Speaker Recognition Systems: A Tutorial

Abimbola A. Fisusi **Thomas K. Yesufu**
Department of Electronic and Electrical Engineering,
Obafemi Awolowo University,
Ile-Ife, Osun State, Nigeria.
bimbofisusi@oauife.edu.ng *tyesufu@oauife.edu.ng*

Abstract
This paper gives an overview of speaker recognition systems. Speaker recognition is the task of automatically recognizing who is speaking by identifying an unknown speaker among several reference speakers using speaker-specific information included in speech waves. The different classification of speaker recognition and speech processing techniques required for performing the recognition task are discussed. The basic modules of a speaker recognition system are outlined and discussed. Some of the techniques required to implement each module of the system were discussed and others are mentioned. The methods were also compared with one another. Finally, this paper concludes by giving a few research trends in speaker recognition for some years to come.

Keywords: *speaker recognition, feature extraction, pattern matching, mel-frequency cepstrum coefficient, vector quantization.*

I. Introduction

Human speech conveys different types of information. The primary information the human speech carries is the meaning of the words being spoken. However, it also carries other information like the language being spoken, the gender and identity of the speaker and the emotional state of the speaker. Speech processing technology exploits the various kinds of information present in speech signal for different applications. Speech processing is a wide field and includes such areas as analysis, synthesis, recognition and coding of speech. Recognition could be either speech recognition, speaker recognition or language recognition. The focus of this paper is the review of the speaker recognition field.

Speaker recognition is the task of automatically recognizing who is speaking by identifying an unknown speaker among several reference speakers using speaker specific information included in speech waves [1]. The goal of speaker recognition is therefore to extract this speaker-specific information, characterize and use them for identification purposes. It is different from speech recognition and language recognition since these concepts deal with recognition of speech (i.e. the words that are spoken) and recognition of language (i.e. the language in which the words or sentences are spoken).

Speaker recognition can be divided into speaker verification and speaker identification. Speaker verification is the use of a machine to verify a person's claimed identity from his voice [2]. In speaker identification, there is no prior identity claim, and the system decides who the person is among the several enrolled speakers. Speaker verification is defined as deciding if a speaker is whom he claims to be. This is different from the speaker identification problem, which is deciding if a speaker is a specific person or is among a group of persons. Speaker verification is an easier task to perform than speaker identification because it is a case of one-to-one matching as compared to the one-to-many comparison of the speaker identification case [3]. Speaker identification can be further divided into two categories, closed-set and open-set problems. The closed set problem is to identify a speaker from a group of known speakers. Alternatively, one may want to decide whether the speaker of a test utterance belongs to a group of known speakers. This is called the open-set problem since the speaker to be identified may not be one of the known speakers [4].

Speaker recognition can also be classified as text-dependent or text-independent [5]. In text-dependent speaker recognition, the recognition system has prior knowledge of the text that will be spoken by the user and it is expected that the user will cooperate. For example, the user speaking one or more specific phrases, like passwords, card numbers, PIN codes. In text-independent systems, the user is free to utter any word for the system to perform the recognition process. The text-independent option is a more difficult task than text-dependent based systems, although it is more flexible because no restriction is placed on what the user can say and this will make it possible to perform recognition during normal conversation.

Human speech signal conveys different types of information which can be classified into high-level and low-level information. High level information includes dialect, context, speaking style, emotional state and gender of

the speaker. Low level information includes the pitch, intensity, bandwidth and short-time spectrum of the speech signal. Earlier speaker recognition projects were implemented using low level information (which are related to the physical traits of the vocal apparatus) [6]. However, high level information have been utilized in some recent projects and with favorable results [6, 7]. Speaker recognition is usually composed of two phases: training phase and testing phase. During the training phase, speech sample is collected from each user and used to build speaker models for the user. These models are stored and used during the testing phase. During the testing phase, speech sample from the users are matched with the stored speaker models created during the training phase and recognition decision is made. Fig. 1 shows the stages involved in the implementation of speaker verification and identification systems. Generally, speaker recognition involves feature extraction, speaker modelling, pattern matching, classification and decision making stages or modules. These modules are discussed in subsequent sections of the paper.

II. Speech Production

Speech is the major means of communication among human beings and it conveys different kinds of information which include the meaning of the word spoken, the emotional state of the speaker, the gender of the speaker and the identity of the speaker. Information like accent and speaking style can also be conveyed by speech.

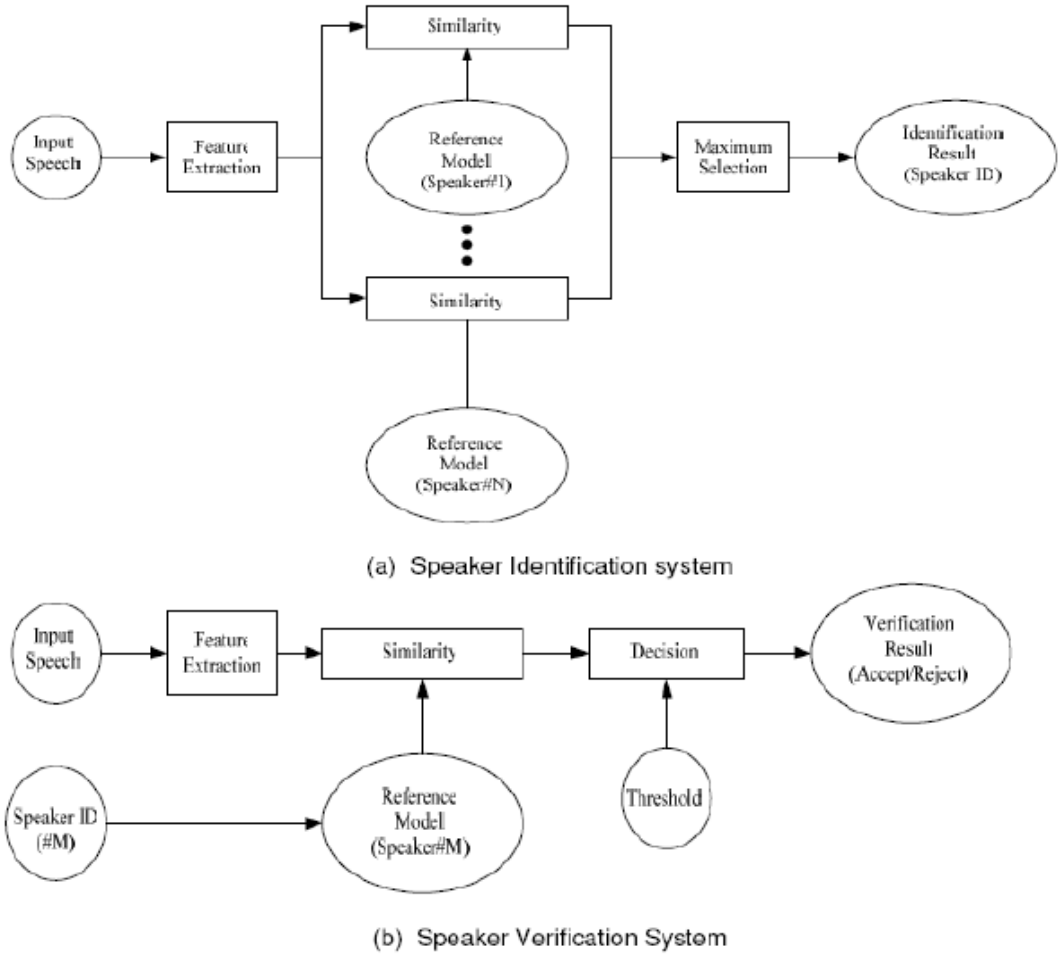


Fig. 1 Block diagram of the different classifications of Speaker Recognition Systems

Speech production begins with a thought in the mind of the speaker. The speaker then converts the thought into words and phrases in line with the language he or she wants to speak. Finally, the brain produces motor nerve commands, which move the vocal organs in an appropriate way [8]. Speech production involves three stages: excitation production, vocal tract articulation and radiation from the lips and/or nostrils. Understanding of the process of speech production is fundamental to the task of speaker recognition.

A. Excitation Production

Sound is an acoustic pressure produced from the compressions and rarefactions of air molecules that originate from movements of human speech organs. Fig. 2 shows a diagram of the human speech organs. The important components of the human speech production system are the lungs (source of air during speech), trachea (windpipe), larynx or its most important part - vocal folds (organ of voice production), nasal cavity (or nose), soft palate or velum (allows passage of air through the nasal cavity), hard palate (enables consonant articulation), tongue, teeth and lips. These components are called articulators; they move to different positions to produce various sounds. Based on their production, speech sounds can be divided into consonants and voiced and unvoiced vowels [8].

Excitation is produced by airflow from the lungs and carried by the trachea through the vocal folds [2]. The vocal folds are a pair of elastic muscles or membranes that extend from the front of the larynx (the thyroid cartilage) to the back (arytenoids) [9]. Excitation can be classified into three major categories: voiced, unvoiced and plosive. The vocal folds can either be in a relaxed state or tensed state. When it is in a relaxed state it is open and allows air pressure from the lungs to pass directly to the vocal tract. In the tensed state, the vocal folds are closed until the air pressure builds up and is high enough to force them apart creating an opening which is referred to as the glottis [9]. After the passage of air through the glottis the pressure on the vocal folds reduces and they are drawn together as a result of the combination of their tension, elasticity and Bernoulli Effect [2]. Voiced excitation is produced as a result of vibratory motion of the vocal folds caused by repeated opening and closing of the vocal folds. The frequency of oscillation is called the fundamental frequency, F0 and it depends upon the length, tension, and mass of the vocal folds [2]. The fundamental frequency is a distinguishing characteristics based on the vocal tract properties.

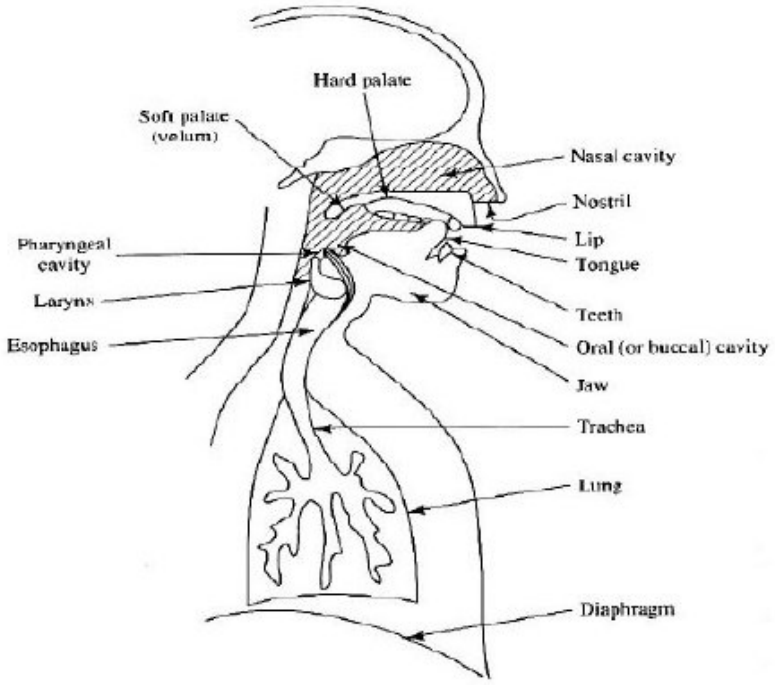


Fig. 2 Human Speech Organs [10]

Unvoiced excitation is produced by forming a constriction at some point in the vocal tract towards the mouth and forcing air through the constriction at a very high velocity to produce turbulence [11]. Constrictions maybe produced by the tongue, teeth or lips. Unvoiced excitation is modeled as white noise.

Plosive excitation is produced when a vocal tract that is completely closed and under pressure experiences a sudden release of pressure.

B. Vocal Tract Articulation

The properties of the speech signal is not determined solely by the nature of the excitation that produced it but also by the geometric dimensions of the three main cavities of the vocal tract [9].

For the purpose of speaker recognition it is more useful to think about the speech production process in terms of an acoustic filtering operation that affects the air that flows from the lungs. There are three main cavities that comprise the main acoustic filter: nasal, oral and pharyngeal cavities. As the acoustic wave passes through the vocal tract its frequency contents are modified by the resonances of the vocal tract. Theses resonances are referred to as formants. Therefore, the shape of the vocal tract can be obtained from the spectral shape of the speech signal [2]. A simple acoustic model of the speech production process is shown in Fig. 3.

C. Speech signal acquisition

Signal processing of speech signal is done in the digital domain. The sound pressure wave is converted to an analog signal with the aid of a microphone or telephone handset. The resulting analog signal is filtered with an anti-aliasing filter in order to limit the signal bandwidth to approximately the Nyquist rate before sampling the signal by an analog to digital converter to obtain a digital signal. The sampling frequency used for speech processing is usually between 8 KHz and 16 KHz [9].

D. Speech Model

A good model of speech that is very useful in speaker identification is the source filter model which is given by the following equation:

$$s(t) = h(t) * p(t) \quad (1)$$

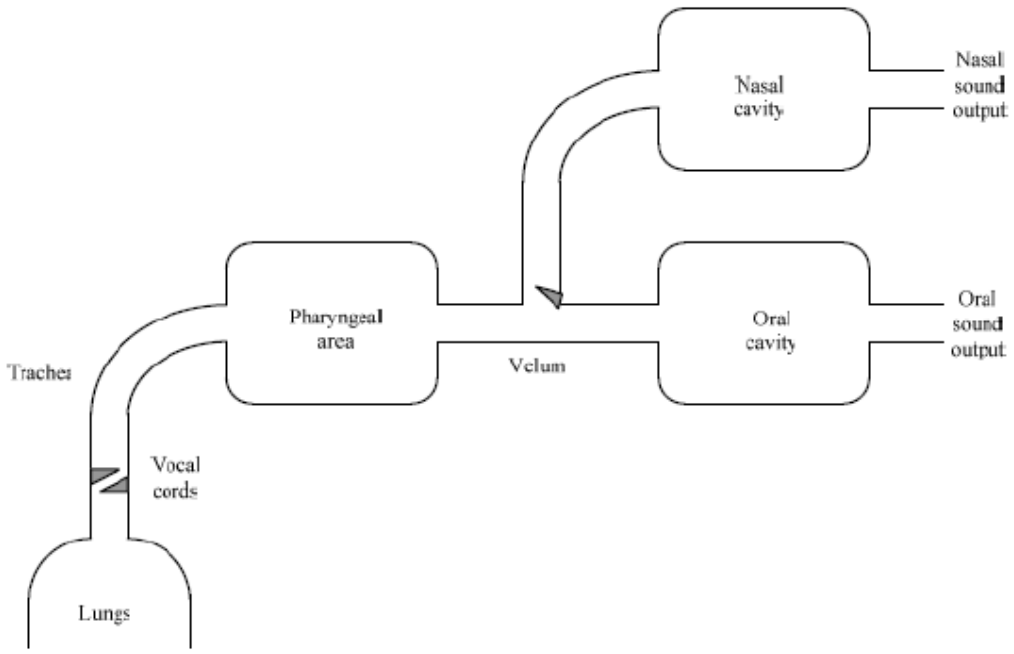


Fig. 3 A simple acoustic model of the speech production process

That is, the speech signal s(t) is the convolution of a filter h(t) and some signal p(t) [12]. Where h(t) is the impulse response of all the things that get in the way of speech emanating from the lungs, e.g. teeth, nasal cavity, lips, etc, while p(t) is the excitation that is refer to as the pitch of speech.

III. Feature Extraction

The main purpose of the feature extraction stage is to extract speaker specific information from speech samples which can be used to build models for the speakers and thus used for the identification task. The feature extraction phase converts the input speech sample into a series of multidimensional vectors each corresponding to a short segment of the input speech sample [13]. This phase converts the speech waveform to some type of parametric representation (at a considerably lower information rate) for further analysis and processing. This is often referred to as the signal-processing front end [14].

A wide range of possibilities exist for parametrically representing the speech signal for the speaker recognition task. Linear Predictive Coding (LPC) and Mel-Frequency Cepstrum Coefficients (MFCC) are the most commonly used features [12]. MFCC is perhaps the best known and most popular [13, 15]. LPC is known to be less expensive but not as effective as MFCC.

A. Linear Predictive Coding (LPC)

Linear Predictive Coding (LPC) is based on the speech production source-filter model with the assumption that the model is an allpole model. It was explained in [8] that the pole-zero system function can be used to represent the speech production process and the assumption to use an all-pole model has two main reasons. Firstly, it simplifies the task of representing the speech signal by merely a linear combination of terms. Secondly, the human ear is deaf to phase information and thus phase information is not so important. The all-pole model can preserve the magnitude spectral dynamics in speech without keeping the phase information. The allpole LPC models a given signal, s_n as a linear combination of its past values and a scaled version of the present input [16].

$$s_n = -\sum_{k=1}^{P} a_k \cdot s_{n-k} + G \cdot u_n \qquad (2)$$

Where s_n is the present output, P is the prediction order, a_k are the model parameters called the predictor coefficients, s_{n-k} are the past outputs, G is the gain scaling factor and u_n is the present input. In speech applications the present input is not known and is therefore ignored. Hence, the LPC approximation $\hat{s}_n$ based on past output values alone is

$$\hat{s}_n = -\sum_{k=1}^{P} a_k \cdot s_{n-k} \qquad (3)$$

The dropping of the present input has simplified the computation since the source (input signal) and the filter (vocal tract) have been decoupled. The source, which corresponds to the vocal tract excitation, is not modeled by the predictor coefficients, a_k and thus some speaker discriminative information in the excitation signal could be lost.

The prediction error is the difference between the real and approximated or predicted output (i.e. $s_n - \hat{s}_n$) and it is also called the prediction residual. In speech applications, the LPC is based on short-term spectra analysis where speech signal is divided into smaller units called frames and a set of prediction coefficient is computed for every frame. These coefficients can be used as features to represent the speech signal and in fact the speakers. In practice, 12-20 coefficients are computed for each frame.

The least-square method is one of the approaches used to compute the prediction coefficient and it selects prediction coefficients to minimize the mean energy in prediction over a frame of speech using methods like the autocorrelation and covariance methods [17]. The cepstrum has proved to be the most effective way of representing speech signal for speech applications, thus the set of prediction coefficients are usually converted to the corresponding Linear Predictive Cepstral Coefficients (LPCC).

B. Cepstrum

Cepstrum computation is used extensively in speaker recognition technology as part of the signal processing used to obtain features from speech samples. It is used in the process of extracting mel-frequency cepstrum coefficients (the most common feature for speaker recognition) from speech

samples. This is the reason why it is discussed before the MFCC method. The word cepstrum is a play on spectrum, and it denotes mathematically:

$$c(n) = \text{ifft}(\log(|\text{fft}(s(n))|)) \quad (4)$$

Where s(n) is the sampled speech signal, and c(n) is the signal in the cepstral domain. Cepstrum computation is a means of discarding the source characteristic of the speech signal since they contain less speaker identification information than the vocal tract characteristics. In practice, the extraction of the two convolved signals from each other is difficult but the cepstrum gives a good approximation for the slow spectral variations, i.e. the envelope structure of the signal, which characterizes the behavior of the vocal tract [12].

Basically cepstrum computation is a deconvolution operation, which decomposes the signal into its source and filter characteristics. Cepstral analysis is used in speaker identification because the speech signal is of the particular form in (1) above, and the "cepstral transform" of it makes analysis incredibly simple.

The convolution of the impulse response of the vocal, h(t) and the source excitation, p(t) will change to multiplication when it is transformed to the frequency domain.

$$c(n) = \text{ifft}(\log(\text{fft}(h(t)*p(t)))) \quad (5)$$

$$= \text{ifft}(\log(H(jw)P(jw))) \quad (6)$$

Then by taking the logarithm the multiplication becomes addition. This results in the desired division into additive components.

$$c(n) = \text{ifft}(\log(H(jw)) + \text{ifft}(\log(P(jw))) \quad (7)$$

Then we can apply the linear operator inverse DFT, knowing that the transform will operate individually on these two parts and that the Fourier transform will put them into different, hopefully separate parts in new, also called *quefrency* axis.

C. MFCC Technique

The technique of computing MFCC is based on the short-term spectrum analysis where the speech signal is divided into frames and feature vectors (MFCCs) are computed from each frame. The stages involved in computing MFCCs are shown in Fig. 4.

1) Frame Blocking

Speech signal is a slowly time varying signal (i.e. quasi-stationary). When it is observed over a sufficiently short period of time (20-30 milliseconds) it has fairly stable acoustic characteristics [8]. This is why the short-time spectral analysis is commonly used to characterize speech signals.

In this stage, the speech signal is divided into smaller units called frames. This is done in such a way that adjacent frames overlap in order to ensure smooth transition from frame to frame. For example, if the first and subsequent frames contain N samples, with the second frame beginning M (M<N) frames after start of the first frame, there will be an overlap of N-M samples between the first and second frames. Furthermore, if the third frame begins 2M samples after the first frame (or M samples after the second frame) it will overlap the second frame by N-M and the first frame by N-2M samples. This process continues until all the speech samples are accounted for within one or more frames. The frame sizes are usually chosen as a power of two to fit the FFT algorithm.

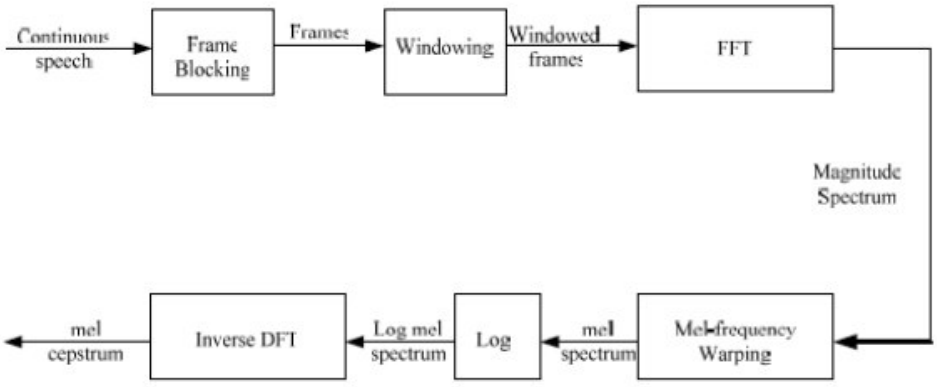


Fig. 4 Block diagram of the MFCC Process

2) Windowing

The next step in computing MFCCs is to apply a window function to each frame of the previous stage. This is done to minimize the discontinuities at the beginning and end of each frame. Each frame is multiplied by a window function which gives greater weight to signals at the centre of the frames and tapers the beginning and end of the frames to zero, so that there is continuity between frames. If the speech signal is defined as $x_i\ (n)$ and N is the number of samples in each frame, then the result of windowing is the signal

$$y_i = x_i(n)w(n), \quad 0 \le n \le N-1 \tag{8}$$

Typically the Hamming window is used, which has the form:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right) \tag{9}$$

where $0 \le n \le N-1$

3) Fast Fourier Transform (FFT)

The Fast Fourier Transform is used to convert each windowed frame of *N* samples from the time domain into the frequency domain. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT) which is defined on the set of *N* samples $\{x_n\}$ by the equation below:

$$X_n = \sum_{k=0}^{N-1} x_k e^{-2\pi jkn/N} \tag{10}$$

where $n = 0,1,\ldots, N-1$.
The result of the FFT stage is the spectrum of the windowed speech data.

4) Frequency Warping

The human perception of the frequency contents of sounds for speech signals does not follow a linear scale. For each tone of actual frequency, *f* in Hz, a subjective pitch is measured on a different scale called the 'mel' scale. The *mel-frequency* scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. For a given frequency, *f* in Hz, the corresponding mel-frequency can obtained from the following approximate formula:

$$mel(f) = 2595 * \log_{10}(1 + f/700) \tag{11}$$

Often, a bank of bandpass filters is used to simulate the mel spectrum. A filter bank that has a triangular bandpass frequency response, and its spacing as well as bandwidth determined by a constant mel frequency interval is suitable. This filter bank is applied in the frequency domain, therefore it simply amounts to taking triangle-shape windows on the spectrum.

5) Log

This step which is based on homomorphic speech processing performs natural logarithm. The purpose of this stage is to separate the convolved source excitation, p(t) and vocal tract response, h(t).

6) Inverse Discrete Fourier Transform

In this final step, the log mel spectrum is converted back to time. The result is called the mel frequency cepstrum coefficients (MFCC). The cepstral representation of the speech filterbank spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients (and so their logarithm) are real numbers, they can be converted to the time domain using the Discrete Cosine Transform (DCT).

D. Comparison of the Feature Extraction Techniques

The MFCC and LPC techniques discussed above both produce cepstrum coefficients as their outputs. However, they have some differences. While MFCC is based on the filtering of the speech spectrum using known properties of the perception of speech by the human ear, LPC is based on the autocorrelation of the speech frame. There is no agreement in literature over which of the technique is better. LPC is computationally less expensive that MFCC but MFCC produce more precise results [1]. This opinion is hinged on the fact that the all-pole model used for LPC is a good model for the voiced regions of speech but unsuitable for unvoiced and transient regions.

IV. Speaker Modelling, Pattern Matching and Decision Making

As mentioned earlier, speaker recognition involves two phases: the training phase and the testing phase. During the training phase speaker models of enrolled speakers are created and stored. In order to recognize a

user, his/her input voice sample is compared to the model of the claimed speaker for a verification system or it is compared to all the stored models if it is an identification task.

The pattern-matching task involves computing a match score, which is a measure of the similarity of the input feature vectors to the stored model or models. Based on a decision algorithm, a decision is made concerning the test input speech. The decision could be to reject or accept it as valid. The decision could also be to request for another test utterance or to make an identification decision if there was no prior claimed identity.

Speaker models are created from feature vectors extracted from the speech samples and they can be either template models or stochastic models [2]. For template models, the match score is obtained by evaluating the distance between the observed speech sample and the model. In stochastic models, the matching score is obtained by a measure of the likelihood of the observed speech sample and the model being from the same speaker. It is probabilistic and more flexible than the template model.

Vector Quantization, Dynamic Time Warping and Nearest Neighbours are examples of template models while Gaussian Mixture Models and Hidden Markov Models belong to the stochastic model classification. Vector Quantization and Gaussian Mixture Model are the most studied techniques and they are discussed as examples of each model classification.

A. Vector Quantization Technique

Vector Quantization (VQ) is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region is called a *cluster* and can be represented by its center called a *codeword*. The collection of all codewords forms a *codebook*.

Fig. 5 shows a conceptual diagram of this mapping process. In the figure, only two speakers and two dimensions of the acoustic space are shown. The circles represent the acoustic vectors for speaker 1 while the triangles are for speaker 2. In the training phase, a speaker-specific VQ codebook is generated for each speaker by clustering his/her training acoustic vectors. The resulting codewords (or centres) are shown in Fig. 4 by circles and triangles for speaker 1 and 2 respectively. The distance from a vector to the closest codeword of a codebook is called a VQ distortion. In the testing phase, input utterance from an unknown speaker is "vector quantized" using each training codebook and the *total VQ distortion* is computed. The speaker corresponding to the VQ codebook with smallest total distortion is identified.

.

Clustering the Training Vectors

The acoustic vectors (MFCCs) extracted from the speakers' speech samples are used to create codebooks for them using the VQ technique. The Linde Buzo Gray (LBG) algorithm [18] is a well known algorithm for performing the VQ process.

The LBG algorithm implements the VQ process through the following iterative procedure:

1. Design a codebook with just a single codeword. This codeword is the centre of all the vectors.
2. Double the size of the codebook by splitting the present codebook into two using the following rule:

$$y_n^+ = y_n(1+\varepsilon) \quad (12)$$

$$y_n^- = y_n(1-\varepsilon) \quad (13)$$

   where y_n is the previous codebook, y_n^+ and y_n^- are the new codebooks, n varies from 1 to the current size of the codebook and ε is a splitting parameter often chosen as 0.01.
3. Next, clusterize the vectors around the codewords using a similarity measure. Each vector is associated with the closest codeword.
4. Obtain new codewords for each cluster by determining the centre of the clusters.
5. Repeat steps 3 and 4 until the average distance falls below a given threshold.
6. Repeat 2, 3 and 4 until a codebook of the desired size is obtained.

B. Gaussian Mixture Model (GMM)

A mixture model is a model in which the components of a distribution are expressed as fractions of the total. The same concept is used for the Gaussian Mixture Model (GMM), here the contribution of each component in terms of weight is used in speaker modeling

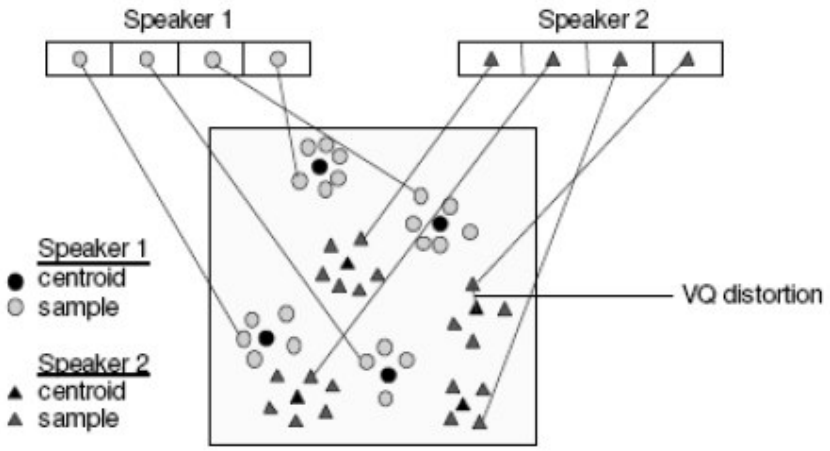


Fig. 5 Conceptual diagram illustrating vector quantization codebook formation [19]

GMM models a probability distribution function as the weighted sum of a number of components and has the form given by the following equation [20].

$$p(\bar{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\bar{x}) \qquad (14)$$

Where M is the number of component densities, $\bar{x}$ is a D-dimensional observed data (random vector), $b_i(\bar{x})$ are the component densities and p_i are the mixture weights for $i = 1,...,M$. Each component density has the form given by the equation:

$$b_i(x) = \frac{1}{(2\pi)^{D/2}\left|\sum_i\right|^{1/2}} \exp\left\{-\frac{1}{2}(\bar{x}-\bar{\mu}_i)'\sum_i^{-1}(\bar{x}-\bar{\mu}_i)\right\} \qquad (15)$$

with mean vector μ_i and covariance matrix $\sum_i$. The mixture weights satisfy the condition:

$$\sum_{i=1}^{M} p_i = 1$$

The complete Gaussian Mixture Model is characterized by the following parameters: the mean vectors, the covariance matrices and the mixture weights from all component densities. These parameters are collectively represented as

$$\lambda = \{\, p_i, \bar{\mu}_i, \Sigma_i \,\} \; ; \mathrm{I} = 1, \ldots, \mathrm{M} \qquad (16)$$

For the identification task, each speaker is represented by a GMM and is referred to by his/her model **λ**. The Expectation Maximization (EM) Algorithm is a commonly used algorithm for determining the parameters of a GMM model. The number of components can be determined by a clustering algorithm. For GMM based on the EM algorithm, an initial model **λ** is first obtained and this model is used in estimating a new model $\bar{\lambda}$ such that $p(X \mid \bar{\lambda}) \geq p(X \mid \lambda)$. The new model becomes the initial model for the next iteration step and the process is repeated until some sort of convergence threshold is reached.

C. Comparison of the Pattern Matching Techniques

VQ and GMM are the best studied techniques for speaker identification. What VQ does basically is to reduce the amount of feature vectors by a clustering process. It is more of a quantifier than a modeler. GMM treats the speech production process as a stochastic process and it produces more accurate speaker model for robust identification [20]. However, VQ outperforms GMM when training data is small and fast training time is required [21, 22]. Stochastic modelling offers more flexible and theoretically meaningful probabilistic scores than the VQ approach that has been in use earlier [2].

V. Applications

Speaker recognition system has many potential applications. They are:

Security Control
Access control to confidential information and facilities through authentication techniques is conventionally through the use of passwords, smart cards or keys that can be stolen, forgotten, borrowed or lost [23]. Speaker

recognition systems do not have this limitation and will therefore be a relevant tool for such an application. Speaker verification systems have been deployed commercially for access control to information, services and computer accounts [5].

Telephone Banking
Speaker verification systems now exist that allow verification of the identity of a bank customer over the telephone before he can perform some banking operations.

Information structuring
Speaker recognition technology can also be used for the purpose of organizing the information content of audio documents. This capability of the speaker recognition system is relevant in the following areas: automatic annotation of audio archives, speaker indexing of sound tracks, and speaker change detection for automatic subtitling. [24]

VI. Conclusion and Future Trend

Speaker recognition is the task of automatically recognizing who is speaking by identifying an unknown speaker among several reference speakers using speaker specific information included in speech waves.

The low level features are the most commonly used feature for speaker recognition. However, because of the effect of noise and channel impairment on these features, exploitation of high level feature will still be a major research focus for some time to come. The encouraging results obtained from projects employing high level features make this a good direction to take. Furthermore, there are research efforts to use the new market paradigm to implement software radio receivers, and classify both source excitations and vocal tract characteristics from speech data based on the successful use of the paradigm to classify paths taken by signals and recover messages from modulated waveforms [25, 26].

VII. References

[1] Gish, H., and Schmit, M., "Text-Independent Speaker Identification",*IEEE Signal Processing Magazine,*1994, pp 18-32.

[2] Campbell, J. P., "Speaker Recognition: A Tutorial", *Proceedings of the IEEE*, vol. 85, no. 9, 1997, pp. 1437-1462.

[3] Fisusi, A., "Development of a Text-Independent Speaker Identification System", An MSc Thesis submitted to the Department of Electronic and Electrical Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria, 2007, 101p.

[4] Chen, K., Dahong, X., and Huisheng, C. (1996): "Speaker Identification Using Time-Delay HMEs", *China International Journal of Neural Systems*, vol. 7, no. 1, 1996, pp. 29-43.

[5] Reynolds, D. A., "An Overview of Automatic Speaker Recognition Technology", *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, 2002, pp 4072-4075,.

[6] Campbell, J. P., Reynolds, D. A., and Dunn, R. B., "Fusing High- and Low-Level Features for Speaker Recognition", *In Proc. Eurospeech in Geneva, Switzerland, ISCA*, 2003, pp. 2665-2668..

[7] Peskin, B., Navratil, J., Abramson, J., Jones, D., Klusacek, D., Reynolds, D. A.and Xiang, B. "Using Prosodic and Conversational Features for High Performance Speaker Recognition: Report from JHU WS'02", *In Proc. International Conference on Acoustics,Speech, and Signal Processing in Hong Kong, China, IEEE*, 2003, pp. IV: 792- 795,.

[8] Deller, J. R., Hansen, J.H.L. and Proakis J. G. "Discrete-Time Processing of Speech Signals", *IEEE Press, New York,*2002*,* pp 99-342.

[9] Nickel, R., "Automatic Speech Character Identification", *IEEE Circuits and Systems Magazine*, vol. 6, no. 4, 2006, pp 8-29.

[10] Huang, X., Acero, A., and Hon, H.-W. "Spoken Language Processing: a Guide to Theory, Algorithm, and System Development", Prentice-Hall, New Jersey, 2001.

[11] Flanagan, J. L., "Speech Analysis, Synthesis and Perception*,"* Springer, New York, 1972.

[12] Kinnunen, T., Kilpeläinen T. and Fränti P., "Comparison of Clustering Algorithms in Speaker Identification", *Proc. IASTED Int. Conf. Signal Processing and Communications (SPC 2000)*, 2000, pp. 222-227.

[13] Siafarikas, M., Ganchev, and T., Fakotakis, N., "Wavelet Packets based Speaker Verification", *Proceedings of the ISCA Speaker and Language Recognition Workshop – Odyssey'2004,* pp. 257–264.

[14] Pan, Y. and Waibel, A., "The Effects of the Room Acoustics on MFCC Speech Parameter," *International Conference on*

Spoken Language Processing 2000 (ICSLP 2000), vol. 4, pp. 129–133.

[15] Zilca, R.D., Navratil, J. and Ramaswamy, N. "Syncpitch: A pseudo pitch synchronous algorithm for speaker recognition," *Proceedings of EUROSPEECH 2003*, pp.2649-2652.

[16] Makhoul, J., "Linear prediction: A tutorial review," *Proceedings of the IEEE*, vol. 63, no. 4, 1975, pp. 561- 580.

[17] O'Shaughnessy, D., "Linear Predictive Coding", *IEEE Potentials,* Vol. 7, no. 1, 1988, pp. 29-32.

[18] Linde, Y., Buzo, A., and Gray, R., "An algorithm for vector quantizer design", *IEEE Transaction on Communication*, vol. 28, 1980, pp. 84-95.

[19] Soong, F.K., Rosenberg, A.E. and Juang, B.H. "A vector quantisation approach to speaker recognition", *AT&T Technical Journal*, Vol. 66-2, 1987, pp. 14-26.

[20] Reynolds, D. and Rose, R., "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE transactions on speech and audio processing*, vol. 3, No1, 1995, pp. 72-83

[21] Do, M. and Wagner, M., "Speaker Recognition with Small Training Requirements Using a Combination of VQ and DHMM", *Proc. of Speaker Recognition and Its Commercial and Forensic Applications*, 1998, pp. 169-172.

[22] Filho, T., Messina, R. and Cabral, E., "Learning Vector Quantization in Text-Independent Automatic Speaker Identification", *5th Brazilian Symposium on Neural Networks*, 1998 , pp. 135- 139.

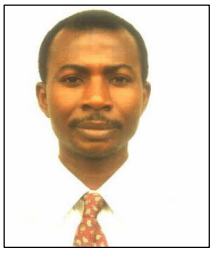
[23] Beumer, G. M., Veldhuis, R.N.J., and Bazen, A.M., "Transparent Face Recognition in Home Environment", *15th Annual Workshop on circuits, Systems and Signal Processing (ProRISC)*, 2004, pp. 225-229.

[24] Bimbot, F., Bonastre, J., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-Garcia, J., Petrovska-Delacretaz, D., and Reynolds, D. A. "A Tutorial on Text-Independent Speaker Verification", EURASIP Journal on Applied Signal Processing, 2004, pp 430-451.

[25] Yesufu, T.K. and Yesufu, O.A., "Dynamic Determinants for Decision-Making in Smart Engineering Systems", In: Dagli, C.H., Buczak, A.L., Ghosh, J., Embrechts, M.J., Ersoy, O. (eds): *Intelligent Engineering Systems Through Artificial Neural Networks*, Vol. 13. ASME Press Series, New York, 2003, pp. 763-768.

[26] Yesufu, T.K. and Yesufu, O.A.: "High Definition Information Scheme for Data Services Network Based on the Market Paradigm", In: Kehinde, L.O., Adagunodo, E.R. and Aderounmu, G.A. (eds): *Proc. of the Int. Conf. on Applications of ICT to Teaching, Research and Administration (AICTTRA 2005)*, Obafemi Awolowo University, Ile-Ife, Nigeria,,Vol. 1, 2005, pp. 95-102.

Abimbola Adeola Fisusi received a B.Sc. in Electronic and Electrical Engineering from the Obafemi Awolowo University, Ile-Ife, Nigeria in 2001. He is on his M.Sc. in Electronic and Electrical Engineering at Obafemi Awolowo University, Ile-Ife, Nigeria. His research interests include signal processing with applications in speaker recognition, information and communication systems engineering and security systems. He is a lecturer at the Department of Electronic and Electrical Engineering, Obafemi Awolowo University, Ile-Ife, Nigeria.



Thomas Kokumo Yesufu received a B.Eng. in Electronics and Communication Engineering from the University of Jos, Nigeria in 1987; an M.Sc. and Ph.D. in Electronic and Electrical Engineering from the Obafemi Awolowo University, Ile-Ife, Nigeria in 1994 and 1999 respectively. He is a lecturer and the National Coordinator of the Cooperative Information Network (COPINE), Obafemi Awolowo University, Ile-Ife, Nigeria. His research Interests include signal processing with applications in radio propagation studies, cryptography and information and communication systems engineering. He has published more than thirty-five (35) book chapters, and conference and journal articles. He is a member of the Institute of Electrical and Electronics Engineers (IEEE) and the Nigerian Society of Engineers (NSE).