TESTING A METHOD FOR MODELLING PRONUNCIATION VARIATION

Judith Kessens, Mirjam Wester, Catia Cucchiarini, Helmer Strik

A²RT, Dept. of Language & Speech, University of Nijmegen
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
{kessens, wester, catia, strik}@let.kun.nl, http://lands.let.kun.nl/

ABSTRACT

In this paper we describe a method for improving the performance of a continuous speech recognizer by modelling pronunciation variation. Although the results obtained with this method are in line with those reported by other authors, the magnitude of the improvements is very small. In looking for possible explanations for these results, we computed various sorts of statistics about the material. Since these data proved to be very useful in understanding the effects of our method, they are discussed in this paper. Moreover, on the basis of these statistics we discuss how the system can be improved in the future.

1. INTRODUCTION

At the Department of Language and Speech of the University of Nijmegen we are working on a Spoken Dialogue System (SDS) that will be employed to automate part of a public transport information service. This system was adapted from a German prototype developed by Philips Research Labs, and was further improved by means of a bootstrapping method [1, 2].

An important component of this SDS is a continuous speech recognizer (CSR). This part of the SDS was also gradually improved through the bootstrappping method, by adding more data. However, since a point was reached at which no further increase in performance could be obtained by increasing the data, new methods of improving the system were sought. Given that the SDS is a mixed-initiative system and that the kind of speech the callers may use is extremely varied, we thought of improving the system's performance by modelling pronunciation variation.

In this paper, the method used for modelling pronunciation variation is discussed in detail (2). Subsequently, the results obtained with this method are presented together with various sorts of statistics about the material (3). In section 4 we discuss how the statistics we computed helped us to understand why the variations in performance were so small, and how this knowledge can be used to improve the system in the future.

2. METHOD AND MATERIAL

2.1. Method

The starting point of the current research was a CSR in which a single pronunciation lexicon was used. For each word only the transcription we thought was most probable (the canonical form) was available. In this experiment we wanted to test to what extent the performance of the CSR could be improved by modelling at least part of the pronunciation variation that is encountered in the material. The approach we adopted in this attempt resembles those used previously with success in [3, 4].

In this approach phonological rules are used to generate pronunciation variants, i.e. to expand the lexicon. The expanded lexicon can then be used during training, recognition (test) or both. During test the old test lexicon is simply replaced by the new one, in order to make it possible to recognize pronunciation variants. During training the pronunciation variants can be used to obtain new acoustic models as follows.

❏ Forced recognition is carried out to determine which variant is realized in the corpus. This way a new transcription of the training corpus is obtained.

❏ The new transcription of the training corpus is used to calculate new phone models.

Our ultimate goal is to find the rules that are optimal in the sense that they produce the greatest increase in performance. The goal of the current research was to test whether the method proposed above was suitable for our purposes. In order to do so we have tested the method with only four phonological rules, as will be explained below.

2.2. Phonological rules

In order to select the initial set of phonological rules a number of criteria were followed. As is well known, variation occurs both within words and at word boundaries. Given the use of a lexicon in our CSR, it was obvious to begin with word internal variation. Therefore, the first criterion was to choose rules of word phonology.

Second, we decided to start with rules concerning those phenomena that are known to be most detrimental to automatic speech recognition. Of the three possible recognition errors, i.e. insertions, deletions and substitutions, the first two have the greatest consequences for speech recognition, because they affect the number of segments present in different realizations of the same word. Therefore, starting with rules concerning insertions and deletions was the second criterion we adopted.

A third criterion was to choose rules that are frequently applied. Actually, frequently applied is amenable to two interpretations. A rule can be frequent either because it is frequently applied whenever the context for its application is met or because the context in which it can be applied is very frequent (even though the rule is applied in only 50% of the cases). Obviously, it is this latter case of 'frequent occurrence' that is most interesting for automatic speech recognition, since in this case it is difficult to predict which variant should be taken as the

canonical form, while in the former case the most frequent form would probably suffice as sole transcription.

A fourth criterion (related to the previous one) we followed was that the rules should regard phones that are relatively frequent in the language, since rules that concern infrequent phones probably have fewer consequences for the recognizer's performance. Finally, we decided to start with rules that have been extensively described in the literature, so as to avoid possible effects of overgeneration and undergeneration due to incorrect specifications of the rules. On the basis of the above-mentioned criteria the following 4 rules were selected.

1. /ə/-deletion: obs + ə + liq + ə → obs + liq + ə
                     Ex: /ɑndərə/ → /ɑndrə/
2. /t/-deletion: obs + t + cons → obs + cons
                     son + t + obs → son + obs
                     word final: obs + t → obs
                     Ex: /snɛlstmoxələk/ → /snɛlsmoxələk/
3. /n/-deletion: syllable final: ə + n → ə
                     Ex: /rɛizən/ → /rɛizə/
4. /ə/-epenthesis: in nonhomorganic clusters in coda position
                     Ex: /mɛlk/ → /mɛlək/

2.3. Material

The CSR used in this experiment is part of an SDS [1, 2]. The speech material was collected with an online version of the SDS, which was connected to an ISDN line. Recordings with high levels of background noise were excluded from the material used for training and testing. The training and test material consisted of 24,676 utterances (81,090 words) and 6,276 utterances (21,106 words), respectively.

The most important characteristics of the CSR are the following. The input signals consist of 8 kHz 8 bit A-law coded samples. 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. Apart from these 14 filterbank coefficients the 14 delta coefficients, log energy, and slope and curvature of the energy are also used. This makes a total of 31 feature coefficients. The CSR uses acoustic models (HMMs), language models (LMs: 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 the online SDS the output of the CSR, and thus the input of the following natural language processing component, is a wordgraph [1, 2]. In the research version it is possible to use the LMs to compute the Best Sentence (BS). Obviously, the error rates for the wordgraph are much lower than those of the BS [1, 2]. Nevertheless, we will use the BS in this article, because they are better suited for the goals of the present research: evaluation of the results is easier and more transparent.

The single variant training lexicon contains 1,436 entries (these are all the words contained in the training corpus). The four phonological rules selected for investigation affect 536 of the 1436 (38%) words in the training lexicon. In a number of cases more than one rule could be applied to one word. On average, 1.3 variants were generated for each of the 536 words. The multiple variant lexicon contains 2,147 entries, 1,436 (67%) of which are canonical.

The test lexicon contains 863 entries, which are all the words present in the online version. The number of out of vocabulary (OOV) words in the test corpus is 298. The four phonological rules concern 354 of the 863 entries in the test lexicon (41%). Also in this case, more than one rule could be applied to one word. On average, 1.3 variants were generated for each of the 354 words. The multiple pronunciation lexicon contains 1,342 entries, 863 (64%) of which are canonical.

2.4. Forced recognition

Forced recognition was imposed through the language models (LMs). For each sentence unigram and bigram LMs were derived on the basis of 100.000 repetitions of the same sentence. After the first forced recognition round, 484 utterances of the training corpus were not correctly recognized. 47 of these utterances turned out to contain obvious transliteration errors which were corrected afterwards. Since the remaining 437 sentences appeared to be problematic for a number of reasons (they contained background noise, disfluencies, unexpectedly long pauses within words and in some cases the loudness level was insufficient) they were removed from the original training corpus and only 24,667 utterances were used for further experiments.

It turns out that forced recognition is a useful tool to identify all sorts of errors and utterances which (for some reason) are problematic for the CSR. These utterances will certainly be examined more closely in the near future. Instead of forced recognition with LMs, as described above, we could have used a standard Viterbi algorithm. Although the main advantage of the Viterbi algorithm is that a forced alignment can be obtained for all utterances, the main disadvantages of this algorithm are (1) that the alignment is not always meaningful, e.g. because the transliteration contains errors, and (2) that it is not possible to find the errors and the problematic utterances.

The resulting training corpus with 24,667 utterances was again used for training and forced recognition. In the 24,566 cases in which forced recognition was successful, the pronunciation variants chosen by forced recognition were substituted for the original (canonical) transcriptions. In the 101 cases in which forced recognition was not successful, the canonical form was chosen. The new transcriptions were subsequently used to train new phone models.

3. RESULTS

Above it has been explained how single (S) and multiple (M) pronunciations during training lead to two different sets of phone models. In addition either single (S) or multiple (M) pronunciations can be used in the test lexicon. This makes a total of four combinations, for each of which we present the sentence and word errors rates (SER and WER, respectively) of the best sentences (BS) in Table 1. As appears from Table 1, there are only slight variations in recognition performance between the various conditions. Nevertheless, it is interesting to analyze these data in more detail, in order to see whether the various tendencies are in line with those reported in the literature. For instance, the worst performance level appears to be obtained

when multiple pronunciations are used for training but not for testing (i.e., when the new phone models are combined with the old lexicon). This is exactly what appeared in [4].

Table 1. SER and WER for the BS of four different CSRs.

CSR	SS	SM	MS	MM
train	S	S	M	M
test	S	M	S	M
SER(%)	32.63	32.39	33.03	32.41
WER(%)	23.63	23.50	23.81	23.50

Furthermore, Lamel & Adda [4] found that using multiple pronunciations for testing gave better results than using single pronunciation lexicons. This is confirmed by our data (compare column 2 with column 3). However, these authors also found that recognition performance improved even further when multiple pronunciations were used both for training and for testing, which is not confirmed by our data: there is practically no difference in performance between column 3 and column 5.

Therefore, on the basis of these results we can conclude that the applied method improves the performance, albeit to a small extent. Moreover, the observed improvements are in line with those reported elsewhere [4]. However, since the magnitude of the changes is considerably smaller than that reported by other authors, it is interesting to consider why this is the case.

A possible explanation for these results would be that during forced recognition the CSR selects the wrong variant. In order to test whether this was the case, we checked for a small amount of words whether the correct pronunciation variant was chosen by looking at and listening to the signals. Since it turned out that in 90% of the 711 words the correct version was chosen, there is no reason to believe that the small increase in performance was mainly due to errors in forced recognition.

Another reason could be that the number of pronunciation variants that can be selected is relatively small. Against this background it is interesting to know how often one of the alternative variants could be chosen and how often it was indeed chosen. Of the 81,090 recognized words in the training corpus 66,590 words had single pronunciations, whereas for 14,500 words (17.9%) alternative pronunciations were available. In 6,363 cases an alternative variant was indeed chosen. This is 7.8% of the total number of words (81,090) and to 43.9% of the 14,500 cases in which an alternative variant could be chosen.

In the test corpus there are 19,962 and 20,011 recognized words for the original phone models and for the new ones, respectively. For 15,556 and 15,640 words, respectively, there were only single pronunciations, while for 4,406 and 4,371 words (22.1% and 21.8%, respectively) an alternative variant could be chosen. In 2,028 (original) and 2,128 (new) cases one of the multiple variants was chosen. This is 10.2% and 10.6% of the total number of words (19,962 and 20,011, respectively) and 46.0% and 48.7% of the 4,406 and 4,371 potential cases.

From these data we can infer that, on average, one of the

alternative variants is chosen in about 45% of the possible cases, and in 8-10% of the total number of words. However, most variants will only differ in one phone from the canonical form. A comparison of the two transcriptions of the training corpus (i.e. the canonical forms versus the transcriptions obtained with forced recognition) reveals that they differ in 6,594 of the total 318,774 phones (2.1%). This seems to be one of the reasons why the effects on recognition performance are far from dramatic.

In order to gain more insight in these data, we compared the four CSRs. First we determined for each CSR which BS contained an error. Subsequently, for four of the six logical combinations of the CSRs (those in which only one factor changes, while the other is kept constant, i.e. SS-SM, MS-MM, SS-MS and SM-MM) the BS containing errors were compared. The results of these comparisons are shown in Table 2.

Table 2. Comparisons of the performances of the four CSRs.

CSR 1	SS	MS	SS	SM
CSR 2	SM	MM	MS	MM
same errors	1630	1592	1089	1066
other errors	364	400	836	844
improvements	54	81	123	123
deteriorations	39	42	148	124
net result	+15	+39	-25	-1

From Table 2 it appears that a considerable number of utterances contain a recognition error in both CSRs, either the same (row 3) or a different one (row 4). Furthermore, there are cases in which a better solution is chosen (improvements, row 5). However, since in an almost equal number of cases a worse solution is chosen (deteriorations, row 6), the two effects balance each other off and the net result (row 7) is small. This neutralization effect explains why no considerable changes in the error rates were observed in Table 1.

It is well-known that including alternative pronunciation variants leads to some sort of trading relation between improving performance (by covering part of the variability in speech) and deteriorating it (by increasing the confusability between the entries in the lexicon).

Based on the fact that only 2.1% of the phones differ between the two transcriptions of the training corpus and the results shown in Table 1, it could be concluded that the use of multiple pronunciations during training has little consequences for the recognition process, for instance, because the acoustic models hardly change. However, comparison of columns 4 and 5 with columns 2 and 3 in Table 2 reveals that varying the phone models produces more changes than varying the test lexicon. A comment on this may be in order.

Using multiple variants for testing simply means that the CSR can choose from among a greater number of possibilities for each word. Put differently, the variations in the system occur at the word level and concern only a limited number of words.

When multiple variants are used for training, on the other hand, they produce different acoustic models. In other words, in this case the variations occur at the phone level. Since all words in the corpus are made up of phones, the effects of variation modelling during training are likely to be more pervasive.

Further inspection of Table 2 also reveals that, in spite of the greater number of changes in columns 4 and 5, the net result is negative, while in columns 2 and 3 it is positive. In other words, the fewer changes in columns 2 and 3 successfully conspire to achieve better recognition results, while the net result of the larger number of changes in columns 4 and 5 is a deterioration.

A final remark concerns the number of utterances in which there is room for improvement. It appears that 4,038 of the 6,276 utterances are recognized correctly in all four systems. Since 1,066 utterances contain OOV words they can never be recognized correctly. Therefore there is only room for improvement in the remaining 1,172 utterances. With this in mind no dramatic changes in recognition performance can possibly be expected.

4. DISCUSSION AND CONCLUSIONS

In the previous section we examined the results of an experiment aimed at determining the contribution of pronunciation variation modelling to improving the performance of our CSR. One of the things we have learned from this experiment is that forced recognition as it was implemented in this method is a useful instrument to identify possible errors in the transliterations and in the lexicons and to spot the utterances that, for some reason, present insurmountable problems to automatic speech recognition. Studying these sentences in further detail is certainly worthwhile. Furthermore, in 90% of the cases this forced recognition procedure selects the correct pronunciation variant.

As far as the main goal of this experiment is concerned, i.e. establishing whether the applied method is suitable for improving the performance of our CSR, we can conclude that there are no reasons to assume that this is not the case. As a matter of fact the observed changes are in line with those reported by other researchers. The only problem seems to be that in our research the variations are very small. In this respect it may be instructive to consider the following facts.

First, the statistics concerning the material may have played an important part in limiting the effect of pronunciation modelling on recognition performance. It should be borne in mind that an alternative variant was chosen in only 8-10% of the cases. Moreover, in most of the cases the alternative transcriptions differed in only one phone from the canonical form. In connection with this, no more than 2.1% of the phones were changed as a result of variation modelling. Furthermore, in only 1,172 sentences was there room for improvement. Finally, another factor that should not be overlooked concerns the phones involved in the rules under study. Since the four rules concern phones that are very frequent in Dutch and in the material under study (in the training corpus /n/, /t/ and /ə/ are the three most common phones), there are so many occurrences of these phones, that the impact of variation modelling is likely to be limited.

If we consider all these aspects, it is not surprising that recognition performance hardly improved. Moreover, it is important to point out that our research is at an early stage and that a number of things that we intend to do have not been done yet. For instance, in this experiment we have confined ourselves to within word variation, whereas modelling variation above the word level may be even more important [5]. Second, since only four rules were investigated, only a small part of the variation in the material could be covered. However, it is our intention to expand the set of phonological rules so as to maximize coverage. Another factor that might be responsible for the limited impact of pronunciation modelling on recognition performance and that we have not controlled yet is overcoverage, that is the fact that the rules selected generate a great number of variants that are not present in the corpus. This was to be expected because no pruning of variants whatsoever was carried out. The reason for this is that in this phase of our research we did not want to exclude variants that might turn out to be useful at a later stage. Since we opted for overcoverage, this should be considered when analyzing the results. It is obvious that in the future we intend to examine pronunciation variants more critically, before including them in the lexicon. More attention will be paid to the variants that are indeed present in the corpus. In addition, the frequency with which they occur will also be investigated, so that a probability count can be attached to each variant. In the light of these considerations it is therefore legitimate to conclude that the results of this experiment are promising, in spite of the limited increase in recognition performance.

5. ACKNOWLEDGEMENTS

6. REFERENCES

[1] H. Strik, A. Russel, H. van den Heuvel, C. Cucchiarini & L. Boves (1996) Localizing an automatic inquiry system for public transport information, *Proc. ICSLP'96*, Philadelphia, pp. 853-856.

[2] H. Strik, A. Russel, H. van den Heuvel, C. Cucchiarini & L. Boves (1997) A spoken dialogue system for the Dutch public transport information service, to appear in *International Journal of Speech Technology*.

[3] M.H. Cohen (1989). *Phonological structures for speech recognition*. PhD dissertation, Univ. of California, Berkeley.

[4] L.F. Lamel & G. Adda (1996) On designing pronunciation lexicons for large vocabulary, continuous speech recognition, *Proceedings ICSLP'96*, Philadelphia, pp. 6-9.

[5] N. Cremelie & J.P. Martens (1995) On the use of pronunciation rules for improved word recognition, *Proceedings EUROSPEECH'95*, Madrid, pp. 1747-1750.