THE HEAD SYSTEM AND ITS APPROACH TO RULE BASED ACOUSTIC-PHONETIC RECOGNITION OF SPEECH

Anders Bækgaard*, Paul Dalsgaard* and Peter Holtse**

0. ABSTRACT

This paper presents an approach towards using acoustic-phonetic knowledge in Automatic Speech Recognition (ASR). Signal processing hardware and algorithms are described, and a method for the representation and application of phonetic, phonological, and linguistic knowledge is outlined. Finally, the HEAD system, in which these algorithms are implemented, is presented.

1. INTRODUCTION

One approach to ASR is based on the Hearsay II (ref 1) architecture which utilises a number of phonetic and linguistic knowledge sources in its recognition process. The fundamental ideas behind the Hearsay II architecture has been used with a Danish speech project, which is aiming for industrial applications implemented into a standard microprocessor system to demonstrate a continuous speech, speaker dependent recognition system (ref 4).

In its first version the system is implemented in a simple CAD demonstration application and is capable of recognising a limited vocabulary (currently about 24 words used with 84 sentences) in a very restricted grammar. However, the flexibility of the ASR system allows both vocabulary and grammar to be extended.

2. THE HEAD SYSTEM

2.1. ARCHITECTURE

The HEAD - Heuristic Experimental Analysis and Development - system is designed in an architecture similar to Hearsay II. Within the system the recognition process can be viewed as consisting of two consecutive phases: Signal processing, involving segmentation, formant extraction and pitch estimation, and knowledge based analysis, involving a series of bottom-up hypothesis generations as well as top-down hypothesis verifications on the basis of general phonetic and linguistic knowledge.

Hypotheses are organised in several layers in a Blackboard, where each hypothesis is a data structure containing information about a certain time interval at the given level of abstraction. The hypotheses are manipulated by a rule base which again is governed by a Control Structure. The whole system is implemented using an expert system shell developed especially for this purpose (ref 3). It is implemented in a standard microprocessor system so as to allow easy interfacing to industrial applications. The currently implemented system uses four abstraction levels corresponding roughly to the linguistic description levels of (acoustic) segment, allophone, word (or morpheme), and phrase level (cf. figure 1). As shown rules may, at each level both generate and verify hypotheses by analysing input at that level.

In general, a rule is stimulated, or invoked, by the generation of a hypothesis or the arisen need for verification, i.e. whenever a hypothesis is generated or needs to be verified an automatic scan is performed through the rule base as a cause of which one or more rules may be stimulated and one rule selected for invocation.

*Speech Technology Centre, Institute of Electronic Systems, Strandvejen 19, DK-9000 Aalborg, Denmark
**Institute of Phonetics, University of Copenhagen, Njalsgade 96, DK-2300 Copenhagen S, Denmark
The exact conditions under which a given rule may be stimulated is denoted the Stimulus Frame of the rule. The Stimulus Frame consists of a structural description of a kernel of hypotheses to be modified and of the environmental conditions to be satisfied for the rule to apply.

Every rule has a dynamically evaluated weight, which functions as a priority indicator for that rule. This priority is used to solve possible rule conflicts. The expression denoting the weight of the rule may be of any complexity, and it may refer to data in the hypothesis stimulating the rule, even before it is invoked. The weight of a rule is an indication of the importance or credibility attached to the rule with a certain stimulating hypothesis. Thus, rules with a high weight will be selected earlier for invocation by the Control Structure than more dubious rules with a lower weight.

Correspondingly, every hypothesis has a weight attached to it as an indication of the credibility of the hypothesis.

Correspondingly, every hypothesis has a weight attached to it as an indication of the credibility of the hypothesis.

![Figure 1. The organisation of the Blackboard in HEAD](image)

2.2. EXPERIMENTAL TOOLS

The HEAD system is not merely a recognition system but a development system used in creating a recognition system. Thus, it incorporates a set of useful facilities for conducting experiments with both signal processing and logical rules. A major design principle behind the HEAD system has been the visualisation of the data present in the system, e.g. speech signals, formants, hypotheses on the Blackboard, rule application tree etc.

3. SPEECH RECOGNITION USING THE HEAD SYSTEM

The HEAD system was developed as part of a Danish speech recognition project involving partners from industry as well as universities. In the following the details of the signal processing and the knowledge bases of the current systems will be described.

3.1. SIGNAL PROCESSING HARDWARE AND ALGORITHMS

The signal processing part of the HEAD system involves primarily segmentation of the input speech signal, formant extraction, and pitch estimation. To ensure real time operation a Speech Acquisition and Processing Card - SAPCA - has been designed (ref 2). The SAPCA utilises three onboard signal processors. The two signal processors - which are
piggy back mounted - will be replaced by 32 bit floating point processors in order to perform the computationally heavier and more demanding tasks of real time segmentation and parameter extraction.

3.2. KNOWLEDGE BASE

The rule based part of the recognition system contains the phonetic, phonological and linguistic rules in the blackboard model described earlier. The system also incorporates a lexicon used for word hypothesising and verification.

3.2.1. HYPOTHESES

Hypotheses are described primarily in terms of Distinctive Features, although segment level hypotheses also contain certain parametric information (e.g. formants, energy etc.). The Distinctive Features are in principle binary, i.e. plus or minus (present or absent). However, a given feature may be left undefined, which is interpreted as a condition which can neither stimulate a rule nor block the application of a rule.

At segment level hypotheses are described by the features like voiced, hiss, etc, and by the parameters formants and energy in frequency bands. At allophone level the features are of the type consonantal, sonorant, stop, compact, grave, syllabic, stress, etc. At word level hypotheses are described by an index into the lexicon which gives access to, among other things, the syntactic category of the word.

3.2.2. RULE TYPES

The four levels of rules of the HEAD system operate as follows:

Segment Rules generate or verify segment hypotheses by analysing the parametric descriptions resulting from the signal processing. The initial set of segment hypotheses are created by a segmentation preprocessor. This set is gradually modified and refined by the segment rules. Also included in the preprocessing is a set of prosody rules which, based primarily upon pitch estimation parameters, attempt to estimate the number and location of syllabic nuclei and to assign appropriate stress markers to these nuclei. The output from these prosody rules is crucial to the recognition strategy employed at higher levels.

The segment hypotheses invoke Allophone Rules which will attempt to combine or split adjacent segments to allophone hypotheses at the blackboard level immediately above.

The Phonological Rules attempt to identify possible words in the input sequence by referring to a special lexicon. The lexical search is concentrated round the stressed syllables as being the linguistically most important places in the utterance as well as the units least susceptible to articulatory reduction.

At the highest level the word hypotheses are combined by Syntax Rules to form syntactically acceptable phrases. All syntactically possible hypotheses are fed as output to the application interface, which is logically not part of the recognition system. Since information is built up in islands around the stressed syllables the grammar needed at phrase level cannot work in a simple left to right fashion as is usual with conventional parsers. To solve this problem a special parsing technique is being tried out which works in a sort of mixture of induction and deduction.

3.2.3. LEXICON

A central part of the recognition system is the Lexicon in which are listed all words or
morphs which the system is capable of recognizing. Each entry in the Lexicon consists of four fields: 1) orthographic text, 2) syntactic properties, 3) a unique identity marker, and 4) a list of possible pronunciations of the entry.

Since the grammar of the recognition system is currently quite simple the syntactic information is correspondingly reduced and consists mainly of flags of the type Verb, Noun etc.

The pronunciation field contains two types of phonetic transcription: A relatively detailed notation (known as DTL - Danish Technical Transcription), which is basically allophonic in traditional phonetic terms, and a very coarse notation, which distinguishes only the most important phonetic categories. The symbols distinguished in the latter notation are known as Archiphones. The list of archiphones is used in the primary search for word candidates, thus avoiding the need for a very detailed acoustic-phonetic analysis before the lexical search can be initiated.

In the HEAD system the Lexicon is actually implemented as a separate Blackboard with all the search and backtracking facilities associated with a Blackboard. This strategy will allow restrictions on the Lexical search space, such as grammatical or environmental conditions, to be implemented in a straightforward fashion as experience with the Lexicon grows.

4. CAD DEMONSTRATION APPLICATION

The HEAD system will as its first application demonstrate a simplified CAD/CAM approach in which simple geometric shapes are moved about on a screen. The linguistic framework for this demonstration has been carefully designed so as to provide some material with a high degree of phonetic contrast as well as a number of words which can be expected to be easily confused. Currently the demonstration application uses speech signals stored in files. The results shows correct recognition for some of the chosen sentences up to word level. Also the system has demonstrated the power of the HEAD tools.

5. FUTURE WORK

The results gained so far are encouraging. Future work on the HEAD system will go in two directions. First of all, work will be continued along the lines outlined in this paper. Secondly, the execution speed of the knowledge based techniques will be increased by transferring the various levels of the Blackboard system into a number of parallel processing microcomputers, e.g. Transputers.

REFERENCES

4/ Bekgaard, Anders, Michael Bundgaard: "HEAD", Internal work document in the Danish Speech Recognition project