516

[redacted]

MULTI-USER REAL TIME WORD RECOGNITION SYSTEM

S. S. VIGLIONE

SPEECH RECOGNITION GROUP
INTERSTATE ELECTRONICS CORPORATION
ANAHEIM, CALIFORNIA

Interstate Electronics Corporation is presently marketing a discrete word recognition system to be used as a voice data entry terminal and capable of handling one to four users with vocabularies of 250 word per user. The recognizer is an acoustic pattern classifier that produces a digital code as an output in response to the received utterance. It consists of a spectrum analyzer, an analog multiplexer and A/D converter, a programmed digital processor, a reference pattern memory, and an output register.

The spectrum analyzer divides the input audio spectrum into 16 frequency bands that cover the useful frequency range. By means of parallel detection and lowpass filtering the resulting 16 analog signals represent a power spectrum that constitutes the feature for speech classification. These 16 continuous signals are multiplexed, sampled at 200 Hz, and converted to digital form with 8-bit precision. Thus, the original utterance arrives at the digital processor as a string of 8-bit binary numbers.

The coding compressor compensates for changes in the rate of articulation and reduces the spectral data generated by each utterance to a fixed-length code for the classifier. It reduces every word, regardless of length, to a 240-bit pattern. As a result, the fixed-length codes can be processed in real time by simple pattern recognition techniques without the need for a great deal of high speed memory. The compression algorithm is essentially an arithmetic process, preserving all the properties that change during an utterance and eliminating those that remain steady.

The word boundary detector serves to establish the start and end of each utterance for the compressor by means of experimentally determined criteria. During the training or adaptation phase of system operation, a number of utterances of each vocabulary word are elicited from the user. The estimator compensates for variations among these utterances to form a single, 240-bit reference pattern that is stored in memory to represent a particular vocabulary word. These 240 bits represent both the tendencies that are common to the five utterances and the small variations that are inevitable from utterance to utterance.

After the system has been trained to a particular user, each new pattern from the coding compressor is compared with a syntactically determined subset of all the previously learned reference patterns in memory. The classification process matches the patterns bit by bit via a Hamming-distance classifier.

The system, configured with a NOVA 3/12 with 32K words of core, is capable of supporting up to four simultaneous voice input channels in conjunction with a variety of standard minicomputer peripherals.

The heart of Interstate's voice data entry system is the control program that organizes the system so that it meets a particular application need. This specification of the control program is done in a high level language that permits users to write their own application software, or to modify control programs delivered with the system.

Using the VOICE (Voice Oriented in Core Executive) software operating system supplied as an integral component of the voice data entry system, the user can specify such applications specific system parameters as:

Configuration parameters, including the vocabulary size, number of users, configuration of input and output devices, and the number and size of internal buffers and data arrays.

The dictionary of vocabulary items to be utilized in an application, along with multiplicity of representations for each vocabulary item.

Dictionary of prompt and error messages. These messages can be displayed for the operator as a guide through a complex data entry sequence. Error messages can be used as a key to enable error correction immediately at the data source.

An action structure associating an appropriate system action with each command that is recognized. Actions may range from simply outputting a code associated with a recognized word to executing a complex computer program that is a function of several previously input commands.

A syntax structure that associates subsets of the dictionary with specific functions to be performed in the application. The syntax structure provides a context for the user, and permits the use of large vocabularies without loss of recognition accuracy.

Interstate Electronics Corporation has under development an advanced word recognition system capable of handling up to eight users simultaneously. A common speech pre-processor will multiplex and condition the inputs from each station. The heart of the preprocessor is a single board array processor programmed via firmware to perform Hamming weighting of the speech data and an FFT spectral analysis from 80Hz to 8000Hz. The FFT output is processed to detect the peaks in the first four formant bands every 12.5ms. The energy in the spectral bands is amplitude normalized, and the detected formant energies are processed to provide 16 time normalized samples for each utterance following word onset and word ending detection. In addition to the peaks in the four formant bands, three broad energy measures corresponding to the energies in F1, F2 and F4 are also computed along with the gross energy in the utterance. The sixteen time normalized samples of these eight parameters form the pattern vector for classification. The classification is implemented with a "minimum distance" classifier, where, for computational simplicity, the vector components, rather than the vector itself, is used in computing the distance metric. During training a threshold is established for each pattern which permits the generation of multiple templates per word if required.

This system uses a remote user subsystem with two 20-α/N character displays for operating prompting and message verification. The user subsystem also contains operator controls for train, one-word selectable train, one-pass retrain as well as test and operate functions. The system is configured with a controlling minicomputer and floppy disc operating under DOS for control and application programs.

Laboratory work is underway to extend the discrete word recognition system to handle word strings and phrases; as well as generalized, small vocabulary word recognition.

DISCUSSION

S. S. Viglione

Q: Dave Hadden: As I understand the VDET and the Voice software were Scope products and I'm sort of curious as to the relationship in design of the VDET relative to the Army WRS.

A: The early version of the VDETS system that was developed by Scope was continued under Army contract in the implementation of the WRS. There are some changes that have occurred during the last year which are not incorporated in the WRS, particularly in the coding and compression algorithm. The Army version is a disk operating system. The VDETS as a stand alone system is a Link Tape operating system. The modified algorithm encodes the 16 filter outputs into a 120 bit pattern, then takes the one's compliment to form a 240 bit pattern for classification. The speech input data sampling rate has been increased from 100 to 200 SPS, and the training algorithm now uses a variable number of input samples as opposed to a fixed training sample size. None the less the VDETS is essentially the same as the WRS. The changes incorporated in the algorithms this past year have been directed to, and have accomplished, a significant improvement in classification performance.

Q: John Allen: Can you explain why you do the one complementing algorithm? It seems that you double the amount of data and that its redundant.

A: The 'ones' compliment is implemented to aid in the correlation scheme used for word classification. The 120 bit binary pattern, representing the incoming word is inverted to create a second 120 bit pattern which is the 'ones' compliment of the original. The resulting 240 bit pattern is then "ANDED" with the reference patterns for each word. As a result of the ANDing operation, the first 120 bits of each reference pattern will be 'one' bits if and only if each bit was consistently a 'one' bit for all training samples. Bits 121 through 240 will be 'one' bits if and only if each bit was consistently a "zero" bit for all training passes. The total number of 'one' bits form the basis for classification.