CONTINUED DEVELOPMENT OF AN IMELDA BASED VOICE RECOGNITION SYSTEM FOR PERSONS WITH SEVERE DISABILITIES

Gary E. Birch[1], Dariusz A. Zwierzynski[2], Claude Lefebvre[2], and David Starks[3]

[1]Neil Squire Foundation
4381 Gallant Avenue
North Vancouver, B.C.
Canada
V7G 1L1

[2]Neil Squire Foundation
Speech Research Centre
National Research Council of Canada
Building U-61, Montreal Road
Ottawa, Ontario
K1A-OR6

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

INTRODUCTION

The Neil Squire Foundation is a Canadian non-profit organization responsive to the needs of individuals who have severe physical disabilities. Through direct interaction with these individuals we research and develop appropriate innovative services and technology. To date the Foundation has worked on a one-to-one basis with well over 3500 persons with disabilities from all over Canada. In almost every case we have been involved in assessing and recommending the most appropriate interface to computer based assistive devices. The concept of direct speech input to control assistive devices is often an appealing option and hence the Foundation has always worked to keep abreast of and, indeed in most cases, test commercially available speech recognition systems. We estimate that we have had well over 1500 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. These limitations, particularly in control applications, have been related to interference from background noise, varying degrees of speaker stress and/or emotion, fatigue and generally unfriendly and complicated user application/training software. Similar findings regarding the use of speech input systems were also reported by researchers at the University of Tennessee, Centre of Excellence for Computer Applications [1]. As indicated in [2], this experience with speech input lead the Foundation to become involved in a joint project with the Speech Research Centre at the National Research Council of Canada (NRC) and the Canadian Marconi Company (CMC) to research and develop a speech input system that will ultimately overcome these limitations. In 1990, the IMELDA system that had been designed and tested on a mainframe computer at NRC was implemented in stand-alone hardware by CMC and the initial results are very promising. The most dramatic test and evaluation to date has been through a project called "Fly-by-Voice".

BACKGROUND

The general design of our speech recognition system is such that the acoustic features are first extracted from the speech signal in the front-end component. Subsequently, these features are passed on to the back-end component where they are compared with stored templates through the technique of dynamic programming.

The front-end processing in our speech recognition system is a mel-scale fast-fourier-transform based spectral filter-bank analysis followed by a linear transformation [3]. This linear transformation, which is called IMELDA, was developed and tested in our laboratory and is based on linear discriminant analysis [3,4]. Comparisons with other systems indicate that it is the "state of the art" for front-end processing in a robust speech recognition system, outperforming other transforms and recognition systems, particularly in degraded speech [2].

OVERVIEW OF THE FLY-BY-VOICE PROJECT

In the "Fly-by-Voice" project a vocabulary was set up that would enable a pilot to fly a Bell 205A helicopter completely by voice commands. This project was undertaken to demonstrate the feasibility of speech recognition in an extremely adverse environment and it was not intended to suggest that flights could be controlled by voice. It does, however, point to the important potential of controlling secondary features in the cockpit, such as the control of navigational information.

The project used four male speakers whose voice was recorded in the helicopter cockpit under two of the most demanding flight regimes: at the hover and in the cruise. The test data consisted of each speaker recording 280 words and digits in both of the test flight regimes. This resulted in a total 560 tokens for each speaker (15 words and 10 digits). The words and digits were spoken in command strings. The reference templates for each speaker were derived from two passes of the vocabulary recorded in the helicopter in quiet and two passes with the main rotor at 70% maximum rpm. The manner and rate of speaking were chosen by each speaker individually to keep the procedure as natural as possible. Three major factors were found to impact the performance of the recogniser. The first factor involved syntax. A closed syntax that structured the sequence of command words with digits performed much better than an open syntax structure. The second factor was related to thresholding the speech input. It was found that performance was greatly improved when a thresholding technique was applied. This technique is based on the average spectral envelope of the speech for each speaker. This approach helps to preserve formant structure by providing more thresholding effect in the lower, noisier channels, and relatively less in the quieter upper channels. The third factor is related to the computation of the IMELDA transform. It was found that computing a user-specific transform, using examples of speech in both quiet and noisy backgrounds, provided the best performance in the severe conditions of a helicopter cockpit.

Overall, the recognition rate on the test data recorded in the cockpit for all four speakers is as follows: Speaker #1: 99.3%; Speaker #2: 99.6%; Speaker #3: 99.3%; Speaker #4: 99.3%. The hardware recogniser was also tested in actual flight by one of the speakers. In this test the recognition performance was 93.8% in hover and 94.8% in the cruise. Reasons for reduction in performance are mostly related to specific problems in integrating the recogniser into the cockpit, such as the presence of 400Hz aircraft power supply harmonics, variable microphone placement and audio-compression noise. For more details on the "Fly-by-Voice" project see [5].

The performance of the recogniser with the high level of background noise and inherent speaker stress in a helicopter cockpit environment is very encouraging for other applications. Of particular interest to the Neil Squire Foundation, is the expected carryover of these performance characteristics to other applications where persons with severe disabilities are using voice to control various aspects of their environment.

EVALUATION OF SPEAKER INDEPENDENCE

Ideally, there are many applications for persons with disabilities where a speaker-independent system would be highly desirable. For example, a speaker-independent system, because it would avoid the need for any special training of the vocabulary by the user, would greatly simplify the set up procedures for a voice actuated home control system.

The laboratory implementation of the IMELDA speech recognition system has demonstrated good speaker-independent capabilities [2]. Until recently, however, the performance in the speaker independent mode has not been investigated using the stand-alone

hardware recogniser.

Work to assess the speaker independent capabilities of the hardware was carried out in three phases. In the first phase one set of data was used containing isolated and co-articulated digits (0-9). One thousand three hundred fifty co-articulated digits were extracted from triplets recorded by 9 male speakers of English from a pre-recorded database. Isolated digits were recorded on a magnetic tape in a quiet laboratory over a Shure SM-10 public address head-mounted microphone. The isolated digits were recorded 4 times for reference material by 14 speakers representing different accents of English (560 digits) and also 1530 isolated and 150 co-articulated digits were recorded by the same speakers for testing material. The digits and words were then processed for each speaker through the hardware recogniser to collect material for reference templates.

All the data were transferred onto a SUN SPARCstation for K-means clustering of the collected data examples. Three classes were computed for each digit, which was motivated by the limit of the vocabulary size in the hardware which at the time of the experiment was 39 tokens. An IMELDA transform and reference templates were computed and ported to the hardware. Two recognition experiments were conducted. The first test used data recorded for testing by the original 14 male speakers whose data were used for the reference material. The other test was truly speaker-independent, using data input live by 10 completely new male speakers. The overall recognition accuracy for all speakers with all digits (isolated and co-articulated) in the first test was 98.5% and 95.9% in the second test.

In the second phase, isolated and co-articulated digits were used along with 14 words appropriate for a home automation control application: help, next, menu, exit, up, scan, select, dial, number, answer, phone, auto, redial, hang (NSF lexicon). The digits were collected in the same way as in PHASE I, while the words were recorded by 14 speakers 4 times: 2 times for training, and 2 times for testing. K-means clustering was again used, with 2 classes computed for digits and 1 class for words. An IMELDA transform was computed with the entire material, including digits and words.

One recognition experiment was conducted. It used data recorded for testing by the original 14 male speakers whose data were used for the reference material. No test with new speakers was conducted, as it was decided to carry it out in PHASE III on a larger lexicon. The overall recognition accuracy for all speakers with all tokens and an open syntax (co-articulated digits and isolated words: 184 x 14 = 2576) was 97.2%.

In the third phase, isolated and co-articulated digits were used in this test along with 24 words: DME, ILS, MLS, VOR, UHF, VHF, channel, set, TACAN, transponder, normal, ident, squawk, emergency, guard, hijack, point, decimal, correction enter, fife, niner, oh, hundred (CMC lexicon). The digits and words were collected in the same way as in Phase II. K-means clustering was again used, with 1 class computed for digits and 1 class for words.

Two recognition experiments were conducted. One used data recorded for testing by the original 14 male speakers whose data were used for the reference material. The other test was truly speaker-independent, using data input live by 6 completely new male speakers. In the test with microphone input, the words from the CMC lexicon were used in combination with digits (e.g. DME 171.05). There were 40 such strings. The overall recognition accuracy for all speakers with all tokens and an open syntax (co-articulated digits and isolated words: 204 X 14 = 2856) was 95.1% in the first test and 81.1% in the second test. The experiment with microphone input in PHASE III (second test) was the most natural and difficult of all tests, as it comprised whole utterances rather than just lists of isolated words. For comparison, one of the speakers, whose voice was used for training, read through the microphone the list of 40 strings of words and co-articulated digits. In the test with tape input, 3 errors were detected for digits (3/156:98.1%), while with microphone 11 errors (11/129:91.4%) were detected. A similar trend was observed for words: 0/48:100% (tape) and 4/75:94.6% (microphone).

The following conclusions may be derived from these evaluations:

(a) even for a small number of speakers (14) and relatively small databases used for training, it has been possible to demonstrate speaker-independent performance at an accuracy rate of 96%. This is a result obtained for isolated and co-articulated digits, spoken by new speakers in an unconstrained way, with 3 classes of averaged digits as the reference templates.

(b) The inclusion of words in addition to digits did not degrade recognition accuracy. Words were on average recognised better than digits with both tape and microphone inputs with only 1 class of averaged words in each test. The performance of digits deteriorated in all conditions as the classes of averaged reference data were reduced from 3 to 2 and eventually to 1 only.

(c) The computation of more classes for both digits and words would improve recognition accuracy. The hardware recogniser will have to be modified to increase the vocabulary size. Work toward this end is currently being carried out and recently the vocabulary has been increased to 64 tokens.

(d) Automatic gain control (AGC) in the hardware should also contribute to increased recognition accuracy with the microphone. The input level from the tape was manually adjusted for each speaker.

FUTURE WORK

Work in the future will continue in several areas. Work on developing a robust keyword activation algorithm which will allow a user to "hands free" switch the recogniser in and out of the recognition mode of operation will be performed. A method of AGC will be integrated into the hardware to allow a wider range of speaker levels and lessen the dependency on microphone positioning. Once the AGC and keyword activation have been implemented then a speaker-independent home control application will be field tested. In addition, the hardware will be redesigned to take advantage of newer DSP technology allowing for a single board implementation with greater working vocabulary capacity. Finally, as changes to the hardware are implemented the user interface application software will continue to evolve both to support the hardware changes and to improve its overall level of user friendliness.

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. Birch, G.E., Zwierzynski, D.A., Lefebvre C., and Starks, D.R., "An IMELDA Based Voice Recognition System: A Step Towards Effective Voice Recognition for Persons with Severe Disabilities", Proceedings of the Canadian Acoustical Association Conference, Edmonton, Alberta, October, 1991 pp. 111-112.

3. Hunt, M. and Lefebvre, C., "A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech", Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, ICASSP-89, Glasgow, Scotland, Vol 27, No. 3, pp. 262-265, 1989.

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

5. Lefebvre, C.. Zwierzynski, D.A., Starks, D.R., and Birch, G.E., Further Optimization of a Robust IMELDA Speech Recogniser for Applications with Severely Degraded Speech, Proceedings of the International Conference on Spoken Language Processing, Banff, Alberta, Canada, October, 1992.