



ILJS-15-038

## Speech Processing Algorithm for Automatic Speaker Recognition-Based Telephone Voice Dialing in Yorùbá

Akintola<sup>1\*</sup>, A. G. and Ibiyemi<sup>2</sup>, T. S.

<sup>1</sup>Department of Computer Science, University of Ilorin

<sup>2</sup>Department of Electrical Engineering, University of Ilorin

### Abstract

The commonly used method for telephone dialing is through the use of the phone key pad. This form of input is cumbersome and not user friendly to some people (drivers, handicaps etc.). These problems have led to a lot of researches in development of natural and user friendly input such as speech. Most of the works in this area are in English, French, Chinese and other languages. This makes very important a research in using standard Yorùbá as an input in telephone voice dialing. The telephone voice dialing in Yorùbá is achieved through speech recognition training phase and recognition phase of the algorithm. The raw speech captured were digitalized and processed to achieve the speech signal which was used in the pattern matching of the system to allow authorized users and disallow the unauthorized users from using the system. The telephone voice dialing algorithm was developed through pre-emphasis filter, framing and windowing and dynamic time warp. With the design, telephone voice dialing in other languages could also be developed following the same procedures.

**Keywords:** Speech preprocessing, Voice dialing, Pre-Emphasis filter, Framing, Windowing, Dynamic time warping.

### 1. Introduction

Human voice is the faculty of uttering articulate sounds or words; the faculty of expressing thoughts by words or articulate sounds; the power of speaking (Beigi, 2011). The human voice consists of sound made by a human being using the vocal folds for talking, singing laughing, crying, screaming etc. Its frequency ranges from 200 to 7000 Hz. Generally, the mechanism for generating the human voice can be subdivided into three parts; the lungs, the vocal folds within the larynx, and the articulators (Stevens, 2000). Adult men and women have different vocal folds sizes; reflecting the male-female differences in larynx size. Adult male voices are usually lower-pitched and have larger folds.

---

Corresponding Author: Akintola, A.G.  
Email: [abimbolar@yahoo.com](mailto:abimbolar@yahoo.com)

The male vocal folds (which would be measured vertically in the opposite diagram), are between 17 mm and 25 mm in length. The female vocal folds are between 12.5 mm and 17.5 mm in length (Thurman & Welch, 2000). A voice command device is device controlled by means of the human voice. By removing the need to use buttons, dials and switches, people can easily operate appliances with their hands full or while doing other tasks (Mayur and Mukul, 2014).

Speech is the vocalized form of human communication. It is based upon the syntactic combination of lexical and names that are drawn from very large vocabularies (Bhusan and Krishna, 2013). Speech perception refers to the processes by which humans are able to interpret and understand the sounds used in language. The study of speech perception is closely linked to the fields of phonetics and phonology in linguistics and cognitive psychology and perception in psychology (AskDefine, 2011). Speech research has applications in building computer systems that can recognize speech, as well as improving speech recognition for hearing- and language-impaired listeners.

Speaker recognition is the computing task of validating a user's claimed identity using characteristics extracted from their voices. (Fabian and Aviel, 2000). There are two major applications of speaker recognition technologies and methodologies. If the speaker claims to be of a certain identity and the voice is used to verify this claim, this is called verification or authentication. On the other hand, identification is the task of determining an unknown speaker's identity. In a sense speaker verification is a 1:1 match where one speaker's voice is matched to one template (also called a "voice print" or "voice model") whereas speaker identification is a 1:N match where the voice is compared against N templates (Thurman and Welch, 2000).

This work main contribution to the development of speech recognition is in the area of voice dialing in Yorùbá. The focus of this work is Yorùbá numerals for telephone voice dialing.

## **2. Materials and Methods**

Speech is the faculty of uttering articulate sounds or words; the faculty of expressing thoughts by words or articulate sounds; also can be the power of speaking (Hyperdictionary, 2015).

Speech recognition (also known as automatic speech recognition or computer speech recognition) converts spoken words to text. The term "voice recognition" is sometimes used to refer to recognition systems that must be trained to a particular speaker. For a machine to interpret speech, a microphone must translate the vibrations of a person's voice into a wavelike electrical signal (Narzul & Karmal, 2014). This signal in turn is converted by the system's hardware into a digital signal. The speech recognition program analyzes the digital signal in order to recognize separate phonemes, the basic building blocks of speech. The phonemes are then recombined into words.

Speech recognition systems used by the general public e.g. phone-based automated timetable information, or ticketing purchasing, can be used immediately – the user makes contact with the system, and speaks in response to commands and questions. However, systems on computers meant for more individual use, such as for personal word processing, usually require a degree of "training" before use. Here, an individual user "trains" the system to understand words or word fragments; this training is often referred to as "enrolment".

### 2.1. Description of the Yorùbá Language

Yorùbá is one of the four official languages of Nigeria and is a member of the Niger-Congo family of languages. It is spoken by about 30 million people in southwest Nigeria, Benin, Togo, the UK, Brazil and the USA (Nigeria at CIA World Factbook, 2012). Yorùbá is a tonal language like many African languages, in learning Yorùbá words one have to take note of not only the sound but also of the tone which is an integral part of each word. It is important to know the tonal sign designated for each word in other to understand the speech and to get the intonation correctly.

#### Alphabets

The Yorùbá alphabet consists of 25 letters as compared to the 26 letter alphabets of the English language.

A	B	D	E	È	F	G	Gb	H	I	J	K	L	M	N	O	Ọ	P	R	S	Ş	T	U	W	Y
A	b	d	e	ẹ	f	g	gb	h	i	j	k	l	m	n	o	ọ	p	r	s	ş	t	u	w	y

## Numerals

The Yorùbá basic numbers (0-9) also none as counting numbers are presented below:

**Table 1:** Yorùbá Numerals

Numbers (English)	Yorùbá Pronunciation	Arabic
Zero	Òdo	0
One	Ení	1
Two	Èjì	2
Three	Ètà	3
Four	Èrìn	4
Five	Àrun	5
Six	Èfà	6
Seven	Èje	7
Eight	Èjo	8
Nine	Èsán	9

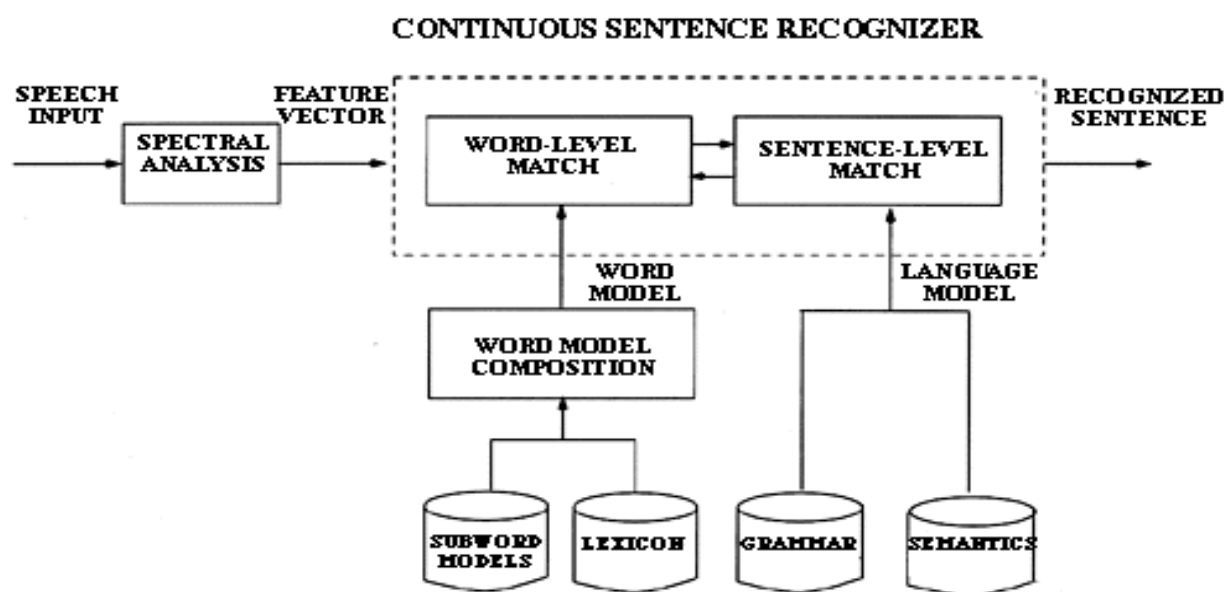
## 2.2 Approaches to Speech Recognition

The earliest approaches to speech recognition were based on finding speech sounds and providing appropriate labels to these sounds. This is the basis of the acoustic-phonetic approach of Vijay & Douglas (1998), which postulates that there exist finite, distinctive phonetic units (phonemes) in spoken language and that these units are broadly characterized by a set of acoustic properties that are manifest in the speech signal over time. Even though the acoustic properties of phonetic units are highly variable, both with speakers and with neighboring sounds (the so-called co articulation effect), it is assumed in the acoustic-phonetic approach that the rules governing the variability are straightforward and can be readily learned (by a machine).

The pattern-matching approach involves two essential steps-namely, pattern training and pattern comparison. The essential feature of this approach is that it uses a well formulated mathematical framework and establishes consistent speech-pattern representations, for reliable pattern comparison, from a set of labeled training samples via a formal training algorithm. A speech-pattern representation can be in the form of a speech template or a statistical model (e.g. a HIDDEN MARKOV MODEL or HMM) and can be applied to a sound (smaller than a word), a word, or a phrase. In the pattern-comparison stage of the

approach, a direct comparison is made between the unknown speeches (the speech to be recognized) with each possible pattern learned in the training stage in order to determine the identity of the unknown according to the goodness of match of the patterns. The pattern-matching approach has become the predominant method of speech recognition in the last decade (Vijay and Douglas, 1998).

The artificial intelligence approach of Vijay and Douglas (1998) attempts to mechanize the recognition procedure according to the way a person applies intelligence in visualizing, analyzing, and characterizing speech based on a set of measured acoustic features. Among the techniques used within this class of methods are uses of an expert system (e.g. a neural network) that integrates phonemic, lexical, syntactic, semantic, and even pragmatic knowledge for segmentation and labeling, and uses tools such as artificial NEURAL NETWORKS for learning the relationships among phonetic events. The focus in this approach has been mostly in the representation of knowledge and integration of knowledge sources. This method has not been used widely in commercial systems. A block diagram of a complete system for large vocabulary speech recognition (Robert and Frank, 2001) based on the pattern matching approach is shown in Figure 1. The first step in the processing is spectral analysis to derive the feature vector used to characterize the spectral properties of the speech input. The second step in the recognizer is a combined word-level/sentence-level matching procedure.



**Figure 1:** Overall block diagram of subwork unit-based continuous-speech recognizer. Source: Besacier *et al.*, (2014).

### 2.3. Voice Dialing

Telephone voice dialing lets you speak a name or number to dial instead of entering a number manually or choosing it from the phone book. There are two types: speaker-dependent (recorded) and speaker-independent (recognition).

#### Speaker-Dependent

With speaker-dependent, voice dial entries (or "tags") must be explicitly created by speaking and recording the name 1-3 times. The phone will only respond to recorded names, and usually only when spoken by the same person who recorded them. The maximum number of voice dial entries is usually limited to a fraction of the size of the whole phone book.

#### Speaker-Independent

With speaker-independent voice recognition, no recording is required. The name can be spoken by anyone, and the phone will automatically match the spoken name with the closest name entered in the phone book. Speaker-dependent is the most common type, especially on older phones. If the type isn't specified, it is usually speaker-dependent.

### 2.4 System Design

Speech recognition is a multileveled pattern recognition task, in which acoustical signals are examined and structured into a hierarchy of subword units (e.g., phonemes), words, phrases, and sentences. Each level may provide additional temporal constraints, e.g., known word pronunciations or legal word sequences, which can compensate for errors or uncertainties at lower levels. This hierarchy of constraints can best be exploited by combining decisions probabilistically at all lower levels, and making discrete decisions only at the highest level.

#### Design Specification

The specifications for the system are:

- Set speech recording parameters to (8 kHz, 16 bit, Mono) in MS Windows environment.
- Collection of digital speech data from 5 (3 males, 2 females) Yorùbá persons, each pronouncing numerals {odo, eni, eji, ẹta, ẹrin, arùn, ẹfa, eje, ẹjo, ẹsan} 20 times.
- This provided 1,000 Yorùbá numeral digit utterance data.
- Collection of digital speech data from the same 5 Yorùbá persons each pronouncing 08034265239, and 08052409098 phone numbers in Yorùbá i.e. [odo, ẹjo, odo, ẹta,

erin, eji, èfa, arùn, eji, ẹta, ẹsan] and [odo, ẹjo, odo, arùn, eji, ẹrin, odo, ẹsan, odo ẹsan, ẹjo] respectively 10 times.

-This resulted in 100 phone number sentences pronounced in Yorùbá.

-These data were burnt on CD.

-Collection of digital speech data from 20 (12 males, 8 females) Yorùbá persons, each pronouncing numerals {odo, eni, eji, ẹta, ẹrin, arùn, èfa, eje, ẹjo, ẹsan} 20 times.

-This yielded 4,000 data samples of Yorùbá numeral digit.

-Collection of digital speech data from the same 20 Yorùbá persons each pronouncing 5 telephone numbers in Yorùbá from each of the 3 selected gsm operators (MTN, GLO, ZAIN).

-This experiment gave us 3,000 phone number utterances.

-Data from this second round of experiments were burnt on DVD.

The systems vocabulary was built using utterances of Yorùbá digits (0-9) and the telephone number utterances recorded as specified above.

**Speech Capture:** This is the state where the system receives input i.e. data (speech).

**Speech Digitization:** This converts the electrical analogue speech signal into its digital equivalent through sampling and quantization.

The analogue signal is sampled at a frequency of 8khz with the sound card having an onboard anti-Aliasing filter. This sampling rate yields 8000 samples per seconds. The recording of 20 Yorùbá speakers was saved by each speakers' name in a separate folder for each person for the training phase. There were 20 Yorùbá speakers for the input speech, each person pronounced one phone number ten times for three different networks namely MTN, Airtel and Globacom networks to test if it recognize the digits or not.

### **Pre-Emphasis Filter**

Pre-emphasis Filter is used to reduce differences in power of different components of the signal. The speech is pre-emphasized with a pre-emphasis filter algorithm below.

$y(n) = x(n) - a \cdot x(n-1)$ . Where:  $x(n)$  - value of input signal at discrete time step  $n$

$y$  - value of output signal at discrete time step  $n$  and  $a$ -constant  $a = 0.95$

### Frame Blocking and Windowing

The speech sample is arranged into overlapping blocks of frames. This is to analyze the speech samples in short intervals of 15-30ms. The frames are overlapped to ensure that all aspect of sample is represented and emphasized. The frame samples are then windowed to cut each small section out for analysis and to handle the problem of signal discontinuity with the aid of an algorithm. The hamming window is used for this work. The equation for hamming window is given below:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), N = \text{Window Length}, n = \text{sample width}$$

### Feature Extraction

Feature extraction is a process that analyzes the windowed speech samples and represents or code it in a more compact vector form. These vectors represent a compact model of each uttered words.

### Feature Matching

The system makes use of the Dynamic time warp algorithm for solving the pattern matching problem during the recognition phase. The standard Dynamic time warp algorithm is:

Initial condition:

$$D(1,1)=0;$$

Recurrence:

$$D(i, j) = \text{Dist}(i, j) + \min[ D(i-1, j), D(i, j-1), D(i-1, j-1) ]$$

**Phases of Speech Recognition:** Speaker recognition system operates in two phases, namely, Training phase and Recognition phase.

### Training Phase

In this phase, the system learns the reference patterns representing the different speech sounds e.g phrase, words, phonemes, that constitute the vocabulary of the application. These sounds were obtained by recording the digits utterances made by the Yorùbá speakers.

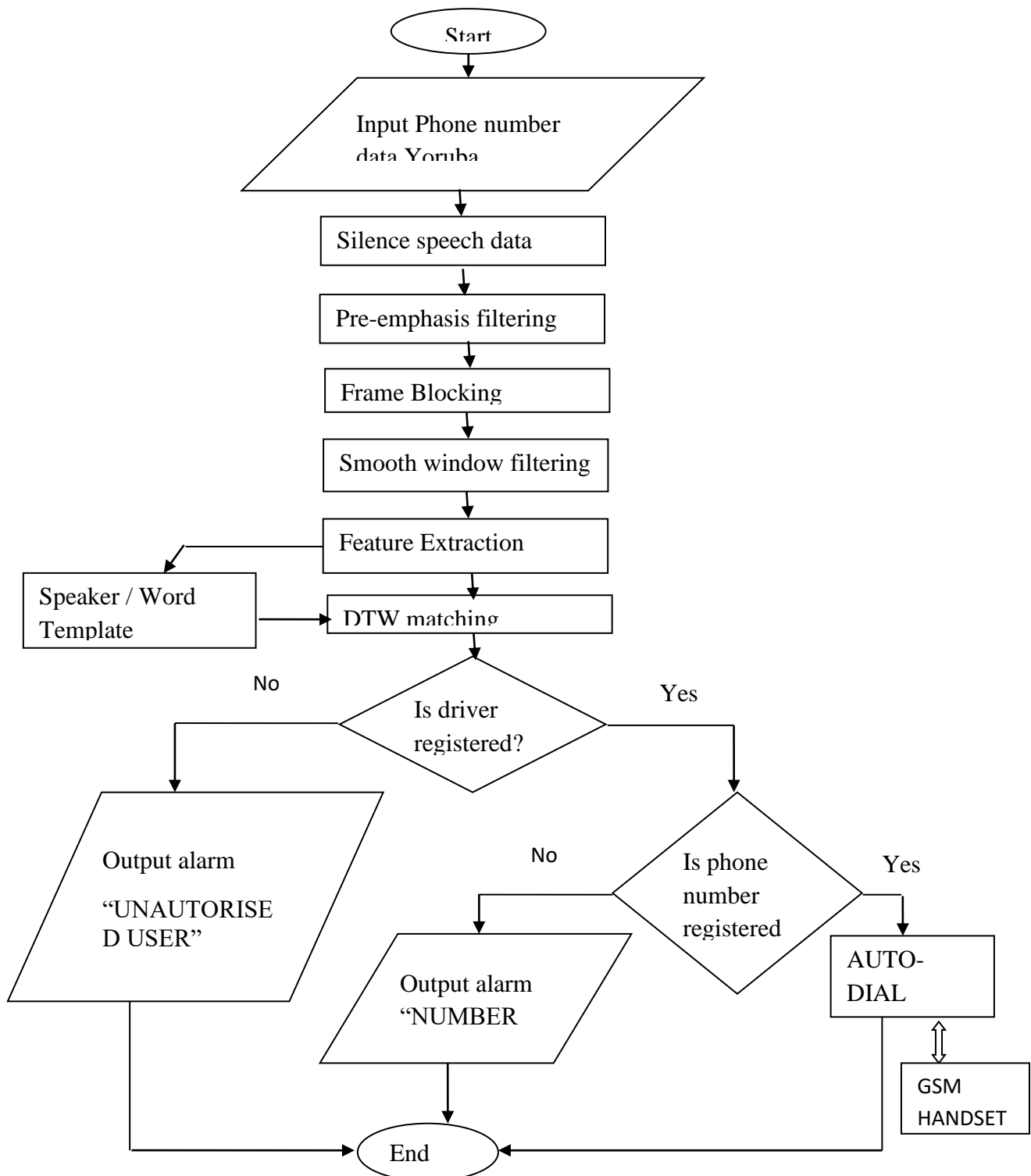
### Recognition Phase

This is the phase during which an unknown input pattern is identified by making reference to the stored samples of the training phase. The phase is carried out with DTW algorithm.



## 2.5. Algorithm Development

The algorithm performs Telephone Voice Dialing in Yorùbá. The core of the system is to recognize the spoken word. This core is divided into three sections: Speech Capture and Digitization, Signal Pre-Processing and Feature Extraction.



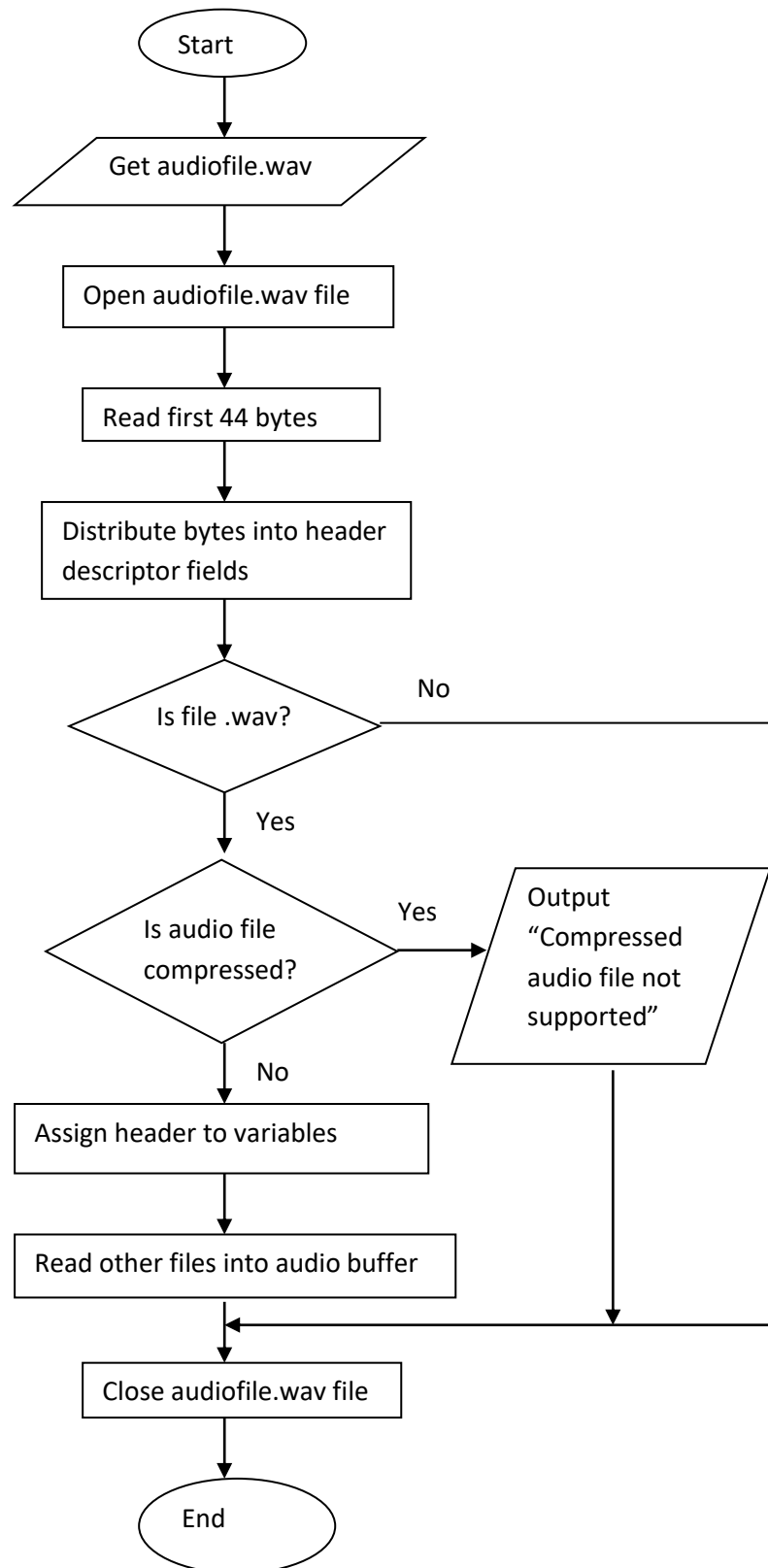
**Figure 2:** Speaker Recognition (voice dialling)

The speech recognizer is made up of modules and sub modules. The modules that make up the speech recognizer include:

1. Data Extraction Module.
  - I. Wave file header descriptor submodule.
  - II. Wave file reader sub module.
  - III. Wave file writer sub module.
2. Signal Pre-Processing Module.
  - I. Pre-emphasis filter submodule.
  - II. Word segmentation submodule.
  - III. Frame blocking submodule.
  - IV. Hamming windowing submodule.
3. Feature Extraction Module.
4. Dynamic time warp Module.

#### **Data Extraction Module**

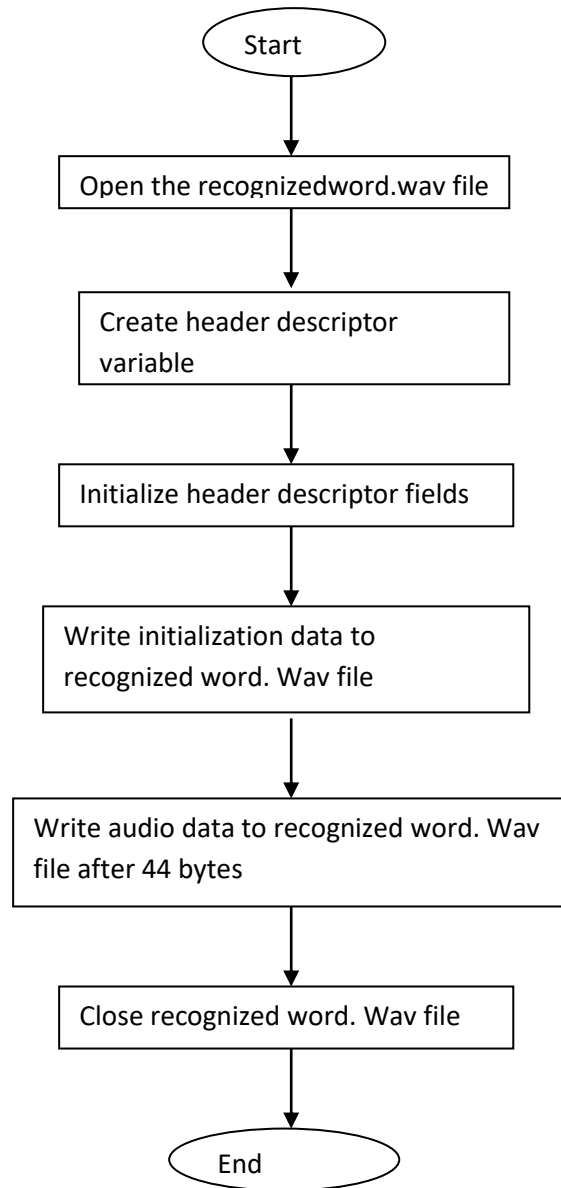
This module involves speech capture and digitization section. It describes the first task performed on the speech samples stored in .wav file format. This module performs the task of extracting the audio header descriptor bytes and the audio samples from the stored files. It also ensures that the file to be read is a .wav file. It has sub modules Wave file header descriptor sub module, Wave file reader sub module, Wave file writer sub module.



**Figure 3:** Wave file reader sub module

**Wave File Writer Sub Module**

This sub module is for writing back the bytes of the header and audio samples recognized words which can be played back. The flow chart is given below in Fig.4



**Figure 4:** Wave file writer sub module

### Signal Pre-processing Module

The module describes the signal processing steps which are; the pre-emphasis filter sub module, frame blocking and windowing sub modules.

#### Pre-Emphasis Filter Sub Module

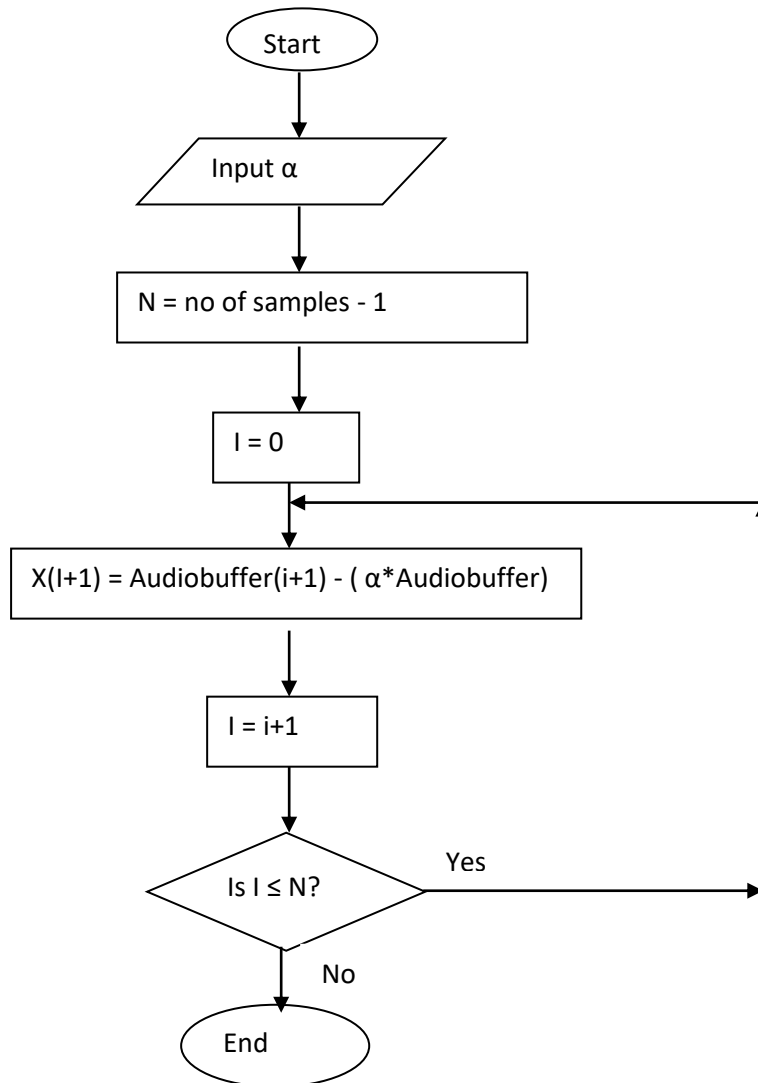


Figure 5: Pre-emphasis filter sub module

### Word Segmentation SubModule

The sub module is for obtaining the beginning and the end point of the digit utterances. It therefore removes the silence speech from the actual speech.

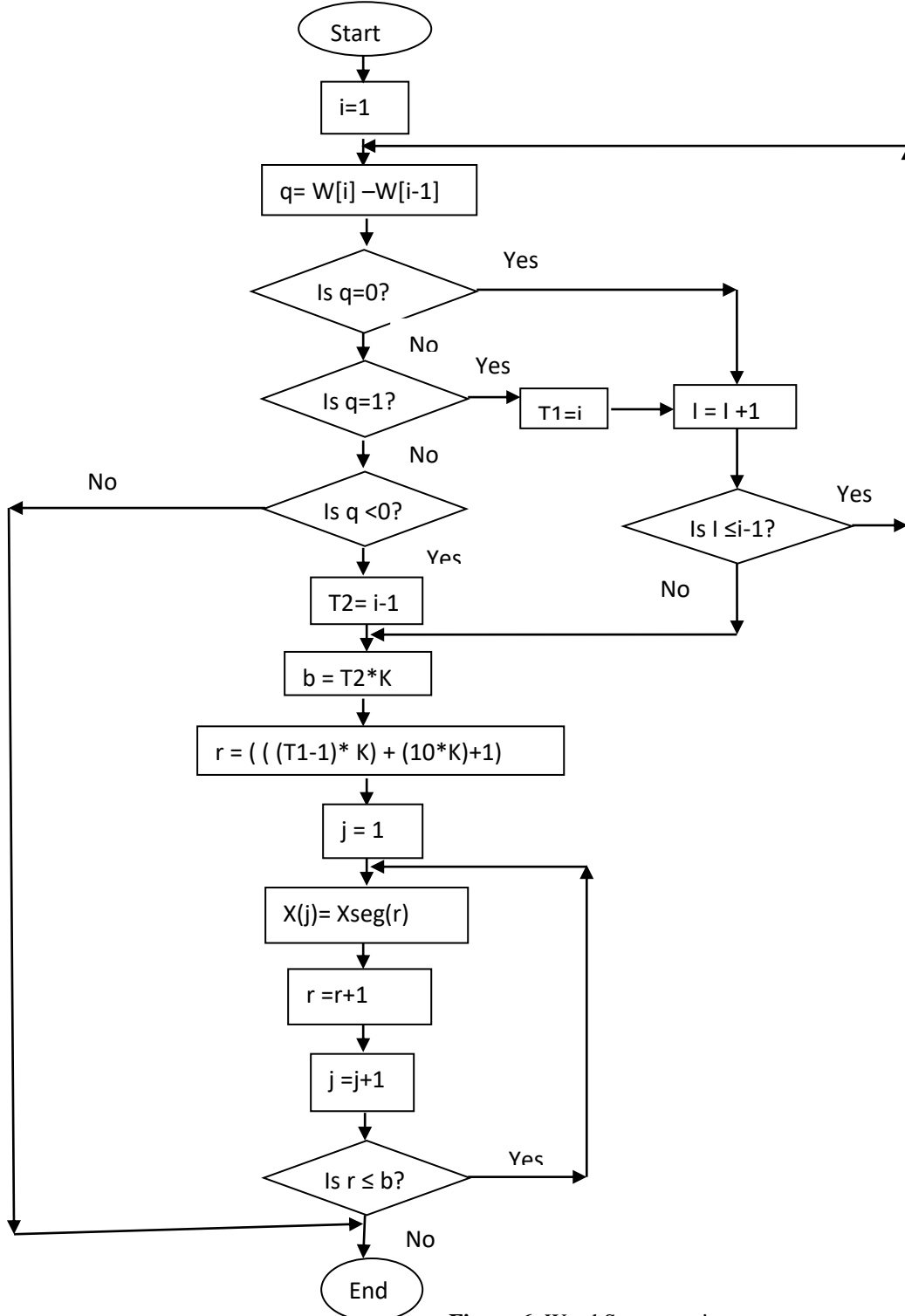
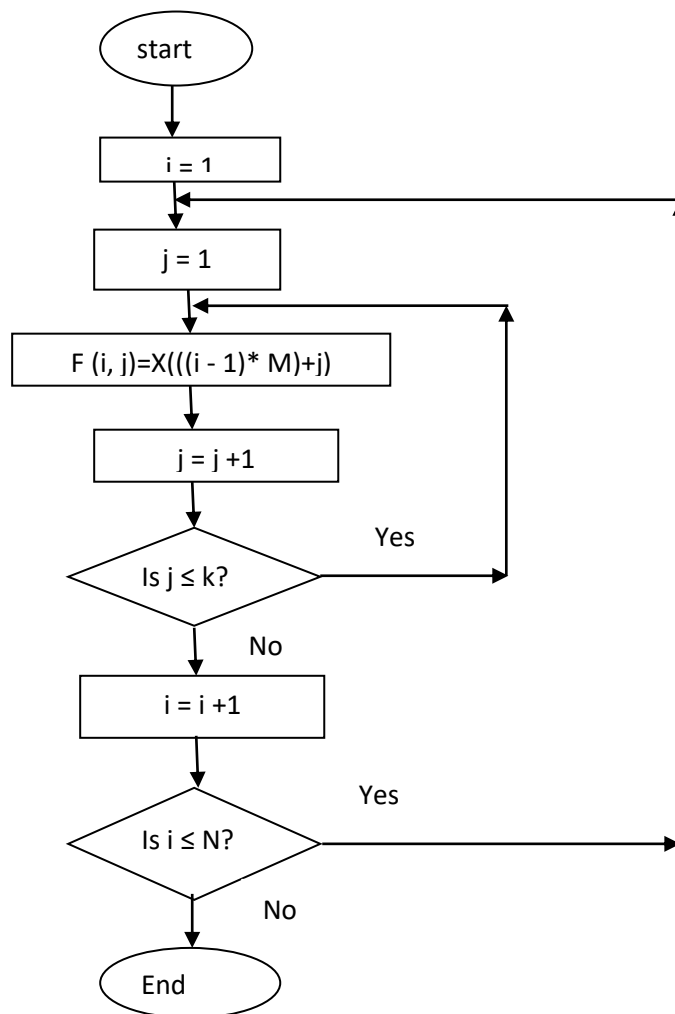


Figure 6: Word Segmentation

**Frame Blocking Sub Module**

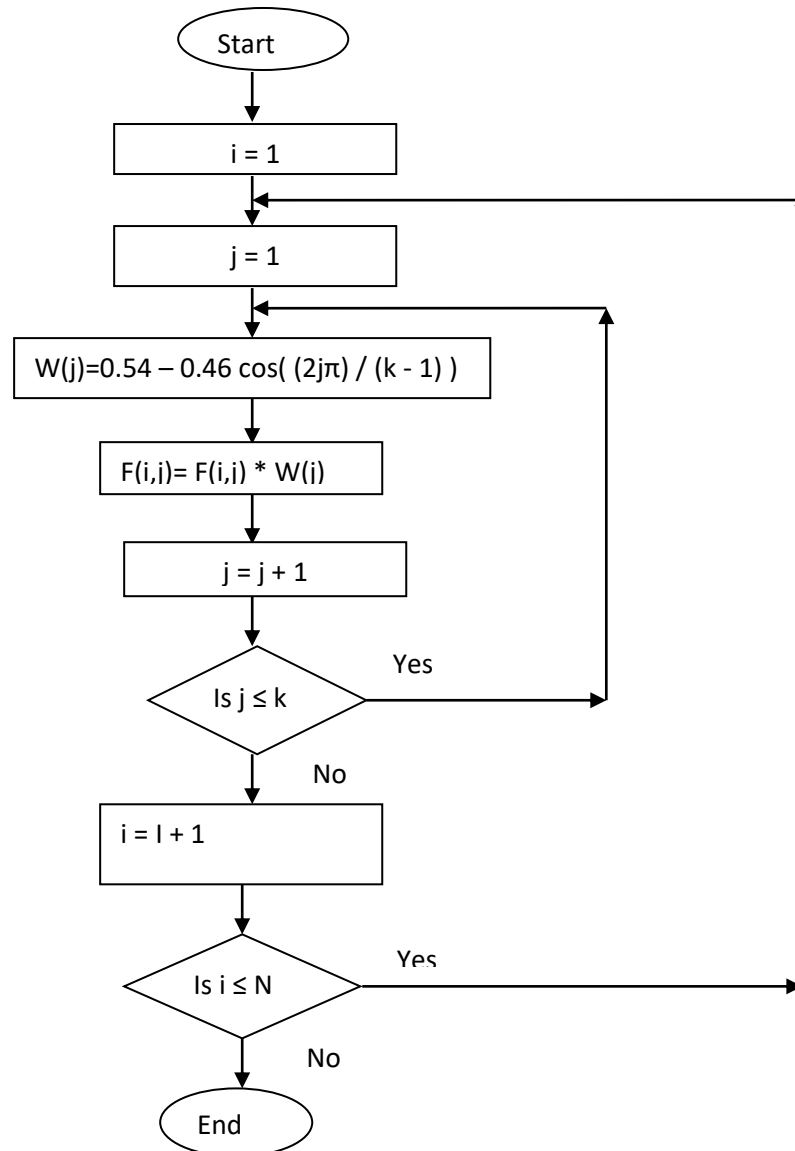
This sub module is for the algorithm to put the samples into blocks of frames for easier computation.



**Figure 7:** Frame blocking algorithm

### Hamming Window Sub Module

The windowing is applied on frame blocks for smoothing.



**Figure 8:** Hamming

### Feature Extraction Module

This describes the feature extraction section. The steps involved in feature extraction include the autocorrelation computation.

### Dynamic Time Warping (DTW) Module







This describes the algorithm used as tool for searching the database of the speech recognizer.

## 3. Results and Discussion




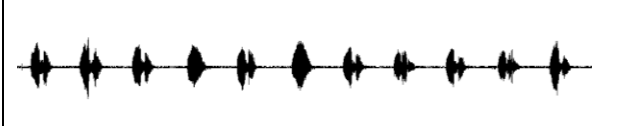
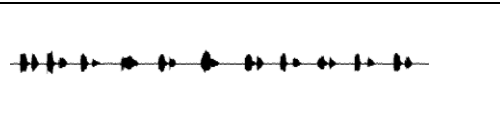

Tables 2 and 3 show waveform of spoken numbers that can be used in telephone voice dialing with variation among speakers.



**Table 2:** Wave form display of mobile number 08034265239 by female and male speakers

Instance	Female	Male
1		
2		
3		

**Table 3:** Wave form display of mobile number 08052409098 by female and male speakers

Instance	Female Utterance	Male Utterances
1		
2		
3		

Yoruba voice recognition is a form of biometrics that can improve communication among indigenous Yoruba people. For example, the aspect of telephone voice dialing. With the developed algorithm, the people who are non English speakers can dictate telephone number and the number will be dialed, it can also be applicable to drivers who do not want to drive and press keypads as well as disabled people.

#### 4. Conclusion

The algorithms for Yorùbá telephone voice dialing algorithm was designed in this paper. The various stages involved in this algorithm development were highlighted step by step. The first step was the capturing of speech signal and digitizing it. The capturing was done with microphone which was connected to a pc. The recorded sound wave further went through a process of filtering, which serves as an equalizer. The filtered sound is later put into a block of frames, this helps to analyze the speech samples in short intervals of 15 to 30 ms, and this is to ensure that none of the samples is deemphasized. Hamming window was used to cut it into section for analyses. The system uses dynamic time warping algorithm to solve the pattern matching of the stored words with the input words. Speech recognition is a developing phenomenon in the whole world that it is of an important benefit if all languages could be designed for such usage.

#### Acknowledgement

The authors acknowledge the comments of the reviewers for their suggestions and comments in improving the quality of the manuscripts.

#### References

- AskDefine (2011): Define Speaking. Retrieve from <http://speaking.askdefinebeta.com/>  
Retrieve date: 20/02/2015
- Bhusan, C. and Krishna, B. (2013): Nepali Text to Speech Synthesis System using ESNOLA Method of Concatenation. *International Journal of Computer Applications* **62**, 25-28.
- Beigi, H. (2011): *Fundamentals of Speaker Recognition*. 1<sup>st</sup> edn., Springer USA.
- Besacier, L., Barnard, E., Karpovc, A. and Schultz, T. (2014): Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*. **56**, 85-100.
- Fabian, M. and Aviel, D. (2000): Keystroke dynamics as a biometric for authentication. *Future generation computer systems*. **16**, 351-359.
- Hyperdictionary: Meaning of Speech. Retrieved from <http://www.hyperdictionary.com/dictionary/speech>. Retrieve date: 20/02/2015
- Mayur, G. and Mukul, S. (2014): Smart Bluetooth And Gsm Based Home Automation System. *Journal of International Academic Research for Multidisciplinary*. **2**, 424-433.

Narzul, I. and Karmal, K. (2014). A process to improve accuracy of voice recognition system by using word correction system. *COMPUSOFT An international journal of advanced computer technology*, **3**(6), 822 – 824.

Nigeria at CIA World Factbook (2012): Yoruba 22% out of a population of 170.1 million (2012 estimate).

Robert, A. and Frank, C. (2001): The MIT Encyclopedia of the Cognitive Sciences.

Stevens, K.N. (2000): Acoustic Phonetics. MIT Press, England.

Thurman, L. and Welch, G. (2000): *Body, mind and voice: Foundations of voice education*. Revised edn., Collegeville, Minnesota.

Vijay, K. and Douglas B. (1998): The Digital Signal Processing Handbook IEEE press US.