AN IMELDA BASED VOICE RECOGNITION SYSTEM: A STEP TOWARDS EFFECTIVE VOICE RECOGNITION FOR PERSONS WITH SEVERE DISABILITIES

Gary E. Birch[1], Dariusz A. Zwierzynski[1], Claude Lefebvre[1], and David Starks[2]

[1]The Neil Squire Foundation
Research and Development Division
4381 Gallant Avenue
North Vancouver, B.C.
Canada
V7G 1L1

[2]Canadian Marconi Company
415 Legget Drive
P.O. Box 13330
Kanata, Ontario
K2K-2B2

1. Introduction

The Neil Squire Foundation has worked individually with well over 3000 adults with severe physical disabilities in assisting them to effectively use computer based technical aids. The Foundation recognizes the significant potential voice recognition represents in solving many human/machine interface problems and therefore, researchers in the Foundation have endeavoured to keep abreast of voice recognition technology. For many severely disabled persons, the ability to use spoken commands to control their environment is a very enticing concept. This interface capability can greatly enhance their ability to pursue a career, obtain an education, or enjoy recreational activities.

2. Background

Through interaction with the Foundation's clients we have had experience with various commercially available voice recognition systems, ranging from medium to high end systems, in various applications. As attractive as voice recognition first appears, the Foundation currently, except in special circumstances, discourages its application. We estimate that we have had well over 200 hours of direct individual involvement with severely disabled persons who were utilizing speech input. Through this experience we have encountered several limitations with currently available systems which have resulted in simply too much frustration for a majority of our clients. Primarily, these include lack of robustness to background noise, change in speaker emotion and speaker fatigue as well as generally complicated and unfriendly user training/application software. Researchers from the University of Tennessee, Centre of Excellence for Computer Applications [1], came to similar conclusions about the problems besetting speech input for persons with severe disabilities. Interestingly, one of the most significant problems cited was the lack of user friendly support software packages.

This experience lead the Foundation to become involved in the development of an IMELDA based voice recognition system that would potentially overcome these limitations. This system is being jointly developed by the Speech Research Centre of the National Research Council of Canada, the Canadian Marconi Company and the Neil Squire Foundation. The theoretical premises and the functional design of the recognition system were originally developed at the NRC. We are also developing a P.C. based user friendly training software package designed to work with the recogniser hardware.

3. IMELDA

Our speech recogniser represents a hybrid system in that it employs a discriminant network in the front-end component and dynamic time warping technique in the back-end component. The front-end discriminant network first extracts the invariant acoustic features from the speech signal. It then passes them on to the back-end component where they are compared with the stored templates through the DTW technique for the final recognition.

The front-end processing is a conventional fast-fourier-transform-based spectral filter-bank analysis followed by a linear transformation. The linear transformation developed and tested at the NRC laboratory is a linear discriminant network [2]. It is very efficient at extracting invariant features for reliable speech recognition, especially when the speech is distorted or spoken in high background noise. The linear discriminant network was called IMELDA, which stands for Integrated MEL-scale linear Discriminant Analysis because it combines various mel-scale spectral representations into a single set of discriminant functions. The mel scale is a frequency scale of the human ear. The computation of the linear discriminant transformations involves the between-class with the within-class covariance information of the spectral features [3].

The original idea behind using LDA to derive a robust linear discriminant transform for speech recognition applications in the NRC Speech Research Centre dates back to a paper written by Dr. Melvyn J. Hunt in 1979 [4]. There, he suggested that the distance measures in speech recognition should be based on within-class variances in speech sounds.

A within-class matrix (W) is derived by non-linearly time-aligning the individual examples of words to their corresponding averaged word models (templates). Two main assumptions underlie the computation of the W matrix. The first is that the aligned frame pairs belong to the same class. This assumption is made possible by non-linearly time-aligning two tokens, differing in their temporal structure, which maximises their phonetic similarity. The second assumption is that template frames represent class centroids, and that the parameters in the corresponding frames of the individual examples of the word in question are distributed about the template values according to a multivariate Gaussian distribution with equal variance in all directions. This implies a spherical symmetry of the probability density functions and hence it is possible to use Euclidean squared distances. Even though it is not possible to consider each template to be a separate class, the estimation for the between-class matrix can be effected by computing the covariance over all the parameters in the frames of all the templates in the vocabulary.

The end result of the IMELDA computation is a reduced set of orthogonal discriminant vectors which focus on the features which most effectively differentiate between speech sounds. In summary, the linear transformation of the output of the mel-scale filter-bank is achieved by multiplying a matrix of discriminant functions by the vector of log energies contained in a frame of the filter-bank output. The transformed feature vector is then used with Euclidean distance calculations to determine the optimal discrimination between template frames.

To test the mainframe implementation of the IMELDA recognition system, we compared its performance to two well respected commercial systems in a digit string recognition test where we voluntarily distorted the speech signal to simulate speech distortions found in noisy environments. In one case, we added white noise to the speech signal such that the SNR was set approximately to 15dB. In the other case, we changed the spectral balance ('tilt') of the speech signal by 6 dB/octave, by processing speech through a pre-emphasis filter. The IMELDA system demonstrated a profound advantage in the tests with distorted speech and it was also superior in speaker independent tests for both clean and distorted speech. [3] (see Table 1).

Male Continuous Word Recognition Percentage Errors (%)

Representation	Speaker-Dependant			Speaker-Independent		
	Quiet	Noise	Tilt	Quiet	Noise	Tilt
IMELDA	0.2	0.6	0.1	1.8	4.1	1.4
commercial.1	1.2	1.2	23.6	9.3	12.0	50.7
commercial.2	N/A	N/A	N/A	N/A	N/A	N/A

Male Isolated Word Recognition Percentage Errors (%)

IMELDA	0.1	0.5	0.0	0.3	1.9	0.4
commercial.1	0.4	1.3	20.0	5.0	10.0	45.2
commercial.2	0.7	16.4	12.4	8.5	26.7	26.5

***Table 1** Continuous and isolated digit recognition test results for 1352 digits by 9 male speakers. The test material was presented in three conditions: undegraded (Quiet), with white noise added to give a 15 dB SNR (Noise), and with a 6 dB/octave spectral tilt applied (Tilt). The commercial.1 system had a finely adjusted noise mask to achieve the good test noise results. Since commercial.2 is an isolated word type, results for continuous-word recognition would not be meaningful.*

4. Hardware

The IMELDA speech recognition algorithms have been implemented in hardware and are packaged in a 10" x 5" x 6" enclosure. This hardware recogniser was designed as a real-time prototype platform so that we could demonstrate IMELDA as well as develop, test, and evaluate the ongoing research efforts. Our goal is to achieve the same level of recognition performance with the hardware as with the mainframe implementation.

Two digital signal processor circuit cards are used. One for the signal processing aspects of IMELDA and the other for the continuous word pattern matching algorithm. This approach allows a high degree of development flexibility. In fact there is 40% spare capacity on one processor. However, the pattern matching algorithm is so computationally expensive that the number of words that can currently be recognised in real-time is limited to 30 words. We will be investigating techniques to increase the active vocabulary size.

Preliminary evaluations of the recogniser indicate that we are approaching the baseline established on the mainframe despite some limitations imposed by the hardware. The hardware recogniser has a lower mathematical dynamic range because it uses fixed point arithmetic rather than floating point and it employs a bandwidth limiting audio codec. The dynamic range was addressed by choosing the Analog Device ADSP-2100A DSP processor with its 40 bit multiply-accumulator which extends the dynamic range during long series of consecutive math operations and by using the block-floating point technique for the FFT implementation of the Mel-scale filterbank. The limited precision actually assisted the 'back-end' processing of the distance calculation in the pattern matching algorithm. This is reflected in Table 2 which compares pattern matching algorithm performance of the mainframe with the hardware.

Isolated Digits for speaker 'aa'

Platform	Quiet	Noise	Tilt
Mainframe	0/80	3/80	1/80
Hardware	0/80	1/80	1/80

***Table 2** Mainframes vs. Hardware results for speaker 'aa'. Number of errors over 80 digits spoken in three different conditions.*

The audio codec limits the speech bandwidth from 200 Hz to 3400 Hz. Fortunately, the IMELDA transform derives most of its information from the vocal tract resonances (formants) which reside within this range. The IMELDA transform for the hardware prototype has been derived from speech processed through this audio interface and, hence, the spectrum is weighted accordingly.

5. User Interface Software

The user interface software is implemented on a P.C. based platform and it communicates to the hardware through a serial link. This software handles the presentation and execution of the various functions of the recognition system in an orderly manner.

The first iteration of this software was intended to be a support tool for the hardware developers as well as a generalized user interface. As such, a certain degree of computer competence was assumed in the first iteration. The issue of how to handle novice computer users has be taken into consideration in the design process, but the user interface features necessary to support them have not been fully incorporated into the current version of the software.

The main functions for the program were determined to be: 1) File creation and handling capabilities for user vocabulary files 2) Edit capabilities for the vocabulary files 3) Syntax specification and editing capabilities 4) "Terminate-and-stay-resident" (TSR) code that accepts recognition results from the hardware and generates the corresponding command level macros for the user's application 5) Training mode for training each of the acoustic templates 6) Define mode for binding acoustic templates to keystroke macros 7) Embedded and isolated word training modes.

The basic flow and presentation for the program were laid out based on the operation of the voice recognition hardware and the experience of the NSF engineers with other commercial voice recognition systems. As a result, the interface was designed so that it operated in a menu driven graphical environment. Essentially, the user lays out word groups on a graphics screen in block elements. The user then specifies the syntax for a vocabulary by connecting word groups together in the order in which he/she wants word groups to be recognized. The user interface software will then generate the appropriate data file to specify the syntax of the vocabulary. This data file is then downloaded to the recognizer and used as part of the recognition process. This syntax definition method greatly simplifies data entry for novice users as it is more intuitive in nature and puts the burden of generating the syntax description file on the interface software.

6. Conclusions and Future Directions

Testing of the mainframe implementation of the IMELDA recognizer has demonstrated that it deals well with simulated noisy backgrounds and tilted speech. These results indicate that an IMELDA based recognizer will be able to address the identified limitations of other commercially available recognizers for persons with severe disabilities related to background noise and speaker variability (potentially related to changing emotion and fatigue). Initial testing of the hardware prototype recognizer has indicated that results comparable to the mainframe implementation can be achieved. The user interface software has been designed and implemented in such a manner as to overcome the limitations of generally complicated and unfriendly support software. Current work is focused on the improvement of the user training software and the performance of the recogniser in harsh environments where there is high background noise. Once these improvements are implemented, future work will involve an evaluation of the hardware and the software interface on a population of persons with severe physical disabilities. Other future work will include: increasing the active vocabulary size, continued development of speaker independence capabilities and implementation of robust keyword activation.

7. References

1. V.A. Thomason, P.S. Chopra, S.M. Farajian, and M.A. Abazid, "Application of Voice Recognition Devices for Computer Access and Programming", Proc. RESNA, ICAART-88, Montreal, Canada, 1988, pp. 370-371.

2. Hunt, M. J. and Lefebvre, C., Distance measures for speech recognition, Aeronautical Note, NAE-AN-57, Ottawa, March, 1989.

3. Hunt, M. J., Evaluating the performance of connected-word speech recognition systems, Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP-88), Vol. 1, pp. 215-218, New York, April, 1988.

4. Hunt, M. J., A statistical approach to metrics for word and syllable recognition, J. Acoust. Soc. Am., Salt Lake City, Vol. 66, pp. S535-536, 1979.