CONSTRUCTING A LARGE SIZE LEXICON FOR A CONTINUOUS SPEECH RECOGNITION SYSTEM

Lars Bo Larsen *, Anders Bakgaard *,
Michael Bundgaard **

*) Speech Technology Centre, Institute of Electronic Systems, Fredrik Bajers Vej 7, DK-9220, Aalborg, Denmark.

**) Institute of General and Applied Linguistics, University of Copenhagen, Njalsgade 80 DK-2300 Copenhagen S. Denmark

0. ABSTRACT

This paper describes the principles and the implementation of a large size lexicon to be used within a continuous speech recognition system, the HEAD system.

The lexicon contains the system application vocabulary and it is a major design criterion that a new vocabulary is entered into the HEAD system at lexicon level only. As the lower levels of the system remains unaltered, no additional training of the system is required when the vocabulary is changed. The lexicon is based on partial allophone information realized as a set of distinctive phonetic features. A lookup consists of a search- and a verification-phase. The search phase is based a number of predefined key archiphones, whereas the verification phase utilizes all available information.

1. INTRODUCTION

The HEAD system [1],[2] is currently being developed in a joint university - industry project at the Speech Technology Centre at the University of Aalborg, Denmark. The system is built around an expert system which has three levels of phonetic expertise. The first (lower) level is based on a neural network technique for extracting the acoustic-phonetic distinctive features from the incoming speech signal. The neural network is a Kohonen Self Organizing Neural Network (SONN) [3]. The SONN produces a set of acoustic-phonetic distinctive features every 10 msecs, which is the signal analysis framing interval. To identify segments in the speech signal, a Multi Layer Perceptron (MLP) is trained on the SONN output to divide the speech signal into segments corresponding to individual speech sounds. The classification and segmentation processes is described in detail in [2],[5] and is hereafter referred to as the allophone classifier. At this point the proposed allophones are inserted as hypothesis on the expert system blackboard to form the input to a lexicon, which is implemented as the middle level of the expert system, and is described in detail in the following sections. The upper level of the HEAD system consists of an island driven parser that combines words, produced by the lexicon into legal phrases in respect to the given application grammar [1].

2. PHONETIC BACKGROUND

This section describes the phonetic background of the lexicon architecture. The basic task of the lexicon is to compare a string of incoming phonetic tokens to a phonetic representation of the system application vocabulary, aiming at the production of potential word candidates. This implies that the phonetic representation must be compatible to the input string. In the ideal case the input consists of a string of correctly identified allophones. In real life, however, there is a degree of uncertainty which has to be taken into account. This is achieved by using a set of correlated distinctive features that together designate an allophone, instead of using allophone symbols as the basic phonetic unit.

Sixteen distinctive features, derived from a speech production model (vocalic, consonantal, silence, plosive, nasal, friction, etc.), are used to denote the corpus of Danish vowels and consonants. E.g. the Danish vowel [i] is represented by the set: [+vocalic, +high, +front, +sonorant, -consonantal, -back, -middle, -low, -round] where "+" implies that the feature is present, and "-" that it is not.

Several advantages are gained by using a set of features to describe an allophone rather than just the allophone label itself. When the allophone classifier fails to produce the correct allophone label for a segment, it will typically mistake two acoustically close allophones, e.g. the fricatives [s] and [f]. The problem is that there is no direct way to represent the adjacency of two allophones, so a minor mistake like this will have to be treated by an extensive error handling procedure. By using a set of distinctive features, the error will in most cases be reduced to a single feature in the set which is easily handled as shown in the following section. In the example given above, [f] and [s] are represented by almost identical feature sets, the only difference being that the [f] set includes [+labial], whereas the [s] set includes [-labial].

The lexicon handles this by not requiring a full match between an input set and a set denoting an allophone in the transcription of a word candidate. It is only required that a sufficient number of features are present and correct to determine a match. This is a very important property, as it allows the lexicon to operate on erroneous and/or incomplete input data, which will be the operating conditions in any real life application.

To speed up the lexicon lookups, a search is first made based only on certain predefined key archiphones (denoted superphones). This is done in order to minimize the matching process by reducing the search space. Instead of a "flat" structure, where the total vocabulary is considered, the lexicon is organized as a number...
of tree structures, thus narrowing the field of possible word candidates for each step in a search process. The branching points of the search structures consist of the superphones. A lexicon look-up includes two phases. First a search based on a coarse description by superphones is made to narrow the search space, and thereafter a detailed match is performed to select the most probable candidates which are then inserted on the system blackboard. The search- and verify phases will be elaborated in the following sections.

3. ARCHITECTURE

In this section the lexicon architecture will be described. The fundamental design requirements of the lexicon are first stated, keeping in mind that the lexicon is to be used in a continuous speech recognition system. The main goal of the lexicon design is flexibility. Firstly, the vocabulary of the lexicon is easily updated or replaced completely. Secondly, the lexicon is adaptable to different allophone classifiers, i.e. the phonetic representation of the lexicon matches the degree of resolution of the input string. This is important as the allophone classifier is improved continuously. Also, the possibility of a drop in classifier performance has to be taken into consideration, for example in connection to a shift from speaker dependent- to speaker independent allophone classification.

Thirdly, as there is no indication in the allophone input string of the beginning and ending points of the words spoken, a lexicon lookup is independent of knowledge of the word limits, in other words: it is possible to initiate a lookup from a point selected at random in the input string. This is a very important feature in a continuous speech recognition system, as word limits are very hard to determine.

To fulfill these requirements the architecture shown in figure 1 has been developed. The central elements are the lexicon generator program and the lexicon search structures. The elements shown below the dashed line run directly in the recognition process, whereas the elements above are used "off-line". The transcription rules has been developed within the Danish Speech Synthesis Project [6] and will not be discussed here.

The demands for flexibility have led to the development of a generator program, the task of which is to generate the complex data structures corresponding to the lexicon vocabulary.

The generator uses three input sources. The system application vocabulary, a list of symbols used in the phonetic transcriptions together with the set of distinctive features that denote each symbol, and a specification of which superphones the lexicon is to be based upon. The vocabulary is represented partly by the orthography of each word, and partly by a phonetic representation of the word. The phonetic representation typically consists of two or three transcriptions of the word, corresponding to different pronunciations (primarily reduced forms) of the word.

The set of symbols used are the SAMPA set (originating from the Esprit SAM project [7]). The set of superphones forms the basis from which the lexicon structures are generated. In fact, the superphones exclusively determine the organization of the lexicon. To see this, it is necessary to take a close look at how the lexicon is organized.

4. SEARCH AND VERIFY PHASES

At the present stage the phonetic classifier is able to distinguish among four vowel classes and 2 consonant classes (labelled I,E,O,Q and N,S) with a close accuracy. The classes are described by a minimum set of features, as e.g. 1 : [+vocal, +sonorant, +front, -low]. These classes will then denote the superphones. (Please refer to [4] for a discussion of the implications of the choice of the superphones). If a new classifier is later installed in the system, new superphones (e.g. the stops: [p,k, b,d]) can be added, and the lexicon regenerated.

The lexicon is arranged as a number of search trees with superphones at the branching points. There is one tree for each superphone. The principle is best understood through the example shown below.
The tree shown in figure 2 has the root "U", and the branches "Q", "I", "S" and "N" on the first level. It is entered when a segment feature set of which includes the set for the "U" superphone (which is \(+\text{voc}, +\text{back}, +\text{round}\)) is selected in the input string. Only the features that define the given superphone are considered at this stage. When a set of distinctive features for a segment in the input string falls into one of the superphone categories it is marked "super" as shown in figure 2. It can then be used as a base for a lexicon lookup. A lookup is initiated by a random "super" segment in the input string, selected by the expert system control structure. It is then matched against the roots of the lexicon search trees, and the tree it matches is entered.

Then a search based solely on the superphones is made. The lexicon tries to match the neighbouring superphones in the input string against the nodes on the next level of the tree. Any transcriptions listed by a node are considered for verification, as will be explained later. The search is continued until no node matches a segment in the input string. In the example given above, the lexicon will first try to match the node "Q" against the first superphone to the right of the "U", which will fail. ("." indicates to the right, "+" to the left of the preceding superphone). The next match will succeed as the "I" node in the tree matches the "T" segment to the left of the "U". The search is now continued following the "T" branch, and will terminate at the "E" node at the second level. Note that the word [viol'Ed] can be found by two nodes in the tree. The reason for this is, that the feature set for the allophone symbol [E] in the transcription matches both the "I" and the "E" superphone sets. Two feature sets are said to match when all features present in the first set are found in the second set. Any additional features present in the second set are not considered in this situation.

By now the following words are candidates for verification: [p'O], [bl'O], [g'ul] and [viol'Ed]. The objective of the verification process is to compute a score for each proposed word candidate. The score is used by the expert system to select the most probable word candidate. Words scoring below a predefined threshold are not inserted into the expert system blackboard, but are stored in an auxiliary blackboard to be reconsidered in case none of the verified words combines to a phrase. The first step in the verification process is to align the word candidate superphones and the allophone string superphones, and compute a matching score for each superphone. At the next step the score of the remaining non superphones are computed, taking into account the possibility of segmentation errors. A matching score for two feature sets is computed by inspecting each feature separately, and adding up the individual scores. For example, the score for the feature [vocal] is +50 if correctly, and -50 if wrongly determined.

5. IMPLEMENTATION

The lexicon is implemented partly as an element of the speech recognition system, and partly as separate programs, as shown in fig. 2. The lexicon search and verify algorithms are implemented as an expert system rule. The advantages of this is that lexicon activity is controlled through the expert system control mechanisms, and further, that all communication to and from the lexicon goes through the system blackboard and is therefore simple and easy to trace.

6. RESULTS

Lexicon performance is very sensitive to the quality of the phonetic classifier output. This makes it hard to give an accurate quantitative evaluation of the lexicon. Also the lexicon has not yet been tested on a vocabulary large enough to derive any significant statistical results.

Qualitatively, the results of the testing show, as expected, that the performance is almost exclusively determined through the choice of the superphones. When chosen correctly, the lexicon as a rule produces the correct word candidates in a given interval of the speech signal. The correct word is in most cases among the highest scoring candidate, thus easing the task of parsing proposed word candidates into phrases [2].

7. DISCUSSION

The present implementation of the lexicon performs satisfactorily in some aspects, as shown in the preceding section. The principles of organizing the lexicon as a number of search trees defined by superphones, and of operating on sets of distinctive features instead of allophones, makes for a very flexible lexicon architecture. This is a great advantage in the process of development.
ing a total speech recognition system, as future changes in other parts of the system can easily be dealt with. This has made it necessary to introduce a lexicon generator to create the search structures from the system application vocabulary.

The concept of superphones has been introduced to narrow down the number of possible word candidates by a search process. This is a necessity when the vocabulary reaches a size, where a detailed consideration of the total vocabulary becomes improbable. There is a danger in this approach, however. As the superphones form the branching points of the search structures, it is essential that the phonetic classifier is able to identify the superphones in the speech signal with a very close precision. Otherwise the results of the search phase will be wrong. This implies, that the superphones must be chosen with great care and must reflect the phonetic classifier’s effectiveness precisely.

Consequently, the superphones are to be chosen conservatively, which leads to a few broadly described superphones. This is contrary to the wish to define as many and as precise superphones as possible, as this will make the search quicker and preciser.

8. REFERENCES

1. A. Bækgaard et al. :
"The HEAD system and its approach to rule based acoustic-phonetic recognition of speech".

2. A. Bækgaard and P. Dalsgaard :
"A system for continuous speech recognition using Neural Networks and Expert System".

3. T. Kohonen et al. :
"Phonetic Typewriter for Finnish and Japanese".

4. M. Bundgaard. :
"Analysis of Large-scale Lexical Corporuses in the Context of Continuous Speech Recognition Systems".

5. P. Dalsgaard. :
"Semi-automatic phonemic labelling of speech data using neural networks".

6. P. Molbæk Hansen. :

7. A.J. Fourcin et al. :