Proceedings of the Institute of Acoustics

USE OF LINEAR DESCRIMINANT ANALYSIS IN ISOLATED AND CONTINUOUS SPEECH RECOGNITION FOR AN AIR TRAFFIC CONTROL TASK

Melvyn J. Hunt, Stephen M. Richardson and *Martin G. Abbott*

Marconi Speech & Information Systems
Airspeed Road, The Airport
Portsmouth, P03 5RE, U.K.

SUMMARY

This paper describes experiments with a speech recogniser using whole-word models and linear discriminants for its spectral representation in the framework of an Alvey demonstrator for an air traffic control task. The demonstrator task and specification are first set out. After a brief account of the speech output component, the implementation of the recognition hardware is described. Three series of recognition tests are then described. In the first, the performance of the new hardware is shown to be superior to the best previous hardware in speaker-independent tests with normal and with speech in which the loudness and voice quality is varied. In the second, the absolute performance of the system is tested on sentences from the demonstrator task. Finally, the third series provides results to allow comparison on a common database with a sub-word modelling system.

1. Introduction

In 1988 Marconi Speech and Information Systems (MSIS) took over the leadership of the DTI-funded Alvey Large-Scale Speech Technology Demonstrator Project. The reborn project has as partners the Centre for Speech Technology Research (CSTR) at Edinburgh University and the Human Sciences and Advanced Technology Unit (HUSAT) at Loughborough University of Technology. In addition, MSIS augmented its own research effort with a small team in a sister organisation, GEC Hirst Research Centre (HRC).

Early in the life of the new project it was decided that instead of having it culminate in a single large demonstrator, a series of demonstrators would be produced in stages spread over the duration of the project. To reflect this revised intention, the project was renamed the Alvey Integrated Speech Technology Demonstrator (ISTD).

The first demonstrator in the series, a system for the U.K. Civil Aviation Authority (CAA) intended to help air traffic controllers, was planned to be based entirely on existing MSIS technology. In the event, however, technology was developed for the demonstrator that, despite exploiting existing MSIS hardware, nevertheless represents a radical new departure for the company. It incorporates an acoustic representation derived using linear discriminant analysis (LDA).

The demonstrator was delivered for user evaluation around Christmas 1989 and some preliminary results were published[1]. However, technical developments have continued since then, and commercial spin-offs from the demonstrator hardware have been produced. Also, while this demonstrator used whole-word modelling, some later demonstrators will be based on sub-word modelling. The task domain of this first demonstrator has recently served as a test-bed for the direct comparison of whole-word modelling with sub-word modelling schemes developed at CSTR and at HRC.

The purposes of this paper are somewhat larger than the title suggests. They are: (i) to describe the task that the first demonstrator was required to fulfill; (ii) to describe the technology that was developed to meet the requirement and the technical evaluations that were carried out on it; and (iii) to describe some recent tests using a common database with two sub-word modelling schemes.

LINEAR DESCRIMINANT ANALYSIS IN SPEECH RECOGNITION FOR ATC

2. The Air Traffic Control Task

All aircraft flying in a particular area of U.K. air space called a *sector* are under the control of an individual air traffic controller. Information on each aircraft in the sector is recorded on a strip of paper known as a *flight strip* (Fig. 1). All the strips for the sector are displayed on a board in a definite order. When an aircraft leaves the sector the corresponding strip is handed to the controller responsible for the sector it has entered. Changes to flight data — for example to the altitude or expected arrival time of the aircraft — are marked by hand on the strip.

The CAA has recently developed an experimental *electronic flight strip* system, in which images of flight strips are presented on a graphics terminal. The data on these flight strips can be updated using a mouse and keyboard. Such a system offers many advantages over the current paper method. For example, several operators can have simultaneous access to the same information, and the data can be recorded in a log automatically.

The purpose of the first ISTD demonstrator was to explore the possibility of using speech as an alternative method of updating the information on the electronic flight strips and controlling various system management parameters.

3. The Task Specification

The task specification called for a "closed community" speaker-independent continuous speech recognition system. We took this to mean that the users of the system would have contributed to its training, but that it should not need to be given the current user's identity.

After analysing the task, HUSAT produced a vocabulary list of around seventy words and a finite-state syntax specifying allowed input sequences for an example containing thirteen strips identified by aircraft call signs. Off-line tests used a modified version (Fig. 2) of the opriginal syntax[1] which excluded irrelevant error-correction options.

To avoid confusions over distorting radio links, air traffic controllers use special pronunciations for certain digits: *tree* for *three*, *fife* for *five*, and *niner* for *nine*. Users of the recognition system were therefore required to adopt these pronunciations. This had an unfortunate effect in some demonstrations, since audiences unfamiliar with air traffic control practice mistakenly took it that the non-standard pronunciations had been introduced to help the recogniser.

The specification also called for spoken output to confirm the information received by the system and to issue warnings. The requirement here was for a system which could use encoded natural speech but which was capable of being upgraded to a text-to-speech system. As the next section will show, the approach taken to speech output matched this requirement literally.

4. The Speech Output Component

Linear Predictive Coding (LPC) is a well established technique for encoding speech signals, and it is well known that the quality of the analysis in voiced sounds can be improved by performing it in synchrony with the quasi-periodic glottal excitation. Usually, the synthesized speech waveform is generated by exciting the filter specified by the LPC analysis with random noise for voiceless sounds and with a fixed waveform, most commonly a single impulse, in voiced sounds. If, however, the filter is excited by the prediction error signal, the LPC *residual*, the original waveform is recovered exactly. This, in itself, is an uninteresting process, because nothing is achieved. However, when the speech is recorded carefully and the analysis is carried out in synchrony with the excitation, the residual in voiced speech is dominated by an impulse at the instant of glottal closure and there is particularly little activity in this waveform in the region around 80% of the distance towards the next excitation point. By modifying the residual at this point — deleting samples or inserting zero-amplitude samples — the duration of the glottal cycle, and hence the fundamental frequency, can be manipulated. Also, by deleting or repeating whole glottal cycles the speech rate can be increased or decreased. Experiments carried out in Canada[2] have shown that a wide range of prosodic modifications can be generated in this way without

any detectable degradation.

Some of the most effective text-to-speech systems have been based on the concatenation of diphones[3]. Much of the limitation in quality in these systems comes from the degradation imposed by the speech coding scheme used, typically conventional LPC. As had been proposed earlier[2], this source of degradation can be removed completely by using the modified residual method just described.

As a first step to a real-time text-to-speech synthesizer of this kind, workers at MSIS developed real-time synthesis with a modified residual on an LSI DSP32C board. In early implementations of the demonstrator the synthesizer was used simply to reconstruct recorded phrases. However, in close collaboration with colleagues at CSTR[4] a diphone inventory has been recorded from the same speaker and the phrases have been constructed from these diphones. We are particularly encouraged that as well as having high intelligibility the speech reconstructed in this way clearly preserves the identity of the speaker. We see this property being important when a large corpus of recordings exists for a speaker who is no longer available and a few more words need to be added to the corpus in the same voice.

5. Real-Time Speech Recognition using Linear Discriminants

In Canada one of us (MJH) had developed acoustic representations using LDA and called IMELDA[5]. The derivation and properties of IMELDA representations are described in another paper in these proceedings[6], and that paper should be read in conjunction with this one.

Two versions of IMELDA have been used in the work described here: a version using only static spectral information called IMELDA-1 and a version incorporating both static and dynamic (spectral change) information called IMELDA-2.

The Canadian IMELDA implementation used the output of an FFT-based simulated mel-scale filter-bank, while MSIS hardware used a true mel-scale digital filter-bank. The published IMELDA results were first replicated on the same database and with the same recognition algorithms. An IMELDA representation was then derived for the MSIS IIR digital filter-bank and the performance of the two versions was compared. Although there appeared at first to be an advantage for the FFT-based version in noise and for the direct filter-bank version in other conditions[7], we now believe this to have been due to slightly different levels of spectral thresholding in the two implementations. More recent comparisons between an FFT-based IMELDA-2 and an IIR-based version (Table 1) show no systematic differences. There are, however, tradeoffs to be made in computational cost and memory requirements between the two methods, and the balance of advantage depends on the number of channels in the filter-bank and the frequency with which the spectrum is to be sampled. Both versions of IMELDA are massively better than a mel-ceptrum representation.

In the Canadian work, spectral frames were estimated every 6.4 ms, and dynamic parameters were derived by linear regression over seven consecutive frames. In the real-time implementation, however, frames were estimated only every 16 ms, and linear regression (which then reduces to simple subtraction) was applied over just three consecutive frames. This change degrades performance somewhat, particularly in noise.

Experiments showed, on the other hand, that the limited precision integer arithmetic used to compute the spectral representation in the MSIS real-time front-end did not seriously degrade performance. However, a more severe limitation is presented by the arithmetic used in the distance calculations in MSIS hardware. In one implementation, nineteen four-bit channels are available. IMELDA-1 had been found to be best with just eight coefficients and IMELDA-2 had been used with just twelve (though we currently believe that seventeen is closer to the optimal number). Moreover, the variance of the first few coefficients is much greater than those of the later ones. Encoding IMELDA coefficients directly such that the first coefficient did not overflow the four-bit range would have meant that the last coefficient would occupy a range of only about two bits, while up to eleven channels would be left unused. The loss of numerical precision that this would entail would be expected to degrade recognition performance seriously.

For IMELDA-1 our solution was to project the representation back into the spectral domain. LDA can be seen as a rotation followed by a scaling, a further rotation and finally a truncation. By adding the inverses of the two rotations, the transform is projected back into the original parameter space. Since rotations do not affect Euclidean distances, the normal LDA representation and its back-projected version have identical pattern classification properties. An IMELDA-1 representation derived from, say, a nineteen-channel filter-bank can be converted back to a nineteen-channel representation with roughly equal variances across the channels. Four bits per channel are then adequate. There is, moreover, a further advantage in that the "spectral IMELDA" shows clear formant structure and can be interpreted visually, unlike a normal IMELDA. It was therefore possible to verify that the processing in the real-time front-end was working properly.

This solution was not available for IMELDA-2, since transformation back to the spectral domain would generate twice as many parameters as there are channels. Instead, a new orthogonal transformation was developed that can spread the information in the twelve parameters over the larger number of channels and make the variance in each channel exactly equal. Since it is an orthogonal transformation, Euclidean distances are unaffected.

IMELDA was first implemented in real-time in a stand-alone recognition system derived from the ASR1000 flyable recogniser[8]. This hardware was later developed into the *MR4* recogniser, the name signifying Marconi's fourth generation continuous speech recogniser. While ASR1000 comprised two separate entities, a flyable recognition unit and a ground-based training station, MR4 combined them into a single box. The introduction of IMELDA added only 6% to the front-end computation, and could easily be accommodated in the ASR1000 hardware. More recently, an IMELDA-based PC-card recogniser has been produced, giving recognition performance comparable to that of MR4 in a much smaller, inexpensive unit.

6. Training the Demonstrator Recognition System

The requirement for closed-community speaker independence could have been met by using multiple templates, one for each speaker. However, this approach could not be extended to true speaker independence, and it would reduce the size of vocabulary that could be used. We therefore decided to allow ourselves just one template per word in the vocabulary, though a few words such as *left* and *right* were assigned templates both with and without a released */l/*.

For our training speech we used one example of each word in the vocabulary spoken in isolation by sixteen male speakers. This arrangement was a result of time limitations and certainly does not give optimal recognition performance. We know that true speaker-independent recognition accuracy is improved by increasing the training population to at least 64 speakers per sex (all tests here are with male speakers) and that continuous speech recognition is much improved with embedded training, which MR4 can perform.

Similarly, it would have been possible to derive an IMELDA transform from the training speech, thus exploiting the small advantage gained from deriving the transform from the same vocabulary on which it is to be tested. However, since it would not always be practicable to derive a new transform for each application, we decided to use the IMELDA transform derived from the digit database used in Table 1. This contains only North-American speakers and includes noise and spectral-tilt degradations, even though such degradations were not to be encountered in the tests.

The syntax was specified using a graphical syntax editor and compiler, in which syntaxes can be drawn and edited on a PC screen with the aid of a mouse.

Early tests of the system showed a need to allow pronunciations of "to" in "descend to" and "climb to" with both a reduced vowel and an unreduced vowel as in the digit "two".

Representation	Speaker Dependent			Speaker Independent		
	Q	N	T	Q	N	T
mel-cepstrum	1.00	18.30	71.20	5.30	28.8	76.40
FFT-IMELDA-2	0.00	1.11	0.07	0.52	2.22	0.89
IIR-IMELDA-2	0.00	0.82	0.00	0.45	2.67	1.34

Table 1. Percentage error rates for quasi-isolated-digit recognition tests with 1348 digits from 9 male speakers. Three test conditions: (Q) undegraded, (N) additive white noise to give 15dB SNR, and (T) 6dB/octave spectral tilt.

7. Performance Evaluations

i) Comparison with Existing Technology

To test whether the technology developed for the demonstrator represented an advance over the best previous technology at MSIS, it was compared directly with an ASR1000 in closed-community speaker-independent and true speaker-independent digit recognition tasks. The templates were derived from the single isolated-word examples of the digits provided by the sixteen speakers for the demonstrator. Five of these speakers were then chosen as test speakers for the closed-community tests and a further five male speakers not included in the training group were chosen for the true speaker-independent tests. The MR4 hardware was tested with an early IMELDA-1 transform as well as with a later IMELDA-2 version.

It was quickly apparent that the error rates for normally spoken isolated digits with IMELDA-based hardware would be too low to measure reliably. Consequently, as well as having each test speaker produce a hundred normally spoken isolated digits, he was asked to produce fifty examples shouted (roughly 12 dB louder) and fifty examples spoken softly, almost whispered (roughly 12 dB quieter). Fifty continuously spoken three-digit groups were also collected from each test speaker.

The purpose of the softly spoken and shouted digits was twofold: first to test the tolerance of the recognisers to level variation; and second to test their tolerance to changes in voice quality. The second of these properties was measured by roughly correcting for these level changes in the loud and soft speech. The shouted speech then resembles that of someone under stress or with noise in the ears, while the soft speech has some of the qualities of a speaker with laryngitis. All these conditions might be expected to occur in a real application.

It would have been possible to train an IMELDA transform for the kinds of speech variations just described, but the aim of the tests was to see whether an IMELDA trained on other kinds of degradations could cope with them.

As Table 2 shows, the performance of MR4 with IMELDA-1 was much better than that of ASR1000 in all conditions, and IMELDA-2 brought further improvements in most conditions.

In many conditions the true speaker-independent performance with the speakers outside the training group was not markedly worse than the closed-community performance. This was a surprising result given the small size of the training group and presumably reflects the speaker-normalising character of IMELDA representations.

Recognizer	Closed-Community Speaker Independent Tests					
	N	Sc	S	Lc	L	C
ASR1000	2.20	6.40	38.80	10.00	12.80	3.92
MR4 IMELDA-1	0.40	2.40	11.60	2.80	0.40	0.69
MR4 IMELDA-2	0.00	0.80	10.80	3.20	0.80	0.26

Recognizer	True Speaker Independent Tests					
	N	Sc	S	Lc	L	C
ASR1000	1.20	8.80	51.20	6.00	39.60	5.88
MR4 IMELDA-1	0.00	2.80	16.80	3.60	10.00	1.75
MR4 IMELDA-2	0.20	1.20	15.60	3.20	9.60	0.53

Table 2. Percentage error rates for closed-community and true speaker-independent digit recognition tests. Five test conditions: (N) spoken normally, (Sc) spoken softly with level compensation, (S) spoken softly, (Lc) spoken loudly with level compensation, (L) spoken loudly, (C) spoken continuously.

ii) Tests with the Demonstrator Sentences

To test performance on the demonstrator task, a total of 250 random sentences averaging nine words in length were recorded by five speakers belonging to the training group. The response of the recogniser when repeatedly presented with these recordings turned out to vary slightly from one occasion to the next, sometimes giving just one and sometimes giving just two substitution errors: corresponding to phrase error rates of 0.4% and 0.8% respectively, and word error rates of 0.04% and 0.09%.

iii) Experiments for Comparison with a Sub-Word Modelling System

As mentioned in the introduction, CSTR's speech recognition activity has been directed primarily towards sub-word modelling. Although this first demonstrator was always intended to use whole-word modelling, it offers a suitable test-bed for comparison of the two approaches.

Unlike many sub-word modelling systems demonstrated elsewhere, CSTR's approach has total separation between the vocabulary from which the sub-word units are derived and that on which they are tested. The system developed for the Alvey project is, however, speaker dependent.

CSTR recorded a total of 99 syntactically valid test sentences for four male speakers for the demonstrator task. Speaker-independent recognition tests were performed using a demonstrator simulation and word models that were created from the 16 MSIS speakers for the first demonstrator test. These results are given in Table 3. The simulation was chosen in preference to the real hardware for reasons of speed in performing the tests. Previous experiments on other databases have shown the simulation to be an accurate representation of the hardware.

Speaker	GSW	HXB	JMR	PMS
Phrase errors (%)	5	59	6	9
Weighted word errors (%)	0.8	14.5	0.8	1.0

Table 3. Speaker-independent recognition results with the CSTR demonstrator sentences. The weighted word error rate counts insertions and deletions as half errors[9].

It is difficult to compare these results with those from CSTR for various reasons, which include: (i) CSTR's tests were speaker-dependent; (ii) the syntactic constraints were applied differently; (iii) CSTR used a 10kHz sampling rate for front-end analysis whereas MSIS used 8kHz.

The poor performance for speaker HXB can be attributed to the large difference between his accent and those used to create the speech models. The lower performance of these speakers as a group compared with the MSIS speakers could be due to their not belonging to the training group — though the digit recognition results in table 2 suggest that this is not a major factor. The fact that the CSTR speakers, unlike the MSIS speakers, lacked experience in using the real-time demonstrator system may be more important.

Acknowledgments

This work was part of the IED ISTD Project in collaboration with HUSAT and CSTR. We are grateful to the National Research Council of Canada for providing access to databases and software and to many colleagues at MSIS for their help with the work.

References

1. Martin Abbott, "The Use of Speech Technology to Enhance the Handling of Electronic Flight Progress Strips in an Air Traffic Control Environment," *Proc. Voice Systems Worldwide*, pp. 126-134, 1990.
2. Melvyn J. Hunt, Dariusz A. Zwierzyński and Raymond C. Carr, "Issues in High Quality LPC Analysis and Synthesis," *European Conference on Speech Communication and Technology*, vol. 2, pp. 348-351, Paris, France, 1989.
3. Isard, S. and Miller, D. A., "Diphone Synthesis Techniques," *IEE Conference Publication no. 258*, pp. 77-82, 1986.
4. J. Verhoeven, *Context-Sensitive Diphones As Units In Speech Synthesis*, Bowness, Windermere, November 1990. Institute Of Acoustics 1990 Autumn Conference, Speech and Hearing
5. Melvyn J. Hunt, Claude Lefèbvre, "A Comparison of Several Acoustic Representations for Speech Recognition with Degraded and Undegraded Speech," *Proc. IEEE Int. Conf. Acoustics, Speech & Signal Processing, ICASSP-89*, vol. S1, pp. 262-265, Glasgow, Scotland, May 1989.
6. Stephen M. RichardsonMelvyn J. Hunt, *A Comparison Of PLP-RPS and PLP-IMELDA Acoustic Representations In Automatic Speech Recognition*, Bowness, Windermere, November 1990. Institute Of Acoustics 1990 Autumn Conference, Speech and Hearing
7. Melvyn J. HuntStephen M. Richardson and Martin G. Abbott, *Use Of Linear Discriminant Analysis In Isolated And Continuous Word Speech Recognition Experiments*, Bowness, Windermere, November 1990. Institute Of Acoustics 1990 Autumn Conference, Speech and Hearing
8. I. Galletti, M. Abbott, "Advanced Airborne Speech Recognizer," *Proc. American Voice Input Output Society*, pp. 127-136, Newport Beech, California, September 1989.
9. Hunt, M. J., "Figures of Merit for Assessing Connected-Word Recognisers," *to be published in Speech Communication.*

LINEAR DESCRIMINANT ANALYSIS IN SPEECH RECOGNITION FOR ATC

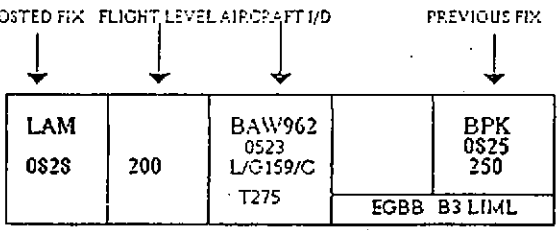

Figure 1. Example flight strip

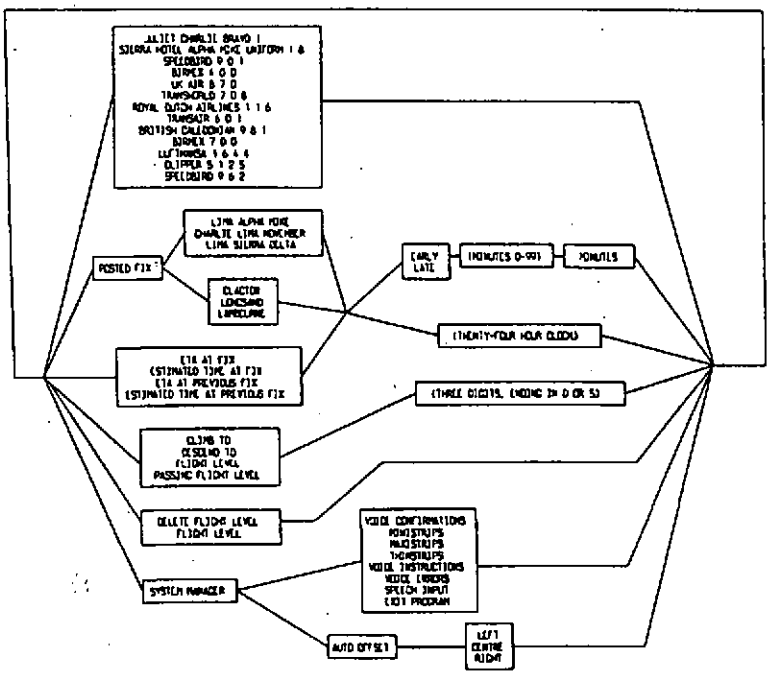

Figure 2. Simplified recogniser syntax