AUTOMATIC SPEECH RECOGNITION USING ARTIFICIAL INTELLIGENCE METHODS

V. Walsh*, H.R. Taylor*

ABSTRACT
An expert system has been created to realize a microcomputer-based speech analysis and speech recognition system which uses a low-cost analog filter bank for extracting the acoustic features of speech. The expert system is used to create high-level speech knowledge and offers knowledge-based speech processing. Two speaker-dependent recognition methods were tested with each method resulting in average recognition scores of 95% on a vocabulary size of ten words and branching factor of ten. The system is in an early stage of development and has the potential for improved recognition accuracy and applications in continuous speech recognition.

INTRODUCTION
The research methods described in this paper use neither Dynamic Time Warping (DTW) nor complex statistical or mathematical modelling techniques. The methods rely, for their effectiveness, on the ability of the researcher to devise, and implement in software, sets of rules, pattern identification and matching routines that had previously been tested on hard-copy printouts of digitized speech signals. The software used for speech analysis and speech recognition is, in effect, an expert system which models the reasoning, pattern-matching, controlling and procedural functions of a human researcher to extract, control and process the acoustic parameters of speech. All processing is based upon a digital representation of an input speech signal derived from a speech data acquisition unit (SDAU) using a 15-channel analog filter bank unit. The software is described in terms of levels according to the nature of the processing task.

LEVEL 1 SOFTWARE: FRAME SEGMENTATION AND PEAK PATTERN DETECTION
The system is able to reliably detect the onset of speech and the end of speech and performance is relatively tolerant of high ambient noise levels. Speech onset is determined by polling for threshold levels at the output of any one of the 15 filter channels of the SDAU. Level 1 software is concerned, firstly, with the segmentation of the subsequently-produced frames of channel levels. A frame is the fundamental processing unit on which all subsequent processing is based. A frame consists of the outputs from each of the 15 channels with new frames being obtained at 10ms intervals.

For most 1-2 syllable words, 3 segments are created from the block of frames which constitutes the digitized speech. The start segment, ASEG, describes the build-up of levels from the analog filters; depending upon the word, its length can vary from 1 frame (i.e. 10ms) to 20 frames. The mid segment, BSEG, usually contains the most frames. The end segment, CSEG, describes the attenuation of levels which occurs at the end of a word. For polysyllabic words, CSEG describes the transition from one set of phonemes to a further set. 6 segments or 9 segments, modelled on the ASEG, BSEG, CSEG rules, are created to apportion the speech frames into useful segments for further processing. For continuous speech, clusters of such tri-segments can be used for speech of any length.

* Department of Electrical Engineering and Electronics
UKIST PO BOX 88 MANCHESTER
On completion of segmentation, the remaining task of Level 1 software is peak detection. Peaks are detected whenever the level in any one channel in the frame is greater than the levels in the adjacent upper and lower channels. These peaks constitute formant information. The pattern which is coded as 3369, for example, denotes peaks in channels 3, 6 and 9 with the major, dominant peak being recorded in channel 3. This pattern is characteristic of the phoneme /ei/ in the words 'nine' and 'five'. The lower the channel number, the lower the frequency covered by the filter in that channel.

**LEVEL 2 SOFTWARE: CREATION OF SYMBOLIC SPEECH KNOWLEDGE (ASEG, CSEG)**

Selected frames from ASEG and CSEG are processed by Level 2 software to extract acoustic features and assign identifying labels to extracted features. The segments are tested for the following features: low frequency dominant peaks (LFD), low frequency peaks, alone (LFA), high frequency dominant peaks (HFD), high frequency peaks, alone (HFA) and various transitions of peaks (LLT, HHT, HLT, LHT). One byte is used to represent this information. Tests are also made for high frequency unvoiced frames (UVA, UVC), frication (FRICA, FRICC) and nasality (ANASAL, CNASAL). The durations of ASEG and CSEG are also stored along with ratios obtained from blocks of channel levels. Selected patterns from ASEG and CSEG (selected to reduce processing overheads and information redundancy) are analyzed and the patterns from these frames are stored.

The significance of using labels to identify acoustic and phonemic features, or coding frequency information as peak patterns, is threefold. Firstly, knowledge about the input speech is now contained in the symbolic form of labels which are readily amenable to processing by any high level language. Each of the two recognition methods described below uses the flexibility, power and ease-of-use of a high level language to perform the recognition task. Hence, these recognition methods can be described as knowledge-based speech processing as distinct, for example, from DTW approaches which make little, if any, use of higher level knowledge from phonetics. Secondly, manipulation of this symbolic knowledge can be of great benefit when sorting groups of words within a larger vocabulary set so as to reduce the processing task. Thirdly, a considerable reduction has been achieved in terms of the number of bytes needed to represent an utterance.

**LEVEL 3 SOFTWARE: CREATION OF SYMBOLIC SPEECH KNOWLEDGE (BSEG)**

ASEG and CSEG contain relatively few frames in most cases. Most of the detailed frequency information for an utterance is consequently enclosed within the BSEG boundaries. To reduce the processing load, every alternate frame in BSEG is processed by Level 1 software to obtain a list of peak patterns. In analysis mode only, Level 3 software processes multiple examples of training words to obtain patterns which are stored in a pattern store. A new pattern, i.e. a previously unrecorded pattern, is stored as it is encountered. This pattern store contains all the recorded pattern variations from the selected group of speakers contributing to a vocabulary database.

The expert system calculates the occurrence frequency of any patterns for different productions of the same word. This incidence factor (I.F.) analysis is used to assign weighting scores to properties and patterns so aiding the researcher in identifying reliable speech parameters. Also, in the analysis mode, Level 3 software derives minimum, maximum and mean values for ASEG and CSEG ratios and durations. ASEG and CSEG properties (ALFD etc) are also treated to I.F. analysis. The system operates here as a self-creating database and spreadsheet.
LEVEL 4 SOFTWARE: RULE CREATION AND AUTOMATIC RULE VALIDATION

Automatic rule validation by the expert system is a process whereby data is extracted so that rules can be tested. Level 4 software is concerned with extracting the properties and data which are necessary for the operation of the recognition rules. The rule 'if P then Q', for example, is created by the programmer. The expert system can validate the rule by checking for instances of P. By an iterative process of rule creation and validation, with amendments made to the rules according to the success of the validation process, sets of rules can be speedily developed in training/analysis mode which are reliable and robust.

LEVEL 5 SOFTWARE: RECOGNITION MODE

In recognition mode, use is made of the routines from lower software levels to extract properties and patterns from the input word, where the input word is from either a microphone, cassette player or a digital file from either of the former sources. To be recognized as a particular word, an utterance must comply with the rules which were established for that word during the analysis mode. Two recognition methods were employed.

Method 1: Speaker-Dependent, Hand-Crafting of Rule Validation Data.
Five training samples of each word in the recognition vocabulary are analyzed and a print-out obtained of the results of analysis. An exhaustive visual analysis is then performed by the researcher to detect reliable and robust properties and patterns. The rules and validation data which are formulated from this human processing are then written (at the computer keyboard) into the appropriate part of the recognition program. These rules are an attempt to encompass within-speaker variability. The rules are a mixture of sorting rules, feature rules and pattern rules. Sorting rules test for properties which will reduce the search space. Feature rules are concerned with high level speech knowledge: the rule set for the word 'six', for example, contains a requirement that there be a high frequency start to ASR if an input word is to be considered as the word 'six'. Pattern rules increment various counters with selected patterns for a word (identified from analysis) having a unique counter. This recognition method uses a depth-first search strategy so that, after sorting has reduced the likely candidates, the rule sets for the selected words are fully tested until either a necessary condition fails to test true or until the rule set tests true and the word for that rule set is chosen, so ending the recognition process.

Method 2: Speaker-Dependent, Automatic Storing of Rule Validation Data.
With this method, five training samples of each word in the recognition vocabulary are analyzed to obtain within-speaker variability concerning patterns and properties. Key patterns, phonetic information and Incidence Factor data are then stored to magnetic disc. A knowledge store is created, thereby, for each word in the vocabulary for use in the recognition mode. In the recognition mode, rules are employed which process weighted scores attached to patterns and other phonetic information extracted from the input word to be recognized and matched against the rule set for a word. A recognition score is therefore acquired for the sets of rules selected by the sorting logic. The highest score from amongst the selected possible words in the vocabulary represents the chosen word. The training and recognition system is fully automatic in the sense that manual entry of validation data into the program, as in Method 1, when changing from analysis mode to recognition mode, is not necessary.

RESULTS

The results from the trials are shown in Table 1. In all cases, the recognition vocabulary consisted of the digits zero and 1-9. A branching factor of ten was used. Trial utterances and test utterances were made in
an office environment using a table-mounted, low-cost unidirectional microphone. Speakers were asked to speak at a distance of around four inches (100mm) from the microphone.

**TABLE 1**

<table>
<thead>
<tr>
<th>SUBJECT</th>
<th>SCORE (max=50)</th>
<th>%SCORE</th>
<th>SUBSTITUTIONS</th>
<th>REJECTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. VW ENGLISH M</td>
<td>50</td>
<td>100</td>
<td>nil</td>
<td>nil</td>
</tr>
<tr>
<td>2. WH USA M</td>
<td>49</td>
<td>98</td>
<td>5/9</td>
<td>nil</td>
</tr>
<tr>
<td>3. AH ENGLISH M</td>
<td>46</td>
<td>92</td>
<td>8/3, 7/6, 7/1/9</td>
<td>nil</td>
</tr>
<tr>
<td>4. RT ENGLISH F</td>
<td>45</td>
<td>90</td>
<td>2/7, 5/9, 2/0</td>
<td>8 (2)</td>
</tr>
</tbody>
</table>

**METHOD 1 RESULTS : AVERAGE SCORE = 95%**

<table>
<thead>
<tr>
<th>SUBJECT</th>
<th>SCORE (max=50)</th>
<th>%SCORE</th>
<th>SUBSTITUTIONS</th>
<th>REJECTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. VW ENGLISH M</td>
<td>49</td>
<td>98</td>
<td>4/1</td>
<td>nil</td>
</tr>
<tr>
<td>2. WH SCOTTISH M</td>
<td>49</td>
<td>98</td>
<td>1/0</td>
<td>nil</td>
</tr>
<tr>
<td>3. DG ENGLISH M</td>
<td>45</td>
<td>90</td>
<td>9/0,7/0,9/5 (3)</td>
<td>nil</td>
</tr>
<tr>
<td>4. RT ENGLISH F</td>
<td>46</td>
<td>92</td>
<td>3/0 (2), 1/5, 0/3</td>
<td>nil</td>
</tr>
</tbody>
</table>

**METHOD 2 RESULTS : AVERAGE SCORE = 95%**

**CONCLUSIONS**

Although both methods performed equally well, method 1 was very time-consuming. Method 2 was fully automatic and seems to be best suited for further development. One advantage of the system is its low cost: the SDAU hardware can easily be accommodated on a single extended card and discrete components and board could be bought for less than £60. Another advantage is the ability of the system, due to its nature as described above, to move into continuous speech recognition using similar methods as used in the whole word recognition: a start has already been made in reliably labelling phonemes as a step in this direction. The knowledge based approach would, we believe, interface smoothly to the syntactic and semantic processing involved in continuous speech recognition. All analysis and recognition programs use less than 24Kbytes of RAM for both the program and variables for a recognition vocabulary of ten.

The main disadvantages of the system are concerned with processing time and accuracy. We believe that accuracy can definitely be improved: Method 2 has only been devised for six days at the time of writing and analysis of the errors has shown that adjusting the weighting factors could reduce such errors. In addition, some of the knowledge used in Method 1 could be incorporated into Method 2 to further enhance accuracy. Our experience has shown, however, that improved software and hardware is no substitute for clear delivery of words or having richly-modulated speech: monotonous or indistinct speech is as difficult to decode by machine as it is by ear.

The computer used is a BBC microcomputer, based on an 8 bit microprocessor with a 2MHz clock. The recognition response time can vary from around 4 seconds to as much as 30 seconds. This slow response is partly due to the limitations of the computer and partly to the programming language. Many routines are performed in 6502 assembly language but the high level controlling language is BASIC which is, of course, an interpreted language. With a 16 bit machine on a 6MHz or 10MHz clock, using compiled code or a non-interpreted high level language, we believe that a real-time response could be achieved.

The hardware and the software described in this paper has been designed, built, tested and developed by the authors and thanks are due to the U.K. Science and Engineering Research Council for supporting this research.