). However, the information about pronunciation variation that can be found in the literature is not exhaustive. There are probably processes which have not yet been described, especially if we consider that the type of speech that we study is spontaneous speech, and that this is still a very under-researched area at the moment (Strik & Cucchiarini, 1999).

Since we are convinced that recognition performance can improve even further if we are able to model more of the variation in pronunciation that is present in our material, we have been looking for alternative ways of obtaining information about pronunciation variation. An alternative could be to derive this information directly from the speech signal. The goal of this research is to investigate whether it is possible to use a CSR for this purpose, or in other words, whether it is possible to derive this information automatically.

To this end, a CSR is used to generate an automatic transcription of the speech material. By analyzing the difference between this automatic transcription (T_{aut}) and a reference transcription (T_{ref}) we can obtain hypotheses about pronunciation variation. A problem with automatic transcriptions, however, is that generating them is not straightforward. It is possible to perform phone recognition by only using the acoustic models, i.e. without the top-down constraints of language model and lexicon, but the problem is that only 63% of the resulting phones are correct (Wester et al., 1998), which is not enough for our purposes.

Therefore, we evaluated a different technique called forced recognition, in which the CSR is constrained in the sense that it is only allowed to choose between different variants of the same word (Kessens et al., 1998). In this way, the variant that most closely resembles the uttered word can be chosen. By choosing from among alternative variants that differ from each other in the representation of one specific segment, the CSR can be forced, as it were, to choose between different transcriptions of that specific segment. We compared the performance of nine experienced listeners with that of the CSR for this specific task and found that the results were very similar,

thus indicating that the CSR can be used to obtain phonetic transcriptions of the speech material by resorting to this kind of hypothesis verification (Kessens et al., 1998).

In this paper, we propose two different methods for obtaining information about pronunciation variation directly from the speech signal. The kind of pronunciation variation which is modeled is variation at the segmental level, because we expect that this kind of variation is most detrimental to speech recognition. In both methods, hypotheses about pronunciation variation were formulated, which were subsequently verified by means of a forced recognition. In the first method, hypotheses about pronunciation variation were obtained by comparing the output of a phone recognition to a reference transcription. Forced recognition was then performed in order to eliminate rule hypotheses which were based on an incorrect output of the phone recognition. A drawback of this method is that not all of the pronunciation variation present in the material will be found, because phone recognition is only partly correct.

For this reason, we investigated a second method to automatically obtain hypotheses about pronunciation variation. In this method, forced recognition was also used, but there was no constraint beforehand on which hypotheses were tested; all deletions were possible. This was achieved by generating variants in which each phone could be deleted. After this forced recognition, we selected the most frequently applied rule hypotheses as the set of deletion rules.

The structure of this paper is as follows. In Section 2, the speech material and the CSR that we used for our experiments is described. This is followed by an explanation of the two methods in Section 3. Next, in Section 4, we will present the results. Finally, in the last section, the results are discussed and it will be explained how we will pursue this research in future.

2. Speech material and CSR

2.1. *Speech material*

The speech material used in this experiment was selected from the VIOS database. For training, 25,104 VIOS utterances (83,890 words) were used. All of the material that was used for obtaining rule hypotheses was selected from the same VIOS database.

2.2. *CSR*

The CSR that we used is part of OVIS (Strik et al., 1997). 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. In the following step, the energy in 14 Mel-scaled filter bands between 350 and 3400 Hz is calculated. Next, a discrete cosine transformation is applied on the log filterbands coefficients. The final processing state is a running cepstral mean subtraction. 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, word-based language models (unigram and bigram) and a lexicon. The acoustic models are continuous density hidden Markov models (HMMs) with 32 Gaussians per state. The topology of the HMMs is as follows: each HMM consists of six states, three parts of two identical states, one of which can be skipped (Steinbiss et al., 1993). In total, 39 HMMs were trained. For each of the phonemes /l/ and /r/, two models were trained, because a distinction was made between prevocalic (/l/ and /r/) and post-vocalic position (/L/ and /R/). For each of the other 33 phonemes context-independent models were trained. In addition, one model was trained for non-speech sounds and a model consisting of only one state was employed to model silence.

3. Method

In Section 3.1 and 3.2, the two methods of automatically obtaining hypotheses about pronunciation variation are described. The first step in both methods is obtaining automatic transcriptions (T_{aut}) of the speech material. The next step is a time-alignment of T_{aut} with T_{ref}, which is described in Section 3.3. In Section 3.4, we explain how the rule hypotheses were formulated on the basis of these alignments. Finally in the last section, we explain the selection criteria that were used to select the rule hypotheses.

3.1. *Method 1: Combination of phone recognition and forced recognition*

For method 1, T_{aut1} was obtained by performing a phone recognition. Instead of using a lexicon containing words, a lexicon containing phones was used. Furthermore, the

recognition process was constrained by using phone language models (unigram and bigram), which were trained on the reference transcriptions of the training corpus.

Next, T_{aut1} and T_{ref} were time-aligned in order to formulate rule hypotheses. Subsequently, a number of rule hypotheses were selected with the selection criteria that are described in Section 3.5. With the selection of rules variants were generated, which were then tested to see if they were valid by carrying out a forced recognition of the training utterances (25,104 utterances). After forced recognition, the number of times a rule was applied was counted, and divided by the total number of times a rule could have been applied to obtain the percentage of a rule's application. All rules with an application of less than 10% were excluded from the set of rule hypotheses.

3.2. *Method 2: All possible deletion rules verified by forced recognition*

For method 2, the deletion rule hypotheses were not constrained prior to forced recognition. We generated all possible variants in which each phone can be deleted. For practical reasons, a minimum number of one phone per syllable was taken, e.g. the following variants were generated for the word "wil" (to want): /wIL/, /wI/, /wL/, /IL/, /w/, /I/, and /L/. Next, with the variants that were generated forced recognition was carried out, and the result is T_{aut2}.

After forced recognition, T_{aut2} and T_{ref} were time-aligned, and the percentage rule application was calculated for each deletion rule. Subsequently, we selected a number of rule hypotheses with the selection criteria that are described in Section 3.5. In this paper, only those rule hypotheses for which the rule application is more than 20% are presented.

3.3. *Time alignment*

The second step after obtaining an automatic transcription of the speech material is a time-alignment of the automatic transcription (T_{aut}) with a reference transcription (T_{ref}). T_{ref} was automatically generated with the text-to-speech system developed at the University of Nijmegen (Kerkhoff & Rietveld, 1994).

In order to time-align T_{aut} with T_{ref}, a DP algorithm was used in which the distance between two phones is not just 0 (when they are identical) or 1 (when they are not identical), but is expressed in a more gradual way. More details about this DP

algorithm can be found in Cucchiarini (1996). In order to keep the CPU time within reasonable bounds, it was necessary to perform a selection of the utterances that could be processed. For method 1, we selected 50,000 utterances (82,101 words). More details about the selection criteria can be found in Wester et al. (1998). For method 2, we selected a set of 11,247 utterances (83,447 words).

3.4. *Formulation of rule hypotheses*

The DP-alignments were used to formulate hypotheses about pronunciation variation in the form of rules: A phone X with left context L and right context R in T_{ref} is replaced by Y in T_{aut}. In this way, three types of rule hypotheses could be obtained ('-' means that no phone is present):

Deletion rule:	{L X R}	→	{L - R}
Insertion rule:	{L - R}	→	{L Y R}
Substitution rule:	{L X R}	→	{L Y R}

3.5. *Selection criteria*

It can be expected that a number of the rule hypotheses that are incorrect or less important for our goals. For this reason, we decided to impose a number of selection criteria, which also, incidentally, led to a manageable number of rule hypotheses. The following selection criteria were used:

1. The rules must have a frequency of occurrence of more than 100. We assume that rules that are less frequently applied are less important.
2. The left (L) and right context (R) must be the same in T_{ref} and T_{aut}. We applied this criterion because we expect that T_{aut} will contain more errors if besides the specific phone, also the context in which it occurs is altered.
3. The left (L) and right context (R) may not be the beginning or end of an utterance. It can be expected that the beginning and end of an utterance contains more acoustic artefacts like noise or truncation of the speech signal.

Table 1. BU and TD statistics for the 17 rule hypotheses of method 1

	SUBSTITUTION RULES	**#BU**	**%BU**	**#TD**	**%TD**
1	{| e: n} → {| I n }	132	25	115	28
2	{@ n |} → {@ R | }	281	7	2882	39
3	{a: x |} → {a: R | }	236	11	506	28
4	{O m |} → {O n | }	226	13	347	16
5	{| m a:} → {| n a:}	341	32	129	11
	INSERTION RULES				
6	{a: - |} → {a: x |}	245	2	1368	27
7	{a: - |} → {a: L |}	127	1	729	14
8	{a: - |} → {a: R |}	880	7	2449	48
9	{y - |} → {y R |}	299	12	339	19
10	{@ - |} → {@ R |}	310	12	615	20
11	{R - |} → {R x |}	309	4	2473	23
	DELETION RULES				
12	{R t |} → {R - |}	125	10	118	16
13	{@ n |} → {@ - |}	183	2	1767	42
14	{f t |} → {f - |}	138	14	68	33
15	{x t |} → {x - |}	437	13	149	13
16	{A t |} → {A - |}	458	19	95	13
17	{i t -} → {i - |}	125	7	76	10

4. Results

In this section, the rule hypotheses, which were obtained with the two methods, are presented. Furthermore, we investigated whether the rule hypotheses that were found are described in the literature.

4.1. *Method 1: Combination of phone recognition and forced recognition*

In Table 1, the absolute (#) and relative (%) number of times a rule is applied are shown. BU denotes the bottom-up counts, which were obtained by phone recognition on 50,000 utterances (82,101 words). TD denotes the top-down counts, which were obtained by forced recognition on 25,104 utterances (83,890 words).

In our previous work (Kessens et al., 1999), we modeled the phonological processes: /n/-deletion, /r/-deletion, /t/-deletion, /@/-deletion, /@/-insertion, and a number of cross-word processes, for instance cliticization, contraction and reduction. One of the selection criteria for choosing to model these processes was that they occur frequently in Dutch. For this reason, we expect that these processes will also be found with the two automatic methods proposed in this paper. Table 1 shows that this is indeed the case. Five of the 17 rules concern previously modeled processes: deletion rule 13 concerns the process of /n/-deletion, deletion rules 14 and 15 relate to the process of /t/-deletion, and deletion rules 16 and 17 concern one of the cross-word processes. Another process, which also has been described in the literature, is deletion rule 12. Goeman (1999) describes that /t/-deletion at word endings mainly occurs after non-sonorant consonants, but that it can also occur after sonorants like, for instance, /R/. To summarize, all of the deletion rules, which is 35% (6/17) of the rules that we found with method 1, can be related to phonological processes that have been described in the literature.

4.2. *Method 2: All possible deletion rules using forced recognition*

For method 2, 15 deletion rules were obtained with a percentage rule application which is larger than 20%. These rules, together with the TD statistics are presented in Table 2. The top-down statistics were obtained by forced recognition on 11,247 utterances (83,447 words).

Table 2. TD statistics for the deletion rules of method 2

	DELETION RULES	#TD	%TD
1	{ d @ R } → { d - R }	126	27
2	{ v @ R } → { v - R }	444	30
3	{ n d @ } → { n - @ }	118	27
4	{ n d I } → { n - I }	107	60
5	{ | d @ } → { | - @ }	176	32
6	{ | h E } → { | - E }	172	26
7	{ s t @ } → { s - @ }	499	30
8	{ @ n t } → { @ - t }	175	23
9	{ @ n | } → { @ - | }	1653	40
10	{ @ R d } → { @ - d }	1095	51
11	{ @ R m } → { @ - m }	125	63
12	{ @ R t } → { @ - t }	346	38
13	{ @ R | } → { @ - | }	105	24
14	{ n t s } → { n - s }	131	27
15	{ i t | } → { i - | }	283	23

Again it can be seen that a number of rules are found which are related to previously modeled phonological processes: rule 8 and 9 concern the process of /n/-deletion, rules 10 to 12 concern the /r/-deletion. Furthermore, we found an example of /r/-deletion (deletion rule 13) which we have not previously modeled, but which could be extension of the process of /r/-deletion. Cucchiarini and van den Heuvel (1999) describe that /r/-deletion may occur if it is in coda position, preceded by a schwa and followed by a consonant. Deletion rule 13 might be an indication that this rule can be extended across words: an /r/ might be deleted at the end of a word if it is preceded by a schwa and if the following word begins with a consonant. Rule 14 is related to the process of /t/-deletion, and finally, rule 15 is related to one of the cross-word processes that we modeled previously. To summarize, 53% (8/15) of the deletion rules which we found using the second method are related to processes that have been described in the literature.

5. Discussion and conclusion

The goal of this research was to investigate whether it is possible to automatically derive information about pronunciation variation from the speech signal. To this end, we used two different approaches in which we employed a CSR to automatically derive this information from the speech signal. The first method was a combination of phone recognition and forced recognition, and in the second method, forced recognition was used in order to determine for each phone whether it has been realised or not. Since previous research has shown that forced recognition is a reliable tool for testing hypotheses about pronunciation variation (Kessens et al., 1998), we expect that this will also hold for the hypotheses about pronunciation variation which we found using each of the two automatic methods. Another reason for expecting the rule hypotheses to be reliable is that we found that 37-53% of the rules are related to Dutch phonological processes that have been described in the literature. Furthermore, we found that the two methods partly overlap, as two of the deletion rule hypotheses are found in both methods. For all of these reasons, we can conclude that there is evidence for assuming that the results of the two proposed methods are useful to automatically obtain information about pronunciation variation.

However, our methods have a number of limitations. One of the limitations in the way in which we derived rule hypotheses is that we only considered each phone with their direct left and right phone neighbours, whereas it is to be expected that in some cases pronunciation variation will extend over a larger number of phones. Therefore, in the future, we will derive rules using a larger context.

As mentioned in the introduction, a drawback of this method is that not all of the pronunciation variation present in the material will be found, because the phone recognition is only partly correct. Therefore, in the future we will try to optimize the phone recognition, e.g. by training phone models and phone language models on the basis of improved transcriptions in which pronunciation variation has been transcibed. A limitation of the second method is that we only considered deletions of phones. In future, we plan to extend the method to substitutions and insertions of phones. However, it will be clear that the more the number of rule hypotheses is expanded, the less reliable forced recognition will be. Therefore, usually some constraints are used during forced recognition. For instance, Cremelie & Martens (1999) constructed a stochastic automaton in order to model all possible pronunciation variants of a word. By using different transition and emission probabilities, and by prohibiting the substitution of phones from different phonemic classes, forced alignment is

constrained. It will be clear that the constraints imposed during forced recognition have to be selected carefully in order to prevent exclusion of relevant information.

Despite the fact that we have some evidence to assume that the two proposed methods of automatically obtaining rule hypotheses provide reliable information about pronunciation variation, we have no proof that the new rules that we found (rules which are not described in literature) actually apply in real speech. Another cause for finding the new rules could be improper acoustic modeling in our CSR. In order to analyze this, we are currently processing the results of an evaluation by expert listeners of the sets of new rules generated by the two methods. Another reason for this kind of evaluation is to find out which of the two methods is most suitable for our purposes. The results will reveal whether the new rules actually apply in real speech. In this way, we hope to show that automatic speech recognition can not only benefit from phonology, but that phonology can also benefit from automatic speech recognition.

6. 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. We would like to thank Catia Cucchiarini and an anonymous reviewer for their useful comments on this paper.

7. References

Booij, G. (1995). *The Phonology of Dutch.* Oxford: Clarendon Press.

Cremelie, N. & Martens, J.-P. (1999). In search of better pronunciation models for speech recognition. *Speech Communication* **29**(2-4), 115-136.

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

Cucchiarini, C. & van den Heuvel, H. (1999). Postvocalic /r/-deletion in Dutch: more experimental evidence. *Proc. of the 14th Int. Congress of Phonetic Sciences (ICPhS'99)*, San Francisco, 1673-1676.

Goeman, T. (1999). *T-deletie in Nederlandse Dialecten; Kwantitatieve Analyse van Structurele, Ruimtelijke en Temporele Variatie* (Ph.D. dissertation, VU Amsterdam).

Kerkhoff, J. & Rietveld, T. (1994). Prosody in Niros with Fonpars and Alfeios. In: P. de Haan and N. Oostdijk (eds.), *Proc. of the Department of Language & Speech, University of Nijmegen* **18**, 107-119.

Kessens, J.M., Wester, M., Cucchiarini, C. & Strik, H. (1998). The selection of pronunciation variants: Comparing the performance of man and machine. *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP'98),* Sydney, 2715-2718.

Kessens, J.M., Wester, M. & Strik, H. (1999). Improving the performance of a Dutch CSR by modeling within-word and cross-word pronunciation variation. *Speech Communication* **29**(2-4), 193-207.

Steinbiss, V., Ney, H., Haeb-Umbach, R., Tran, B.-H., Essen, U., Kneser, R., Oerder, M., Meier, H.-G., Aubert, X., Dugast, C. & Geller, D. (1993). The Philips research system for large-vocabulary continuous-speech recognition. *Proc. of the 3d Conf. on Speech Comm. and Techn. (Eurospeech'93),* Berlin, 2125-2128.

Strik H., Russel A., Van den Heuvel, H., Cucchiarini, C. & Boves, L. (1997). A spoken dialogue system for the Dutch public transport information service. *Int. J. of Speech Technology* **2**(2), 119-129.

Strik H. & Cucchiarini, C., (1999). Modeling pronunciation variation for ASR: a survey of the literature. *Speech Communication* **29**(2-4), 225-246.

Wester, M., Kessens, J.M. & Strik, H. (1998). Two automatic approaches for analyzing the frequency of connected speech processes in Dutch. *Proc. of the Int. Conf. on Spoken Language Processing, Student Day (ICSLP'98),* Sydney, 3351-3356.