

Speaker identification using multimodal neural networks and wavelet analysis

Item Type	Journal article
Authors	Aggoun, Amar;Almaadeed, Noor;Amira, Abbes
Citation	Almaadeed, N., Aggoun, A., and Amira, A. (21015)Speaker identification using multimodal neural networks and wavelet analysis, IET Biometrics, 4 (1), pp. 18-28
DOI	10.1049/iet-bmt.2014.0011
Publisher	IET
Journal	IET Biometrics
Download date	2026-04-17 06:52:10
Link to Item	http://hdl.handle.net/2436/620913

Speaker identification using multimodal neural networks and wavelet analysis

Noor Almaadeed^{1,2}, Amar Aggoun³, Abbas Amira^{2,4}

¹Department of Computer Engineering, Brunel University, Kingston Lane, Uxbridge, Middlesex UB8 3PH, UK

²Department of Computer Science and Engineering, College of Engineering, Qatar University, Doha, Qatar

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

⁴Department of Engineering and Computer Science, University of the West of Scotland, Paisley, UK

E-mail: n.alali@qu.edu.qa

Abstract: The rapid momentum of the technology progress in the recent years has led to a tremendous rise in the use of biometric authentication systems. The objective of this research is to investigate the problem of identifying a speaker from its voice regardless of the content. In this study, the authors designed and implemented a novel text-independent multimodal speaker identification system based on wavelet analysis and neural networks. Wavelet analysis comprises discrete wavelet transform, wavelet packet transform, wavelet sub-band coding and Mel-frequency cepstral coefficients (MFCCs). The learning module comprises general regressive, probabilistic and radial basis function neural networks, forming decisions through a majority voting scheme. The system was found to be competitive and it improved the identification rate by 15% as compared with the classical MFCC. In addition, it reduced the identification time by 40% as compared with the back-propagation neural network, Gaussian mixture model and principal component analysis. Performance tests conducted using the GRID database corpora have shown that this approach has faster identification time and greater accuracy compared with traditional approaches, and it is applicable to real-time, text-independent speaker identification systems.

1 Introduction

The task of speaker recognition can comprise speaker identification (i.e. identifying the current speaker) or speaker verification (i.e. verifying whether the speaker is who he claims to be) [1]. There are two types of speaker identification: text-dependent (the speaker is given a specific set of words to be uttered) and text-independent (the speaker is identified regardless of the words spoken) [2]. This paper proposes a novel approach towards building a text-independent speaker identification system (SIS).

A digital speech signal in its crudest form comprises frequency values sampled at consistent time intervals. It must be pre-processed to extract feature vectors that represent unique information for a particular speaker irrespective of the speech content. A learning algorithm generalises these feature vectors for various speakers during training and verifies the speaker's identity using a test signal during the test phase. In practice, no two digital signals are the same even for the same speaker and the same set of words. The amplitude and pitch in a speaker's voice can vary from one recording session to another. Environmental noise, the recording equipment, the speed at which the speaker speaks and the speaker's various psychological and physical states increase the complexity of this task. Text-independent speaker identification allows the speaker to speak any set of words during a test. For such versatile systems, there is a need for a general feature

extraction strategy to extract text-independent features from a speech signal.

The classical Mel-frequency cepstral coefficients (MFCC) method is likely the most popular feature extraction strategy used to date. This method is utilised herein for comparison with wavelet analysis. Linear predictive coding (LPC) has immensely aided text-dependent identification tasks [3]. Both MFCC and LPC use a global approach for speech analysis and are, therefore, susceptible to additive noise in the speech [4]. In this paper, we employed MFCC for comparison and relied heavily on wavelet-analysis strategies for feature extraction.

There are essentially two broad categories for methods to develop learning algorithms based on extracted speech features: generative and discriminative models. Generative methods are widely used and include stochastic models such as the hidden Markov model (HMM) [5], the Gaussian mixture model (GMM) [6] and template-based models (e.g. vector quantisation) [7]. The goal of a generative model is to symbolise the distribution space of the stored data generated from a particular class. This training process ignores competing data and considers only related data. In contrast, discriminative models shape the discriminative areas of a distribution. The primary purpose of this method is to reduce classification errors in the stored data as much as possible. Unlike generative models, data from all competing classes are also considered. Major discriminative models include polynomial classifiers [8], the support

vector machine [9], the multilayer perceptron and artificial neural network (ANN) [10] methods, such as the general regressive NN (GRNN) [11], probabilistic NN (PNN) and radial basis function NN (RBF-NN) models [12].

To date, no single biometric system has been developed that can claim to identify or verify a speaker in all varieties of environments. Accurate classification of a speaker is a challenge when inter-class differences exceed intra-class differences, which primarily arise from a text-independent approach or noisy data. In an attempt to resolve this problem, two or more biometric techniques can be combined in a single system to improve the effectiveness of identification. This information fusion can be generated at different levels for multimodal biometrics. Information fusion is information that is merged from disparate sources with different conceptual, contextual and typographical expressions. In multimodal biometrics, this is possible at the sensor, feature, score or decision levels [13, 14]. In sensor-level fusion, the core data from multiple sensors are combined for each modality which reduces classification error. In feature-level fusion, the speaker information received from multiple sources undergoes a feature extraction step, and this information is fused logically. In score-level fusion, a score is assigned to each individual biometric system, and these scores are used to make decisions for the final classification. In decision-level schemes, the final decision to accept or reject an individual system is generated via a voting procedure (e.g. majority, AND, OR etc.). Many researchers, like Nefian *et al.* [15], tend to lean towards the early fusion approaches for audio-visual speech recognition. A speaker verification system based on audio-visual hybrid fusion from a set of features that are cross-modal was proposed in [16]. For a personnel authentication system based on face and voice, Chetty and Wagner [17] also developed a feature-level fusion to check the liveness, and presented test results performed on the VidTIMIT and UCBN databases.

The fusion performed in this paper involved the decision-level scheme. Different wavelet feature extraction techniques and decision-level schemes were investigated using three popular classifiers for text-independent, open-set speaker identification. The selected architectures were GRNN, PNN and RBF-NN. These NNs are fast, reliable and efficient for non-linear and complex data. Compared with back-propagation NNs (BPNN), which require a long training period, these networks are instantly trained and produce immediate results when applied to a test signal. Combining multiple ANNs enhances the generalisation capability and increases the identification rate. It also reduces the false accept rate (FAR) for a given false reject rate (FRR), and vice versa [18]. This motivated us to develop a novel identification system, namely the multimodal NN (MNN).

This paper is organised as follows. Section 2 describes wavelet feature extraction methods and the basics of NN. In Section 3, we introduce the proposed fusion system with a detail justification of the feature extraction and NNs that were chosen. Section 4 presents a comprehensive analysis of the performance and test results for this scheme, and finally Section 5 presents conclusions and recommendations for future work.

2 Overview of system components

Wavelet and wavelet packet analysis have been proven as effectual signal processing techniques for a variety of

digital signal processing problems. They have also been used in many different methods in feature extraction plans designed for the task of speech or voice identification.

A speech signal contains a huge amount of data. For example, a 1 s speech signal consists of ~50 000 floating-point values in a single linear vector. The performance enhancement of a SIS requires a careful selection of suitable features from the raw set available, which is usually somewhat redundant. The most relevant and significant information must be chosen from the original feature space using an appropriate feature selection scheme. The first block in the SIS is the feature extraction block. In this phase, the rough audio signal is pre-processed to extract only the distinguishing features for analysis from the entire signal. The feature extraction techniques used are discrete wavelet transform (DWT), wavelet packet transform (WPT), wavelet sub-band coding (WSBC) and MFCC. Section 2.1 presents a review of the basics of these techniques.

NNs are the most common approach to learning non-linear or complex training spaces. NNs are vastly applied in numerous data analysis and speaker identification schemes, as well as classification tasks [19]. For an ANN, there is no need to predict the transfer function between the input and output ahead of time, and this is one of its greatest advantages. In Section 2.2, we provide a comprehensive description of different NNs in context of our proposed SIS.

2.1 Wavelet analysis and feature extraction

Wavelet transforms [20–22] have been studied comprehensively in the recent times and widely utilised in various areas of science and engineering. Under the class of wavelet analysis, a mother wavelet is processed on dilation and translation. Many signals of interest can be represented with wavelet decompositions, in general. The fundamental idea behind wavelets is to analyse a given signal according to a scale [10]. The wavelet successively decomposes the given signal into a set of smaller signals at multiple levels and analyses each piece of the signal at different frequencies with different resolutions. For instance, good time resolution with poor frequency resolution at high frequencies and high-frequency resolution with poor time resolution at low frequencies are more suitable for samples with short-duration high-frequency components and long-duration low-frequency components, respectively. The window width is altered as the transform is computed for each spectral component. Wavelets are well suited for approximating data with sharp discontinuities. An example of a signal in the wavelet domain and a short-time Fourier transform is illustrated in Fig. 1.

Wavelets are a class of functions used to localise a given function for both space and scaling. A family of wavelets can be constructed from a function called a mother wavelet, which is confined in a finite interval with a zero average. A set of wavelets are formed from the mother wavelet by translating and scaling the mother wavelet. The wavelet tree and methodology that have been used for speaker or speech recognition include the DWT, WPT and the Mel-scale and sub-band coding algorithms WSBC, which were first utilised for speaker identification in [23, 24]. The advantage of WSBC is that it models the human auditory system and thus decreases the number of parameters for the entire WPT, which reduces the time required for speaker identification. Finally, the wavelet with irregular decomposition algorithm, along with other wavelet

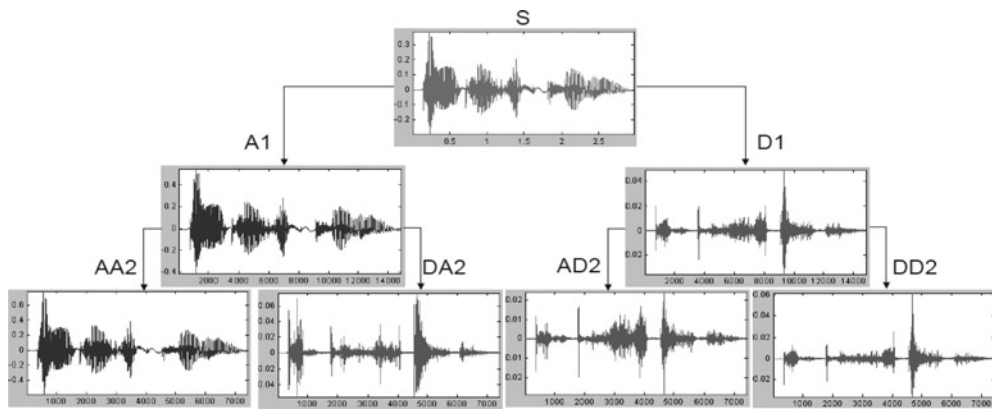


Fig. 1 Wavelet decomposition of signal *S* into detailed and approximate components

analyses, were first tested and proposed in [23]. The primary idea was to irregularly prune the decomposition tree generated by WPT for enhanced accuracy. A feature extraction scheme derived from the wavelet eigenfunction was proposed in [25], and a text-independent SIS was proposed in [26] based on an improved wavelet transform, which relies on the kernel canonical correlation analysis. WPT, which is analogous to DWT in some ways, obtains the speech signal using a recursive binary tree and performs a form of recursive decomposition. Instead of performing decomposition only on approximations, it decomposes the details as well. WPT, therefore, has a better feature representation than DWT [25]. This is why WPT is used as part of our proposed MNN as laid out in Sections 3 and 4.

GMM is extensively used for classification tasks in speaker identification [1, 19]. It is a parametric learning model that assumes that the process being modelled has the characteristics of a Gaussian process whose parameters do not change over time. This assumption is valid because a signal can be assumed to be stationary over a Hamming window. GMM tries to capture the underlying probability distribution governing the instances presented during the training phase. Given a test instance, the GMM tries to estimate the maximum likelihood that the test instance has been generated from a specific speaker’s GMM. The GMM with the maximum value of likelihood owns the test instance and is declared to belong to the respective speaker. In this paper, we employ GMM for the classification task.

2.2 Neural networks

An NN 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 has an associated weight. These weights are adjusted during training to map the training samples to the known target concepts. At the end of training, a tuned weight matrix is produced, which corresponds to a complex function that maps the input to the output.

The most common NN types include BPNN and feed-forward networks. The training input is passed through the network a number of times to adjust the weights accordingly. The iterative data training process requires multiple passes through the network for correct training. This requires a large amount of time before the network converges to a fine-tuned weight matrix. Therefore ANNs are notorious for long training times and over- or under-fitting training data. The use of combined multiple NNs is an excellent means to apply machine learning under

high dimensionality or strict decision conditions [10, 27]. The combination of multiple NNs eliminates the poor performance from over- or under-fitting the training data with individual NNs wherein each network has a different level of generalisation capability. The combination of multiple NNs resolves the higher identification rate problem but complicates the method by increasing the training time. Below we briefly describe the architectures of some of the NNs that we implemented as part of our proposed MNN.

The PNN has an input layer where the input vectors are inserted (in this case, the audio feature vectors). The network also includes one or more hidden layers with multiple neurons that are connected through weighted paths. Additionally, it includes one or more output neurons depending on the number of different classes. PNN is a statistical classifier network that applies the maximum a posteriori hypothesis to classify a test pattern *X* as Class *C* if the following applies

$$P(X_i|C_i)P(C_i) \geq P(X_i|C_j)P(C_j) \forall j \quad (1)$$

Here, $P(C_i)$ is the prior probability of speaker *i* that is determined from the training feature vectors. $P(X_i|C_i)$ is the conditional probability that this pattern is generated from class *C_i* assuming that the training data follow a probability density function (PDF). A PDF is estimated for each speaker class. As shown in Fig. 2, the third layer from left

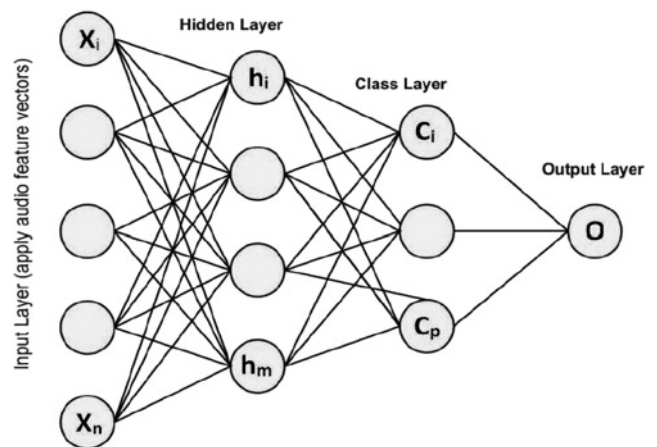


Fig. 2 Architecture of a PNN: input layer, hidden layer, class layer and output layer [12]

© Massachusetts Institute of Technology, 1989

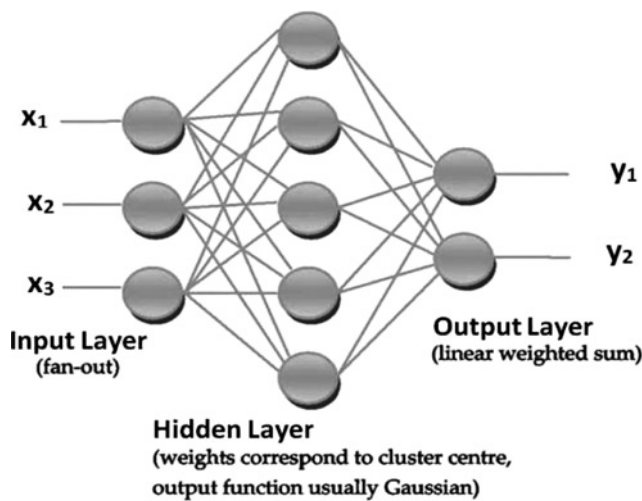


Fig. 3 Structure of an RBF NN [12]

© Massachusetts Institute of Technology, 1989

to right is the class layer and the last layer is the output, which represents the winning class.

On the other hand, RBF networks [12] employ RBFs directly to each input value without associating a weight line from the input to this RBF layer as shown in Fig. 3. It consists of three layers of neurons: input, hidden and output. The hidden layer neurons represent a series of centres in the input data space, as shown in Fig. 3. The RBF is given by

$$y_m = f_m(\mathbf{x}) = \exp[-|\mathbf{x} - \mathbf{c}_m|^2 / (2\sigma^2)] \quad (2)$$

Here $|\mathbf{x} - \mathbf{c}_m|^2$ is the square of the distance between the input feature vector \mathbf{x} and the centre vector \mathbf{c}_m for the current RBF node.

The network output is a weighted sum from these RBF nodes and is calculated as follows

$$z_j = \left(\frac{1}{M}\right) \sum_{m=1}^M (u_m y_m) \quad (3)$$

These networks have many uses, such as time series prediction, classification and system control. In the context of speaker identification, the RBF-NN utilises the projection of an eigenface space to compute the NN input features. There are two main categories of learning: the supervised learning and the unsupervised learning. The RBF-NN has both a supervised and unsupervised component to its learning.

Next is the GRNN, which is based on a general regression. These networks, first proposed in 1991 [27], are widely used in many identification tasks. Fig. 4 shows the block diagram of the GRNN architecture. It is a one-passing learning algorithm, which can be used for estimating continuous variables such as some transient content in speech signal. GRNN has a structure similar to PNN and RBF networks but are based on general regression as proposed by [28]. In contrast to PNNs and RBF-NNs, a GRNN uses a PDF based on a normal distribution.

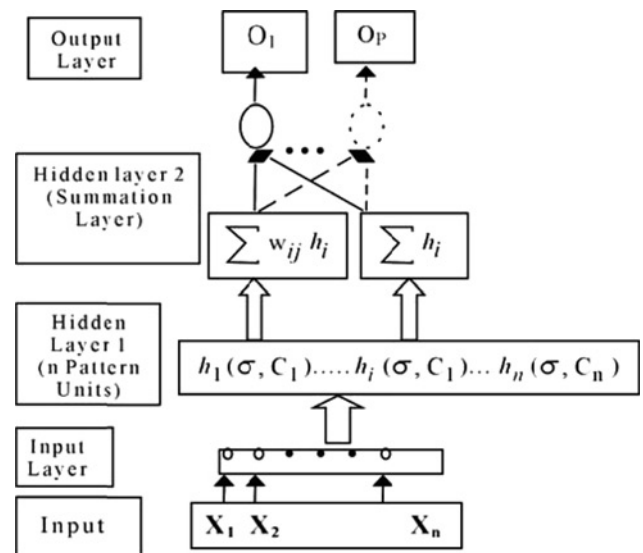


Fig. 4 GRNN architecture [27]

© IEEE, 1991

3 Proposed MNN

This section presents a novel approach using multiple NNs (PNN, RBF-NN and GRNN) to classify an SIS using a wavelet-based selection method. The proposed system consists of feature extraction and modelling blocks that use multiresolution analysis and decision fusion with an MNN, respectively.

Speaker identification is an expert system based on the single biometric of voice data. It first extracts the audio from the raw audio data. An audio stream consists of thousands of values in the range $[-1, 1]$ that are sampled at a regular interval. An 8 kHz sampling rate means that 8000 such values vary each second when a speaker's audio is recorded. These raw values only tell us about the amplitude variations in the speech and do not convey any explicit information about the speaker. Since we are using text-independent speaker identification, we must extract distinguishing speech features that describe a speaker's orientation or, more specifically, the qualities of the speaker's glottal tract which are independent of the language being used. Therefore, if the same speaker speaks a different set of words next time, our system should identify the speaker. Therefore, we must transform the raw signal into a parametric representation.

Usually, short-time spectral analysis techniques, such as LPC and MFCC, are used to transform the raw signal into a parametric representation containing the most important characteristics of the signal [29]. MFCC, originally developed for speech recognition systems, employs both logarithmically spaced filters and Mel-scale filters, which are less susceptible to noise and variations in the physical conditions of the speaker. Herein, we have used MFCC to capture the most phonetically important characteristics for speaker identification from the audio signal. We select the widely accepted MFCC features for our research because of their demonstrated superior performance. However, the MFCC feature vector describes only the power spectral envelope of a single frame, but not the information in its dynamics. To incorporate the ongoing changes over multiple frames, the first and second derivatives of the

features can be computed, which are known as the delta and delta-delta coefficients, respectively [30]. These dynamic features of cepstral coefficients are often employed to improve speech recognition performance [31, 32]. As a pre-processing step on the audio signal, we perform pre-emphasis to compensate for the high-frequency falloff, and then use the short-term analysis technique using windowing.

In the proposed MNN, we use multiple ANNs for classification using wavelet-based feature extraction methods, namely: DWT, WPT, WSBC and irregular decomposition. The prominent features extracted through these methods are fed into a learning model wherein the target concept is modelled and mapped to the training samples for classification. This system employs the following three different classifier architectures in parallel: GRNN, PNN and RBF-NN. The architectures of these NNs have been described in Section 2.2. In this section, we present the major highlights of the text-independent SIS based on bootstrap aggregating these equally robust but fast learners, which are chosen for the reasons stated below.

BPNNs, RBF-NNs, GRNNs and PNNs can be easily differentiated from each other on the basis of structure, training strategy, samples requirement, training time, accuracy and suitability for various types of data. PNNs, RBF-NNs and GRNNs require just a fraction of the samples, as well as much lesser training times, compared with BPNN. These NNs are more adaptive in converging quickly to a decision surface as more neurons can be added at runtime to aid the results compared with BPNNs, which have a fixed number of neurons in the hidden layers. PNNs,

RBF-NNs and GRNNs are more suitable for low-dimensional data like that the different wavelet-analysis methods yield through DWT, WPT, WSBC or irregular decomposition. Therefore PNNs, RBF-NNs and GRNNs are the best candidates for bagging as they are simultaneously strong and fast learners. Furthermore, the work in [23] reported that the back-propagation algorithm over-fits training data and has a higher error rate than RBF-NN. These reasons were the primary motivations for developing a scheme that resolves the under- and over-fitting problems and minimises the training time.

The proposed system architecture for speaker identification is illustrated in Fig. 5. A system with text-independent speaker identification methods was constructed using MNN with majority vote, including the GRNN, PNN and RBF-NN models. The voting is conducted as follows

$$\begin{aligned} \text{VoteCount}(X_i|C_i) &= \text{GRNN_Output}(X_i|C_i) + \text{PNN_Output}(X_i|C_i) \\ &+ \text{RBFNN_Output}(X_i|C_i) \end{aligned} \quad (4)$$

Each test sample is passed through each of the three NNs. If any two of the networks classify the given test sample as belonging to the same speaker from the training data, then the test sample is declared to belong to that speaker. However, where each network classifies the given test sample as a different class, the sample is considered to be 'not identified'.

During the training phase, feature vectors extracted from the training data are fed into each of the networks in parallel. These networks require only one pass through the data in contrast to the multiple epochs/iterations that are used in BPNN. The size (i.e. number of neurons) of the input layer is equal to the number of MFCC features. Each neuron takes in streams of data as inputs that arise from the consecutive frames. Some of the advanced NNs have the size of the input layer enlarged to two or three adjacent frames [33] in order to obtain 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 [34], leading to an extremely large input layer size. However, in both the above cases, the computational times are affected because of 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 [33] (in addition, the network may over fit the data and start modelling random noise).

For testing, the extracted feature vectors from the test signal are fed to all the ANNs in parallel and three classification outputs are calculated corresponding to the three classifiers used here. The majority voting scheme is employed for the classification results of the ANNs. The class that obtains two out of three votes is taken to be the final classification result. During the test phase, the procedure used was almost the same as the one in the training phase. The test speaker's file is pre-processed to extract wavelet features. These features are classified individually by the trained PNN, GRNN and RBF-NN. The majority voting scheme ensures equal weight and the final classification is made on the premises.

In contrast to the BPNN and feed-forward networks, none of these networks requires iterative training, which takes a considerable amount of time. Additionally, each of these

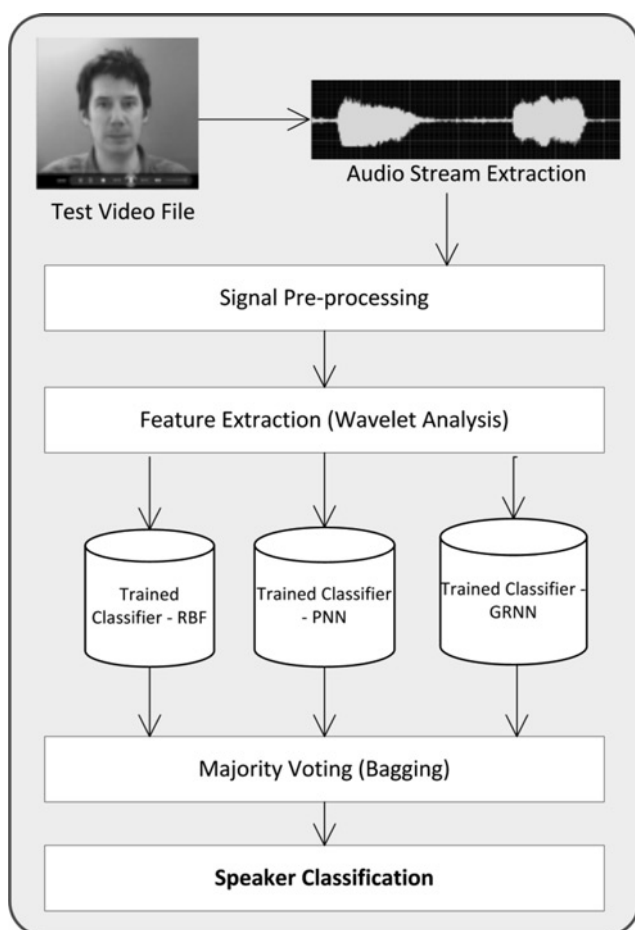


Fig. 5 Proposed system for speaker identification

networks focuses on a different probing level to fit the training data. One of them focuses on the training data completely (over-fitting), the second learns the training data with an error margin (under-fitting) and the third lies between the previous two and thus helps to increase the ability to generalise the overall system for both known and unknown signal instances. Moreover, the combination of these networks with a majority voting scheme helps to overcome the under- and over-fitting problems. This approach improves the classification accuracy of the overall system. Only the fusion of the PNN, RBF-NN and GRNN networks in the voting scheme is capable of reducing the training time and obtain a higher accuracy than those of the BPNN and feed-forward networks. However, such a method would still be faster than methods that use BPNN and feed-forward networks.

4 Results and analysis

In this section, we describe the outcomes of the comprehensive testing performed on the GRID corpus [35]. Below we first describe the experimental procedure in Section 4.1, and then present the accuracy and computational effectiveness of the proposed MNN in Sections 4.2–4.5. We provide a comprehensive analysis and comparisons with some other existing systems in Section 4.6.

4.1 Evaluation methods

The identification experiment was performed using the GRID speech corpus [35]. GRID is a multi-speaker audio-visual sentence database that supports joint computational-behavioural studies in speech perception. GRID consists of high-quality audio and video recordings of 1000 sentences spoken by 18 male and 16 female speakers. It uses a fixed and simple grammatical structure <command:4> <colour:4> <preposition:4> <letter:25> <number:10> <adverb:4>, where the numbers in brackets indicate the number of choices at each point, for example, ‘bin blue at A1 again’ or ‘place green by D2 now’. Speakers produced such sentences at a normal speaking rate and were asked to complete the sentence in 3 s. The reason we chose GRID for our studies is that these sentences control for differences in speaking style and syntax, and the existence of many keyword repetitions allows for cross-condition comparisons of acoustic properties. Different gross phonetic classes (nasal, vowel, fricative, plosive and liquid) were used as the initial or final sounds of filler words in each position [36], thereby allowing a wide range of phonetic features to be captured.

The 10-fold cross-validation experiments were used to test all 34 speakers in the GRID database using different values of the spread. The spread denotes how closely the NN should fit the training data. The default value range for spread is between 0 and 1, with 1 being the most generalised fitting to the training data with relatively lower accuracy. A spread of 0 is a complete close fit to the training data and produces maximum accuracy. We can say 1 under-fits the training data, whereas 0 over-fits the training data. The spread is also known as the radius of a neuron. With larger spread, neurons at a distance from a point have a greater influence. There is a trade-off in choosing different values of spread between 0 and 1. This variable was chosen as the base variable and 30 different values were assigned to it. Therefore, they resulted in 30 different experiments on the

same data from the 34 speakers in GRID. Since all the utterances recorded in the GRID corpus have the same length and sampling rate, they transform to the same number of frames, same MFCC output vector length and same number of neurons in the input layer of the subsequent NN. The averaged identification results are presented in the subsequent sections. The test results presented in this section were collected on a computer with a 2.8 GHz Intel Core 2 Duo processor and 4 GB of memory.

4.2 Audio and feature extraction

A speech signal contains a massive amount of data. For example, a 1 s speech signal consists of ~25 000–50 000 floating-point values in a single linear vector. GRID database files have a fixed sampling rate of 25 kHz. An audio signal is usually segmented into frames of 10–30 ms with some overlap [37]. Each frame has to be multiplied with a Hamming window in order to keep the continuity of the first and the last points in the frame. Overlapping windows allow analysis centred at a frame point. In our case, the audio signal is divided into 15 ms frames using Hamming windows with a 10 ms overlap to smooth out the frequencies at the edges of each frame or window. An audio signal is constantly changing, but we assume that on short-time scales the audio signal does not change much. If the frame is too short, we do not have enough samples to obtain a reliable spectral estimate; if it is too long, the signal changes too much throughout the frame. A 15 ms window at 25 kHz (for GRID database) transforms to 375 samples, which is enough to obtain a reliable spectral shape. Although some researchers tend to choose a larger frame size, several others [38–40] have found 15 ms frame sizes more useful than longer ones depending on the database and methodology applied. We employed different frame sizes in our studies and 15 ms turned out to be the best choice.

The resultant frames are further processed using logarithmically spaced filters and Mel-scale filters, Fourier transforms and cosine transforms to produce an MFCC vector for each frame. The number of filters in the Mel-scale filter is adjusted to control the number of MFCC features. The delta and delta-delta features are computed using linear regression formulas. These additional features have the capability of performing better than an MCCC-only implementation, but usually incur enormous numerical burden. Later in this section, we perform some tests to find out the appropriate balance of these features.

The feature extraction block of this system consists of the following algorithms: DWT, WPT, WSBC and irregular decomposition. All feature vectors are linear vectors of

Table 1 Summary of feature extraction vectors for wavelet analysis

Input	Feature extraction scheme	Output vector length
1 s long audio signal (GRID) recorded at 44.1 kHz	discrete wavelet transform	8
	wavelet packet transform	64
	WPT in Mel-scale (WSBC)	6
	irregular decomposition	57
	MFCC	20 × 450

length ≤ 64 , as summarised in Table 1. During the scope of this research, experimentation and testing were performed with all of these approaches, selecting one at a time. In this phase, the rough audio signal is pre-processed to extract only the distinguishing features from the entire signal for analysis. Only one of the above strategies is used at a time, and the programme allows the user to select the feature extraction strategy during the training and testing phase. Each feature yields a different set of parameters and a unique training data on which the programme trains itself. The testing phase includes the feature extraction strategy to generate consistent results. Experimental results show that WPT generates the most accurate results as described in Section 4.3.

One of our goals was to establish the optimum number of MFCC features for our case. Fig. 6 summarises the results obtained with the various numbers of MFCC features used. These results were collected for 20 files per speaker using 10-fold cross-validation. This experiment indicates that varying the number of MFCC features gradually from 10 to 20 has a notable impact on the overall accuracy of the system, but the gain starts decaying beyond 20 features. However, the number of MFCC features is directly proportional to computation time. Therefore, 20-MFCC features were found to be the best trade-off value between the computation time and the overall accuracy.

We have also incorporated the delta and delta–delta MFCC features into our framework to check whether they offer any additional performance gain. As stated before, adding 20 features each from the delta and delta–delta MFCCs will result in a total of 60-element feature vector, thereby tremendously affecting the computation time. In Table 2, we provide a performance summary on changing the numbers of MFCC, delta MFCC and delta–delta MFCC features, where the total number of features is 20 or close. It can be seen that the all-MFCC case performs better than any of the other cases. If we use all three types of features in equal amounts, we need a total of 48 elements to get close to the 20-MFCC case. These results imply that these time derivative features are not good substitutes for having more MFCC features, especially when the computational burden has to be accounted for. The classification accuracy presented in [41] also supports a similar observation, where different feature sets perform unevenly and the differential

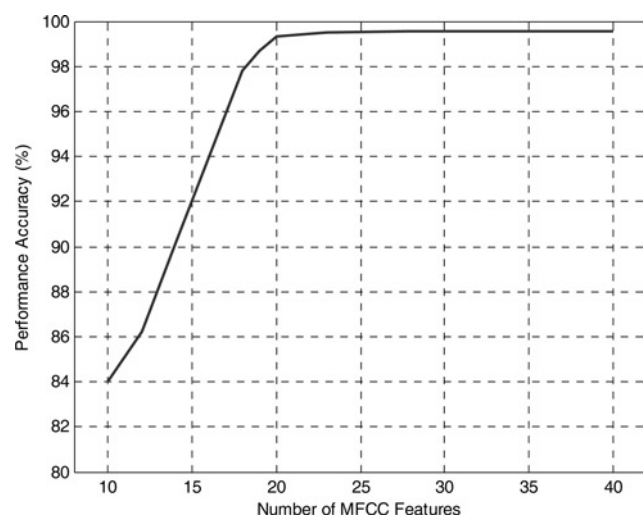


Fig. 6 Accuracy (%) of speaker identification using MFCC features and GMM

Table 2 Comparison of MFCC, delta MFCC and delta–delta MFCC

Number of features from			Total number of features	performance accuracy, %
MFCC	Delta MFCC	Delta–delta MFCC		
20	0	0	20	99.3
0	20	0	20	52.5
0	0	20	20	36.9
10	10	0	20	91.1
10	0	10	20	86.4
0	10	10	20	77.5
6	6	6	18	66.6
7	7	7	21	72.3
8	8	8	24	76.2
16	16	16	48	98.9

features do not perform very well. Therefore, we choose to use 20-MFCC features and none of the delta or delta–delta features in the remainder of our experiments.

We also experimented to improve the accuracy of the model with respect to the number of Gaussian mixtures allowed per GMM. The objective is to choose the best mixture components to achieve high discrimination accuracy. Theoretically, too few mixture components can produce a GMM model which does not accurately model the distinguishing characteristics of a speech distribution. However, too many components can reduce performance when there are a large number of model parameters relative to the available training data and can also result in excessive computational complexity [42].

In the tests performed in [43], the results show that, as the number of Gaussians in GMM increase from 2 to 32, the average speech entropy in each Gaussian decrease while the average speaker entropy remains near constant. We varied the order of this mixture gradually from 1 to 16 to find the most appropriate value. Fig. 7 summarises the results. It can be seen that increasing the order does not necessarily increase the system accuracy; as a matter of fact, there are some uneven fluctuations at certain values. The large orders caused very high computational expense, but did not seem to yield good performance for the small amount of available training data. The mixture component selection is limited by the amount of training data. Model order selection becomes more important with smaller amount of training data. On further investigation, we found that many of the mixtures reduced to single points, as they did not have enough values to carry on further computation. However, the above experiment shows that 1 and 2 Gaussian mixtures provide the optimum accuracy for voice.

4.3 Identification accuracy

The same testing criteria were applied to GMM, BPNN and principal component analysis (PCA) for comparison with the proposed MNN. The wavelet packet analysis (8-level) generated a 97.5% identification rate, which is a large improvement compared with MFCC (with 20 feature vectors), which had a 77.5% identification rate for these experiments. The findings in [23] suggest that irregular decomposition generates better results than DWT, WPT and WPT in Mel-scale algorithms. In contrast, we found that both WPT in Mel-scale and irregular decomposition were less accurate compared with WPT when tested with the GRID speech database. This difference is because [23] used a limited set of five sentences for each speaker, which

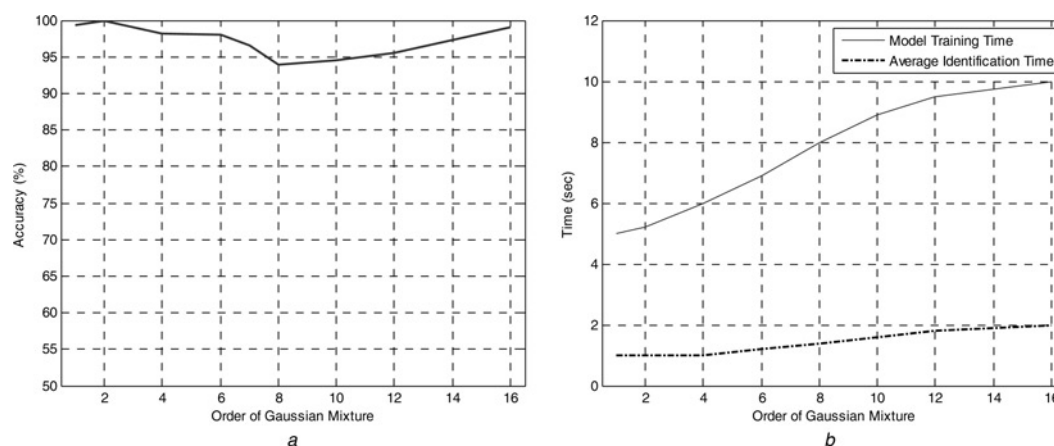


Fig. 7 Effect of the Gaussian mixture order on

a Accuracy

b Model training and average identification times

generated a text-dependence for the training data, whereas in our training set, the user speaks up to 1000 different sentences yielding a text-independent data set.

The performance results are summarised in Table 3 and show that our proposed system yields the most accurate results (97.5%) for text-independent speaker identification compared with an established set of algorithms including GMM, PCA, the parallel classifier model in [10] and BPNN. Note that in [10] text-dependent classification was used.

It is noteworthy that MFCC produces a two-dimensional (2D) matrix, whereas BPNN is by nature designed to cater for 1D input. Therefore there is a mismatch between these algorithms and they cannot be used together without losing a large part of the information. Hence, there is no data for MFCC and parallel BPNN in Table 3. Moreover, it can be observed that MNN outperforms the other classifiers for most feature extraction schemes as expected (in most columns of Table 3, MNN results in the highest identification rate). The only exception is MFCC; in this case, GMM gives a better result than MNN, which is because of the following reason. When GMM is fitted to a smoothed spectrum of speech, an alternative set of features can be extracted from the signal. In addition to the standard MFCC parameterisation, complementary information is embedded in these extra features. Combining GMM means with MFCC by concatenation into a single feature vector can therefore improve identification performance. This is the reason why MFCC performs the best with GMM. However, the best result in this table is achieved using MNN when used with WPT.

4.4 Receiver operating characteristics

We also calculated the receiver operating characteristic (ROC) curve illustrating the performance superiority of our proposed system. The ROC curve shows the true positive

rate (TPR) as a function of the false positive rate (FPR) for different values of the spread. The fraction of true positives out of the total actual positives is known as the TPR, and the fraction of false positives out of the total actual negatives is called the FPR. The FPR is the same as the complement of specificity (i.e. one minus specificity), also known as the FRR. TPR is also known as sensitivity, and is the complement of the FAR. The ROC curve is a graphical plot that can illustrate a binary classifier's performance with variation of the discrimination threshold. TPR and FPR depend on the size of the enrollment database and the decision threshold for the matching scores and/or number of matched identifiers returned. Therefore, a ROC curve plots the rate of accepted impostor attempts against the corresponding rate of true positives parametrically as a function of the decision threshold. The results can be changed by adjusting this threshold. The 30 experiments we conducted produced different combinations of the FPR and TPR. These two values were plotted to generate a ROC curve for DWT, WPT, WSBC, irregular decomposition and MFCC for a comparison of accuracy with the proposed MNN as shown in Fig. 8. This curve shows that the ROC curve for WPT lies very close to the upper left boundary and has more area under it compared with DWT, WSBC, MFCC and irregular decomposition. The ROC curve for MFCC with the same data lies closest to the diagonal and shows the least effective accuracy as compared with the rest of the pre-processing algorithms. A speech identification system would be far from usable if the TPR is too low or the FPR is too high. The goal is to operate with low values of FPR and high values of TPR; therefore the upper-left portion of this figure is the practical region of operability.

The variations of FAR and FRR with different sets of threshold values for WPT, WSBC and DWT are shown in Fig. 9. Since WSBC and DWT are the closest competitors of WPT (as depicted in the ROC of Fig. 8), we only

Table 3 Accuracy rate of the MNN compared with other algorithms

	DWT, %	WPT, %	MFCC, %	WSBC, %	Irregular decomposition, %
GMM	35.80	38.60	83.30	36.50	33.26
MNN	84.70	97.50	77.50	94.40	80.80
BPNN	40.38	41.47	21.20	34.48	32.07
PCA	—	—	82.90	—	—
parallel BPNN [10]	61.25	65.43	—	58.67	56.85

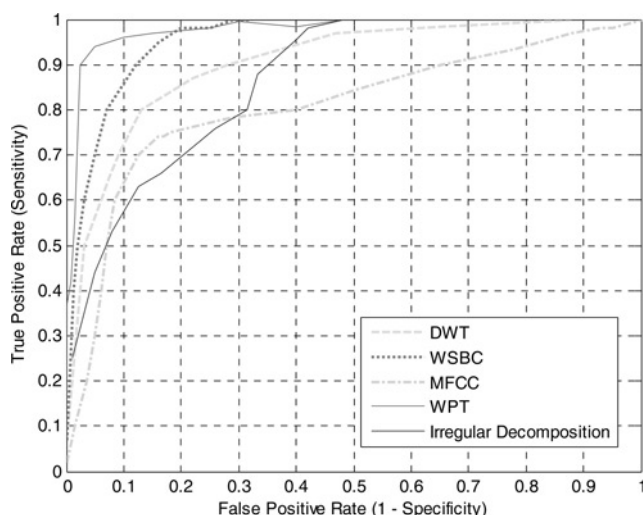


Fig. 8 ROC curves for various wavelet-analysis algorithms tested using the proposed MNN

compared their performances to WPT and omitted the other algorithms for legibility. For each of the algorithms, we plotted the FAR–FRR pairs for different thresholds, and placed the equal error rate (EER) performance (when FAR equals FRR) in the middle. Note that the threshold values are not of interest here, and they are different for different algorithms; we normalised them to align their EER values. Fig. 9 shows that WPT can achieve a much lower EER (about 5%) than the other two schemes. If we inspect this figure along any vertical line, we can see that the FAR and/or FRR for WPT are less than those of the other schemes. Both Figs. 8 and 9 suggest that WPT combined with the proposed fusion system of MNN outperforms DWT, WSBC, MFCC and irregular decomposition.

4.5 Operational speed

The proposed system has 2-fold advantages in terms of accuracy and speed. PCA, because of its dual nature (a classifier and a dimensionality reduction algorithm), is compatible with MFCC as the feature extraction strategy. Table 4 shows the training time and identification time of the state-of-the-art algorithms in speaker identification

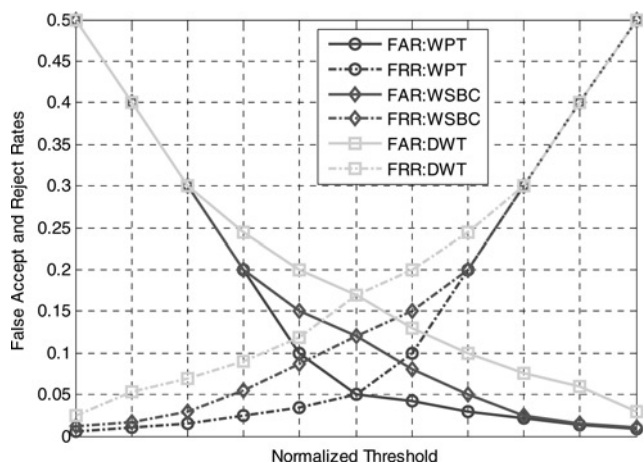


Fig. 9 FARs and FRRs of the proposed MNN compared with different schemes

Table 4 Average training and identification times for different algorithms

	GMM	MNN	3-BPNN [10]	PCA	BPNN
average training time, s	5.8	0.8	120	2.8	90
average identification time, s	2.5	0.05	0.10	1.5	0.8

compared with our MNN. The proposed MNN takes only 0.05 s on average for the identification phase of a test signal. This is the fastest identification time ever seen in a text-independent SIS. Training is the phase where all the time is spent in traditional systems, but the proposed system takes care of it as it employs the instantaneously adaptable classifiers in parallel with no training time.

We have also performed some simulations to test the effect of changing the size of the NN 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. However, as shown in Table 5, the increased size can severely affect the identification and training times. Hence, we conclude that the size of the input layer is best set equal to the number of MFCC features.

4.6 Performance improvement

The performance results described in Table 3, performance times in Table 4 and the ROC curve in Fig. 8 comprehensively validate that the proposed system is more sophisticated because it outperforms other systems both in accuracy and performance times. Below we analyse the reasons and trends of our MNN and then compare them with some state-of-the-art techniques.

In the research results reported in [1], it was suggested that irregular decomposition yields better results than DWT, WPT and WPT in Mel-scale algorithms. On the contrary, we found that both WPT in Mel-scale and irregular decomposition were less accurate than WPT. Pawar *et al.* [1] used a limited, text-dependent set of five sentences for each speaker, whereas our training set is truly text-independent with the user speaking up to 1000 different sentences. One of the main reasons why our system works more efficiently is because of the application of the majority voting scheme during the parallel combination of three classifiers, which are all fast and robust. These classifiers may suffer from inadequacies such as under- or over-fitting problems when used alone, but mitigate each other’s shortcomings when combined in the proposed manner. In summary, the proposed system owes its performance improvement to: (a) bootstrap aggregating of multiple classifiers for a better

Table 5 Average training and identification times for MNN with different input layer sizes

Size of the input layer of NN	20	40	60
average training time, s	0.8	3.5	8.1
average identification time, s	0.05	0.32	0.85

Table 6 Performance comparison with state-of-the-art speaker identification approaches

References	Algorithm, database	Performance accuracy, %
[44]	32-mixture GMM, UBM, MFCC, spectro-temporal modulation, GRID corpus	91.7
[45]	exemplar-based sparse representation, sparse discriminant analysis, dot-scoring, GRID	95.5
[46]	GMM-UBM (mixed-UBM and multi-conditioned GMMs), GRID corpus	85
[47]	GMM speech prior, single mixture HMM, speech segregation, GRID corpus	96.3
[48]	GMM-UBM (independent of pre-processing algorithm), TIMIT	96.8
[49]	LPC, K-means TI digits_1, TI digits_2 and TIMIT databases	96.37
[50]	MFCC, parametric neural network CSLU speaker recognition corpora	90.6
proposed multimodal neural network	wavelets, MNN GRID corpus	97.5

hypothesis in the decision space; (b) careful selection of multiple combined ANN instances of the same class that complement each other by tackling the under- and over-fitting problems; and (c) selection of the most suitable feature extraction strategy (i.e. WPT). Instead of using other recently popular methods like BPNN methods, we explored the more adaptive instantly trained class of NNs, which substantially improved the classification accuracy and reduced the identification time.

The recent development in human identification in other areas of the world has been inspiring and competitive. For any novel method to succeed, the comparative analysis of performance with the state-of-the-art methods is deemed necessary. Table 6 shows the reported performance of some of the existing methods in line with the performance accuracy of our system. It shows that our system outperforms several other published systems that achieve some of the best identification rates available in the literature. In the first four rows of Table 6, we provide comparisons with systems that used the same corpus as ours (GRID). For the purpose of comparison across different databases (e.g. CSLU and TIMIT), in the last three rows of the table, we also show results based on some other widely used databases. A short description of these systems follows.

In [44], an algorithm which distinguishes speech from non-speech based on spectro-temporal modulation energies is proposed and evaluated in robust text-independent closed-set speaker identification simulations. An exemplar-based representation and sparse discrimination was proposed in [45] that outperformed the baseline GMM-universal background model (UBM) and HMM-based systems with a large margin. The GMM-UBM system in [46] has shown an average 85% identification accuracy on GRID corpus when a mixed-UBM and multi-conditioned GMMs are utilised. The work in [47] presents a novel fragment-based speaker identification

approach that allows the target speaker to be reliably identified across a wide range of signal-to-noise ratios by treating segregation and recognition as coupled problems. The system in [48] is mainly based on the verification using the likelihood ratio test. The likelihood functions used some effective GMMs that are relatively simple and easy to implement. For speaker representation, it employed the UBM, from which speaker models were derived using Bayesian adaptation. The verification accomplishment was further enhanced using score normalisation. Their performance was successfully tested in several NIST speaker recognition evaluations. In [49], a robust perceptual features and iterative clustering approach are proposed for isolated digits and continuous speech recognition and speaker identification, and its evaluation is performed on clean test speeches. A new SIS based on a modified NN was proposed in [50], namely the multiple parametric self-organising map (M-PSOM). It attempted to reduce the acceptance of impostors while maintaining a high accuracy for identification. Most of the prior systems would rely on a single NN for an entire SIS, but the M-PSOM utilises parametric NNs for the individual speakers to record and depict their distinctive acoustic signatures. This paper demonstrated that this method outperforms many other competitive methods like wavelets, GMM, HMM and vector quantisation. Our proposed approach outperforms all these published systems in terms of accuracy, and therefore proves itself as one of the best candidates for speaker identification.

5 Conclusions

Conventional approaches to speaker identification with slow identification and poor accuracy are inadequate in a real-world setting. We have been motivated by these shortcomings to conceive and implement a novel approach in this paper. We developed a novel approach that combines multiple NNs with wavelet analysis to construct a method that outperforms classical GMM, BPNN and PCA in both identification time and accuracy. Through comprehensive testing using the GRID database, the system described herein is 97.5% accurate with a 50 ms identification time where WPT is the feature extraction method. Our real-time approach is directly applicable to industrial devices for security and authentication. This method lays the foundation for further research in speaker identification for real-time systems. In the future, to further develop the approach described herein, we will combine real-time facial recognition with speaker identification to generate a more robust system that is applicable for the industry. Moreover, we will combine audio and visual features at the feature level with MNNs to further improve the accuracy.

6 References

- 1 Pawar, R.V., Kajave, P.P., Mali, S.N.: 'Speaker identification using neural networks'. Proc. World Academy of Science, Engineering and Technology, 2005, no. 7, pp. 429–433
- 2 Rabiner, L., Juang, B.H.: 'Fundamentals of speech recognition' (Prentice-Hall, 1993)
- 3 Kinsner, W., Peters, D.: 'A speech recognition system using linear predictive coding and dynamic time warping'. Proc. Annual Int. Conf. IEE, Engineering in Medicine & Biology Society, New Orleans, LA, 4–7 November 2006, no. 3, pp. 1070–1071
- 4 Benesty, J., Sondhi, M., Huang, Y.: 'Springer handbook of speech processing' (Springer, 2007)

- 5 Abdalla, M.I., Ali, H.S.: 'Wavelet-based Mel-frequency cepstral coefficients for speaker identification using hidden Markov models', *J. Telecommun.*, 2010, **1**, (2), pp. 16–21
- 6 Suvama Kumar, G., Prasad Raju, K.A., Rao, M., *et al.*: 'Speaker recognition using GMM', *Int. J. Eng. Sci. Technol.*, 2010, **2**, (6), pp. 2428–2436
- 7 Kekre1, H.B., Kulkarni, V.: 'Speaker identification by using vector quantization', *Int. J. Eng. Sci. Technol.*, 2010, **2**, (5), pp. 1325–1331
- 8 Campbell, W.M., Assaleh, K.T., Broun, C.C.: 'Speaker recognition with polynomial classifiers', *IEEE Trans. Speech and Audio Processing*, 2002, **10**, (4), pp. 205–212
- 9 Wang, J.C., Yang, C.H., Wang, J.F., Lee, H.P.: 'Robust speaker identification and verification', *Taiwan IEEE Computational Intelligence Magazine*, 2007, **2**, (2), pp. 52–59
- 10 Shukla, A., Tiwari, R., Hemant Kumar, M., Kala, R.: 'Speaker identification using wavelet analysis and modular neural networks', *J. Acoust. Soc. India (JASI)*, 2009, **36**, (1), pp. 14–19
- 11 Revada, L.K.V., Rambatla, V.K., Ande, K.V.N.: 'A novel approach to speech recognition by using generalised regression neural networks', *IJCSI Int. J. Comput. Sci. Issues*, 2011, **1**, pp. 483–489
- 12 Moody, J., Darken, C.J.: 'Fast learning in networks of locally-tuned processing units', *Neural Comput.*, 1989, **1**, (2), pp. 281–294
- 13 Hall, D.L., Llinas, J.: 'Handbook of multi-sensor data fusion' (CRC Press, UK, 2011)
- 14 Ross, A., Jain, A.: '*Information fusion in biometrics*', *Pattern Recognit. Lett.*, 2003, **24**, (3), pp. 2115–2125
- 15 Nefian, A., Liang, L., Pi, X., Liu, X., Murphy, K.: 'Dynamic Bayesian networks for audio-visual speech recognition', *EURASIP J. Adv. Signal Process.*, 2002, **11**, pp. 1274–1288
- 16 Chetty, G., Wagner, M.: 'Audio visual speaker verification based on hybrid fusion of cross modal features', in *Pattern Recognition and Machine Intelligence*, (Springer, Berlin, 2007)
- 17 Chetty, G., Wagner, M.: 'Investigating feature-level fusion for checking liveness in face-voice authentication'. *Int. Symp. on Signal Processing and its Applications*, 2005, vol. 1
- 18 Arora, S., Bhattacharjee, D., Nasipuri, M., Malik, L., Kundu, M., Basu, D.K.: 'Performance comparison of SVM and ANN for handwritten Devnagari character recognition', *IJCSI Int. J. Comput. Sci.*, 2010, **7**, (3), pp. 1–10
- 19 Xiang, B., Berger, T.: 'Efficient text-independent speaker verification with structural Gaussian mixture models and neural network', *IEEE Trans. Speech Audio Process.*, 2003, **11**, (5), pp. 447–456
- 20 Mallat, S.: 'A wavelet tour of signal processing' (Elsevier, UK, 1999)
- 21 Lung, S., Chen, C.: 'Further reduced form of Karhunen–Loeve transform for text independent speaker recognition', *Electron. Lett.*, 1998, **34**, (14), pp. 1380–1382
- 22 Vetterli, M., Kovacevic, J.: 'Wavelets and subband coding' (Prentice-Hall, New Jersey, 1995)
- 23 Wu, J.D., Lin, B.F.: 'Speaker identification using discrete wavelet packet transform technique with irregular decomposition', *Expert Syst. Appl.*, 2009, **36**, (2), pp. 3136–3143
- 24 Deshpande, M.S., Holambe, R.S.: 'Speaker identification using admissible wavelet packet based decomposition', *Int. J. Inf. Commun. Eng.*, 2011, **6**, (1), pp. 20–23
- 25 Lung, Y.: 'Feature extracted from wavelet eigenfunction estimation for text-independent speaker recognition', *Pattern Recognit.*, 2004, **37**, pp. 1543–1544
- 26 Lung, Y.: 'Improved wavelet feature extraction using kernel analysis for text-independent speaker recognition', *Digit. Signal Process.*, 2010, **20**, (5), pp. 1400–1407
- 27 Specht, D.F.: '*A general regression neural network*', *IEEE Trans. Neural Netw.*, 1991, **2**, (6), pp. 568–576
- 28 Amrouche, A., Rouvaen, J.: 'Efficient system for speech recognition using general regression neural network', *Int. J. Intell. Technol.*, 2006, **1**, (2), pp. 183–189
- 29 Lu, W., Sun, W., Lu, H.: 'Robust watermarking based on DWT and non-negative matrix factorization', *Comput. Electr. Eng.*, 2009, **35**, (1), pp. 183–188
- 30 Ye, J.: 'Speech recognition using time domain features from phase space reconstructions'. PhD thesis. Marquette University Milwaukee, Wisconsin, 2004
- 31 Furui, S.: 'Speaker-independent isolated word recognition using dynamic features of speech spectrum', *IEEE Trans. ASSP*, 1986, **34**, (1), pp. 52–59
- 32 Wilpon, J.G., Lee, C.H., Rabiner, L.R.: 'Improvements in connected digit recognition using higher order spectral and energy features'. *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, Toronto, Canada, 1991
- 33 Rottland, J., Neukirchen, C., Willett, D., Rigoll, G.: 'Large vocabulary speech recognition with context dependent MMI-connectionist/HMM systems using the WSJ database'. *EUROSPEECH*, 1997
- 34 Hamzah, R., Jamil, N., Seman, N.: 'Filled pause classification using energy-boosted Mel-frequency cepstrum coefficients'. *Proc. Int. Conf. on Robotic, Vision, Signal Processing & Power Applications*, 2014, pp. 311–319
- 35 'The GRID audio corpus for speech recognition'. Available at <http://www.dcs.shef.ac.uk/spandh/gridcorpus>
- 36 Cooke, M., Barker, J., Cunningham, S., Shao, X.: 'An audio-visual corpus for speech perception and automatic speech recognition', *J. Acoust. Soc. Am.*, 2006, **120**, (5), pp. 2421–2424
- 37 Holmes, W.: *Speech synthesis and recognition*, (CRC Press, UK, 2001)
- 38 Gelbart, D.: 'Ensemble feature selection for multi-stream automatic speech recognition'. Technical Report No. UCB/ECS-2008-160, University of California at Berkeley, December 2008
- 39 Mirhassani, S.M., Ting, H.N.: 'Fuzzy-based discriminative feature representation for children's speech recognition', *Dig. Signal Process.*, 2014, **31**, pp. 102–114
- 40 Morris, A., Bloothoof, G., Barry, W., Andreeva, B., Koreman, J.C.: 'Human and machine identification of consonantal place of articulation from vocalic transition segments'. *EUROSPEECH*, 1997
- 41 Li, D., Sethi, I., Dimitrova, N., McGee, T.: 'Classification of general audio data for content-based retrieval', *Pattern Recognit. Lett.*, 2001, **22**, (5), pp. 533–544
- 42 Morris, A., Wu, D., Koreman, J.: 'GMM based clustering and speaker separability in the TIMIT speech database', *IEICE Trans. Fundam. Syst.*, 2005, **85**, pp. 1–8
- 43 Reynolds, D.: 'Robust text-independent speaker identification using Gaussian mixture speaker models', *IEEE Trans. Speech Audio Process.*, 1995, **3**, (1), pp. 72–83
- 44 Chi, T.S., Lin, T.H., Hsu, C.C.: 'Spectro-temporal modulation energy based mask for robust speaker identification', *J. Acoust. Soc. Am.*, 2012, **131**, (5), pp. 368–374
- 45 Gemmeke, J., Virtanen, T., Hurmalainen, A.: 'Exemplar-based sparse representations for noise robust automatic speech recognition', *IEEE Trans. Audio Speech Lang. Process.*, 2011, **19**, (7), pp. 2067–2080
- 46 Saeidi, R., Mowlae, P., Kinnunen, T., Tan, Z., Christensen, M., Jensen, H., Franti, P.: 'Signal-to-signal ratio independent speaker identification for co-channel speech signals'. *Proc. IEEE Int. Conf. Pattern Recognition*, 2010, pp. 4545–4548
- 47 Barker, J., Ma, N., Coy, A., Cooke, M.: 'Speech fragment decoding techniques for simultaneous speaker identification and speech recognition', *Comput. Speech Lang.*, 2010, **24**, (1), pp. 94–111
- 48 Reynolds, D., Quatieri, T., Dunn, R.: 'Speaker verification using adapted Gaussian mixture models', *Digit. Signal Process.*, 2000, **10**, (3), pp. 19–41
- 49 Revathi, A., Ganapathy, R., Venkataramani, Y.: 'Text independent speaker recognition and speaker independent speech recognition using iterative clustering approach', *Int. J. Comput. Sci. Inf. Technol.*, 2009, **1**, (2), pp. 30–42
- 50 Gomez, P.: 'A text independent speaker recognition system using a novel parametric neural network', *Int. J. Signal Process., Image Process. Pattern Recognit.*, 2011, **1**, pp. 1–16