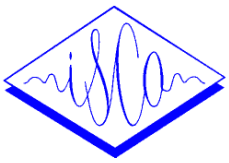

2nd European Conference on
Speech Communication and Technology
EUROSPEECH '91
Genova, Italy, September 24-26, 1991

Optimizing Lexical Fast Search in a Large Vocabulary Isolated Word Speech Recognition System

H. Drexler, R. Roddeman, L. Boves & H. Strik

Nijmegen University, Dept. of Language and Speech, Phonetics Section
P.O.Box 9103, NL-6500 HD Nijmegen, The Netherlands
e-mail : drexler@afac.kun.nl

abstract

This paper describes methods developed to improve the performance of a preselection algorithm in a large vocabulary isolated word recognizer. First we investigate ways for optimizing the phonetic representations of the words in the lexicon (the base forms of which are obtained from a grapheme-to-phoneme converter) in such a way that the inherent limitations of the front-end are taken into account and where possible are remedied. Then an attempt is made to improve the quality of the symbol strings produced by the acoustic front-end.

keywords: Isolated-Word-Recognition preselection lexicon

Introduction

The work described in this paper was done in the framework of the ESPRIT project POLYGLOT. One of the aims of that project is to adapt a speaker adaptive large vocabulary isolated word recognizer, originally developed for Italian (Billi et. al., 1989), to a number of other European languages, including Dutch. The result should be a multi-lingual recognizer that employs the same hardware and software for all languages.
The systems runs on an MS-DOS PC that uses one or two special purpose plug in boards. After signal processing, resulting in vectors of 20 LPC Cepstrum coefficients and two energy values for each 10 ms speech frame, the acoustic distance is computed to a set of some 15 phonetic templates. These templates are essentially the only speaker dependent information in the system. The templates are easy and fast to train for each new speaker.
The trellis of template labels with acoustic distance scores is then submitted to a dynamic programming procedure that outputs the most likely string of phonetic units that describe the word to be recognized. This phase is naturally called Phonetic String Build Up (PSBU).
The string formed by PSBU is then used for a fast lexical access that retrieves the 100 or so most likely word candidates (Billi et. al., 1986). The Dynamic Programming string match used in this Preselection phase relies on knowledge about phoneme confusions, that is built during a speaker independent training phase.
In the next stage, called Fine Phonetic Analysis (FPA), the word candidates produced during preselection are sorted and only the five or so best scoring candidates are retained.
Finally, a language model combines word frequencies, a bigram model and some deterministic linguistic knowledge and the acoustic probability of each candidate in a single probabilistic score.
Up to now, work has been concentrated on PSBU and Preselection. In the remainder of this paper we will describe the problems encountered in attempting to adapt the knowledge sources available for Italian to the Dutch language and the solutions that we have experimented with.

Phonetic String Build Up (PSBU)

The speech is picked up by a table-mounted microphone and digitized with a sampling frequency of 16 kHz and 12 bit resolution. After end-point detection, the speech signal is submitted to an autocorrelation LPC analysis. LPC-20 is used and the frame-rate is 100 Hz. The LPC coefficients are converted to 20 cepstral coefficients. Energy and Log-Energy are also computed for each frame. Every 10 ms frame of the input is compared to a set of spectral templates, represented in the form of cepstral coefficients; each template is identified by means of a phonetic label. Obviously, this procedure limits the set of templates to represent quasi-stationary states. Especially highly dynamic speech sounds like plosives cannot be directly represented in this framework. Moreover, continuants that have highly variable acoustic characteristics (like the /r/, that in Dutch can be a dental or an uvular trill, as well as a /ə/-like vowel) or that have spectra that closely resemble vowel spectra (like the /l/) are not included in the set of spectral templates. In fact, only some 14 templates are used; most monophthong vowels are present; in addition, there are three fricatives and one nasal template. This procedure limits the set of templates to those that can be detected reliably, but at the same time it leads to an extremely crude description of the phonetic make-up of the words to be recognized.
For each 10 ms frame the WLR distance between the analysis frame and all spectral templates is computed and stored in a lattice. A dynamic programming algorithm is then used to search the optimal sequence of

labels through the lattice. This sequence optimizes a score based on acoustic (WLR) distance, the frequency of the phoneme corresponding to the label in the language, phoneme-pair frequency, and phoneme durations. During the search adjacent frames with identical labels will be merged. Sequences of identical labels that do not exceed the minimum duration for the corresponding phoneme are appended to neighbouring phonemes.

The result of this procedure is a string of phonetic elements. It should be clear that this string gives an extremely crude representation of the phonetic make-up of the word to be recognized, because it can only contain the limited number of phonetic elements covered by the set of template labels. Specifically, no plosives and no /l,r/ sounds are present in the PSBU strings. Yet, these strings must be used to select the most likely word candidates from the lexicon of the recognizer.
Information about phoneme and phoneme-pair frequency was obtained from two sources. First, the lexica originating from a previous ESPRIT project *Linguistic analysis of European languages* were used. In addition, use was made of the CELEX lexical database that is available in Nijmegen. We derived the model for prototype durations from a number of labelled speech databases. One was available from the Dutch national SPIN-ASSP program on text-to-speech conversion; it contains about 15 minutes continuous read speech of one professional speaker. A number of smaller data bases available in Nijmegen were also used. Finally, a small data base containing the training words was recorded and segmented for the purpose.

Lexicon

The lexica contain word models generated by a rule based program for grapheme- to- phoneme conversion. This program was developed in Nijmegen as part of a system for automatic text- to- speech conversion. Thus, it produces an accurate phonemic representation of the citation forms of the words. It predicts assimilation and reduction phenomena in poly-syllabic words. The word models generated in this way were not adjusted to the restrictions posed by the set of phonetic labels in the recognizer. Specifically, the word models are not adapted to reflect the fact that the PSBU strings are limited to 14 phonetic elements. In the base line system there is just one entry in the lexicon for each word, so no between-speaker pronunciation variation is accounted for.
We have used two lexica to test the preselection performance in this paper. The first lexicon contains 2000 words; the other contains 8000 words. The 2000 words comprised in the first lexicon are a subset of the words in the 8000 word lexicon. The lexica contain the 200 most frequent words of the language, a set of 300 words chosen to include as many of the diphones and triphones occurring in the language as possible, another set of 500 phonetically representative words, plus the 1000 c.q. 7000 words that follow the 200 most frequent words in the CELEX data base (of course, words already present in the sets of phonetically representative words were skipped).
The 200 most frequent words plus the set of 300 phonetically representative words are used to train the recognizer. The set of 500 phonetically representative words is used to test Preselection performance.

Training

The major part of the training in our recognizer concerns the confusion matrix. Training starts with a bootstrap confusion matrix that is adapted in an iterative process. It appeared that the result of the training depends heavily on the initial matrix. Uniform matrix entries seem to lead to better results than a matrix initialized on the basis of phoneme confusions reported in the literature.
During the first training phase the optimal alignment between complete PSBU strings for the training words and the word models in the lexicon are found using the present version of the confusion matrix. Then the number of substitutions between phonetic labels in the PSBU strings and phonemes in the word models are counted. The results are used to update the confusion matrix. It is not possible to prove that this procedure must converge; nor is it possible to define a useful criterion function that allows one to decide when to stop the training. In general, five iterations seem to be sufficient.
In the second training phase the same procedure is repeated, but now the PSBU strings are divided into shorter substrings that are used in independent training subsessions. The matrices resulting from those partial trainings are finally combined using deleted interpolation methods.
As said before, the confusion matrix is trained using the 200 most frequent words in the language (most of which are very short) and 300 words chosen to contain the most frequent diphones and triphones in Dutch. The training material was spoken by five female and five male subjects. The training material was recorded in a normal office environment, using the exact same hardware that is used during actual use of the recognizer.

Preselection Results

Preselection performance of the system described above was tested with recordings made of the same ten speakers used for training. The recordings of the test set were made under the same condition as described for the training set. The test set consisted of 500 phonetically representative words. The solid lines in Figs. 1 and 2 show the results of the preselection performance of the base-line system. The horizontal axis in the figures show the size of the preselection set as a proportion of the size of the lexicon. The vertical axis corresponds to the proportion of words in the test set that had the correct lexicon word in the preselection set. It can be seen that the rate at which the inclusion of the correct word in the preselection set increases with increasing size of that set diminishes quickly if the size of the set exceeds 100 words. It can also be seen that the performance of the base line system is clearly insufficient.

To improve preselection performance we have increased the amount of phonetic detail in the PSBU strings; also, the word models in the lexicon were adapted to the limitations caused by the set of 14 phonetic labels. In the next sections we explain the procedures using the plosives as an example.

Adapting the models

First, we tried to improve the way in which word initial plosives are handled. These sounds are notoriously difficult to discriminate from abrupt word onsets. Thus, it is not surprising that in most cases nothing in the PSBU strings suggests the presence of an initial plosive. However, sometimes the burst is detected and represented by a fricative label. This is different from what happens to word medial plosives; almost always their presence gives rise to the inclusion of the fricative label in the PSBU string. What we have now are context dependent confusions: a word initial plosive tends to be deleted (or to be confused with a null-label) whereas word medial plosives are confused with fricatives. The large number of deleted word initial plosives result in a confusion matrix that is harmful to word medial plosives: the penalty for aligning the PSBU string of a word containing a plosive (present in the disguise of a fricative) with a lexical model without that plosive is to low. Increasing that penalty is not possible, since it would decrease the possibility that lexical models with initial plosives are chosen as a likely candidate.
As a solution of that problem of context dependent confusions we introduce context dependent labels; i.e., we represent initial plosives by a separate symbol in the word models. The meaning of this symbol can be described as:

> Most of the time I am nothing, but sometimes I am the burst of an initial plosive that is represented in the PSBU string by a fricative label.

If we had not adapted the models, the absence of initial plosives in the PSBU strings would increase the deletion probabilities of all plosive symbols in the word models. The local behaviour of initial plosives would affect the global deletion probabilities of plosives. Deleting all initial plosives from the word-models would prevent this, but would lead to misalignments if the initial plosive is represented by a fricative. Introducing a separate symbol in the word models solves both problems. The effect of the adapted models on preselection performance is shown by the dashed lines in Figs. 1 and 2. In the 2000 word lexicon the performance improves as long as the preselection sets are small. The quickly diminishing gain of the adaptation with larger sets is probably due to the fact that it addresses only one of the problems. Of course, words that did not get their correct lexical form included in the preselection set for other reasons than problems with plosives will gain very little. In the 8000 word lexicon the improvement is much more impressive. It seems that the 8000 word lexicon needs the extra discriminative power of the adaptation more than the 2000 word lexicon because the density of the word models in the space spanned by the 8000 word lexicon is higher.

Adapting the PSBU strings

Most of the time word medial plosives appear in the PSBU string as fricatives. The alignment algorithm maps those fricatives onto plosives in the word models. This means that the discriminative power between fricatives and plosives is considerably weakened.
Since we have two Energy scores for each 10 ms frame, it is possible to detect word medial plosives by looking for dips in the Energy envelope. In fact, we can segment the sequences of 10 ms frames in subsequences separated by Energy dips, consider those dips as the presence of a plosive and build PSBU strings accordingly by applying the dynamic programming algorithm to the subsequences and concatenating the resulting PSBU substrings. The resulting PSBU strings now will contain a label corresponding to plosives. Of course, doing so means that the confusion matrix for preselection must be retrained. That was done using the same procedure as described above.
At word endings a special problem arises. Long bursts of word final plosives appear in the PSBU string as fricatives. We constructed an algorithm which looks at the energy contour of the input signal and decides whether word final 'fricatives' that are immediately preceded by a dip symbol might rather correspond to the burst of a final plosive; the decision is based on the observed length of the burst. If the frication is likely to correspond to a plosive release, the dip symbol and the fricative symbol are replaced by a single prototype representing a final plosive. Successful detection of final plosives is dependant on both the performance of the dip detector and the final burst detector. This performance will be lower than that of the dip detector alone. We use a separate symbol for final plosives in the word models for the same reason we use a separate symbol for the initial plosives. Preselection results again improve thanks to this modification as can be seen from the dotted lines in Figs. 1 and 2. Again, the improvement is largest for small preselection sets.

Adaptation of the lexicon

Although the adaptations described above each led to noticeable improvements, the performance of preselection is still not adequate. This is probably due to the large discrepancies that can exist between the extremely crude phonetic descriptions given by the PSBU strings on the one hand and the fairly traditional phonemic representation of the words in the lexicon on the other. Thus, it was decided to analyse the strings that PSBU produces for the words that were not included in the preselection sets to see if the lexical representations of these words can be adapted to reflect the phonetic discrimination power of PSBU. First results obtained with the 2000 word lexicon are shown as the dash-dot line in Fig. 1. The performance now approaches a level that can be considered acceptable. Up to now, the additional

adaptations of the lexicon were done by hand. Since that results in a time consuming procedure that, moreover, requires highly skilled personnel, we have started an investigation into the possibility to adapt the grapheme-to-phoneme converter so that it generates phoneme strings that are better adapted to the needs of PSBU and preselection.

Conclusions

We have shown that harmonization of strings built by an acoustic classifier with word models improves preselection results. We described automated procedures to detect plosives and to adjust the word models accordingly. For applications it will be necessary to analyze the remaining discrepancies carefully in order to fine-tune the system.

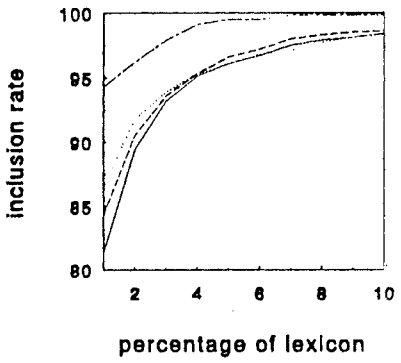

Figure 1: Preselection performances on the 2000 word lexicon

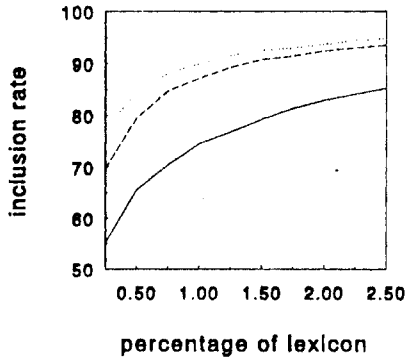

Figure 2: Preselection performances on the 8000 word lexicon

References

Billi, R., Arman, G., Cericola, D., Massia, G., Mollo, M., Tafini, F., Varese, G., Vittorelli, V. (1989) A PC-Based Large Vocabulary Isolated Word Speech Recognition System. In *Proceedings of EUROSPEECH 1989*, Vol 2, pp 157-160

Billi, R., Massia, G., Nesti, F. (1986) Word Preselection for Large Vocabulary Speech Recognition. In *Proceedings of ICASP 1986*, Vol 2, pp 65-68