# Student Authentication for Oral Assessment in Distance Learning Programs

Barry Hayes and John V. Ringwood, *Senior Member*, *IEEE*

**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 a telephone-based oral examination which can, in turn, be used to confirm a student's ability in relation to submitted assignments and online test results. The prototype low-cost system is shown to be sufficiently accurate to act as an effective deterrent against plagiarism.

**Index Terms**—Distance learning, e-learning, plagiarism, oral assessment, authentication.

## 1 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, both in 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 programs permits greater flexibility, it is also evident that online 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 program.

- *B. Hayes is with Intel (Ireland) Ltd., Leixlip, Co. Kildare, Ireland. E-mail: barry.hayes@gmail.com.*
- *J.V. Ringwood is with the Department of Electronic Engineering, NUI Maynooth, Maynooth, Co. Kildare, Ireland. E-mail: john.ringwood@eeng.nuim.ie.*

*Manuscript received 11 Nov. 2008; revised 30 Dec. 2008; accepted 5 Jan. 2009; published online 8 Jan. 2009.*
*For information on obtaining reprints of this article, please send e-mail to: lt@computer.org, and reference IEEECS Log Number TLT-2008-11-0099.*
*Digital Object Identifier no. 10.1109/TLT.2009.2.*

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. While some biometric measurement requires relatively complex hardware and processing capability (e.g., fingerprinting, iris scanning), many useful biometric measurement systems can be based on relatively common hardware devices, such as microphones and cameras, with associated modest data processing requirements.

In e-learning applications, the most obvious biometric characteristic to use is voice, since voice is easily available via telephone communication. While a combination of voice and image/video is likely to produce a more reliable authentication system, such a combination is likely to considerably restrict the student clientele due to the relatively narrow deployment of video telephony equipment. This study therefore focuses on the use of biometric verification via speech records where the student, upon registration, provides a voice "signature" against which verification can be made immediately prior to oral assessment. 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.

Ultimately, any student verification system can be cracked, given sufficient resources and determination. For example, an "alias," who is possibly an expert in the subject area, could provide the requisite voiceprint at registration and be the respondent at any potential oral examinations. However, this would require an "alias" who is willing to:

- be available for the full duration of the e-learning program,
- cover all of the topics in the e-learning program, and
- complete all the assignments for the true student.

1939-1382/08/$25.00 © 2008 IEEE Published by the IEEE CS & ES

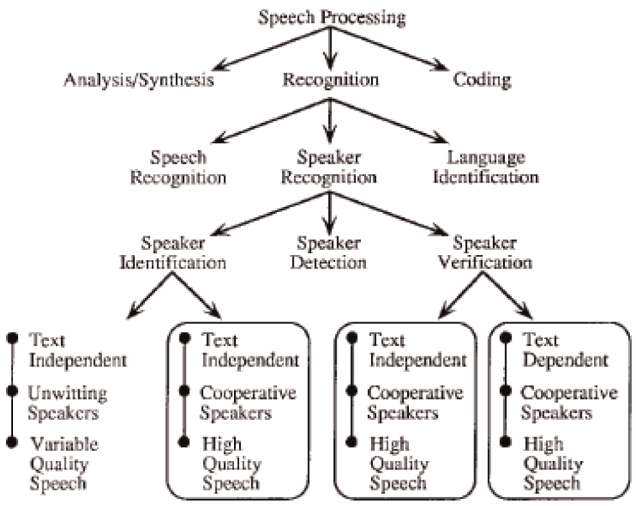


Fig. 1. Speech and speaker recognition and verification [6].

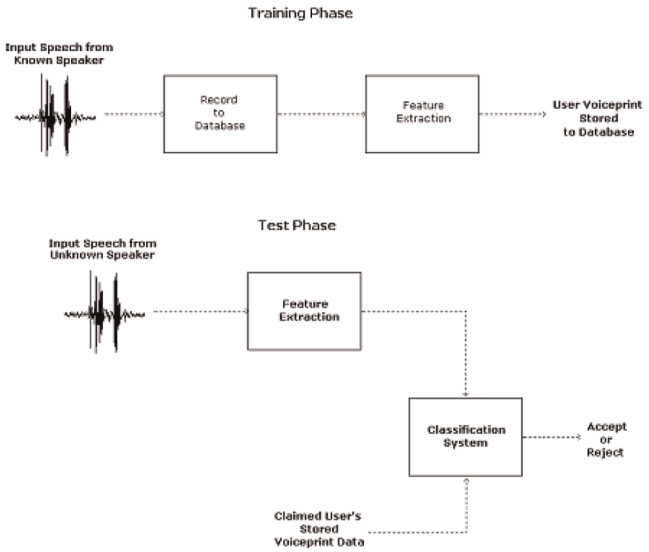


Fig. 2. Speaker verification system.

Ultimately, we feel that a voice authentication system, coupled with a provision for oral examination, provides an appropriate measure of deterrent for an e-learning program. It remains, then, to assess if a voice authentication system can provide an appropriate level of fidelity in confirming that the person undergoing oral assessment is the same as the student who registered for the program. Such an assessment is the main focus of the paper, which is laid out as follows: Section 2 discusses the general problem of speaker verification, alongside the related problems of speech and speaker recognition. Section 3 briefly describes the graduate program the system is designed to work with, while Section 4 details the hardware and software requirements of the speaker verification system. Sections 5, 6, and 7 deal with the technical aspects of the data logging and organization, preprocessing, and feature extraction, while the speaker verification performance is documented in Section 8. Conclusions are drawn in Section 9.

## 2 Speech and Speaker Recognition and Verification

When humans communicate with each other, we can generally recognize immediately both what is being said (provided that we are familiar enough with the language being used) and who is speaking (provided that we are familiar enough with the speaker). For our purposes, it is very important to make the distinction between these two different tasks. Speech recognition [7] is the task of recognizing what is being said; essentially, it is the conversion from speech to text. Speech recognition has been studied extensively since the 1950s, and while many commercially available applications exist, a universal speech recognition system is not yet available that will work for an unlimited vocabulary and for all speakers [8].

Speaker recognition [6], on the other hand, has to do with recognizing 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, recognize attributes such as the gender of the speaker, the accent with which the speaker is talking, their emotional state, etc. [9]. 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 [10], 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 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 [11], 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., normalization, noise removal, time-matching, and feature extraction) are similar for both tasks. The relationship between speech and speaker recognition and verification is summarized in Fig. 1 [6], while the specific tasks involved in speaker verification are detailed in Fig. 2.

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.

It is therefore important that the features extracted from the speech input are robust to these intraspeaker changes. Some applications (see, for example, [12]) describe a method of refining the model after each session to improve overall performance in the long-term.

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. The input speech content will be very different, for instance, if the speech is recorded in a quiet, isolated room compared to the same speech recorded on a busy street. Another phenomenon which should be considered is the Lombard effect [13], 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. The microphone used in each session and the specific transmission channel are also major sources of variation. In an ideal setup for speaker recognition, the same high-quality microphone should be used in training and at each subsequent session. However, for the purposes of our telephone-based system, a wide range of different telephone handsets could potentially be used. While channel effects present significant difficulties, poor quality microphones introduce nonlinear distortion into the signal and cause "phantom" formants, which tend to occur at sums and multiples of the real formant frequencies. Much research has been carried out to examine the effect of various microphone types on the verification task [13].

The technical error sources described above become most problematic for a speaker verification system when the conditions between the training and test phases are *mismatched*. The use of multistyle training, where the user must generate utterances in a range of environments using a variety of microphones, can greatly improve robustness [14]. The trade-off between recognition accuracy and the user-friendliness of a given speaker recognition system is universally acknowledged [15]. This trade-off has potential implications for restricting the range of speaker sources in an effort to achieve suitable accuracy and accessibility.

## 3 Educational Program Context

The speaker verification system described in this paper is designed to be used in conjunction with a graduate program offered by the Electronic Engineering Department 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 eight taught modules, each rated at 7.5 credits on the European Credit Transfer System (ECTS), and a research project of 30 ECTS credits. The full-time program covers a calendar year, with four taught modules per 12-week academic semester, with the project completed over the summer months. In part-time mode, two taught modules are taken per semester, with the project completed over the two summer periods. Further information on the ME program 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 online. For remote students, presentations are delivered using some form of slide management environment, along with an audio track. The Moodle [16] environment is used to manage electronic course content, as well as providing facilities for assignment collection and management of online examinations. Discussion forums are also available, along with a range of other features to enhance class/instructor interactivity. The remote ME program is designed specifically so that no on-campus attendance is mandatory, either for teaching or assessment, thus maximizing the geographical scope and flexibility for remote students. For students outside Ireland, VoIP is an attractive option for oral examination due to its relatively low cost and, therefore, needs to be facilitated in the speaker recognition system.

### 3.1 Operational Procedure

The authentication system is designed to be used as follows:

1. The system is installed and commissioned by appropriate technical staff, which is easily accomplished by staff with general software skills.
2. Annually, the authentication system is updated (by the technical staff), in consultation with administrative staff, to update the list of students (list of current students or those whose applications have been accepted).
3. Upon registration, each student on the ME program is required to deposit a "voiceprint," consisting of a standard phrase (see Section 7.2), in the student database. The voiceprint logging is facilitated by automated response software, as described in Section 5.1, and records are stored for use in conjunction with the student authentication system. Note that each student's registration is not deemed to be completed until their voiceprint has been successfully logged, sending a clear message to students that plagiarism is taken seriously and counteracting measures are in place. Note that the rationale behind the logging of the voiceprint is clearly explained to students.
4. Preprocessing is now carried out (offline), via the execution of a macro utility, on the student voiceprints to minimize the wait time during the use of the live authentication system. The preprocessing procedure is articulated in Section 6.

   The system is now ready for use.
5. In the ME program, each module is coordinated by a "module coordinator" (MCO) who has overall module responsibility in the case of multiple instructors on the module. Each MCO makes an assessment of the risk of plagiarism for that module, based on:

- the nature of the material,
- the degree to which group assignments are employed, and
- the degree to which assignments are individualized.

Based on this assessment, the MCO establishes a sample size appropriate for oral testing. In any event, a minimum of 1/10th of the student enrollment for each module is recommended for oral testing.

6. Oral testing, with the support of the student authentication system, is carried out in the last quarter of the semester (each teaching semester is 12 weeks long).
7. For each oral test, the call is initiated by the MCO at a time agreed with the student. Following the initial contact, the student is switched to the voice authentication system, which requests a voiceprint (using the same standard phrase as in Step 3). The voiceprint is then preprocessed and submitted to the classifier (see Section 8), which returns a level of confidence (LoC) that the student being examined is the same student which registered for the program. The computation due to preprocessing and classification may take up to 15-20 seconds. Following the authentication result, the MCO should proceed as follows:

   - If LoC $> 0.7$, the oral test proceeds, since authentication is confirmed.
   - If $0.3 \leq$ LoC $\leq 0.7$, the MCO can recheck authentication.
   - If LoC $< 0.3$, authentication has failed and the situation is reported to the Departmental Disciplinary Committee.

## 4 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 percent with minimization 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 64 Kbps, 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.

The interface to students, therefore, whether they are calling from a landline, mobile phone, or VoIP channel, is a single telephone number. For the system developer, the IVR software provides a set of .wav speech files which can be used both for the training of the verification system and in the operational "student verification" mode.

## 5 Data Collection and Database Organization

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

### 5.1 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 [13] and the YOHO Voice Verification Corpus [17]. These databases contain a huge number of reference voice samples which 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 diminishes 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 [11] and Lamel and Gauvain [18] 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 flavors" from the MIT database [13], "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

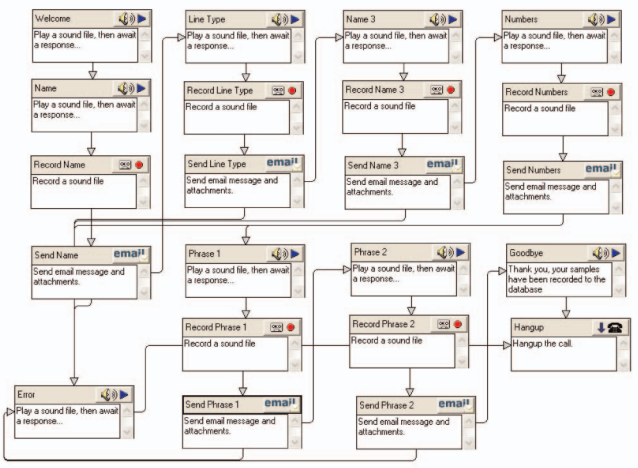


Fig. 3. Automatic call logging with VoiceGuide.

males and females, as well as from a range of age groups and accents.

The VoiceGuide software package has a graphical user interface which allows interactive voice programs to be designed quickly and easily. In order to collect training data, a series of prompts were recorded using a standard PC microphone for use in the script. Each subject dials the line, with the script being triggered to start when an incoming call is received. The script answers the call automatically and asks the caller to record their name and the type of telephone line they are calling from, and then to repeat a number of utterances which are recorded for use as voice samples. Each recorded file is stored to the hard drive under a predefined filename and also sent via an e-mail client to a dedicated e-mail account. Files can then be retrieved either from the hard drive or downloaded from the email account remotely, which allows the test subjects to record voice samples at their own convenience and also allows the developer to access all of the recorded material remotely at any time. The VoiceGuide interface for automatic call logging is shown in Fig. 3.

### 5.2 The Speaker Verification Database

Table 1 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 organization, 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:

$$\text{IntialSurname\_Line type\_Call number/SegmentRepetition}$$

TABLE 1
Database Records for Each Test User

| # | 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 |

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").

## 6 Audio Preprocessing

A number of issues associated with the raw voice signal exist which need to be resolved by preprocessing. The signal 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, intraspeaker and intersession variability should be minimized.

### 6.1 Amplitude Normalization

Amplitude normalization was performed by attenuating/amplifying the audio segments to a range of $\pm 1$ (normalized units). The effect is demonstrated in Fig. 4.

### 6.2 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 percent speech and 60 percent 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 recognizing unvoiced phonemes [19]. The method used in our approach is based on the minimum-statistics algorithm outlined in [20]. 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.

### 6.3 Start and Endpoint Detection

The nonspeech 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 normalization (discussed in the following section) easier to carry out and can be done by manually

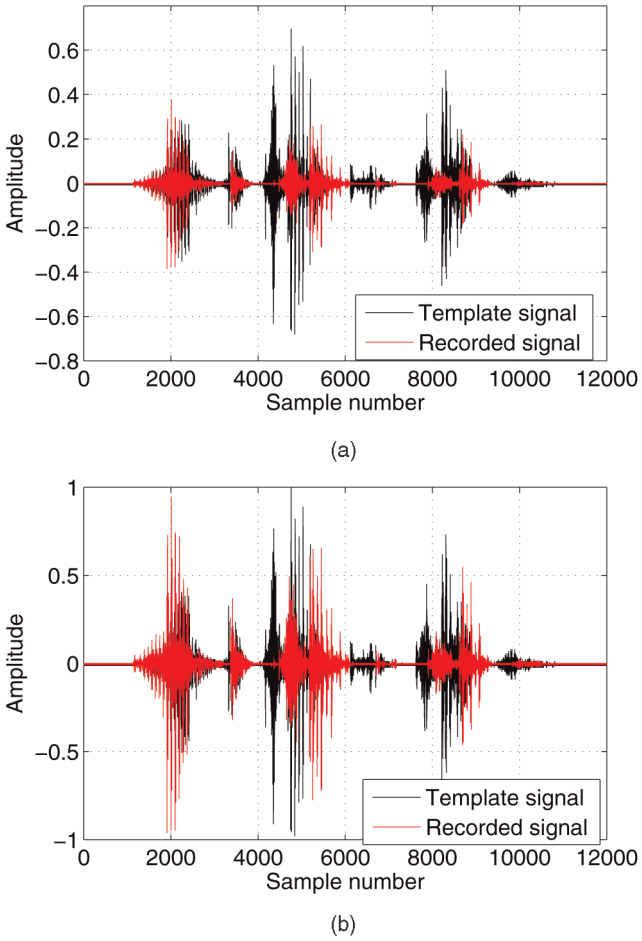


Fig. 4. Amplitude normalization. (a) Before amplitude normalization. (b) After amplitude normalization.

deleting the unwanted parts of the waveform using sound editing software. However, manual deletion is tedious and time consuming, especially when there are a large number of audio files to process. An automatic method of efficiently removing the nonspeech 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 predefined 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. By way of example, Fig. 5b shows the peaks in each windowed segment after the first iteration of the algorithm, with Figs. 5a and 5c showing the corresponding raw and processed signals, respectively. The dashed line in Fig. 5b (at a threshold of 0.025) represents the threshold below which the signal is classified as not belonging to the speech part of the waveform. In the case of Fig. 5b, both the first and last windowed segments are removed and the algorithm continues, applying the window and the threshold to the signal again. Some experimentation was required to

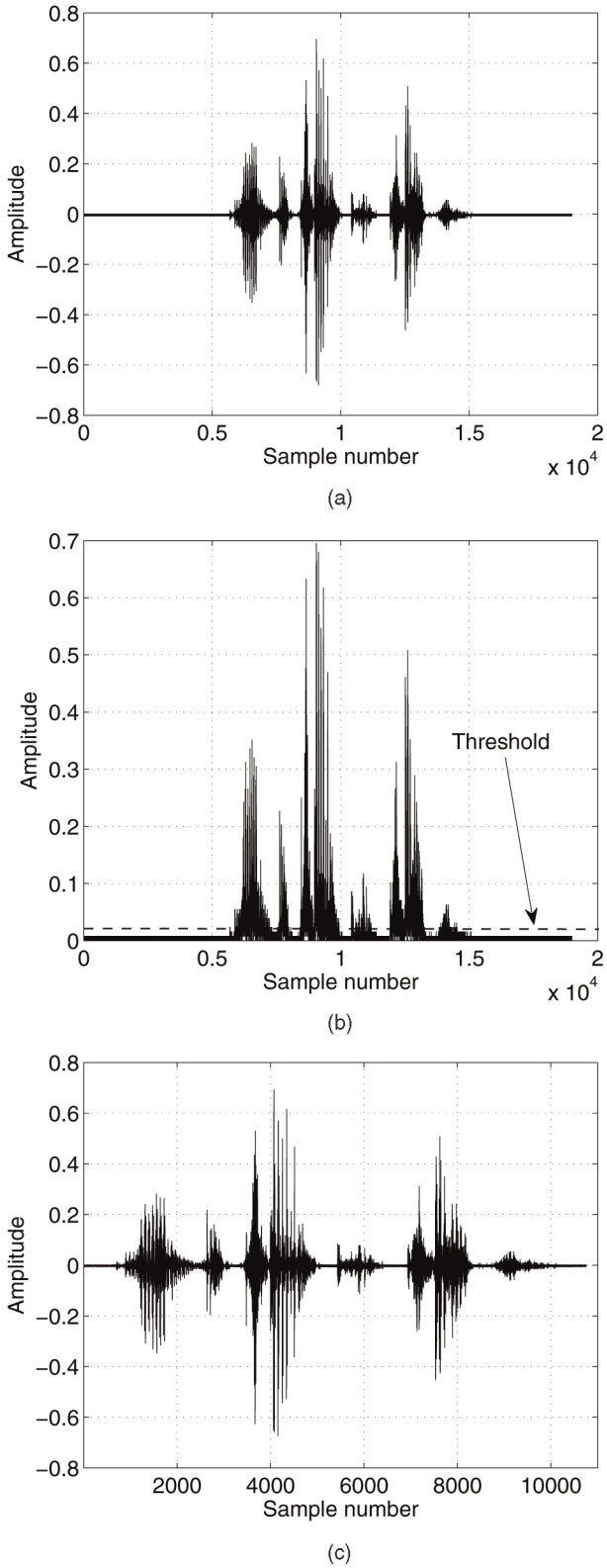


Fig. 5. Start and endpoint detection. (a) Raw signal. (b) Peak calculation. (c) Processed signal.

find the optimum window length and the amplitude threshold (these values were set at 100 samples and 2 percent of maximum amplitude, respectively).

This detection method works on the assumption that the intensity of the spoken phrase is significantly greater than

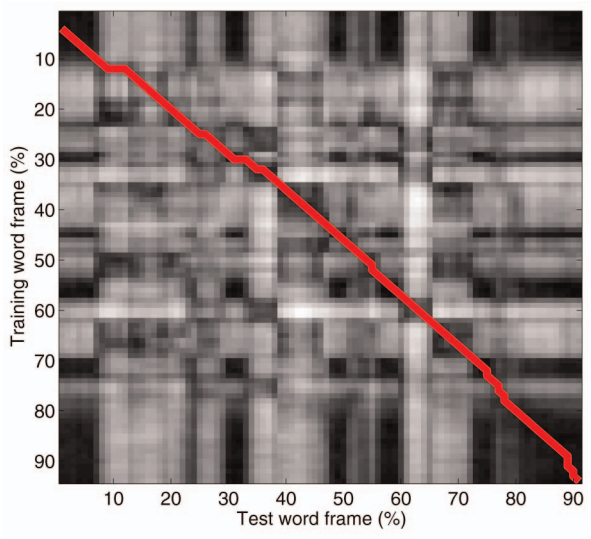


Fig. 6. Match matrix for dynamic time warping.

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 nonspeech sounds [19], are available, but the algorithm suggested works effectively for samples with a good signal-to-noise ratio.

### 6.4 Time Normalization

In practice, speakers generally vary their speed of talking in a nonuniform manner [9], 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 preprocessing stage [21]. 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 optimization algorithm can be used to calculate a nonlinear 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 20 ms with an overlap of 25 percent is used to calculate the STFT in this application.

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 Ellis [22]). This path is indicated by the red line in Fig. 6. 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. 7 shows the results of the time normalization algorithm.

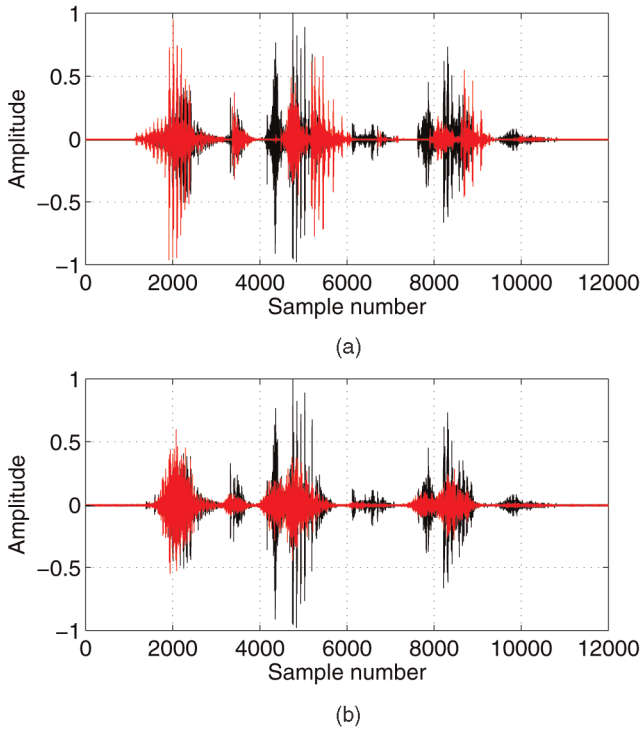


Fig. 7. Time normalization. (a) Before time normalization. (b) After time normalization.

## 7 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 [8]. We must carefully choose the correct features for our application; any subsequent speaker classification system can only be as good as the features presented to the system. The speech signal contains a number of high-level properties, such as accent, dialect, emotion, and speaking style. These features tend to be robust to most types of noise, but are very difficult to extract effectively from a short speech sample without the use of a human expert. The speech signal also contains low-level properties, such as intensity, pitch, formant frequencies and bandwidths, and spectral coefficients. These features are easier to extract and to use to build up a model of the speaker to be verified. Ideally, the features selected need to meet the following criteria:

- easy to measure,
- stable over time,
- high interspeaker variation,
- low intraspeaker variation, and
- robust against noise and distortion.

Speech production can be modeled by the source-filter model proposed in [23]. 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. Sometimes, the vocal tract is modeled as a series of tubes through which the sound flows, each with its own resonant frequency. Such a model is analogous to a time-varying acoustic filter,

TABLE 2
Spectral Distances Based on LP and MEL Cepstral Coefficients

| S # | M/F | Line | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | M | M | **1.96** | 16.64 | 4.76 | 6.32 | 18.96 | 6.27 | 6.46 | 3.19 | 8.09 | 10.41 | 11.54 | 6.18 | 5.84 | 5.69 | 4.17 |
| **2** | M | V | 6.89 | **1.46** | 7.42 | 5.44 | 2.48 | 8.71 | 5.32 | 5.79 | 11.66 | 7.71 | 6.47 | 21.68 | 3.55 | 8.45 | 16.7 |
| **3** | M | L | 4.02 | 5.9 | **1.18** | 3 | 3.91 | 7.58 | 2.97 | 2.49 | 12.97 | 3.17 | 2.63 | 7.73 | 2.52 | 2.97 | 5.26 |
| **4** | F | M | 4.94 | 5.3 | 3.53 | **1.9** | 3.58 | 4.85 | 3.11 | 2.73 | 8.92 | 4.12 | 6.13 | 4.45 | 3.11 | 3.82 | 16.83 |
| **5** | M | V | 14.59 | 3.8 | 7.68 | 4.65 | **1.92** | 4.75 | 6.59 | 5.28 | 10.88 | 4.51 | 6.32 | 14.5 | 6.43 | 5.5 | 11.16 |
| **6** | F | M | 12.71 | 13.85 | 10.12 | 7.74 | 5.8 | **2.17** | 7.3 | 6.42 | 7.35 | 4.55 | 12.34 | 9.01 | 10.4 | 4.76 | 11.88 |
| **7** | M | L | 2.9 | 4.46 | 2.32 | 3.11 | 2.85 | 5.27 | **2.06** | 2.85 | 10.85 | 4.14 | 11.08 | 4.54 | 3.58 | 3.37 | 5.42 |
| **8** | M | L | 3.24 | 8.97 | 1.87 | 3.18 | 3.97 | 5.94 | 2.4 | **1.76** | 8.43 | 3.27 | 3.55 | 6.93 | 3.87 | 2.98 | 7.47 |
| **9** | F | M | 10.65 | 20.46 | 10.62 | 10.23 | 9.6 | 6.71 | 13.32 | 10.25 | **3.11** | 8.45 | 12.39 | 13.02 | 10.47 | 7.51 | 19.88 |
| **10** | F | L | 8.15 | 7.41 | 8 | 5.42 | 3.3 | 3.95 | 5.63 | 3.36 | 7.74 | **1.1** | 6.22 | 6.15 | 4.16 | 3.31 | 12.81 |
| **11** | M | M | 11.66 | 6.19 | 4.96 | 5.55 | 4.92 | 8.91 | 6.59 | 6.82 | 16.67 | 8.36 | **4.00** | 14.75 | 4.24 | 7.07 | 13.56 |
| **12** | F | M | 5.49 | 9.45 | 4.73 | 3.7 | 5.12 | 9.86 | 5.97 | 4.52 | 8.18 | 4.29 | 8.08 | **1.89** | 3.77 | 5.32 | 9.17 |
| **13** | M | V | 23.04 | 4.82 | 3.9 | 6.34 | 3.81 | 12.72 | 20.9 | 25.99 | 18.43 | 11.28 | 4.33 | 24.59 | **1.71** | 8.36 | 15.73 |
| **14** | F | L | 6.43 | 6.69 | 5.11 | 5.43 | 5.7 | 3.92 | 6.87 | 4.95 | 7.97 | 6.5 | 7.69 | 9.14 | 9.45 | **2.81** | 8.22 |
| **15** | M | L | 9.08 | 15.4 | 5.68 | 12.09 | 12.86 | 9.38 | 8.2 | 5.95 | 18.31 | 9.76 | 18.93 | 11.89 | 10.64 | 5.13 | **2.68** |

which shapes the sound produced as it is produced by the speaker. Speaker recognition is based on the idea that each speaker has their own unique "filter," the characteristics of which can be used for identification.

### 7.1 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. Specifically, for the digital signal, $s(k)$, the cepstrum is evaluated [24] as:

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

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

Equation (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.

#### 7.1.1 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:

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

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

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

#### 7.1.2 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 to 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 [9]. The locations of the center frequencies of the filters are given by:

$$f_{\text{mel}} = \frac{10^3 \;\; log_{10}(1 + \frac{f_{\text{linear}}}{10^3})}{log_{10}(2)}. \tag{4}$$

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 (5):

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

where $N$ is the total number of cepstral coefficients.

### 7.2 Feature Selection

Table 2 shows the cosh spectral distances [25], calculated inter and intraspeaker for a combination of mel and LP cepstral coefficients. The "M/F" column denotes the gender of the speaker, while the "Line" column indicates whether the speaker is using a landline (L), mobile network (M), or

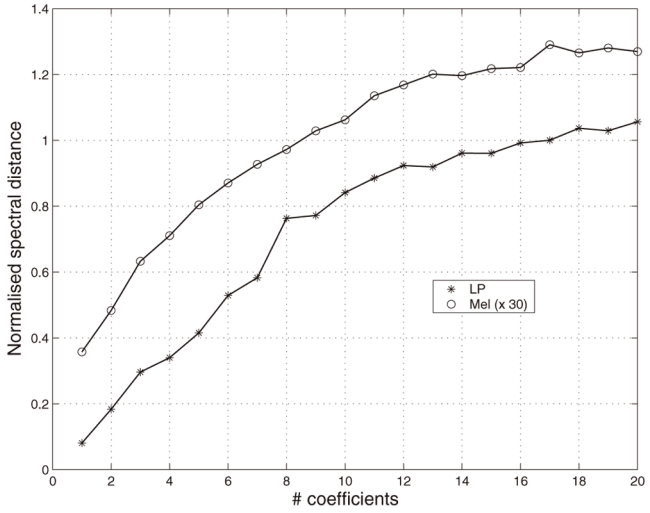


Fig. 8. Spectral metric variation with number of coefficients.

VoIP connection (V). The distance metric suggests that the mel and LP coefficients can provide good discrimination between intra and interspeaker comparisons and are reasonably insensitive to gender and phone line types. At this stage, it was also noted that, in terms of the spectral distance metric, the phonetically-balanced phrases ("chocolate fudge" and "mint chocolate chip") provided better discrimination than the user's name or number sequences.

In order to select the appropriate number of LP and/or mel coefficients to use, the interspeaker minus intraspeaker cosh spectral distance [25], averaged over a number of speakers, was used as a selection criterion. Fig. 8 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.

## 8 Speaker Classification

While the cosh spectral distance can provide a good level of discrimination for inter and intraspeaker 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. The structure of the classifier is shown in Fig. 9.

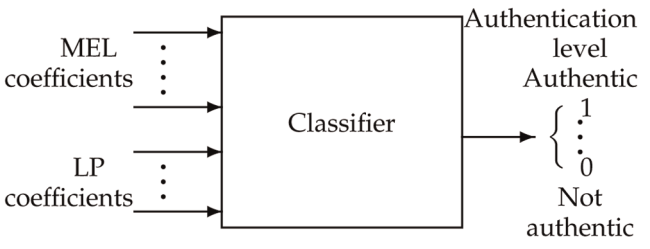


Fig. 9. Classifier structure.

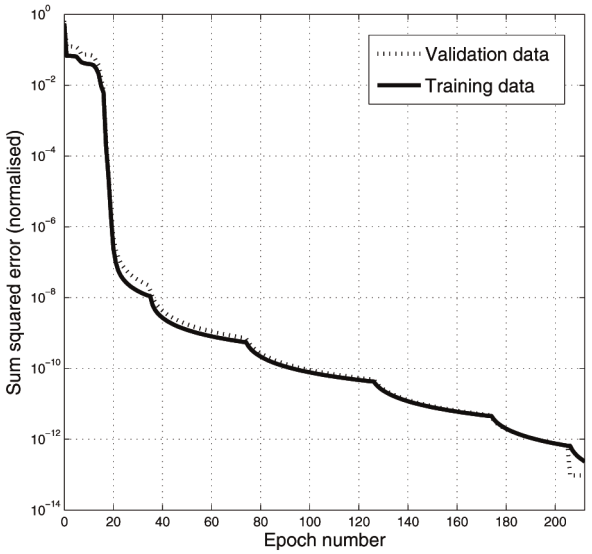


Fig. 10. MLP training evolution.

### 8.1 Data Set 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 multilayer perceptron (MLP) was used to implement the classifier. MLPs demonstrate good global approximation abilities, with a relatively small neuron count, particularly for a significant number of inputs (21 in our case). The MLP 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 generalization, i.e., training was terminated as soon as the validation performance began to deteriorate. Fig. 10 shows the training evolution for the final MLP used. 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 optimized for $N_{L1} = 3$ and $N_{L2} = 4$.

### 8.2 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 3 shows the aggregate performance of the four classifiers over the full set of test vectors.

TABLE 3
Classification Accuracy

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

Note that errors resulting in false positives (an imposter is impersonating the true student) and false negatives (the real student is incorrectly classified as an imposter) are equally weighted. If desired, the error weighting can be adjusted to eliminate either false positives or false negatives, as desired.

## 9 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 programs. The potential for an oral examination gives the program director a level of confidence in the authenticity of submitted work and provides an effective level of deterrent against plagiarism. While the system described in this paper uses a significant amount of signal processing technology, the final system has a very low capital requirement and can be easily implemented at a cost of approximately $300. Commissioning and maintenance of the system is straightforward. 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 percent) 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 (a speaker who is not the real student is incorrectly classified as the real student) or false negatives (a real student is incorrectly classified as another person). We also anticipate that the classifier accuracy can be improved with the addition of further training data; in particular, the class of "not the speaker" can be built up from legacy data built up over a number of semesters/years. The system was designed to operate with a graduate (masters) program in electronic engineering, delivered via e-learning, though some other potential applications may be found. For example, user authentication is a significant issue in telephone banking, though it is likely that a higher classification accuracy may be required (with correspondingly higher cost and complexity) in such a situation.

## Acknowledgments

The authors would like to thank Denis Buckley and John Maloco of the Electronic Engineering Department 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 Engineering Department 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," *J. 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 J. Educational Technology,* vol. 34, no. 4, pp. 467-477, 2003.
[6] J. Campbell, "Speaker Recognition: A Tutorial," *Proc. IEEE,* vol. 85, no. 9, pp. 1437-1462, 1997.
[7] C.-H. Lee, F. Soong, and K. Paliwal, *Automatic Speech and Speaker Recognition.* Springer, 1996.
[8] T. Kinnunen, "Spectral Features for Automatic Text Independent Speaker Recognition," PhD Thesis, Univ. of Joensuu, Finland, 2003.
[9] W. Holmes, *Speech Synthesis and Recognition,* second ed. Taylor and Francis, 2001.
[10] H. Gish and M. Schmidt, "Text-Independent Speaker Identification," *IEEE Signal Processing Magazine,* vol. 11, no. 4, pp. 18-32, 1994.
[11] J. Naik, "Speaker Verification: A Tutorial," *IEEE Comm. Magazine,* vol. 21, no. 1, pp. 42-48, 1990.
[12] A. Park and T.J. Hazen, "Asr Dependent Techniques for Speaker Identification," *Proc. Seventh Int'l Conf. Spoken Lang.,* pp. 1337-1340, Sept. 2002.
[13] R. Woo, A. Park, and T. Hazen, "The MIT Mobile Device Speaker Verification Corpus: Data Collection and Preliminary Experiments," *Proc. IEEE Odyssey: Speaker and Language Recognition Workshop,* pp. 1-6, June 2006.
[14] W. Hueng and B. Rao, "Channel and Noise Compensation for Text Dependent Speaker Verification over Telephone," *Proc. Int'l Conf. Acoustics, Speech, and Signal Processing,* vol. 1, pp. 337-340, May 1995.
[15] F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-García, D. Petrovska-Delacrõtaz, and D.A. Reynolds, "A Tutorial on Text-Independent Speaker Verification," *EURASIP J. Applied Signal Processing,* vol. 2004, no. 4, pp. 430-451, 2004.
[16] M. Dougiamas and P. Taylor, "Moodle: Using Learning Communities to Create an Open Source Course Management System," *Proc. World Conf. Educational Multimedia, Hypermedia and Telecomm.,* pp. 171-178, http://go.editlib.org/p/13739, 2003.
[17] A. Higgins, J. Porter, and L. Bahler, *YOHO Speaker Authentication—Final Report.* ITT Defense Comm. Division, 1989.
[18] L.F. Lamel and J.L. Gauvain, "Speaker Verification over the Telephone," *Speech Comm.,* vol. 31, nos. 2-3, pp. 141-154, 2000.
[19] L. Jian-bin, Y. Ji-Kun, Z. Hui, and N. Zhong-Xia, "Two-Stage Speech/Non-Speech Classification of Telephone Signals," *Proc. Int'l Conf. Comms, Circuits and Systems,* pp. 490-492, 2006.
[20] R. Martin, "Noise Power Spectral Density Estimation Based on Optimal Smoothing and Minimum Statistics," *IEEE Trans. Speech and Audio Processing,* vol. 9, no. 5, pp. 504-512, 2001.
[21] V. Vuckovic, "Dynamic Time-Warping Method for Isolated Speech Sequence Recognition," *Proc. Int'l Conf. Telecomm. in Modern Satellite, Cable and Broadcasting Services (TELSIKS),* pp. 257-260, Sept. 2001.
[22] D. Ellis, "Matlab Audio Processing Examples," Lab for Recognition and Organization of Speech and Audio, Columbia Univ., http://labrosa.ee.columbia.edu/matlab, 2006.
[23] G. Fant, *Acoustic Theory of Speech Production.* Mouton, 1960.
[24] L. Rabiner and R. Schafer, *Digital Processing of Speech Signals.* Bell Lab., 1978.
[25] J. Gray, A. and, J. Markel, "Distance Measures for Speech Processing," *IEEE Trans. Acoustics, Speech, and Signal Processing,* vol. 24, no. 5, pp. 380-391, Oct. 1976.

**Barry Hayes** received the electrical and electronic engineering degree from University College Cork in 2005 and the MS degree in electronic engineering from the National University of Ireland (NUI), Maynooth, in 2008. He is currently employed in industry as a semiconductor process engineer at Intel Ireland's headquarters near Dublin. He is a member of the Institute of Engineers of Ireland.

**John V. Ringwood** received the electrical engineering diploma from the Dublin Institute of Technology and the PhD degree in control systems from Strathclyde University, Scotland, in 1981 and 1985, respectively. He was with the School of Electronic Engineering at Dublin City University from 1985 to 2000 and, during that time, held visiting positions at Massey University and the University of Auckland, New Zealand. He is currently a professor of electronic engineering with the National University of Ireland (NUI), Maynooth, and is the associate dean for engineering with the Faculty of Science and Engineering. He was the head of the Electronic Engineering Department at NUI Maynooth from 2000 until 2005, developing the department from a greenfield site. He has acted as a consultant to a number of companies in the power, servomechanism, and process industries. His interests cover a number of areas, including time series modelling, control of wave energy systems, control of plasma processes, biomedical engineering, and e-learning. He is a chartered engineer, a senior member of the IEEE, and a fellow of the Institution of Engineers of Ireland.