Authenticating student work in an e-learning programme via speaker recognition

Barry Hayes
Institute for Energy Systems
University of Edinburgh
Edinburgh EH9 3JY, Scotland
Email: b.hayes@ed.ac.uk

John Ringwood
Dept. of Electronic Engineering
National University of Ireland, Maynooth
Maynooth, Co. Kildare, Ireland
Email: john.ringwood@eeng.nuim.ie

***Abstract*—The past decade has seen the proliferation of e-learning and distance learning programs across a wealth of discipline areas. In order to preserve maximum flexibility in outreach, student assessment based exclusively on remotely submitted work has become commonplace. However, there is also growing evidence that e-learning also provides increased opportunity for plagiarism, with obvious consequences for learning effectiveness. This paper reports on the development of a prototype student authentication system, designed for use with a graduate e-learning program. The proposed system can be used to authenticate telephone-based oral examination which can, in turn, be used to confirm a student's ability in relation to submitted assignments and on-line test results. The prototype low-cost system is shown to be sufficiently accurate to act as an effective deterrent against plagiarism.**

***Index Terms*—E-learning, plagiarism, speaker recognition**

I. Introduction

The past decade has seen the proliferation of distance learning programs facilitated by e-learning environments. Much has been written on the requirements of effective e-learning systems and the special needs of remote students, in both trying to recreate the positive aspects of the traditional classroom environment as well as addressing particular difficulties and opportunities posed by e-learning-mediated programs. It is widely accepted that assessment forms an integral part of the learning experience [1], which is especially true of e-learning systems, where peer-pressure or face-to-face teacher-student interaction [2] may be absent. While it is clear that the asynchronous form of assessment usually associated with e-learning distance education programmes permits greater flexibility, it is also evident that on-line assessment, coupled with the ease of access and communication of electronically-held information, permits a greater possibility of plagiarism and cheating [3], [4], [5].

In order to achieve an effective and useful e-learning assessment methodology, some balance must be reached between the quality and integrity of assessment versus the implied workload on staff and students. This compromise, in turn, must be balanced by the probability of plagiarism or impersonation. To this end, we propose a system largely based on submitted assignment reports, where the provision for oral examination on any of the assessment material is reserved. The outstanding problem, given that the remote student is unlikely to have been met by university staff, is to ensure that the oral examination is being conducted with the student who originally registered for the programme.

Biometric identification has recently gained much attention, where unique, invariable biological characteristics of a person (fingerprints, voice, face, handwriting, etc.) are used to authenticate a user; the idea being that if the user him/herself is the key, the possibility of stealing or duplicating the key no longer exists.

In e-learning applications, the most obvious biometric characteristic to use is voice, since voice is easily available via telephone communication. In order to provide the maximum flexibility, speaker verification across a variety of communication channels, including land-lines mobile networks and voice over internet protocol (VoIP), should be allowed.

The paper is laid out as follows: Section II discusses the general problem of speaker verification, alongside the related problems of speech and speaker recognition. Section III briefly describes the graduate programme the system is designed to work with, while Section IV details the hardware and software requirements of the speaker verification system. Sections V, VI and VII deal with the technical aspects of the data logging and organisation, pre-processing and feature extraction, while the speaker verification performance is documented in Section VIII. Conclusions are drawn in Section IX.

II. Speaker recognition and verification

Speaker recognition [6] is concerned with recognising characteristics of the user - i.e. who is speaking. When we communicate with each other, even if we are not familiar with the person, we can in most cases recognise attributes such as the gender of the speaker, the accent in which the speaker is talking, emotional state etc. [7]. This paper focuses solely on the identity of the user. Speaker recognition is usually further divided into two separate tasks, speaker identification and speaker verification.

In speaker identification [8], the task is to determine which person the voice belongs to i.e. which one out of a set of possible categories is present. These categories may be a closed set, where it is assumed that the speaker must belong to one of the categories, and the task is to determine which category the speaker most likely belongs to. The categories

978-1-4244-4398-7/09/$25.00 ©2009 IEEE

may be an open set, where the possibility also exists that the speaker does not belong to any of the categories and should therefore be rejected.

Finally, in the speaker verification task [9], one already "knows" who the speaker is. The verification task is to match the input speech utterance to the voiceprint (i.e. the sample(s) taken during enrolment/registration), and return a decision which either accepts or rejects the speaker. This verification task is a 1:1 matching problem, and can be viewed as a special case of the open-set speaker identification problem. While speaker identification and speaker verification are quite different problems in terms of classification, many of the algorithms used in the front-end of the system (i.e. normalisation, noise removal, time-matching, and feature extraction) are similar for both tasks.

A number of challenges exist in the development of any speaker verification system. One of the major practical difficulties relates to intra-individual variation. A speaker's voice can change significantly from session to session for a number of reasons:

- Physical state(e.g. head cold, tiredness),
- mental/emotional state (happiness, nervousness, depression etc.), and
- other long-term changes due to aging and physiological condition.

Technical error sources also degrade system performance, arising from either environmental noise or channel noise. Background noise during the recording, for example, noise from an office environment (coughs, keyboard clicks, footsteps, people speaking nearby, etc.) contributes to environmental noise. Another phenomenon, which should be considered, is the Lombard effect [10], which describes the way in which the user will naturally change their style of speech to compensate for noisier environments. The acoustics of the room in which the speech is recorded in a given session can be another source of error (reverberations, echoes, etc.) and can add unwanted components to the input signal.

III. Educational programme context

The speaker verification system described in this paper is designed to be used in conjunction with a graduate programme offered by the Electronic Engineering Dept. at the National University of Ireland (NUI), Maynooth, Ireland. The Master of Engineering (ME) in Electronic Engineering is offered on both a full-time and part-time basis and also on an in-house and e-learning (remote) basis. The ME consists of 8 taught modules, rated each at 7.5 credits on the European Credit Transfer System (ECTS), and a research project of 30 ECTS credits. The full-time programme covers a calendar year, with 4 taught modules per 12-week academic semester, with the project completed over the summer months. In part-time mode, 2 taught modules are taken per semester, with the project completed over the 2 summer periods. Further information on the ME programme is available at: http://www.eeng.nuim.ie/courses/ postgraduate/me_ee.html.

The mode of assessment is tailored to suit each taught module, but generally consists of some mixture of written assessment and examination. Some components require the development of presentations. For remote students, written assignments are submitted electronically and the examinations are on-line.

IV. Hardware and software requirements

The broad intention was to build a system which could be run on a personal computer (PC). In order for the system to be operationally effective, the system would need to be able to perform a verification, prior to oral examination, in no more than 20 seconds and have a verification accuracy of better than 90%, with minimisation of false negatives i.e. the true person is not misclassified as an imposter. In view of these system specifications, a hardware system was assembled, consisting of a PC based on a Pentium 4 processor running at 1.6 GHz with 1GB of RAM.

In terms of telephone line and interfacing requirements, the system was to be based in NUI Maynooth, which runs an Ericsson digital Private Branch Exchange (PBX). This PBX supports primarily digital extensions, with (simulated) analogue lines available on request. The simulated analogue line was selected, in view of:

- The difficulty in getting information on the exact Ericsson protocol used on the digital lines,
- the long lead time and comparatively high installation and rental costs on a dedicated (true) analogue line, though this would provide better quality, and
- the flexibility of using an analogue line with a range of PC telephone interface cards.

A number of proprietary telephone interface cards are available and the Dialogic D/4PCIUF Combined Media board telephony board was found to be a relatively cheap, reliable solution which could be integrated easily with the proposed interactive voice response (IVR) software. As a front end, the IVR software allows interactive voice programs to be designed quickly and easily using a graphical user interface (GUI). The VoiceGuide IVR software was selected, which records telephone signals to a 64Kbps, 8 kHz, PCM-coded .wav file, and provides facilities to automatically answer an incoming call, prompt the user for test utterances and record the samples to a specified location on a computer hard drive.

V. Data collection and database organisation

In assembling a speaker verification system, care must be taken in the specification of the voice sequences which are to be recorded, if the verification system is to operate successfully.

A. Data collection

A number of telephone speech databases are available for the purposes of speaker identification and verification, such as the MIT Mobile Device Speaker Verification Corpus [10] and the YOHO Voice Verification Corpus [11]. These databases contain a huge number of reference voice samples which

#	Segment	Use/properties
1	User's name	For reference
2	Line type	Identification of channel type
3	Name x 3	Name repetitions for verification
4	Numbers x 3	Number sequence 6-6-1-0 for verification
5	Phrase I x 3	Phrase 'Chocolate fudge' for verification
6	Phrase II x 3	Phrase 'Mint chocolate chip' for verification

TABLE I
DATABASE RECORDS FOR EACH TEST USER

can potentially be used for the development and testing of new speaker verification algorithms. However, these databases typically use high-quality handsets and often have no channel effects, which diminish their utility in this application.

In order to create a dedicated voice sample database, decisions needed to be made regarding the source type and the type of utterances which would be useful for speaker verification. We decided not to limit the system to any one type of telephone line; the samples were to be collected across three categories - landline (standard plain old telephone system [POTS] line); mobile line (GSM); and VoIP (from a voice-over-internet provider such as Skype). Naik *et al* [9] and Lamel and Gauvain [12] use three different types of utterances:

- The student's name,
- a sequence of numbers, and
- phonetically-balanced phrases.

Three utterances and three channel types were used as a starting point for our database. Two of the "ice-cream flavours" from the MIT database [10], "chocolate fudge" and "mint chocolate chip", were chosen as phonetically balanced phrases, as they are both short and phonetically rich. Voice samples for the database were sought from both males and females, as well as from a range of age groups and accents.

B. The Speaker Verification Database

Table I gives a list of the samples recorded for each user. Each user was referenced by name, grouped into one of the three line type categories (landline, GSM mobile line or VoIP), and assigned a call number based on whether it was the user's first, second or third call to the line to record samples.

In terms of file and directory organisation, speech segments are classified, in order of hierarchy, according to:

- The name of the user,
- the call number for that user (note that users can call multiple times),
- the type of phone channel being used,
- the nature of the speech segment i.e. name, number or phrase, and
- the repetition index for that utterance.

using a file with format:

IntialSurname_Line type_Call number/SegmentRepetition

being created for each segment. For example, if J. Bloggs makes his first call to the system via a VoIP line, the file jbloggs_v_1/phrase11.wav will be created for the first utterance of Phrase I ("Chocolate fudge").

VI. AUDIO PRE-PROCESSING

A number of issues associated with the raw voice signal exist which need to be resolved by pre-processing. The signal

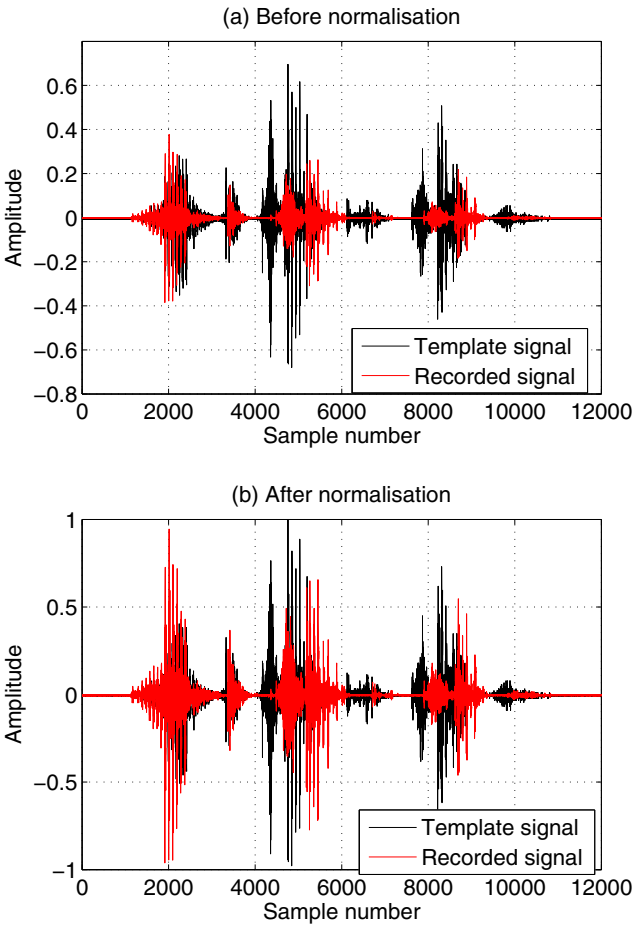

Fig. 1. Amplitude normalization

is corrupted with both channel and environmental noise, the time signatures may be different between successive utterances, a certain vowel sound may be shorter or longer in a given instance and utterances can vary considerably in volume. In addition, periods of silence occur at the beginning and end of each spoken phrase. These "silent" parts contain only background noise; a means of detecting where the speech audio begins and ends is needed in order to remove them. In general, intra-speaker and inter-session variability should be minimised.

A. Amplitude normalization

Amplitude normalization was performed by attenuating/amplifying the audio segments to a range of ± 1 (normalised units). The effect is demonstrated in Fig.1.

B. Filtering

The effectiveness of the spectral subtraction method relies on an accurate estimation of the noise power in the signal. A typical speaker's audio sample consists of 40% speech and 60% speech pauses [19]. If a voice activity detection algorithm can be employed to separate the speech from the speech pauses in the audio file, an accurate estimate of the background noise can be obtained easily. However, voice activation algorithms

are difficult to implement, and generally have issues recognising unvoiced phonemes [13]. The method used in our approach is based on the minimum-statistics algorithm outlined in [14]. For each frequency band, the smallest power spectral density estimate of the input signal, observed in a sufficiently large number of consecutive frames, contains the noise component only. These minima are tracked in a sliding window covering several frames to give an estimate of the noise magnitude spectrum.

C. Start and endpoint detection

The non-speech parts at the beginning and end of each voice file need to be removed; these parts generally only contain noise, and do not carry any useful information about the speaker. Trimming away long silent parts in the audio also makes the time normalisation (discussed in the following section) easier. An automatic method of efficiently removing the non-speech parts of the waveform is required.

One approach is to write code to simply truncate the file where the audio level drops below a certain amplitude. However, simple truncation was found to remove small parts of speech for some files in our database. The approach taken instead was to window the signal (for example, take every 100 samples as a frame) and find the amplitude peak in each frame. If the amplitude peak of the first or the last frame is below a pre-defined threshold, the corresponding samples are removed from the audio segment. This procedure is carried out iteratively until the peaks in both the first and last frames are above the threshold. Some experimentation was required to find the optimum window length and the amplitude threshold (these values were set at 100 samples and 2% of maximum amplitude respectively).

This detection method works on the assumption that the intensity of the spoken phrase is significantly greater than that of the noise background and that any noises in the silent regions, such as coughs or breathing sounds are either of low enough volume or short enough duration not to be confused with speech. More sophisticated methods of segmentation, based on the spectral properties of speech and non-speech sounds [13], are available, but the algorithm suggested works effectively for samples with a reasonable signal-to-noise ratio.

D. Time normalization

In practice, speakers generally vary their speed of talking in a non-uniform manner [7] e.g. a vowel sound may last longer in one sample than in the next. The impact of these time variations on speaker verification performance can be reduced by time-aligning the samples at the pre-processing stage [15]. If a consistent test phrase is used, we can attempt to match the word from the test phrase to the corresponding sample from the training phase and an optimisation algorithm can be used to calculate a non-linear timescale distortion to a word in order to achieve the best match to a template word at all points. In such *dynamic time-warping*, the short-term Fourier transform (STFT) is first calculated for both samples and a "local match" score matrix is constructed as the cosine distance between the STFT magnitudes. A window length of 20ms with an overlap of 25% is used to calculate the STFT in this application.

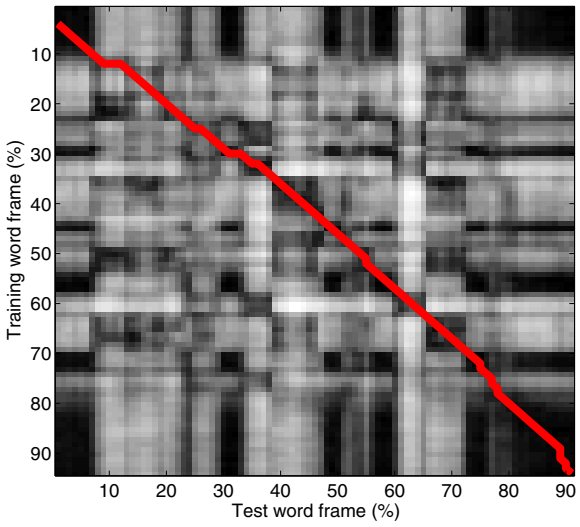

Fig. 2. Match matrix for dynamic time warping

We employ dynamic programming to find the optimum path between the opposite corners of the cost matrix; i.e. the one which has the minimum total difference between the two patterns (the code used for the dynamic programming algorithm was taken from [16]). This path is indicated by the red line in Fig. 2. Note that this line follows the dark stripe which runs roughly diagonally through the scores matrix. The dark areas represent a small distance between the two patterns, while the brighter areas represent larger distances.

Fig.3 shows the results of the time normalisation algorithm.

VII. Feature Extraction

Feature extraction transforms a raw speech input into a set of feature vectors - a compact and effective representation which is designed to be more stable and discriminative than the original signal [17]. The speech signal contains low-level properties, such as intensity, pitch, formant frequencies and bandwidths, and spectral coefficients. Ideally, the features selected should be easy to measure, stable over time, have a high inter-speaker and low intra-speaker variation and be robust against noise and distortion.

Speech production can be modelled by the source-filter model proposed in [18]. The mechanism which produces speech sounds is made up of two components: The source, which produces the airstream coming up from the larynx, and the filter, which represents the vocal tract.

A. Cepstral analysis

In speech applications, the main purpose of cepstral analysis is to separate the source, or excitation component, in speech from the filter component. The cepstrum essentially involves the 'spectrum of a spectrum', though some applications use the inverse Fourier transform as the final transform element.

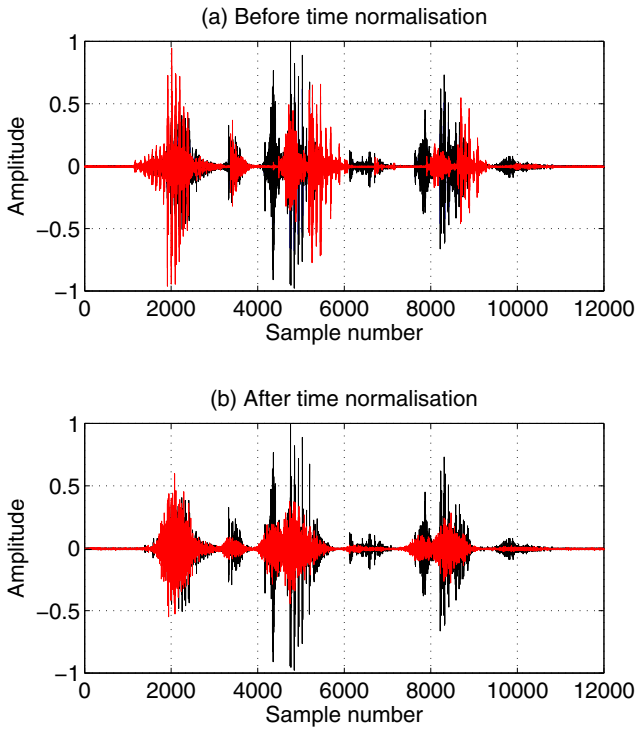

Fig. 3. Time normalization

Specifically, for the digital signal, $s(k)$, the cepstrum is evaluated [19] as:

$$c_x(n) = F^{-1}(log_{10}|F(s(k))|) \quad (1)$$

where $F(\,)$ and $F^{-1}(\,)$ denote the Fourier and inverse-Fourier transform, respectively.

Eq. (1) converts the speech signal into a pseudo-time domain known as quefrency. In this domain, the convoluted slow-varying (the vocal tract filter) and the fast-varying (the excitation or source) components are separated. By retaining only the first few cepstral coefficients , we focus on components of the spectral envelope which contain useful (and relatively consistent) features about the speaker.

1) LP Cepstral Coefficients: Linear predictive coding (LPC) is also based on the source-filter model, with the filter constrained to be an all-pole filter. The analysis performs a linear prediction so that the next sample is predicted by using a weighted sum of past samples:

$$\hat{s}_k = \sum_{i=1}^{p} a(i) s_{k-i} \quad (2)$$

where p is the predictor order and $a(i)$ are the filter coefficients, which can be calculated using correlation analysis. In practice, raw LPC coefficients are rarely used as features, due to the high correlation between adjacent coefficients. Instead, complex cepstrum coefficients are often used, which can be computed easily from the LP coefficients using:

$$c(n) = \begin{cases} a(n) + \sum_{j=1}^{n-1} \frac{j}{n} c(j) a(n-j) & , \quad 1 \le n \le p \\ \sum_{j=1}^{n-1} \frac{j}{n} c(j) a(n-j) & , \quad n > p \end{cases} \quad (3)$$

2) Mel-frequency Cepstral Coefficients: Another popular technique for extracting useful features from speech is to use a filterbank-based cepstral representation. A bank of 15-20 channels, or bandpass filters, whose bandwidth and spacing increase with frequency is generally used, motivated by studies of the human ear. The filterbank represents power logarithmically, which is of phonetic significance - the lower formants are emphasized more. The distribution in each of the channels tends to be Gaussian [7]. The locations of the center frequencies of the filters are given by:

$$f_{\text{mel}} = \frac{1000 \;\; log_{10}(1 + \frac{f_{\text{linear}}}{1000})}{log_{10}(2)} \quad (4)$$

The mel-frequency cepstral coefficients (MFCCs) are obtained using the following procedure:

1) The FFT is applied to the signal, or a windowed version of it,
2) Spectral power values are then mapped onto the mel scale using (4),
3) The logarithm is taken of the spectral powers, and
4) The final spectral representation is obtained using the discrete cosine transform (DCT) which, for P channels, is computed as in eq. (5).

$$c_j = \sum_{j=1}^{P} S_j cos\left(n(j - \frac{1}{2})\frac{\pi}{P}\right) \quad , \quad n = 1, 2, \ldots, N \quad (5)$$

where N is the total number of cepstral coefficients.

B. Feature selection

In order to select the appropriate number of LP and/or mel coefficients to use, the inter-speaker minus intra-speaker cosh spectral distance [20], averaged over a number of speakers, was used as a selection criterion. Fig. 4 shows the metric for various numbers of LP and mel coefficients. Given the presence of 'elbow points' at coefficients 8 and 13 for LP and mel respectively, with resulting diminished contribution of each additional coefficient thereafter, little benefit in choosing more than 8 LP and 13 mel coefficients is apparent.

VIII. Speaker Classification

While the cosh spectral distance can provide a good level of discrimination for inter- and intra-speaker comparisons, cosh spectral distance is a general metric and better discrimination can likely be obtained using a bespoke classifier trained using a supervised learning technique.

A. Dataset and Training

In total, 160 data records were used, broken down into training, validation and test as follows: 28/7/40. Overall, 60 male and 15 female records were used, with the distribution over landline/mobile/VOIP of 37/24/14.

A multi-layer perceptron (MLP) was used to implement the classifier. MLP's demonstrate good global approximation abilities, with a relatively small neuron count, particularly for a significant number of inputs (21 in our case). The MLP

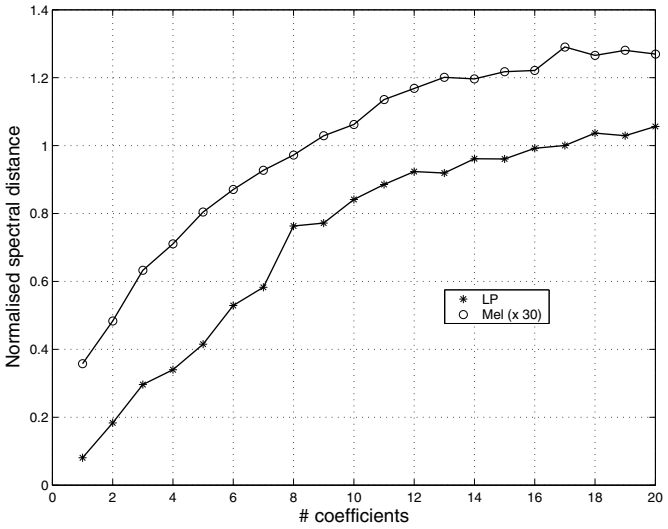

Fig. 4. Spectral metric variation with # coefficients

Total samples	Correctly classified	False positives	False negatives	Accuracy
160	156	2	2	97.5%

TABLE II
CLASSIFICATION ACCURACY

was trained using the Levenberg-Marquardt algorithm, which uses a second-order gradient (Hessian) estimate to achieve fast convergence. Early stopping of the training, based on examination of a performance measure on a validation data set, was used to ensure good generalisation. A number of MLP network structures were examined of the form 21-N_{L1}-N_{L2}-1, with N_{L1} and N_{L2} corresponding to the number of neurons in hidden layers 1 and 2, respectively. Following 20 trials for each network configuration (to eliminate sensitivity to initial conditions), the network structure was finally optimised for $N_{L1} = 3$ and $N_{L2} = 4$.

B. Classifier results

In order to assess the performance of the classifier, four sample student 'models' were constructed, based on the 'mint chocolate chip' phrase, giving a total of 160 (4 x 40) test vectors. Table II shows the aggregate performance of the four classifiers over the full set of test vectors.

IX. Conclusion

A student authentication system, based on telephone speech, has been developed which is designed to effectively eliminate potential plagiarism associated with the submission of assignments for e-learning programmes. The potential for an oral examination gives the programme director a level of confidence in the authenticity of submitted work and provides an effective level of deterrent against plagiarism. The final system has a very low capital requirement and can be easily implemented at a cost of approximately $300. The technical requirements on the student's side are minimal and the system can cope with land, mobile and VoIP lines, without any requirement for a high-quality handset. The final classification accuracy achieved (of 97.5%) is sufficient for the system to be a very effective deterrent and the classifier can be biased, if desired, in order to eliminate false positives (speaker who is not the real student is incorrectly classified as the real student) or false negatives (real student is incorrectly classified as another person).

Acknowledgement

The authors would like to thank Denis Buckley and John Maloco of the Electronic Engineering Dept at NUI Maynooth for their assistance with the project. The sample voice records used in the project were generously provided by the staff and students of the Electronic Eng. Dept. of NUI Maynooth.

References

[1] M. Thorpe, "Assessment and 'third generation' distance education," *Distance Education*, vol. 19, no. 2, pp. 265–286, 1998.
[2] A. Rovai, "Online and traditional assessments: what is the difference?" *The Internet and Higher Education*, vol. 3, no. 3, pp. 141–151, 2000.
[3] J. Cordova and P. Thornhill, "Academic honesty and electronic assessment: tools to prevent students from cheating online—tutorial presentation," *Journal of Computing Sciences in Colleges*, vol. 22, no. 5, pp. 141–151, 2007.
[4] F. Graf, "Providing security for elearning," *Computers and Graphics*, vol. 26, no. 2, pp. 355–365, 2002.
[5] J. Underwood and A. Szabo, "Academic offences and e-learning: individual propensities in cheating," *British Journal of Educational Technology*, vol. 34, no. 4, pp. 467–477, 2003.
[6] J. Campbell, "Speaker recognition: A tutorial," *IEEE Proceedings*, vol. 85, no. 9, pp. 1437–1462, 1997.
[7] W. Holmes, *Speech Synthesis and Recognition*, 2nd ed. Taylor and Francis, 2001.
[8] H. Gish and M. Schmidt, "Text-independent speaker identification," *IEEE Signal Processing Magazine*, vol. 11, no. 4, pp. 18–32, 1994.
[9] J. Naik, "Speaker verification: a tutorial," *IEEE Communications Magazine*, vol. 21, no. 1, pp. 42–48, 1990.
[10] R. Woo, A. Park, and T. Hazen, "The MIT mobile device speaker verification corpus: Data collection and preliminary experiments," *IEEE Odyssey: Speaker and Language Recognition Workshop*, pp. 1–6, June 2006.
[11] A. Higgins, J. Porter, and L. Bahler, *YOHO Speaker Authentication - Final Report*. ITT Defense Communications Division, 1989.
[12] L. F. Lamel and J. L. Gauvain, "Speaker verification over the telephone," *Speech Commun.*, vol. 31, no. 2-3, pp. 141–154, 2000.
[13] L. Jian-bin, Y. Ji-Kun, Z. Hui, and N. Zhong-Xia, "Two-stage speech/non-speech classification of telephone signals," in *Proc. Intl. Conf. on Comms, Circuits and Systems*, Guilin, China, 2006, pp. 490–492.
[14] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. on Speech and Audio Processing*, vol. 9, no. 5, pp. 504–512, 2001.
[15] V. Vuckovic, "Dynamic time-warping method for isolated speech sequence recognition," in *Proc. Intl. Conf. on Telecomms. in Modern Satellite, Cable and Broadcasting Services (TELSIKS)*, Nis, Yugoslavia, Sept. 2001, pp. 257–260.
[16] D. Ellis, "Matlab audio processing resources, lab for recognition and organization speech and audio," Columbia University, US.
[17] T. Kinnunen, "Spectral features for automatic text independent speaker recognition," PhD Thesis, University of Joensuu, Finland, 2003.
[18] G. Fant, *Acoustic Theory of Speech Production*. The Hague: Mouton, 1960.
[19] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals*. Bell Laboratories, 1978.
[20] J. Gray, A. and J. Markel, "Distance measures for speech processing," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 24, no. 5, pp. 380–391, Oct 1976.