OVERVIEW OF SPEAKER VERIFICATION STUDIES TOWARDS TECHNOLOGY FOR ROBUST USER-CONSCIOUS SECURE TRANSACTIONS

J.Bruce Millar, Fangxin Chen, Iain Macleod, Shuping Ran,
Hong Tang, Michael Wagner and Xiaoyuan Zhu.

TRUST Project
Research School of Information Sciences and Engineering
Australian National University

ABSTRACT

The aim of this paper is to provide background material relating to speaker verification work of the Technology for Robust User-conscious Secure Transactions project. It describes the philosophy of this work as it is expressed in the design and collection of a speech data corpus, and the selection of speech analysis and speaker modelling techniques. It then summarises the range of experiments that have been performed using these data and techniques to explore many issues pertinent to effective speaker verification. This paper sets the scene for six other papers appearing in these proceedings.

INTRODUCTION

Since mid-1993 a concerted effort to examine issues relating to speaker verification techniques, for use as part of a sophisticated computer security system, has been pursued at the TRUST (Technology for Robust User-conscious Secure Transactions) Project at the Australian National University. The TRUST project has designed a speech data corpus, developed interactive software for collecting this corpus and analytic software for preprocessing the data into mel-scaled cepstral coefficients. Further software has been developed for modelling speakers using either vector quantisation or hidden Markov models, deriving measures of similarity or difference between utterances from different speakers, and estimating the probability that an incoming utterance is or is not consistent with a given speaker model.

This paper aims to provide background information on the design criteria of much of this work, to summarise results obtained so far, and to reference current publications of the project.

DESIGN AND COLLECTION OF SPEECH DATA CORPUS

A corpus of speech data was designed to suit the initial requirements of the project. The data was required to build models of the speech of individual speakers which would be typical of utterances spoken to command a "Windows-based" transaction processing system. The major criteria to be met by these data were:- (1) speech from one individual should be collected over a period of days if not weeks; (2) in length and phonetic content the utterances used should approximate speech material used in computer interaction; (3) the material should allow the building of both text-independent and text-dependent models of each person's speech; and (4) that in the first instance this corpus should support studies of speaker identification but be extendable to studies of speaker verification.

Accordingly, a set of computer commands extended by some short sentences was established. As a set these commands were phonetically rich with respect to Australian English by virtue of achieving complete first order phonemic coverage and maximising second order phonemic coverage within lexical and syntactic constraints. The 30 items, comprising 310 phonemes, are listed in Table 1. Their first order distribution in broad classes is illustrated in Figure 1.

An initial set of 25 speakers was chosen on the basis of fluent English speaking competency although some speakers were slightly accented. Over time a further 20 speakers were added. Details of accenting and other factors relating to speech performance were recorded for all speakers. Speakers were required to attend three recording sessions spaced approximately one week apart.

My name is:		The System Access Number is: 175-093	
My office phone number is: 249-6898		Open Data Manager	
New window	Edit file	Page set up	Text format
Line number	Show ruler	View preferences	Undo typing
Grammar check	Change font type	Save file	Append text
Send mail	Read only	Resume task	Row height
Help index	Open footer	Print merge	Insert table
Print preview	Measure width	Fill down	Compare with
Clear screen		Move cursor	

Table 1: Utterances comprising the 30 item speech corpus

The speech recording environment was an acoustically treated room in which the ambient sound level was 42-43 dBC. The background sound predominantly comprised the rumble of air-conditioning units in an adjacent room and fan noise from the computer used to control the experiment.

The speaker sat in front of a keyboard, mouse, and screen via which prompts were presented. The speaker was fitted with a headmounted Sennheiser 5058P condenser microphone whose position (10mm from left corner of mouth) and orientation (45 degrees above horizontal) was carefully monitored. This position was chosen to minimise breath noise and maximise the nasal radiation component.

The speaking task was presented to the speaker via visual tokens which appeared in a window on the screen. The speaker controlled the time of appearance of each successive token by pressing a screen button labelled 'start' using the mouse. The speaker was instructed to speak the words displayed as soon as they appeared and to press a 'stop' button when they had been spoken. If the speaker was aware that s/he had made an error, s/he could press a retry button which caused the utterance to be ignored and the token to be resequenced for later presentation. A practice run was provided at the start of each session. This sequence was repeated until the list of tokens was exhausted. The speakers were encouraged to take breaks if they were feeling tired and were provided with a drink of fruit juice to sip in such breaks.

A complete session comprised the production of 5 repetitions of the 30 utterance items. The 150 utterance prompts were presented in random order. In the second and third session 10 further utterance prompts, each presented twice, were added to provide some phonetic variance not present in the main corpus.

Signal conditioning was performed using a pre-amplifier (Revox A77) and low-pass filter (Rockland 432). The lowpass cutoff characteristics were set to -3dB at 9.0 kHz with a roll-off of 48dB/oct. Signal levels

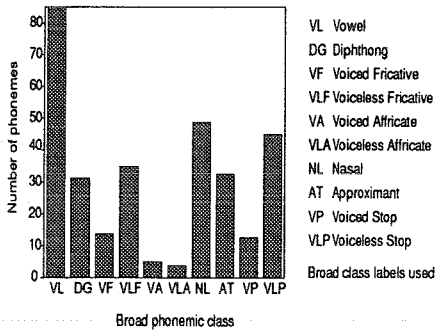

Figure 1: Distribution of phonemes in spoken corpus design

were monitored on an oscilloscope to provide maximum amplitude within the constraints of input to a Pro-Audio-16 sound capture board (set to sample at 20000 samples/s) in a 486/50 IBM compatible computer where each utterance was stored in an individual file.

Most of the experiments carried out within the project have used this data corpus, but some use has been made of the TIMIT data corpus when either larger numbers of speakers (Wagner et al., 1994) or phonemically labelled data (Zhu et al., 1994c) has been required to investigate a particular issue.

TECHNIQUES FOR SPEECH PROCESSING

Each utterance file was processed by end-point detection to remove extraneous background before and after the utterance, by digital filtering to constrain frequency components prior to modelling them, then by Fourier analysis followed by mel-cepstral analysis.

End-point detection used a dual energy threshold method to facilitate inclusion of low energy onsets and offsets of utterances. Using this technique the complete utterance, including all its phonetic components but excluding all but a small amount of the surrounding background signal, was passed on for further processing. The results of this process were checked visually and auditorily, and then manually resegmented where necessary to maintain consistent and accurate end-pointing.

Digital filtering removed frequency bands in which extraneous sound components may interfere with the speech energy. The default settings for this filtering were a high-pass filter at 60Hz and a low-pass filter at 4800Hz (4th order Butterworth). These filters delimit the frequency band with the major contribution to speech intelligibility and have been widely used in other work on speaker recognition.

The 60-4800Hz bandlimited signal down-sampled to 10000 samples/s was subjected to 95% pre-emphasis and Fourier analysis (order 8) on 25.6ms frames with a 10ms frame advance, providing 128 spectral magnitude coefficients. This linear spectral representation was then transformed onto the mel-scale by creating 40 bands of width 110 mel and separation 55 mel between 55 and 2310 mel. A final representation of 20 mel-frequency cepstral coefficients (MFCC) was computed by applying a cosine transform to the contents of the 40 mel-scale bands.

TECHNIQUES FOR SPEAKER MODELLING

Once the speaker space is estimated by a vector of acoustic parameters, this vector defines the instantaneous nature of the speaker. The speaker model is built from a succession of these vectors representing a complete utterance. To date the motivation for much of the work of the project has been the assessment of a variety of speaker modelling techniques on the chosen dataset. The major speaker modelling techniques described in the literature for text independent material include long-term statistics of acoustic features, vector quantisation of the acoustic feature space, and ergodic hidden Markov modelling of acoustic feature sequences.

Long-term statistics can be applied directly only in situations where extended test data utterances of sufficient phonetic diversity are available to match against the long-term statistical models. This condition is not met within the test data constraints of our project.

Vector quantisation (VQ) modelling of the occupancy of the speaker space allows comparison with the model using data of limited phonetic diversity, although phonetically diverse data is required to train an adequate text-independent (TI) model. The 'VQ modelling' class of techniques is therefore well suited to the needs of the TRUST project in which training data can be plentiful but testing data may be very limited.

Ergodic hidden Markov modelling (EHMM) of the acoustic feature sequences is a natural extension of the VQ modelling technique, allowing some of the temporal characteristics of the speech acoustics to be encoded in the model parameters. Because EHMM models are potentially more complex, they can be expected to require larger amounts of data to train them effectively. As there is no inherent limit to the amount of training data that can be provided in this project, the potential advantages of these more complex models are also being explored.

ASSIGNING PROBABILITY OF IDENTITY

The utility of speaker models to assign a probability to the identity of the speaker of test utterances can be evaluated using a speaker identification or a speaker verification paradigm. VQ or HMM modelling can be used in either paradigm, generating VQ-distortion or HMM-maximum-likelihood measures respectively.

In speaker identification studies, where an incoming utterance is compared to each speaker model, the speaker of the model giving the minimum distortion or the maximum likelihood value is assigned to that utterance. The probability of speaker identification is then scored as the total number of correct assignments for a range of utterances (or repetitions of a fixed utterance) by a range of speakers. In this case the utterances and speakers used to test the model and the data used to train the speaker models need to be defined to enable comparison of modelling techniques.

In speaker verification studies the VQ distortion measure or the HMM likelihood measure must be compared to a threshold value which determines whether the incoming utterance should or should not be assigned to the speaker of the model against which it is compared. Speaker verification is scored in terms of how frequently utterances of the modelled speaker are rejected (type-I error) and how frequently utterances of other speakers are erroneously accepted by the model (type-II error). The overall performance of the system is most parsimoniously described by the equal error rate (EER), which is achieved when the threshold is set so as to obtain equal type-I and type-II error rates.

EXPERIMENTAL DIRECTIONS

Having initially collected the data corpus, speaker identification studies were performed in order to assess the performance of different speech analysis and speaker modelling techniques operating on that corpus. It was also important to estimate the adequacy of the corpus design so that ongoing data collection could be planned. Not only are the quantity and phonetic quality of the corpus important but also its ecological relevance. To investigate the necessary diversity to be covered in future data collection, studies in the noise robustness of the techniques used are also being performed.

Early in 1994 the focus of the work moved from identification studies to verification studies with the additional complexity of threshold determination and the need to estimate type-II errors from a small population size.

SPEAKER IDENTIFICATION

Analysis of the performance of speaker recognition techniques using the TRUST data corpus initially used a speaker identification paradigm while the number of speakers represented in the corpus was relatively small.

Initial studies examined the complexity of speaker modelling techniques required to give adequate performance. Cross-session training and testing using variance-weighted VQ codebooks and mixture-Gaussian VQ codebooks indicated very little difference between the performance of these two speaker modelling techniques and that codebook 'free-variables' (number of codewords or twice the number of mixtures) in excess of 128 gave no advantage (Zhu et al., 1994a). Using these techniques, closed-set recognition accuracies of 97-98% for the 25 initial speakers were achieved. Further evaluation of models involving mixture-Gaussian modelling using the ergodic HMM architecture failed to show any advantage of dividing the mixtures between HMM states and modelling the transitions between them.

The work of Furui (1981), indicating that speaker recognition results deteriorated with the temporal separation of the test data from the training data, were confirmed for the TRUST data corpus comparing test data that was recorded one week after the training data with test data recorded four weeks after the training data (Zhu et al., 1994a).

Acknowledging the varied utterance length in our data corpus, we examined the dependence of recognition accuracy on utterance length. Some utterances were artificially shortened to give utterances of 0.3s, 0.5s, 0.7s, 1.0s, 1.5s, 2.0s and 2.5s in duration, ranging from single syllables to 15 syllable sequences. While utterances of at least 2.0s gave near perfect recognition, utterances of 1.0s gave recognition performances which were equivalent to the average of all utterances in the data set. Quite

sharp deterioration of performance occurred when utterances of less that 0.7s were used. This result indicated that short single syllable commands may give performance close to 90% for this closed-set speaker identification task rather than the 97-98% reported on average.

ADEQUACY OF TRAINING DATA

An early test was performed to evaluate the efficacy of the 30 token data corpus in terms of its ability to represent the text-independent behaviour of our speakers (Zhu et al., 1994a). For this test, the 10 additional utterances collected in the second and third recording sessions were used as test data against speaker models built on the main data corpus. The 10 additional utterances were paired with a subset of the main corpus, selected on the basis of similar utterance length to the additional utterances, to reduce any duration effect as noted above. The use of the 10 tokens which did not occur in the training set increased the error rate from about 2.5% to just over 7% indicating that a training data corpus with greater phonetic coverage could be advantageous for true text-independent operation.

The data corpus was also designed with repetitions of the same tokens to enable the investigation of text-dependent speaker recognition on utterances of the type used. The amount of text-dependent data required to support adequate performance was investigated using a 6-state left-right HMM with four mixtures per state (Zhu et al., 1994b). Significant increases in text-dependent speaker identification performance were found for increasing training data up to 6 repetitions of each token, whereas further increases to 8 and 10 repetitions were not significant. As 5 repetitions were available from each recording session, the effect of within-session or cross-session training data was explored using testing data from a third session. For 4-repetition training, cross-session (2 from each session) data showed significant benefit for all utterances (averaged across speakers) and a positive, but non-significant, benefit for all speakers (averaged across utterances).

We had assumed that a left-right HMM would be preferable to an ergodic HMM for modelling fixed utterances. We found however that these two architectures gave very similar performance; analysis of variance was required to show that the LR-HMM benefit of some 0.5% was in fact statistically significant.

PHONETIC SPECIFICITY

In order to explore the relative merit of techniques which vary in the phonetic specificity of the modelling involved, speech data from the TIMIT corpus of American English was used as it has well-validated phonemic labelling required by one of the techniques. Mixture-Gaussian VQ, ergodic HMMs, and phone-based left-right HMMs were used on the same TIMIT data. Text-independent speaker recognition performance was consistently better with the latter technique (Zhu et al., 1994c). This result indicated that speaker recognition performance can be improved when combined with accurate speech recognition. It should be note that this work used a different parametric input (LPCC) to that indicated above and of course a different data corpus

NOISE ROBUSTNESS

Robustness of speaker recognition to noise is a key factor in many application environments. Accordingly, innovative techniques for reducing the impact of noise have been successfully evaluated. The impact of several types of noise from the NOISEX-92 noise data corpus was evaluated at three different levels of signal-to-noise ratio. For example, using multispeaker babble noise at SNR=13dB, speaker identification performance of 97-98% on clean speech had degraded to 61% but was restored to 92% following the use of a novel noise reduction filter (Tang et al., 1994).

SPEAKER VERIFICATION

A major problem for speaker verification is the difficulty in choosing an appropriate threshold. The current approach in the literature is to use a 'cohort' of speakers to provide a local environment for each client speaker which can be used to normalise distortion or likelihood measures against such variability. The use of the standard cohort method on cross-session testing for 12 male and 12 female clients using our data corpus showed an advantage, but it was not statistically significant across all these client speakers. However when the range of impostors tackled with this method was reduced to those who were actually close to acceptance by the client model, the EER was significantly reduced from 5.5% to 1.3% (Chen et

al., 1994).

TYPE-II ERRORS

The nature of the impostor population and the structure of the measurement space linking them to client models is a major unknown in the field of speaker verification. An attempt has been made to understand more about the general characteristics of the inter-speaker distance space by using the TIMIT corpus, which has 462 training speakers drawn from 8 dialect types of American English. The standard analysis described above was applied to these data and a VQ codebook derived for a client speaker from each dialect group. The client speakers were chosen to equally represent male and female and to have a spread of mean fundamental frequency uniformly spanning 87Hz to 263Hz. The study showed that while there was no significant contribution from dialect region to VQ-distortion, there was a significant contribution from relative mean fundamental frequency. Asymmetries in the nature of this significant contribution have been reported but not yet fully understood (Wagner et al., 1994).

CONCLUSIONS

The overall conclusions of the work reported above is that in computer systems that incorporate speech recognition functions already, the speech medium is very suitable for transparent assertion of user identity. This has been demonstrated under laboratory conditions for short utterances typical of commands to a 'windows' graphical user interface. The results so far lay a firm foundation for a range of further studies which will address issues of robustness with respect to both client variance over time and intrusion by a wider range of impostors.

In 1995 the project will move to examine speech data derived from 'live use' of computer systems. This will represent a quantum leap forward in terms of the volume of data for the training and testing of speaker models as well as optimising the ecological validity of the style of speech used.

ACKNOWLEDGEMENT

This work has been carried out on behalf of the Harry Triguboff AM Research Syndicate.

REFERENCES

References to the wider literature on speaker recognition have been omitted from this paper but are available in the reference lists of the papers produced by project staff which are listed below.

Chen,F., Millar,B., and Wagner,M. (1994) Hybrid threshold approach in text-independent speaker verification, Proc. Int. Conf. Spoken Language Processing, Yokohama, 1855-1858.

Furui,S. (1981) Cepstral analysis techniques for automatic speaker verification, IEEE Trans. ASSP, Vol.29, No.2, 254-272.

Tang,H., Zhu,X., Macleod,I., Millar,B. and Wagner,M. (1994) A dynamic-window weighted-RMS averaging filter applied to speaker identification, Proc. Int. Conf. Spoken Language Processing, Yokohama, 1603-1606.

Wagner,M., Chen,F., Macleod,I., Millar,B., Ran,S., Tridgell,A., and Zhu,X. (1994) Analysis of type-II errors for VQ-distortion based speaker verification, Proc. ESCA workshop on Automatic Speaker Recognition, Identification, and Verification, Martigny, 83-86.

Zhu,X., Gao,Y., Ran,S., Chen,F., Macleod,I., Millar,B., Wagner,M. (1994a) Text-independent speaker recognition using VQ, mixture Gaussian VQ and ergodic HMMs, Proc. ESCA workshop on Automatic Speaker Recognition, Identification, and Verification, Martigny, 55-58.

Zhu,X., Macleod,I., Millar,B. (1994b) The effect of training data on the performance of an HMM-based speaker recognition system, Proc. Int. Conf. Systems, Control, Information - Methodologies and Applications, Wuhan, China (to appear).

Zhu,X., Millar,B., Macleod,I., Wagner,M., Chen,F., Ran,S. (1994c) A comparative study of mixture-Gaussian VQ, ergodic HMMs, and left-to-right HMMs for speaker recognition, Proc.IEEE International Symposium on Speech, Image Processing and Neural Networks, Hong Kong, 618-621.