RECOGNITION OF CONTINUOUS SPEECH USING NEURAL NETS AND EXPERT SYSTEM

Anders Baekgaard, Paul Dalsgaard
Speech Technology Centre
Aalborg University
Fredrik Bajers Vej 7
DK-9220 Aalborg, Denmark

0. ABSTRACT

A system for recognising continuously spoken sentences is presented. The system has a vocabulary of approx. 35 words and a grammar specifying a few thousand sentences. The system operates in three stages. In the first stage, cepstrum vectors are computed in real time and used as input to a self organised neural network. The output of the network is mapped to a continuous valued acoustic phonetic distinctive feature vector for each frame of the speech signal. These vectors are in the second stage processed by a multi layer perceptron which is trained to estimate segment boundaries. The output from this stage is a discrete valued acoustic phonetic distinctive feature for each segment of the speech signal (allophones). The third stage contains an expert system, which processes the allophones using a lexicon and a parsing system.

1. INTRODUCTION

The HEAD system is currently being developed in a joint university - industrial project at the Speech Technology Centre at Aalborg University, Denmark. The system aims at speaker dependent recognition of continuous speech. For the present version of the system a CAD like application has been setup with a vocabulary of approximately 35 words (excluding pronunciation variants) and a grammar specifying a few thousand sentences. A new system will be defined in the near future, in which the application is a traffic inquiry system; it will contain a vocabulary with approximately 300 words. This system will be speaker adaptive or speaker independent.

The present system aims at near real time execution for the lower levels of the recognition process. The system is tuned to telephone quality input, i.e. sampling at 8 kHz and filtered to telephone bandwidth at 300 - 3400 Hz.

A previous version of the HEAD system used a rather traditional input component in which formants, energy, pitch, segmentation etc. were estimated followed by rule based allophone hypothesis. However, practical experiments have demonstrated that this approach was not sufficiently robust, even for a single speaker. In contrast, the emerging technology of simulated neural networks have shown evidence, that the above mentioned preprocessing might be more robust using neural network technique.

2. SYSTEM ARCHITECTURE

An essential concept in the system is phonetic distinctive features. The features are chosen to reflect dimensions in a speech production model. The distinctive features are trinary: a feature may be present or absent, or the feature is undetermined (irrelevant). For example, for the allophones having the feature +voc, the feature fric is irrelevant. Using distinctive feature description of an allophone allows various degrees of precision. If one feature is omitted in a description of an allophone, the description is not incorrect, but inexact. For example, the allophone [i] is described using the features [+voc, +front, -low -mid, +high, -round]; [e] is described as [+voc, +front, -low, +mid, +high, -round]. If the feature mid is omitted, the remaining features describe both the allophones [i] and [e]. Therefore, the philosophy is to define an allophone description model that allows various degrees of precision since an exact description is unrealistic. Figure 1 shows distinctive feature diagrams for Danish allophones (SAMPA notation [4]) used in the system.

The system consists of three major components: the input component, the phonetic decoding component and the expert system component, as shown in figure 2.

Input Component
The input component samples the input signal at 8 KHz and computes autocorrelation, LPC and cepstrum coefficients at 10 kHz.
3. NEURAL NET PROCESSING

During the stimulation/calibration process of the neural network each neuron is assigned a weighted average cepstral vector and a corresponding acoustic phonetic distinctive feature vector in a feature map. This vector is 16 dimensional, and each dimension specifies a value for one of the 16 phonetic features. The value is in the range between -1.0 and 1.0.

During use, the input to the neural network is a cepstrum vector. The distance is measured between this input vector and all neurons in the network, and the 4 neurons with the smallest distance are chosen for further processing.

The 4 neurons have a corresponding vector in the feature map, and these are averaged and weighted according to their distance to the incoming cepstrum vector such that the output from this component is a vector of continuous valued phonetic feature vectors. The details are described in [4].

4. PHONETIC DECODING

The phonetic decoding component processes the phonetic feature vectors described in the previous section. The output of the component is allophones described by 16 discrete valued phonetic distinctive features. These allophones are determined by a) segmentation of the sequence of phonetic feature vectors and b) calculating an average phonetic feature vector from some of the frames within each segment and c) discretising these vectors according to a threshold value.

MLP Segmentation
The segmentation is performed by a multi layer perceptron (MLP) with 160 input neurons, 32 neurons in the first hidden layer, 8 neurons in the second hidden layer, and one output neuron. The input to the MLP is a set of phonetic feature vectors which are contained in a sliding window of width 10 frames. The output neuron is trained to fire when the center of the input window is at a segment boundary. The perceptron is trained using the backpropagation algorithm, and the training material contains approximately 10000 speech signal frames and approximately 1000 segment boundaries known from manual placements by an expert phonetician. Experiments have shown, that it is difficult to estimate segment boundaries corresponding to the phases of stop sounds ([b], [d], [g], [p], [t], [k]), and consequently, the perceptron is trained to treat stop sounds as belonging to one segment. At present, there remains experiments to show whether diphthongs (and triphthongs) are best treated as within single segments or two ( or three) segments. Occasionally the perceptron proposes more than one segment boundary in the region around the correct boundary. Hence, the segment limits are processed by an algorithm that collapses very short segments (1 frame) with the nearest segment.

It is particularly important to stress, that the perceptron operates in a purely phonetic domain, and is independent of speaker, speech signal intensity etc., as long as the phonetic feature vectors presented to the perceptron conform to the speech production model.

Allophone generation.
When segment boundaries are given, the corresponding segment is assigned an allophonic description by

1. Eliminating 20% of the frames of the segment near the limits
2. Computing an average phonetic feature vector from the remaining frames.
3. Discretising the elements of this vector by comparing its continuous value to a threshold:
   - > 0.2: distinctive feature present
   - < -0.2: distinctive feature not present
   - otherwise: distinctive feature undetermined
The effect of dividing the range into three subranges of unequal length is that features tend to be present or absent rather than undetermined. The threshold found by experiments is a compromise which on one side insures that only the most probable features (i.e. with the largest magnitude) are determined, and on the other side, insures that a sufficient number of features are determined to specify an allophone with acceptable precision. Experiments indicate, that each feature may have an individual threshold.

The above described technique for determining allophones has shown successful only for stationary sounds (monophthongs, nasals, fricatives, liquids).

Figure 3 shows details from the above described processing in the so called feature gram. The abscissa axis maps frame numbers, the ordinaty axis maps the individual elements of the phonetic feature vectors. The elements have values which lie in the range from -1.0 to +1.0 (positive values in black, negative hatched). Drawn in the feature gram are long vertical lines showing segment boundaries determined by the MLP. Above the feature gram, a reference segmentation and labelling is shown (SAMPA notation). The figure shows that the stationary sounds labelled as [f], first phase of the diphthong [Au], [i] and [i] are almost exactly described by the feature vectors, whereas the stop sounds and the liquids are not correctly described.

5. EXPERT SYSTEM COMPONENT

The expert system component contains the following major modules: allophone rules, lexicon access (phonological rules), syntax rules and a control structure.

The expert system is built as a black board oriented architecture [5]. The black board is structured into three levels: the allophone level, the word level, and the phrase level [2] corresponding to different types of hypotheses.

Allophone rules

The allophone rules initialises the black board with the allophones generated by the phonetic decoding component. These rules also perform some repairment of the segmentation, eg. collapsing adjacent segments with identical phonetic distinctive features.

The recognition rate for the individual allophones is very uneven distributed on the individual allophones. For example, the fricatives have a high recognition rate, while the liquids have a relatively low recognition rate. For this reason, the concept of super phones is introduced: a super phone is an allophone that possesses certain combinations of distinctive features that are recognised with high reliability [6]. The super phones are chosen to cover several allophones, eg. the super allophone "S" is described as [-voc, +cons, +fric, -plos] and covers the allophones [f], [s] and [S]. The super phone "I" covers the allophones [i] and [e] and [E].

When an allophone hypothesis is inserted on the black board, a rule decides if the allophone possesses the features of a super phone and marks the hypotheses accordingly. It should be noted that the precision of an allophone hypothesis is not reduced when it is marked as super phone.

Lexicon Access

The lexicon contains for each word in the application grammar several entries describing expected transcription (pronunciation) variants of the word (typically 2-4). Each transcription is specified as a sequence of allophones, each described by phonetic distinctive features. The lexicon is organised as a set of trees, one tree for each super phone. The nodes of the trees specifies a super phone, a direction to the left or to the right of the root of the tree, and set of word entries (if any). The lexicon is accessed in the following manner: a super phone hypothesis is selected from the black board and the corresponding lexicon tree is picked. The tree is traversed by, for each node investigating if the black board contains the super phone specified in the node in the position
adjacent to the previous found super phone (the term adjacent accounts for super phone only, non-supers are ignored at this time). Whenever the super phone specified in the node (including the root) is found on the black board, the word entries at this node is saved for later verification. The verification includes the following steps: a match score for the super phones of the word is computed (despite the super phones are correct, the corresponding allophones may not be). The allophones between the super phones of the word is matched up against the allophone hypotheses, and a score is computed. The total score is used in the weight calculation for the word hypothesis.

The tree structures are computed only once for a particular application vocabulary by an offline generator program. For details, see [3].

Parser

Parsing in a speech recognition system is, compared to traditional parsing systems, a much more difficult task since a) input is not deterministic and b) input is not presented in a left to right manner. The parsing system described here handles these problems by a) allowing alternative parses for a given input and b) parsing is done where ever input is present (island driven). The parser operates in a black board oriented environment where word hypotheses may "pop up" at any place.

The application grammar is specified using an extended BNF notation, and is context free. No restrictions are imposed. During system initialisation, the application grammar is read by the parser, and a series of transformations are performed resulting in

![Diagram of Parser Architecture](image)

Fig. 4. Parser Architecture

a data structure, grammatical rules, figure 4.

The parser can operate in a bottom-up manner. When a word hypothesis or a phrase (production) hypothesis is inserted on the black board, the control structure invokes the parser. It inspects the hypothesis (denoted the trigger), and by examining the grammatical rules (productions), it finds a set of productions in which the trigger may occur. For each of these productions, predictive parsing is used, i.e. the black board is searched for the symbol (word hypothesis or production hypothesis) adjacent to the right of the trigger, then the symbol to the left of the trigger etc. The parser allows an overlap or a gap of 30 msec for adjacent hypotheses. If the search fails, an entry is made on the parse board describing the actual context (i.e. production, timings, left and right positions in the production), and the symbol (including timing) for which the search failed. The parsing may be resumed at a later time by the following mechanism: when the control structure invokes the parser with a trigger, the parse board is scanned for entries containing the trigger (allowing 30 msec for overlap or gap for the start time or end time).

Although the parser operates in bottom-up and in a predictive manner, top down word hypothesation is not implemented yet.

Control structure

The control structure guides the recognition process. It is organised as a meta black board containing plans and a set of planning rules. It makes long term plans (such as: find phrase hypothesis that covers the entire utterance) and short term plans (such as: find word hypothesis in a given time interval). Rules handle long term plans by decomposing them into shorter term plans.

6. CONCLUSION

It is believed that the system presented in this paper has some novel characteristics: 1) The use of continuous valued phonetic features which allow various degrees of precision in the description of allophones. 2) Segmentation by use of a multi layer perceptron which operates in a pure phonetic domain. 3) The use of super phones which are arch phones that are recognised with high reliability.

There is not yet any confident performance figures due to lack of sufficient training material, but in near future more training material will be established. At present the system is capable of recognising several sentences in a test corpus of 84 sentences, and it is believed that when the system matures, good performance will be achieved.

REFERENCES