J Sign Process Syst (2016) 82:345–356
DOI 10.1007/s11265-015-1005-5

Text-Independent Speaker Identification Using Vowel Formants

Noor Almaadeed[1] · **Amar Aggoun**[2] · **Abbes Amira**[1,3]

Received: 16 February 2014 / Revised: 22 March 2015 / Accepted: 17 April 2015 / Published online: 5 May 2015

Abstract Automatic speaker identification has become a challenging research problem due to its wide variety of applications. Neural networks and audio-visual identification systems can be very powerful, but they have limitations related to the number of speakers. The performance drops gradually as more and more users are registered with the system. This paper proposes a scalable algorithm for real-time text-independent speaker identification based on vowel recognition. Vowel formants are unique across different speakers and reflect the vocal tract information of a particular speaker. The contribution of this paper is the design of a scalable system based on vowel formant filters and a scoring scheme for classification of an unseen instance. Mel-Frequency Cepstral Coefficients (MFCC) and Linear Predictive Coding (LPC) have both been analysed for comparison to extract vowel formants by windowing the given signal. All formants are filtered by known formant frequencies to separate the vowel formants for further processing. The formant frequencies of each speaker are collected during the training phase. A test signal is also processed in the same way to find vowel formants and compare them with the saved vowel formants to identify the speaker for the current signal. A score-based scheme allows the speaker with the highest matching formants to own the current signal. This model requires less than 100 bytes of data to be saved for each speaker to be identified, and can identify the speaker within a second. Tests conducted on multiple databases show that this score-based scheme outperforms the back propagation neural network and Gaussian mixture models. Usually, the longer the speech files, the more significant were the improvements in accuracy.

Keywords Vowel formants · Speaker identification · Vowel recognition · Linear predictive coding · Mel-frequency Cepstral coefficients

✉ Noor Almaadeed
n.alali@qu.edu.qa

[1] Department of Computer Science and Engineering, College of Engineering, Qatar University, Doha, Qatar

[2] Department of Computer Science and Technology, University of Bedfordshire, University Square, Luton LU1, 3JU, UK

[3] Department of Engineering and Computer Science, University of the West of Scotland, Paisley, UK

1 Introduction

The term *Speaker Recognition* [1] consists of *Speaker Identification* – the identification of the speaker speaking the current utterance – and *Speaker Verification* – the verification from the utterance of whether the speaker is who he claims to be. There are two types of speaker recognition, *Text-dependent* – in which the speaker is given a specific set of words to be uttered – and *Text-independent* – in which the speaker is recognised irrespective of what one is saying [2]. The current approach is aimed at *Text-independent Speaker Identification.*

A digital speech signal is a discrete-time signal sampled from a continuous-time signal that has been quantised upon analog-to-digital conversion. Each sample is represented by one or more bytes (e.g. one byte for a 256-level quantisation). This digitised discrete-time signal consists of different frequency values which represent the audio signal. It must be pre-processed to extract feature vectors that represent individual information for a particular speaker regardless of the content of the speech itself. A learning algorithm generalises these feature vectors for various speakers during training and verifies the speaker identity for a test signal during the test phase.

No two digital signals are the same, even with the same speaker and the same set of words, due to the variation in amplitude and pitch in a speaker's voice over different recordings. The environmental noise, the recording equipment, the speed at which the speaker speaks, and the varying psychological and physical states of the speaker, increase the complexity of this task. Text-independent identification further requires that the speaker is free to speak any set of words during testing. Therefore, the need arises for a generalised feature extraction strategy to extract text-independent features from a speech signal.

An audio formant refers to the frequency peaks in a speech signal. These peaks appear with different frequencies in a speech signal and are also called resonant frequencies. These frequencies resonate according to the vocal tract of the speaker. Vowel formants refer to the frequencies associated with vowel sounds in a language. It is well known that the two or three of the lowest vowel formants are sufficient to distinguish between vowels in most cases [3]. Fundamental frequency estimation is an essential requirement in systems for pitch-synchronous analysis, speech analysis/synthesis and speech coding. It has been reported that fundamental frequency can improve performance of a speech recognition system for a tonal language [4] and of a speaker identification system [5]. These formants correspond closely to the acoustic resonance frequencies created by a speaker's vocal tract and carry unique information specific to the speaker [6]. The relevance of the individual formants and resonances have been widely studied [7–10]. Vocal resonances are altered by the articulators to form distinguishable vowel sounds, and the peaks in the vowel spectra are the vocal formants. The term formant refers to peaks in the harmonic spectrum of a complex sound. They are usually associated with the formations of the speaker's vocal tract and they are essential components in the intelligibility of speech. The distinguishability of the vowel sounds across vowels as well as across users can be attributed to the differences mainly in their first three formant frequencies [11]. An illustration is shown in Fig. 1 for two different speakers (left and right). The tracheal air pressure from the lungs passes through the glottis to create sound with the source spectrum as shown in the topmost plot. The vocal tract response specific to the speaker (the middle plot) attenuates this spectrum and results in the output sound (bottom plot) from which the formants can be extracted.

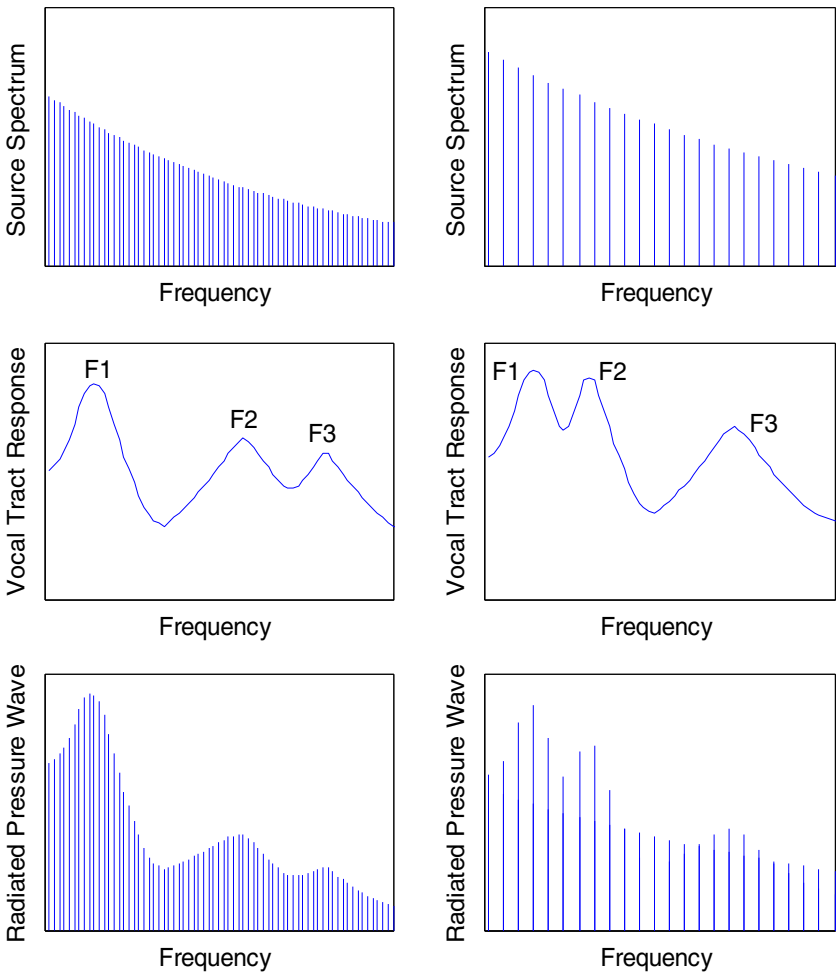


Figure 1 Resonating frequencies for different speakers.

The use of the classical Mel-frequency Cepstral Coefficients (MFCC) [12, 13] is likely the most popular feature extraction strategy extensively used thus far for speaker identification systems. However, MFCC does not contain much pitch information. Speech or speaker recognition systems transmit MFCC feature vectors directly to the speech recogniser. Fundamental frequency and spectral envelope derived from MFCC vectors are two necessary components for speech reconstruction [14]. Because feature vectors are a compact representation, optimised for discriminating between different speech sounds, they contain insufficient information to enable reconstruction of the original speech signal [15]. In particular, valuable speaker information, such as pitch, is lost in the transformation. It is therefore not possible to simply invert the stages involved in MFCC extraction to re-create the acoustic speech signal [12].

Therefore, in order to extract the vowel formants, the standard Linear Predictive Coding (LPC) scheme is used [16]. With LPC, all of the formants are extracted, with each formant portrayed in terms of three or more degree formants. These formants are filtered with a vowel formant filter that separates the vowel formants from the consonant formants.

During the training phase of the system, after processing the training signals, a vowel formant database is created that stores unique vowel formants for each speaker. To distinguish one speaker from another, vowel formants are tracked in the test file and are compared with the vowel formants database. A score-based scheme is employed that assigns the current signal to the speaker with the highest number of matching formants for the current test signal. This scoring scheme also follows a penalty rule, according to which, if a formant does not match the current vowel in hand from the test file, the speaker of that vowel formant is penalised with a negative score.

MFCC and LPC have both been analysed for comparison in the extraction of vowel formants. It is observed that LPC is more efficient in this task. This method is the most powerful way of estimating formants and is computationally the most efficient [17]. The reasons lie in the close resemblance of this strategy to the human vocal tract. LPC gives a recognition rate higher than MFCC and needs much less computational time. The algorithm to perform LPC on a speech signal is much simpler than that for MFCC, which has many parameters to be adjusted to smooth the spectrum, performing a processing that is similar to that executed by the human ear. But LPC is easily performed by the least squares method using a set of recursive formula [18].

For identification, the proposed score-based strategy has been compared with the *Back Propagation Neural Network* (BPNN) [19] and the *Gaussian Mixture Model* (GMM) [20]. GMM is a robust model for text-independent speaker identification as reported in [20]. A GMM is a parametric learning model and it assumes the process being modeled has the characteristics of a Gaussian process whose parameters do not change over time. When employing GMM over a segmented stream of speech signal, it is important that we assume that the frames are independent. This is a reasonable assumption since, generally, the text-independent systems are modeled as statistical speech parameter distribution models, which use GMM as the model of each speaker model as well as the universal background model (UBM) [20–22].

Although speech is non-stationary, it can be assumed quasi-stationary and be processed through the short-time Fourier analysis. In speech processing the short-time magnitude spectrum is believed to contain most of the information about speech intelligibility. The duration of the Hamming window function is an important choice. When making the quasi-stationarity assumption, we want the speech analysis segment to be stationary. We cannot make the speech analysis window too large, because the signal within the window will become non-stationary. On the other hand, making the window duration too small also has its disadvantages. If it is too small, then the frame shift decreases and thus the frame rate increases. This means we will be processing a lot more information than necessary, thus increasing the computational complexity. Also, making the window duration small will cause the spectral estimates to become less reliable due to the stochastic nature of the speech signal. Finally, a pitch pulse (typically with a frequency between 80 and 500 Hz) usually occurs every 2 to 12 ms. If the duration of the analysis window is smaller than the pitch period, then the pitch pulse may or may not be present. Hence, the shape and duration of the Hamming window is an important design criterion.

One of the most common classes of neural networks is the feed-forward network [19]. Back propagation refers to a common method by which these networks can be trained. Training is the process by which the weight matrix of a neural network is adjusted automatically to produce desirable results. Though back propagation is commonly used with feed-forward neural networks, it is by no means the only training method available for the feed-forward neural network. Back propagation works by calculating the overall error rate of a neural network. The output layer is then analyzed to see the contribution of each of the neurons to that error. The neurons' weights and threshold values are then adjusted, according to how much each neuron contributed to the error, to minimise the error next time. There are mainly two training parameters, the learning rate and the momentum, that can be passed to the back propagation algorithm to customise its output. The weights in a feed-forward neural network are adjusted according to the square errors between the actual outputs and the desired outputs. For a BPNN, these errors are propagated layer-by-layer into the input layer in the backward direction. The training input is passed through the network a number of times to adjust the weights accordingly. The iterative process of training the network requires multiple passes through the network to train it

correctly. A BPNN lets us recognise complex patterns and supports any number of training epochs to produce a learnt classifier for the unseen data. But this procedure has a cost associated with how much learning is required to perform classification within a predictable accuracy on the unseen data. Learning, although can be a great way of improving performance, requires computer resources for computation. BPNN not only requires long training time but also a huge number of instances/patterns to become fully trained for classification of the unseen data. It is observed that our score-based strategy outperforms both BPNN and GMM.

This paper is organised into six sections. Section 2 describes the scheme of feature extraction through vowel formants and the steps to create the formants database. Section 3 highlights the steps involved in the score-based scheme for speaker identification. Section 4 compares the performance results of these approaches on different speech databases, and finally, Section 6 concludes with a discussion of the effectiveness of the proposed scheme.

2 Feature Extraction

2.1 Formant Extraction Through LPC

The fundamental idea behind speech formants is the assumption that an audio signal is produced by a buzzer at the end of a tube that closely resembles the actual means of sound production in humans. The glottal portion produces the sound with the help of our breath pressure and acts as the buzzer, whereas the human vocal tract combined with the mouth constitutes the tube.

Audio speech can be fully described by the combination of its frequency graph and its loudness [1]. With this assumption, together with the vocal tract and mouth comprising the tube, the human voice is considered to consist of resonating frequencies called formants [23, 24]. LPC processes a signal in chunks or frames (20–30 ms) to extract these resonating frequencies or formants from the remainder of the noisy signal through inverse filtering [2, 25]. LPC analyses the speech signal by estimating the formants, removing their effects from the speech signal, and estimating their intensity and frequency. The process of removing the formants is called inverse filtering, and the remaining signal after the subtraction of the filtered modeled signal is called the residue.

For a given input sample x[n] and an output sample y[n], the next output sample y′[n] can be predicted with the following equation,

$$\mathrm{y}'[\mathrm{n}] = \sum_{k=0}^{q}(a_k \mathrm{xn}-\mathrm{k}) + \sum_{k=1}^{q}(b_k \mathrm{y}[\mathrm{n}-\mathrm{k}]) \quad (1)$$

The coefficients a_k and b_k above correspond to the linear predictive coefficients. The difference between the predicted sample and the actual sample is called the *prediction error* given as,

$$e[n] = y[n]-y'[n] \quad (2)$$

and hence, $y[n]$ can be written as,

$$y[n] = e[n]-\sum_{k=1}^{q} b_k\, y[n-k] \quad (3)$$

The linear predictive coefficients b_k are estimated using an autocorrelation method that minimises the error using least-square error reduction [26, 27].

2.2 Vowel Formant Filtering

In human speech, there are consonant formants and vowel formants, and there are noise reverberations. Of all of these sound types, we are only interested in the vowel formants. There are twelve vowel formant sounds in the English language, as concluded by a study at the Dept. of Phonetics and Linguistics, University College London [28]. These vowel formants, together with their first, second, and third formant frequency ranges, are listed in Table 1. The vowel formants are filtered using these ranges.

Vowel formants and frequencies were first exhaustively studied and formulated by J.C. Wells in the early sixties [23]. This was one of the few approaches researchers in the speaker identification field started investigating. In speech synthesis [24, 29], digital filters are often used to simulate formant filtering by the vocal tract. It is well known [30] that the different vowel sounds of speech can be simulated by passing a “buzz source” through only two or three formant filters. In principle, the formant filter sections are in series, as found by deriving the transfer function of an acoustic tube [31].

The basic process of vowel formant extraction is shown in Fig. 2. The speech signal, *s(n)* is first modulated by a windowing function, *w(n).* The modulation is typically amplitude modulation (i.e. multiplication) and the most commonly used windowing function is the Hamming window. The resultant signal *x(n)* is fed to the LPC [32]. After performing linear predictive coding, the LP coefficients $\overline{\alpha} = \left[\alpha_1\ \alpha_2\ \alpha_3 \ldots\ \alpha_p\right]$ are computed such that the error $e(n)=s(n)-\sum_{k=1}^{p}\alpha_k s(n-k)$ is minimised. The vector $\overline{\alpha}$ is appended with zeros and the spectral envelope is extracted using discrete Fourier transform (DFT), from which the peak signal is detected. The formants can then be acquired upon analyzing the peak signal amplitude and frequency. As described in Section 2.1, LPC performs an inverse filtering to remove the formants and extract the residue signal. It effectively synthesises the speech signal by reversing its process of formation as depicted in Fig. 1. The synthesised speech signal can then be examined to find the resonance peaks from the filter coefficients as well as the

Table 1 Vowel formant frequencies in English language [28].

Vowel	Formant	Mean frequency (Hz)	Std. dev.
/i/	1	285	46
	2	2373	166
	3	3088	217
/I/	1	356	54
	2	2098	111
	3	2696	132
/E/	1	569	48
	2	1965	124
	3	2636	139
/æ/	1	748	101
	2	1746	103
	3	2460	123
/A/	1	677	95
	2	1083	118
	3	2340	187
/Q/	1	599	67
	2	891	159
	3	2605	219
/O/	1	449	66
	2	737	85
	3	2635	183
/U/	1	376	62
	2	950	109
	3	2440	144
/u/	1	309	37
	2	939	142
	3	2320	141
/V/	1	722	105
	2	1236	70
	3	2537	176
/3/	1	581	46
	2	1381	76
	3	2436	231

prediction polynomial. This is a robust method since it allows the extraction of formant parameters through simple peak detection as shown in Fig. 2.

There are several methods of filtering vowel formants. It can be done by passing it through a series of bandpass filters in the audio frequency domain and systematically varying the filter width and slope [33, 34], by low-pass or high-pass filtering in the temporal modulation domain [35, 36], or by varying the number of audio-frequency channels in the context of cochlear implant simulations [37, 38]. In [36], it has been demonstrated that both low-pass and high-pass filtering in the temporal modulation domain were analogous to a uniform reduction in the spectral modulation domain.

Figure 3 shows a high-level flow chart for vowel extraction from speech signal. Vowels are highly periodic and have distinctive Fourier representations. We passed the test samples through an auto-regressive filter, and then calculated the formant frequencies from the spectral envelope of the LPC filtered vowel. The purpose of the auto-regressive model on each window is to get the transfer function of the vocal tract and output the spectral envelope of each voice sample. The next step is to filter out the consonants by checking the frequency response of the LP filter representing the consonant sounds. Consonants usually have significantly lower magnitudes than vowel sounds. A smoother is then used to eliminate anomalies and then output each vowel, from which the formants are extracted.

2.3 Vowel Database Construction

Vowel formants individually lie in specific frequency ranges, but every speaker has a unique vocal tract and produces vowel formants that are unique. During the training phase, the system is presented with speech files produced by different speakers. The speech files for each speaker are preprocessed with LPC, and subsequently, these formants are filtered to extract only the vowel formants. These vowel formants are saved for each speaker name in a database. This database is a Matlab [39] file to be used during the testing phase of the system.

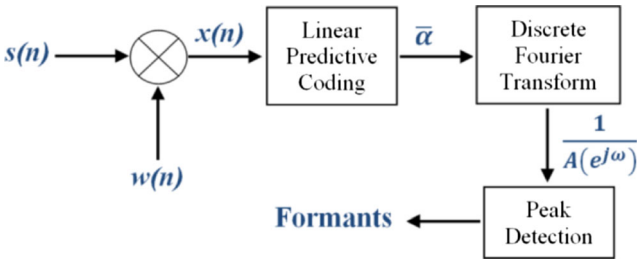


Figure 2 Process of formant extraction.

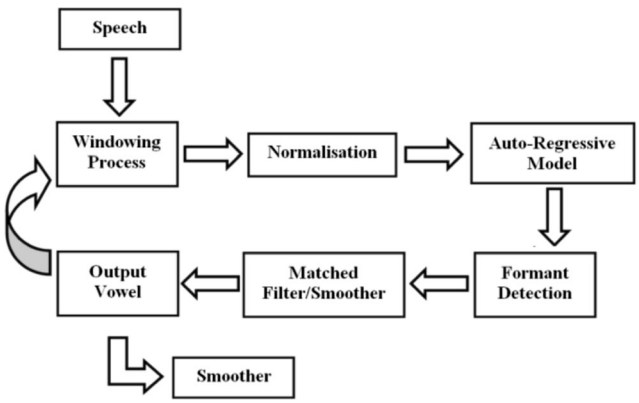


Figure 3 High-level flow chart for vowel extraction from speech signal.

3 Score-Based Scheme for Speaker Identification

The testing phase of the system requires the test signal to be preprocessed with LPC and vowel formants filtering to extract the unique vowel formants along with their first, second and third formant frequencies to be compared with the vowel formant database constructed in the training phase of the system.

As a preliminary effort, different strategies for comparing the test vowel formants with the known vowel formants in the database were tested. These strategies included the following:

1. Both first and second formants,
2. All first, second and third formants,
3. Both first and third formants,
4. Both second and third formants, and
5. Averaging and comparison with least distance.

Extensive testing of these enumerated schemes against known results revealed that these strategies are not powerful enough to yield a good accuracy, as vowel formants for the same vowel often overlap in different speakers. Sometimes only one of the three formant values overlaps, and sometimes two values overlap, with the only difference being in the frequency of the third formant. This challenging complexity is attributed to the text-independent nature of the system, in which we have a speaker speaking the same vowel but in a different word with a slightly different formant track.

To handle this type of situation, a score-based scheme was conceived that awards a positive score if all three formants are matched and penalises a speaker with a negative score otherwise. For a given speech signal, the three test formants are compared against the vowel formant stored in the database for each speaker. For example, if $D_{x,k,1}$ is the first formant frequency of vowel formant k stored in the database for speaker x and $T_{k,1}$ is the first formant frequency of k in the test speech, we conclude that they are "matched" when the difference between the two, i.e. $T_{k,1}-D_{x,k,1}$ is below a certain threshold $\varepsilon_{k,1}$. In this case, we say that the difference, diff$(T_{k,m}-D_{x,k,m})$ is zero. It is almost impossible to get an absolute zero difference. Therefore, this thresholding is an indirect form of quantisation. The vowels had to be first recognised before performing this comparison, the method of which is detailed in Section 2.2.

This quantity is computed for all three frequencies $D_{x,k,1}$, $D_{x,k,2}$, $D_{x,k,3}$ and the test score is,

$$\sum_{m=1}^{3} \text{diff}\left(T_{k,m}-D_{x,k,m}\right) = 0 \rightarrow Score\left(S_{k,x}\right) = 1 \tag{4}$$

$$\sum_{m=1}^{3} \text{diff}\left(T_{k,m}-D_{x,k,m}\right) > 0 \rightarrow Score\left(S_{k,x}\right) = -1 \tag{5}$$

The thresholds $\varepsilon_{k,m}$ are selected experimentally. These two scores aid in calculating the net score of each speaker x, against the test vowel as,

$$\text{Identification(k)} = \text{arg. Max}\left(\text{score}\left(S_{k,x}\right)\right) \text{ for k} = 1\ldots..\text{n} \tag{6}$$

The proposed system consists of a formant extraction component coupled with a vowel formant filtering component and the formant database, as shown in Fig. 4.

Our aim is to identify which acoustic parameters of the vowels (formants) depend more on the individual characteristics of the speaker and less on the linguistic variables. According to literature, high formants (F3 and F4) usually convey individual information, while F1 and F2 are dependent on vowel quality [40–42]. Fundamental frequency (F0) is the most complex acoustic cue, being related in many languages to vowel quality. All these formants can play an important role in speaker identification to different extents [43, 44].

A neural network consists of multiple perceptrons combined in multiple layers beginning with the input layer, followed by one or more hidden layers and ending at the output layer. Each perceptron, which has multiple inputs, has a weight vector associated with it. This weight vector pertains to its set of inputs, and the weights across all perceptrons are adjusted during training to map the training samples to the known target concepts. During the training phase, feature vectors extracted from the training data are fed into each of the networks in parallel over multiple epochs. Once the training is complete, they require only one pass of the input data to get the output. The size (i.e. number of neurons) of the input layer is equal to the number of LPC features. Each neuron takes in streams of data as inputs that arise from the consecutive frames. Some of the advanced neural networks have the size of the input layer enlarged to two or three adjacent frames [45] in order to get a better context dependency for the acoustic feature vectors. The number of input layers can also be chosen by multiplying the cepstral order with the total frame number [46], leading to an extremely large input layer size. But in both the above cases, the computational times are affected due to

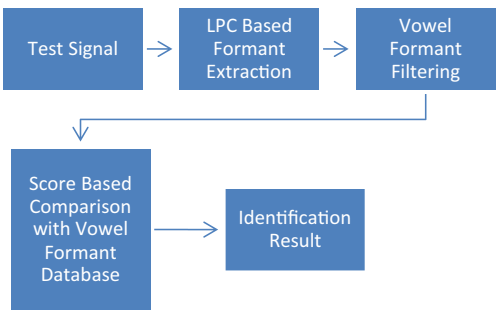


Figure 4 Process flow diagram for the proposed system for speaker identification.

the increased number of hidden layers and states. If an inadequate number of neurons are used, the network will be unable to model complex data, and the resulting fit will be poor. If too many neurons are used, the training time may become excessively long [45]. In addition, the network may over fit the data due to the large number of hidden nodes. If overfitting occurs, the network simply starts memorising the training data, thereby causing poor generalisation. We have performed some simulations to test the effect of changing the size of the neural network input size. A higher number of input layer neurons causes the number of hidden layer neurons to go up (and therefore, the number of states), which eventually increase the identification and training times. As for performance accuracy, there was no significant change observed since the NNs are designed to utilise the closed set of data in the most optimal manner. Hence, we concluded that the size of the input layer is best set equal to the number of LPC features.

4 Results and Analysis

In this section, we compare the score-based scheme proposed in this paper against BPNN and GMM using several databases such as YOHO, NIST, TI_digits1 and TI_digits2.

The YOHO database [47] contains a large scale, high-quality speech corpus to support text-dependent speaker authentication research, such as is used in secure access technology. The data was collected in 1989 by ITT under a US Government contract. The number of trials is sufficient to permit evaluation testing at high confidence levels. In each session, a speaker was prompted with a series of phrases to be read aloud. Each phrase was a sequence of three two-digit numbers. NIST speech databases are part of an ongoing series of evaluations conducted by NIST [48]. The telephone speech in this corpus is predominantly English, but also includes other languages. All interview segments are in English. Telephone speech represents approximately 368 h of the data, whereas microphone speech represents the other 574 h. TI_digits1 and TI_digits2 [49] contain speech which was originally designed and collected at Texas Instruments, Inc. for the purpose of designing and evaluating algorithms for speaker-independent recognition of connected digit sequences. There are 326 speakers each pronouncing 77 digit sequences. Each speaker group is partitioned into test and training subsets. For the training set, we picked a total of 25–30 s of speech per speaker. For some longer speech files, just one file was enough for the training required for one speaker, whereas, for shorter file sizes, multiple files were used (e.g. 20 speech files each 3 s long). We used as many users' data as the database would have, so that we have a good estimate over a large population. Similarly, after separating the training files, we used all the remaining files for testing purposes. The speech segments that we selected for the training phase did not seem to have any noticeable effect on the performance. We tested this by choosing different sets of training samples randomly and obtained somewhat similar results at the end. We also varied the total size of the training samples. Below 25 s, the training was not sufficient but increasing above this point barely made any difference.

We first present some results comparing GMM and GMM-UBM to support the choice of GMM-UBM in the subsequent experiments. The number of mixtures in the Gaussian mixture is an important parameter when employing GMM or GMM-UBM. Gaussian mixtures are combinations of a finite number of Gaussian distributions. They are used to model complex multi-dimensional distributions upon learning the parameters of the mixture through various methods. A mixture of Gaussians can be written as a weighted sum of Gaussian densities, which increases the number of distributions incorporated [50, 51]. The use of Gaussian mixture models for modeling speaker identity is motivated by the interpretation that the Gaussian components represent some general speaker-dependent spectral shapes and the capability of Gaussian mixtures to model arbitrary densities [22]. The number of mixtures can dictate the modeling capability of a GMM. Therefore, in Fig. 5, we show how this number affects the accuracy of the speaker identification system. GMM-UBM clearly outperforms GMM when used for analysing vowel formants. However, Table 2 shows that the training and identification times are somewhat higher in the case for GMM-UBM due to the added complexity. Nevertheless, because of its superior performance, we choose GMM-UBM as one of the baseline schemes which we will compare our proposed scheme against. We use a single input implementation for GMM. The above results were obtained using the YOHO database.

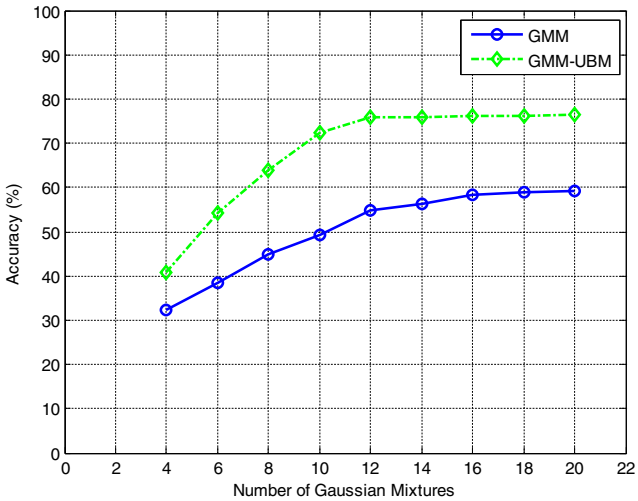


Figure 5 Accuracy of GMM and GMM UBM with varying number of Gaussian mixtures.

Table 2 Training and identification times for GMM and GMM-UBM with different number of speakers.

Number of speakers	GMM		GMM-UBM	
	20	100	20	100
Training time	68	251	110	430
Identification time	2.4	7.8	3.6	11.5

In the above experiments, we also conclude that increasing the number of Gaussian mixtures indefinitely does not necessarily increase the system accuracy. Upon further investigation, we found that many of the mixtures could be reduced to single points, as they did not have enough values to carry on further computation. The above experiment shows that 12 Gaussian mixtures provide the optimum accuracy for the vowel formant.

We next present the results for all four databases and compare out proposed scheme to both BPNN and GMM-UBM. The performance accuracy results have been averaged and are summarised in Table 3.

The same vowel formants were analysed with BPNN [52] for identification against the same training files and their extracted formants for comparison with the proposed score-based scheme. The same training vowel formants were also supplied as inputs to GMM-UBM, and the test sets were tested against these mixtures. The experiments revealed that the vowel formants with the score-based strategy are not only more accurate in identification but also more scalable, as highlighted next.

An identification algorithm is critically evaluated for its accuracy against the test data for a number of speakers. The context of evaluation becomes more critical if the algorithm aims to be applicable for industry devices for biometric security and identity management [53]. However, as the number of speakers increase, traditional algorithms start declining in accuracy. This trend has been an important consideration during the design and testing of the current system to ensure that it is a scalable model. The performance graph in Fig. 6 shows the performance statistics as the number of speakers is gradually increased from 10 to 110 in increments of 10. It is to be noted that both GMM and BPNN start decreasing in accuracy as the

Table 3 Performance comparison (percentage accuracy) of the proposed score-based strategy, BPNN and GMM-UBM algorithms for different databases.

Database	Score-based scheme formants	Formants with BPNN	Formants with GMM-UBM
YOHO	94.23 %	62.14 %	75.84 %
NIST	92.15 %	54.51 %	72.73 %
TI_digits1	96.87 %	57.58 %	69.42 %
TI_digits2	97.34 %	59.12 %	73.59 %

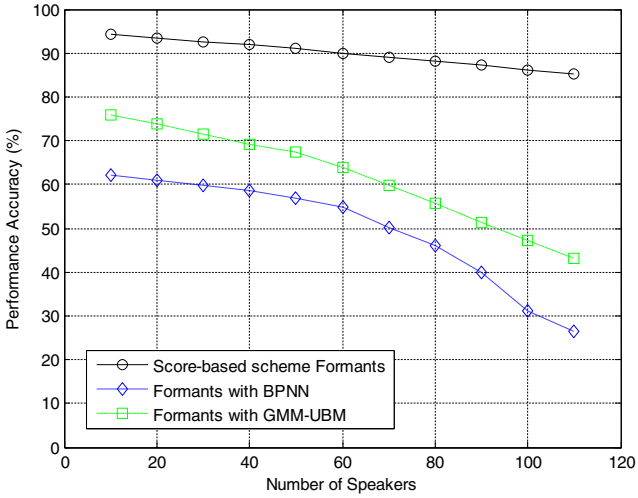


Figure 6 Performance statistics (percentage accuracy) of the three algorithms with varying number of speakers.

number of speakers increases, whereas the proposed score-based scheme shows a fairly stable accuracy that is barely affected by increasing the number of speakers.

Next we present the receiver operating characteristic (ROC) curve for the three algorithms using the YOHO database in Fig. 7. It shows the distribution of the area under the curve when plotted with the results of the tests. It clearly shows the maximum area is covered with score-based scheme as compared to the other schemes.

During these tests, the identification and training times for the score-based strategy was also observed, as shown in Tables 4 and 5. Although BPNN requires less identification time compared with the proposed score-based scheme for most databases, it has a much higher training time and poor accuracy.

It is to be observed that the score-based scheme does not require any training other than saving the filtered vowel formants in the database, which in this case is a Matlab file. Note that the identification time does not include the preprocessing

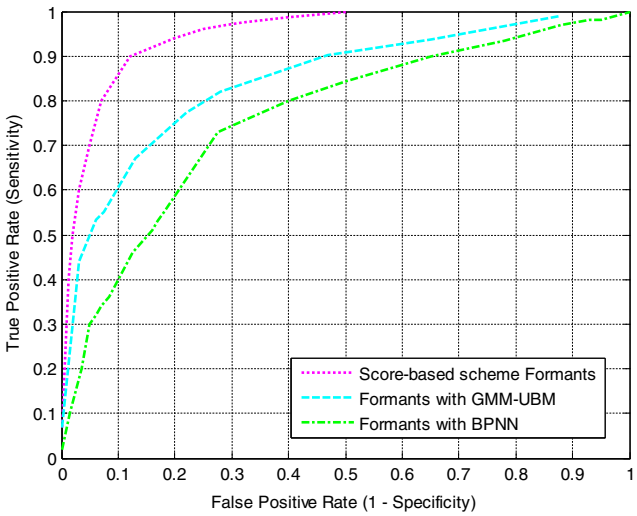


Figure 7 ROC curve for the three algorithms for YOHO.

Table 4 A comparison of the average training time (sec) for different databases.

	Score based scheme	Formants with BPNN	Formants with GMM
YOHO	5.6	72.5	110
NIST	7.5	84.9	135.1
TI_digits1	4.5	68.4	85.6
TI_digits2	4.4	67.4	74.5

time with LPC and vowel formant filtering, as that time is considered to be common for all of the algorithms tested. If the number of speakers is increased, the identification time is expected to increase for any algorithm. As shown in Fig. 8, our proposed scheme has much less identification times compared to the other schemes.

Next we present the effect of the length of speech files. The databases we used have different lengths for speech files. By trimming them down to 3, 2 and 1 s and testing our system using these speech signals, we obtained the accuracy results presented in Table 6. As expected, the longer speech segments provide the best results. All other results presented in this paper are based on 3 s long speech signals.

Finally, we present some results on using different number of formants. Using more formants do not necessarily increase the overall accuracy rates. Usually, additional formants provide a tighter threshold for comparison purposes and helps reduce false accept rates (FAR). But they also cause false reject rates (FRR) to increase, thereby reducing the overall accuracy rate. This phenomenon is demonstrated in Fig. 9 where the FAR and FRR results are presented for 3, 4 and 5 formants. Moreover, using more formants drastically increases the training and identification times of the algorithm, thereby reducing the scope of the speaker identification system for many time-stringent applications.

The results presented in this section clearly show that the proposed scheme outperforms the BPNN and GMM classifiers. Both these classifiers are widely used and regarded as efficient schemes in many speaker identification implementations. However, in a vowel formant based scheme, they fail to perform at a desired level due to the following reasons. BPNN has the problem of entrapment in local

Table 5 A comparison of the average identification time (sec) for different databases.

	Score based scheme	Formants with BPNN	Formants with GMM
YOHO	0.15	0.08	3.6
NIST	0.21	0.25	4.3
TI_digits1	0.14	0.12	2.8
TI_digits2	0.14	0.11	2.9

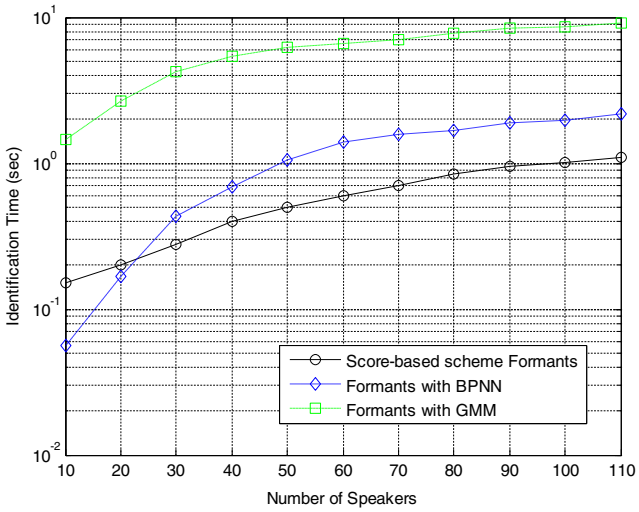


Figure 8 Identification times (sec) of the three algorithms with varying number of speakers for YOHO.

minima, and the network should be trained with different initial values until the best result is achieved. The number of hidden layers and neurons in each layer are required to be determined. If the number of layers or neurons is inadequate, the network may not converge during the training; if the number of the layers or neurons is chosen to be too high, this will diminish the effectiveness of the network operation. This often causes this algorithm to be biased by a specific resonant frequency while completely disregarding the others. The proposed score-based scheme tackles this problem better by exploiting information received from all frequencies. A score-based scheme allows the speaker with the highest matching formants to own the current signal. Furthermore, we choose LPC as the accompanying feature extraction strategy of our novel scheme, which is the best strategy due to its resemblance with the functioning of the human vocal tract. Another difficulty of BPNN lies in its use of the back propagation algorithm that is too slow for practical applications, especially if many hidden layers are employed. The appropriate selection of training parameters in the BP algorithm is sometimes difficult. As for the GMM algorithm, its main limitation is that, it can fail to work if the dimensionality of the problem is too high. This causes the GMM to suffer badly when the number of speakers increases. Another disadvantage of the GMM algorithm is that the user must set the number of

Table 6 Accuracy vs speech file lengths (across databases and/or trimmed speech versions).

Speech file length:	1 s	2 s	3 s
YOHO	83.78	89.46	94.23
NIST	72.54	85.28	92.15
TI_digits1	78.56	91.25	96.87
TI_digits2	80.12	91.89	97.34

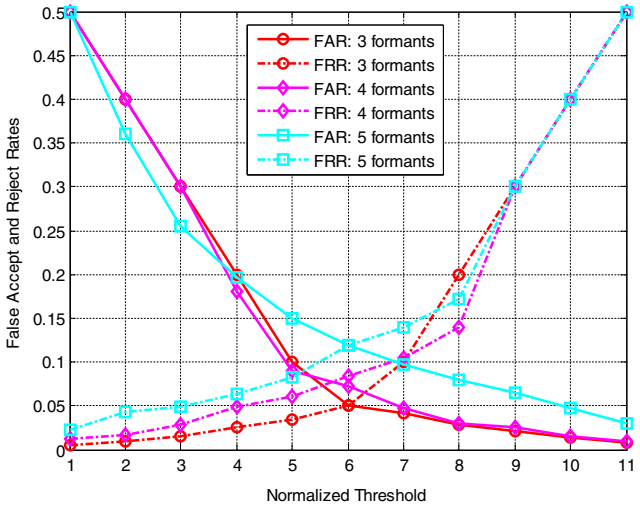


Figure 9 FAR and FRR for different number of formants.

mixture models that the algorithm will try and fit to the training dataset. In many instances the user will not know how many mixture models should be used and may have to experiment with a number of different mixture models in order to find the most suitable number of models that works for their classification problem.

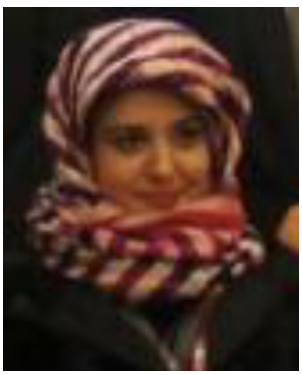
5 Conclusions

Biometric authentication is a multi-disciplinary problem and requires sound knowledge in machine learning, pattern recognition, digital signal processing, image processing and several other overlapping fields such as artificial intelligence and statistics. The objective of this research is to investigate the problem of identifying a speaker from its voice regardless of the content (text-independent). This paper investigates the combination of LPC-based vowel formants with a score-based identification strategy. For comparison, two other combinations of LPC-based vowel formants have been tested with BPNN and GMM. Comprehensive testing on the YOHO, NIST, TI_digits1 and TI_digits2 databases reveals that the proposed scheme outperforms BPNN and GMM-based schemes. It has been observed that the proposed scheme requires very little training time other than creating a small database of vowel formants. Therefore, the proposed scheme is time-wise more efficient as well. The results also show that increasing the number of speakers makes it difficult for BPNN and GMM to sustain their accuracy. Both of these models start losing accuracy, whereas the proposed score-based methodology remains much more stable, making it scalable and suitable for large-scale implementations. In the future, we want to continue further with the current approach to speaker identification and combine it with real-time face recognition to make it more robust and applicable for industry usage. We aim to combine audio and visual features as a feature-level fusion in multimodal neural networks to further improve the accuracy through use of two biometric features.

References

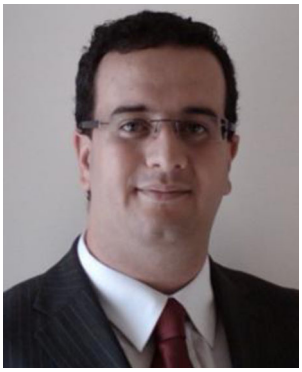
1. Kinsner, W., & Peters, D. (1988). A speech recognition system using linear predictive coding and dynamic time warping. *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 3*, 1070–1071. doi: 10.1109/IEMBS.1988.94689.
2. Campbell, J. P., Jr., Tremain, T. E., & Welch, V. C. (1990). The proposed federal standard 1016 4800 bps voice coder. *Speech Technology, 5*, 58–64.
3. Aalto, D., Huhtala, A., Kivela, A., Malinen, J., Palo, P., Saunavaara, J., Vainio, M. (2012). *Vowel formants compared with resonances of the vocal tract.*
4. Chang, E., Zhou, J., Di, S., Huang, C., Lee, K. F. (2000). *Large vocabulary mandarin speech recognition with different approaches in modeling tones*. Proceedings of ICSLP.
5. Martin, A., & Przybocki, M. (1999). The NIST 1999 speaker recognition evaluation an overview. *Digital Signal Processing, 10*(1), 1–18.
6. Kivela, A., Kuortti, J., Malinen, J. (2013). *Resonances and mode shapes of the human vocal tract during vowel production.* Proceedings of 26th Nordic Seminar on Computational Mechanics.
7. Sambur, M. (1975). Selection of acoustic features for speaker identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 23*(2), 176–182.
8. Rose, P. (2006). Technical forensic speaker recognition: evaluation, types and testing of evidence. *Computer Speech & Language, 20*(2), 159–191. **Elsevier**.
9. Becker, T., Jessen, M., Grigoras, C. (2008). Forensic speaker verification using formant features and Gaussian mixture models. *Interspeech*, 1505–1508.
10. McDougall, K. & Nolan, F. (2007). Discrimination of speakers using the formant dynamics in British English. *Proceedings of the 16th International Congress of Phonetic Sciences*, (pp. 1825–1828).
11. Grieve, J., Speelman, D., & Geeraerts, D. (2013). A multivariate spatial analysis of vowel formants in American English. *Journal of Linguistic Geography, 1*(1), 31–51. **Cambridge Univ. Press**.
12. Shao, X. & Milner B. (2004). *Pitch prediction from MFCC vectors for speech reconstruction.* Proc. of the 2004 International Conference on Acoustics, Speech, and Signal Processing, (pp. 97–100). Montreal.
13. Afify, M., & Siohan, O. (2007). Comments on vocal tract length normalisation equals linear transformation in cepstral space. *IEEE Transactions on Audio, Speech and Language Processing, 15*(5), 1731–1732.
14. Chazan, D., Cohen, G., Hoory, R., Zibulski, M. (2000). *Speech reconstruction from Mel frequency cepstral coefficients and pitch*. Proceedings of ICASSP.

15. Shao, X. & Milner B. (2004). *MAP prediction of pitch from MFCC vectors for speech reconstruction*. Proc. ICSLP.
16. Atal, B. S. (2006). The history of linear prediction. *IEEE Signal Processing Magazine, 23*(2), 154–161.
17. Aliyu, A. O., Adewale, E. O., Adetunmbi, O. S. (2013). Development of a text-dependent speaker recognition system. *Journal of Development, 69*(16).
18. Das, B. P. (2012). *Recognition of isolated words using features based on LPC, MFCC, ZCR and STE, with neural network classifiers*. Jadavpur University Kolkata.
19. Yuanyou, X., Yanming, X., & Ruigeng, Z. (1997). An engineering geology evaluation method based on an artificial neural network and its application. *Engineering Geology, 47*(1), 149–156.
20. Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital Signal Processing, 10*(3), 19–41.
21. Wagner, M. (2012). *Speaker identification using glottal-source waveforms and support-vector-machine modelling. SST 2012*. Sydney: Macquarie University.
22. Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing, 3*(1), 72–83.
23. G. S. University. (2013). http://hyperphysics.phyastr.gsu.edu/hbase/music/vowel2.html. Hyperphysics education website. Accessed 12 Dec 2013.
24. Flanagan, J. L., & Rabiner, L. R. (1973). *Speech synthesis*. Stroudsburg: Dowden, Hutchinson, and Ross, Inc.
25. Rabiner, L. R., & Schafer, R. W. (1978). *Digital processing of speech signals* (Prentice-Hall signal processing series). Englewood Cliffs: Prentice-Hall.
26. Benesty, J., Sondhi, M. M., & Huang, Y. (2008). *Springer handbook of speech processing*. Berlin: Springer.
27. Ramachandran, R., Zilovic, M. S., & Mammone, R. (1995). A comparative study of robust linear predictive analysis with application to speaker identification. *IEEE Transactions on Speech and Audio Processing, 3*(2), 117–125.
28. Rabiner, L. R., & Juang, B. (1993). *Fundamentals of speech recognition*. Englewood Cliffs: Prentice Hall.
29. Wells, J. C. (1962). *A study of the formants of the pure vowels of British English*. (unpublished master's thesis). University of London.
30. Keller, J. B. (1953). Bowing of violin strings. *Communications Pure and Applied Mathematics, 6*, 483–495.
31. Fant, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton.
32. Markel, J. D., & Gray, A. H. (1976). *Linear prediction of speech*. New York: Springer Verlag.
33. Keurs, M. T., Festen, J. M., & Plomp, R. (1993). Effect of spectral envelope smearing on speech reception II. *The Journal of the Acoustical Society of America, 93*, 1547.
34. Baer, T., & Moore, B. C. J. (1994). Effects of spectral smearing on the intelligibility of sentences in the presence of interfering speech. *The Journal of the Acoustical Society of America, 95*, 2277.
35. Drullman, R., Festen, J. M., & Plomp, R. (1994). Effect of reducing slow temporal modulations on speech reception. *The Journal of the Acoustical Society of America, 95*, 2670.
36. Drullman, R., Festen, J. M., & Houtgast, T. (1996). Effect of temporal modulation reduction on spectral contrasts in speech. *The Journal of the Acoustical Society of America, 99*, 2358.
37. Fu, Q. J., & Nogaki, G. (2005). Noise susceptibility of cochlear implant users: the role of spectral resolution and smearing. *Journal of the Association for Research in Otolaryngology, 6*(1), 19–27.
38. Liu, C., & Fu, Q. J. (2007). Estimation of vowel recognition with cochlear implant simulations. *IEEE Transactions on Biomedical Engineering, 54*(1), 74–81.
39. Mathworks. (2013). http://www.mathworks.com. Accessed 12 Dec 2013.
40. Stevens, K. (1971). Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds. *Proc. 7th Intern. Congr. Phon. Sc.*, Montreal, (pp 206–227).
41. Atal, B. S. (1972). Automatic speaker recognition based on pitch contours. *JASA, 52*, 1687–1697.
42. Nolan, F. (1983). *The phonetic bases of speaker recognition*. Cambridge: Cambridge University Press.
43. Hollien, H. (1990). *The acoustics of crime. The new science of forensic phonetics*. Nueva York: Plenum.
44. Kuwabara, H., & Sagisaka, Y. (1995). Acoustic characteristics of speaker individuality: control and conversion. *Speech Communication, 16*, 165–173.
45. Rottland, J., Neukirchen, C., Willett, D., Rigoll, G. (1997). *Large vocabulary speech recognition with context dependent MMI-connectionist/HMM systems using the WSJ database*. EUROSPEECH.
46. Campbell, J., & Higgins, A. (1994). *YOHO speaker verification LDC94S16. Web download*. Philadelphia: Linguistic Data Consortium.
47. Hamzah, R., Jamil, N., Seman, N. (2014). *Filled pause classification using energy-boosted mel-frequency cepstrum coefficients*. In Proc. Int. Conference on Robotic, Vision, Signal Processing & Power Applications.
48. NIST Multimodal Information Group. (2008). *NIST speaker recognition evaluation test set LDC2011S08. Web download* (p. 2011). Philadelphia: Linguistic Data Consortium.
49. Leonard, R. G., & Doddington, G. (1993). *TIDIGITS LDC93S10. Web download*. Philadelphia: Linguistic Data Consortium.
50. Lee, D.-S. (2005). Effective Gaussian mixture learning for video background subtraction. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(5), 827–832.
51. Lee, D-S., Hull, J. J., Erol, B. (2003). A Bayesian framework for Gaussian mixture background modeling. *Proceedings of International Conference on Image Processing, 3*.
52. Templetom, P. D. & Guillemin, B. J. (1990). Speaker identification based on vowel sounds using neural networks. *Proceedings of the 3rd International Conference on Speech Science and Technology*, Australian Association, 280–285.
53. Miles, M. J. (1989). *Speaker identification based upon an analysis of vowel sounds and its applications to forensic work.* (Unpublished master's thesis). University of Auckland, New Zealand.



Noor Almaadeed received her Ph.D. from Brunel University, London, UK, in 2014. Her areas of research include speech signal detection, speaker identification, audio/visual speaker recognition, etc. She received her Bachelor's in Computer Science from Qatar university in 2000, and Master of Science in Computer and Information Sciences from City University, UK, in 2005. She is an Assistant Professor in the Department of Computer Science & Engineering in Qatar University. She was awarded the Qatar Education Excellence Day Platinum Award -New PhD Holders Category 2014-2015.

Prof. Amar Aggoun is a Reader in Information and communication Terchnologies, Brunel University, London, and the Head of Computer Science and Technology Department, University of Bedfordshire. He received his Ph.D. from University of Nottingham in Image/Video processing in 1991. His areas of research include 3DTV, video signal processing and compression, 3D computer graphics, audio-visual delivery, computer architectures, and computer vision systems.



Prof. Abbes Amira took academic and consultancy positions in UK and overseas, including his current positions as a Professor in Computer Engineering at Qatar University, Qatar and a full professor in visual communications and leader of the Visual Communication Cluster at the University of the West of Scotland (UWS), UK. He took other academic positions at the University of Ulster-UK, Qatar University-Qatar, Brunel University-UK and Queen's University Belfast-UK. He received his Ph.D. in Computer Engineering from Queen's University, Belfast, United Kingdom, in 2001. His areas of research include image and vision systems, embedded systems, image/video processing and analysis, pattern recognition, and connected health and security.