Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.6, No.7, 2015

An HCI Speech-Based Architecture for Man-To-Machine and Machine-To-Man Communication in Yorùbá Language

Akintola A. G.[1*], Ibiyemi T. S.[2] Adewole K. S.[1]
[1]Department of Computer Science, University of Ilorin, Ilorin, P.M.B. 1515, Ilorin, Nigeria
[2]Department of Electrical and Electronic Engineering, University of Ilorin, Ilorin, P.M.B. 1515, Ilorin.

Abstract
Man communicates with man by natural language, sign language, and/or gesture but communicates with machine via electromechanical devices such as mouse, and keyboard. These media of effecting Man-To-Machine (M2M) communication are electromechanical in nature. Recent research works, however, have been able to achieve some high level of success in M2M using natural language, sign language, and/or gesture under constrained conditions. However, machine communication with man, in reverse direction, using natural language is still at its infancy. Machine communicates with man usually in textual form. In order to achieve acceptable quality of end-to-end M2M communication, there is need for robust architecture to develop a novel speech-to-text and text-to-speech system.
In this paper, an HCI speech-based architecture for Man-To-Machine and Machine-To-Man communication in Yorùbá language is proposed to carry Yorùbá people along in the advancement taking place in the world of Information Technology. Dynamic Time Warp is specified in the model to measure the similarity between the voice utterances in the sound library. In addition, Vector Quantization, Guassian Mixture Model and Hidden Markov Model are incorporated in the proposed architecture for compression and observation. This approach will yield a robust Speech-To-Text and Text-To-Speech system.
Keywords: Yorùbá Language, Speech Recognition, Text-To-Speech, Man-To-Machine, Machine-To-Man

1. INTRODUCTION

Human Computer Interaction (**HCI**) also called Man-To-Machine Interaction is a form of communication which entails the study, planning and design of communication between people and computers. It is often regarded as the intersection of computer science, behavioral sciences, and several other fields of study (Karry, 2008). It studies both human and machine in conjunction with obtaining knowledge supports from both sides. Examples of supports from machine side include the Operating Systems, Computer Graphic, enabling environment, while the human support entails linguistics, social sciences, and communication rules.

The design in HCI can be illustrated from two focal positions. Originally researchers involve in the design of prototype based studies like theories from the cognitive and the social sciences; ethnographic fieldwork; users with special needs. The prototypes are then designed. Subsequently, it is evident that contemporary HCI is not solely an academic discipline but also a field which is reaching out to and involving consultants, researchers and designers from industry. Their projects may result in objects whose application scope is used by the general public, outside of the walls of research laboratories (Fullman, 2003).

An end-to-end communication with the system in a speech-based interaction deals with speech recognition and Test-To-Speech (TTS) which enables humans to communicate with the system in a more natural way than the use of electromechanical devices such as mouse, keyboard, joystick, printer.

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 (Wikipedia, 2013). Speech research has applications in building computer systems that can recognize speech, as well as improving speech recognition for hearing- and language-impaired listeners (Akintola, 2011).

Speech recognition (also known as automatic speech recognition or computer speech recognition) converts spoken words to text. Speech recognition is the ability of machines to respond to spoken commands. Speech recognition enables “hands-free” control of various electronic devices, a particular boon to many disabled persons and the automatic creation of “print-ready” dictation. Before any machine can interpret speech, a microphone must translate the vibrations of a person’s voice into a wavelike electrical signal. This signal in turn is converted by the system’s hardware; example is a computer’s sound card which is responsible for analog to a digital signal. It is the digital signal that a speech recognition program analyzes in order to recognize separate phonemes. The phonemes are then recombined into words.

A Text-To-Speech (TTS) synthesizer is a computer based system that can read text aloud automatically, regardless of whether the text is introduced by a computer input stream or a scanned input submitted to an Optical Character Recognition (OCR) engine (Sasirekha & Chandra, 2012).

2. LITERATURE REVIEW

2.1 Speech

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. Each spoken word is created out of the phonetic combination of a limited set of vowel and consonant speech sound units (Bhusan & Krishna, 2013). Speech starts with the intention to communicate. There are many man-made sounds that may or may not involve any intention to communicate such as a sigh and a sneeze. The goal of such sounds is typically to cause understanding or response in a listener (Thurman & Graham, 2000).

2.2 Description of the Yorùbá Language

Yorùbá language is native to Nigeria, Togo and Benin. It is spoken by about 42 million people in south west Nigeria, Togo, Benin, Brazil, UK and USA (Akintola, 2011). It is one of the three official languages of Nigeria and also a member of the Niger-Congo language family.

Yorùbá is a tonal language like many African languages. Therefore, the meaning of a word is in the tone. Sounds in many languages are produced when alphabets are combined together. In tonal languages like Yorùbá, same alphabets can be combined together to give different meanings. These words are called homographs. The tonal sign put them aside and not the spelling.

Yorùbá sounds can be classified into three major kinds, namely: consonants, vowel and tonal sounds.

Consonants

The Yorùbá consonants are 18 in number and are drawn from the 25 letters of the Yorùbá alphabets. The consonants are: B, D, F, G, GB, H, J, K, L, M, N, P, R, S, Ṣ, T, W, and Y.

Vowels

The Yorùbá vowels are 7 in number and are also drawn from the 25 letters of the Yorùbá alphabets. The vowels are: A, E, Ẹ, I, O, Ọ, and U.

Syllabic Nasal

There also exists in the language a syllabic nasal phoneme. They occur before other consonants in syllable junctions. The syllabic nasal phoneme is represented as N or M. The homorganic allophones of the syllabic nasal phoneme are: [m], [M], [n], [ñ], [Ñ], and [Ñm].

2.2.1 Tone

Yorùbá is a tonal language with three level tones: high, low, and mid. Every syllable must have at least one tone. Tones are marked by use of the acute accent for high tone (‹á›), the grave accent for low tone (‹à›), mid is unmarked. Examples:

- H: ó bẹ́ 'he jumped'; síbí 'spoon'
- M: ó bẹ 'he is forward'; ara 'body'
- L: ó bẹ̀ 'he asks for pardon'; ọ̀kọ̀ 'spear'. (Wikipedia, 2014)

2.2.2 Alphabets

The upper and lower Yorùbá alphabets which comprises of both the consonants and vowels are;

A B D E Ẹ F G Gb H I J K L M N O Ọ P R S Ṣ T U W Y

A B D E ẹ f G gb H i j k L m N O ọ P r S ṣ t u w y

2.2.3 Phonology

The three possible syllable structures of Yorùbá are consonant + vowel (CV), vowel alone (V), and syllabic nasal (N). Every syllable bears one of the three tones: high ‹◌́›, mid ‹◌̄› (generally left unmarked), and low ‹◌̀›. The sentence 'ń ò lọ', which means 'I did not go', provides examples of the three syllable types.

2.2.4 Numerals

The Arabic numerals (0-9) used for counting have Yorùbá equivalent which are: Òdo, Ení, Èjì, Ètà, Èrìn, Àrun, Èfà, Èje, Èjo, Èsán.

2.3 Speech Recognition

The practice of enabling a computer to identify and respond to the sounds produced in human speech is a form of speech recognition. The computer translates speech spoken by man to text. It is also known as Automatic Speech Recognition (ASR) or Speech-To-Text system which is a way of Man-To-Machine Communication.

Speech recognition system consists of the following:

- A microphone, for the person to speak.
- Speech recognition software.
- A computer to take and interpret the speech.
- A good quality soundcard for input and/or output

2.4 Text-To-Speech (TTS)

Text-To-Speech also known as Speech Synthesis is the computer production of human speech. It is the process of generating spoken words by machine from written input.

Speech is often based on concatenation of natural speech i.e units that are taken from natural speech put together to form a word or sentence. Concatenative speech synthesis according to (Sproat & Olive, 1999) has become very popular in recent years due to its improved sensitivity to unit context.

Rhythm also is an important factor that makes the synthesized speech of a TTS system more natural and understandable; the prosodic structure provides important information for the prosody generation model to produce effects in synthesized speech (Sasirekha & Chandra, 2012).

2.5 Related Work

The earliest attempts at man-to-machine communication was made in 1950s (Anusaga & Katti, 2009), it was based on finding speech sounds and providing appropriate label for it. Since then many researchers have been working on how to improve the technology.

Ibiyemi and Akintola (2012) used Mel's Frequency Cepstral Coefficients (MFCC) for feature extraction and Vector Quantitization for the data compression in a telephone voice dialing system that was limited to digits. MFCC was also used for extraction by Akintola (2011) in a speech recognition system and data compression was not done in any form; thereby resulting in a large database. Linear predictive coding (LPC) and artificial neural network (ANN) combined in speech recognition was proposed by Wijoyo and Thiang (2011). In their research, LPC was used for feature extraction and ANN for recognition method. A dictionary of common words in Arabic language was developed to enhance the system, though made the system to be dependent on a large database. In Ibiyemi and Akintola (2012), recognition phase was implemented by simple Euclidean distance measures which result in the recognition of words. According to Mohammed, Sayed, Abdulnaiem, and Moselhy (2013), MFCC produces best result out of the forms of feature extraction in speech recognition system (Ibiyemi & Akintola, 2012; Akintola, 2011; Wijoyo & Thiang 2011).

On the other hand, Test-To-Speech TTS is still very much at infancy as researchers are working round the clock to have a better algorithm. A TTS system developed by the establishment of corpus-based synthesis unit database that includes nasals, tones, stops and sadhi rules (Sher, Hsu, Chiu, & Chung, 2010), subsystems of the system includes text-input system, text-to-sound convert system, training of basic synthesis units, and the acoustic wave play system. The system has a multiple accent corpus-based database which was developed using combination of basic phonemes of vowels, consonants and tones from MLT (Modern Literal Taiwanese) books. It has limited speech input but uses large database to develop the MLT. A concatenative synthesis and bell lab approach (combination of phonetics and linguistic structure) to speech synthesis relies on designing and creating the acoustic inventory of the language by taking real recorded speech, cutting it into segments and concatenating these segments back together during synthesis (Christogiannis, Varvarigou, Zappa, & Vamvakoulas, 2000). The synthesizer then produces a concatenative system, based on a set of prerecorded acoustic inventory elements that represent all the possible phone-to-phone transitions of the language. An Arabic system that uses a rule-based hybrid system which is combination of formant and concatenative speech techniques reduces the vocabulary independence and can handle all types of input text (Zekki, Kalifa, & Naji, 2010). The system omits some vowels of the language in use and also did not take intonation into consideration.

The use of concatenative synthesis bypasses most of the problems encountered by articulatory and formant synthesis techniques (Sher, Hsu, Chiu, & Chung, 2010). Most developed systems make use of very large database that can slow the system down and also require lots of memory space. The issue of incorrect labeling due the large database can also lead to poor quality of the system.

In Sak, Gungor, and Safkan (2006), their proposed system contains front-end which comprised of text analysis and phonetic analysis. The unit selection algorithm is based on Viterbi decoding algorithm of the best-path in the network of the speech units using spectral discontinuity and prosodic mismatch objective cost measures in place of Hidden Markov Model (HMM). The back-end is the speech waveform generation based on the harmonic coding of speech. The Harmonic coding enabled the system to compress the unit inventory size by a factor of three. Though, the system used transplanted prosody which does not take intonation in to consideration, where generated prosody would have been more effective for the same purpose.

3. METHODOLOGY

This study is specifically to design and develop a Man-To-Machine and Machine-To-Man system. It is broken into two modules which are the Speech Recognition Module and the Text-To-Speech Module.

The Speech recognition module 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. While Text-To-Speech module draws words gotten from the speech recognizer and converts it back to speech through text analysis, natural language processing and digital signal processing. Vector Quantitization, Guassian

Mixture Model and Hidden Markov Model are applied to have better results. Concatenative Synthesis approach of TTS is used to form words by combining syllables.

3.1 Data collection

Speech data (**Yorùbá speech corpus**): The data collection at this stage involves adequate training and testing data of Yorùbá speech samples.

3.1.1 Yorùbá Character Generation: This is a distinct catalog of characters (Yorùbá Alphabet, counting numbers and special symbols) recognized by the computer hardware and software. Each of the character corresponds to a defined number in the upper ASCII character set (128 - 255). The English characters and other control characters make use of the lower ASCII set (0 - 127) leaving the remaining for any other languages such as French and German.

The character set defines 110 characters (the remaining 18 for later expansion) 128-237 decimal in hexadecimal.

Examples:

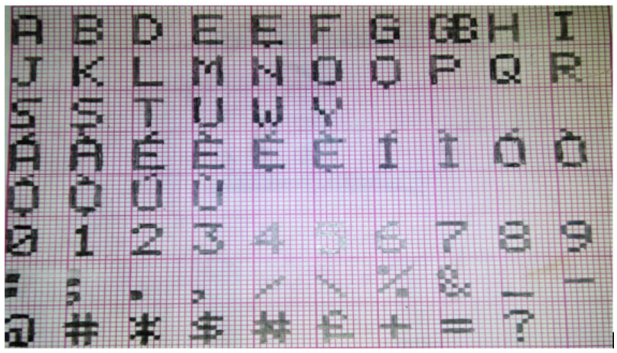

Figure 1: Generated Yorùbá Characters

3.1.2 Phonemes: Yorùbá phonemes are the perceptually distinct units of sound that distinguish a word from another. Table1 below shows the phonemes and pronunciation for Yorùbá alphabets.

Table 1: Thirty (30) Yorùbá phonemes

S/No.	Phoneme	Pronunciation
1	/b/	B
2	/d/	D
3	/f/	F
4	/g/	G
5	/g͡b/	Gb
6	/h/	H
7	/dʒ/ or /ɟ/	J
8	/k/	K
9	/l/	L
10	/m/	M
11	/n/	N
12	/k͡p/	P
13	/r/	R
14	/s/	S
15	/ʃ/	ṣ
16	/t/	T
17	/w/	W
18	/j/	Y
19	/a/	A
20	/e/	E
21	/ɛ/	ẹ
22	/i/	I
23	/o/	o
24	/ɔ/	ọ
25	/u/	U
26	/ã/	An
27	/ɛ̃/	ẹn
28	/ĩ/	In
29	/ɔ̃/	ọn
30	/ũ/	Un

3.1.3 Syllable: Yorùbá syllable is a unit of pronunciation having one vowel sound, with or without surrounding consonants, forming the whole or a part of a word. Table 2 below shows all possible forms of Yorùbá language syllables.

Table 2: 201 (Two hundred and one) Yorùbá Syllables

A	E	Ẹ	I	O	Ọ	U	Ba	Da	Fa
Ga	Gba	Ha	Ja	Ka	La	Ma	Na	Pa	Ra
Sa	ṣa	Ta	Wa	Ya	Be	De	Fe	Ge	Gbe
He	Je	Ke	Le	Me	Ne	Pe	Re	Se	ṣe
Te	We	Ye	Bẹ	Dẹ	Fẹ	Gẹ	Gbẹ	Hẹ	Jẹ
Kẹ	Lẹ	Mẹ	Nẹ	Pẹ	Rẹ	Sẹ	ṣẹ	Tẹ	Wẹ
Yẹ	Bi	Di	Fi	Gi	Gbi	Hi	Ji	Ki	Li
Mi	Ni	Pi	Ri	Si	ṣi	Ti	Wi	Yi	Bo
Do	Fo	Go	Gbo	Ho	Jo	Ko	Lo	Mo	No
Po	Ro	So	ṣo	To	Wo	Yo	Bọ	Dọ	Fọ
Gọ	Gbọ	Họ	Jọ	Kọ	Lọ	Mọ	Nọ	Pọ	Rọ
Sọ	ṣọ	Tọ	Wọ	Yọ	Bu	Du	Fu	Gu	Gbu
Hu	Ju	Ku	Lu	Mu	Nu	Pu	Ru	Su	ṣu
Tu	Wu	Yu	N	M	An	Ẹn	Ọn	Un	Ban
Dan	Fan	Gan	Gban	Jan	Kan	Lan	Han	Yan	Pan
Ran	San	ṣan	Tan	Wan	Bẹn	Dẹn	Fẹn	Gẹn	Gbẹn
Hẹn	Jẹn	Lẹn	Pẹn	Rẹn	Sẹn	ṣẹn	Tẹn	Wẹn	Yẹn
Bọn	Dọn	Fọn	Gọn	Gbọn	Họn	Jọn	Kọn	Lọn	Pọn
Rọn	Sọn	ṣọn	Tọn	Wọn	Yọn	Bun	Dun	Fun	Gun
Gbun	Hun	Jun	Lun	Pun	Run	Sun	ṣun	Tun	Wun
Yun									

3.1.3 Homographs: Yorùbá homographic words are two or more Yorùbá words spelt the same way but not pronounced the same and have different meanings. Yorùbá language makes use of tones to differentiate these words. Table 3 below shows some of the homographic words in Yorùbá with corresponding syllable, English meaning and phoneme pronunciation.

Table 3: Yorùbá Homographic Words

S/NO.	Word	Homographs	Syllable	Meaning	Pronunciation
1	Aba	Abá	A/bá	Attempt	Abá
		Àbá	À/bá	Mat	Àbá
		Abà	A/bà	Barn	Abá
		Aba	A/ba	Staple, Incubation	Aba
2	Abe	Abẹ́	A/bẹ́	Bottom	abɛ́
		Abẹ	A/bẹ	Razor	abɛ
3	Abo	Abo	A/bo	Female	Abo
		Àbò	À/bò	Refuge	Àbò
4	Aja	Ajá	A/já	Dog	adʒá
		Ajà	A/jà	Attic	adʒà
5	Ajẹ	Àjẹ́	À/jẹ́	Sorcerer	àdʒɛ́
		Àjẹ̀	À/jẹ̀	Oar, Paddle	àdʒɛ̀
6	Ala	Àlá	À/lá	Dream	Àlá
		Àlà	À/là	Boundary	Àlà
7	Apa	Apà	A/pà	Arm	akpà
		Àpa	À/pa	Prodigal	àkpa
		Apá	A/pá	Mark, Sign	akpá
8	Ara	Ara	A/ra	Body	Ara
		Ará	A/rá	Relative	Ará
		Àrá	À/rá	Thunder	Àrá
		Àrà	À/rà	Fashion	Àrà
9	Baba	Baba	Ba/ba	Father	Baba
		Bàbà	Bà/bà	Guinea Corn	Bàbà
		Bàbá	Bà/bá	Great thing	Bàbá
10	Dana	Dáná	Dá/ná	Make fire	Dáná

		Dánà	Dá/nà	Robbery	Dánà
		Dána	Dá/na	Pay dowry	Dána
11	Ede	Èdè	È/dè	Dialect	Èdè
		Edé	E/dé	Lobster	edé
		Èdé	È/dé	Buffalo	Èdé
12	Ere	Ère	È/re	Gain	Ère
		Eré	E/ré	Game	Eré
		Èrè	È/rè	Statue	Èrè
		Erè	E/rè	Snake	Erè
13	Ewu	Èwú	È/wú	A day pounded yam	Èwú
		Ewu	E/wu	Danger	ewu
		Ewú	E/wú	Grey hair	ewú
14	Ẹfon	Ẹfọn	Ẹ/fọn	Mosquito	ɛ̀fɔn
		Ẹfọ́n	Ẹ/fọ́n	Arrow	ɛfɔ́n
		Ẹfọn	Ẹ/fọn	Buffalo	ɛfɔn
15	Ẹgba	Ẹgba	Ẹ/gba	Whip	ɛg͡ba
		Ẹgbà	Ẹ/gbà	Two thousand	ɛg͡bà
		Ẹgbà	Ẹ/gbà	Bracelet	ɛ̀g͡bà
		Ẹgbá	Ẹ/gbá	Yorùbá Tribe	ɛ̀gbá
16	Ẹrin	Ẹrín	Ẹ/rín	Laughter	ɛ̀rĩ́
		Ẹrin	Ẹ/rin	Four	ɛ̀rĩ
17	Ẹtu	Ẹtù	Ẹ/tù	Guinea Fowl	ɛtù
		Ẹtù	Ẹ/tù	Gun Powder	ɛ̀tù
		Ẹtu	Ẹ/tu	Antelope	ɛtu
18	Ẹwa	Ẹwa	Ẹ/wa	Ten	ɛwa
		Ẹwà	Ẹ/wà	Beauty	ɛwà
		Ẹ̀wà	Ẹ̀/wà	Beans	ɛ̀wà
19	Giri	Gìrì	Gì/rì	Convulsion	gìrì
		Gírí	Gí/rí	Promptly	gírí
		Girì	Gi/rì	Suddenly	girì
20	Gba	Gbà	Gbà/	Receive	g͡bà
		Gbá	Gbá/	Sweep	g͡bá
21	Gbo	Gbó	Gbó/	Bark, Ripen	g͡bó
		Gbò	Gbò/	To affect	g͡bò
22	Iba	Ìba	Ì/ba	Few	Ìba
		Ìbà	Ì/bà	Respect	Ìbà
		Ibà	I/bà	Fever	Ibà
23	Ibo	Ìbò	Ì/bò	Plant	Ìbò
		Ibo	I/bo	Where	Ibo
24	Idi	ÌdÌ	Ì/dÌ	Bundle	ÌdÌ
		Idi	i/di	Bud	Idi
		Ìdí	Ì/dí	Waist, Reason	Ìdí
25	Igba	Ìgbà	Ì/gbà	Time	Ìg͡bà
		Igba	I/gba	Two thousand	Ig͡ba
		Igbá	i/gbá	Calabash	ig͡bá
		Ìgbá	Ì/gbá	Locust beans	Ìg͡bá
		Igbà	i/gbà	Rope for climbing	ig͡bà
26	Ika	Ìkà	Ì/kà	Cruelty	Ìkà
		Ìka	Ì/ka	Finger	Ìka
27	Iko	Ìkó	Ì/kó	Hook	Ìkó
		Ìkò	Ì/kò	Delegate	ìkò
		Ikó	I/kó	Cough	ikó
28	Obi	Òbí	Ò/bí	Parent	Òbí
		Obì	O/bì	Kolanut	obì
29	Ogun	Ogún	O/gún	Inheritance	ogṹ
		Ògún	Ò/gún	God of iron	ògṹ

		Ógún	Ó/gún	Medicine	ógṹ
		Ogùn	O/gùn	Twenty	ogũ̀
		Ogun	O/gun	War	ogũ
30	Ojo	Òjò	Ò/jò	Rain	òjò
		Ojo	O/jo	Fear	ojo
		Òjó	Ò/jó	Name	Òjó
31	Okun	Òkun	Ò/kun	Sea	òkũ
		Okùn	O/kùn	Rope	okũ̀
		Okun	O/kun	Strength	okũ
32	Orun	Orùn	O/rùn	Sun	orũ̀
		Orun	O/run	Sleep	orũ
		Orún	O/rún	Scent	orṹ
33	Ọka	Ọkà	Ọ/kà	Corn	ɔkà
		Ọkà	Ọ/kà	Child's disease	ɔ̀kà
		Ọká	Ọ/ká	Snake	ɔká
34	Ọkọ	Ọkọ̀	Ọ/kọ̀	Canoe	ɔkɔ̀
		Ọkọ̀	Ọ/kọ̀	Spear	ɔ̀kɔ̀
		Ọkọ́	Ọ/kọ́	Hoe	ɔkɔ́
		Ọkọ	Ọ/kọ	Husband	ɔkɔ
35	Ọrun	Ọ́rún	Ọ́/rún	Bow	ɔ́rṹ
		Ọrùn	Ọ/rùn	Neck	ɔrũ̀
		Ọrún	Ọ/rún	Hundred	ɔrṹ
		Ọrun	Ọ/run	Heaven	ɔrũ
36	Ọwọ	Ọwọ̀	Ọ/wọ̀	Honour	ɔ̀wɔ̀
		Ọwọ́	Ọ/wọ́	Flock of birds	ɔ̀wɔ́
		Ọwọ̀	Ọ/wọ̀	Broom	ɔwɔ̀
		Ọwọ́	Ọ/wọ́	Hand	ɔwɔ́

3.2 System Design

Figure 2 shows the architecture, modules and interfaces for the proposed system to satisfy the requirements for Speech Recognition and Text-To-Speech.

3.3 Sound library

The sound library houses the recorded vowels, phonemes, syllables and homographs pronunciation. The total number of sounds in the library is three hundred and thirty-three (339).

The phonemes and their pronunciations are thirty (30) as shown in Table 1. All possible forms of syllable as derived from Table 2 are two hundred and one (201). This comprises of vowels (V), consonant vowel (CV) and nasal stops (M and N). The Thirty-Six (36) lexis which gave rise to (108) homographic words are also inclusive in the library.

3.4 Pseudo Code for Speech-To-Text

```
Speech_To_Speech ()
{       Step 1: Input speech
Step 2: Call WavReader()
Step 3: Call wordsegmentation()
Step 4: Call preEmphasisFilter()
Step 5: Call FrameBlocking()
Step 6: Call HammingWindow()
Step 7: Call DTW()
Step 8: Call StoreTemplate(word_{k,I,r})
Step 9: End
}
```


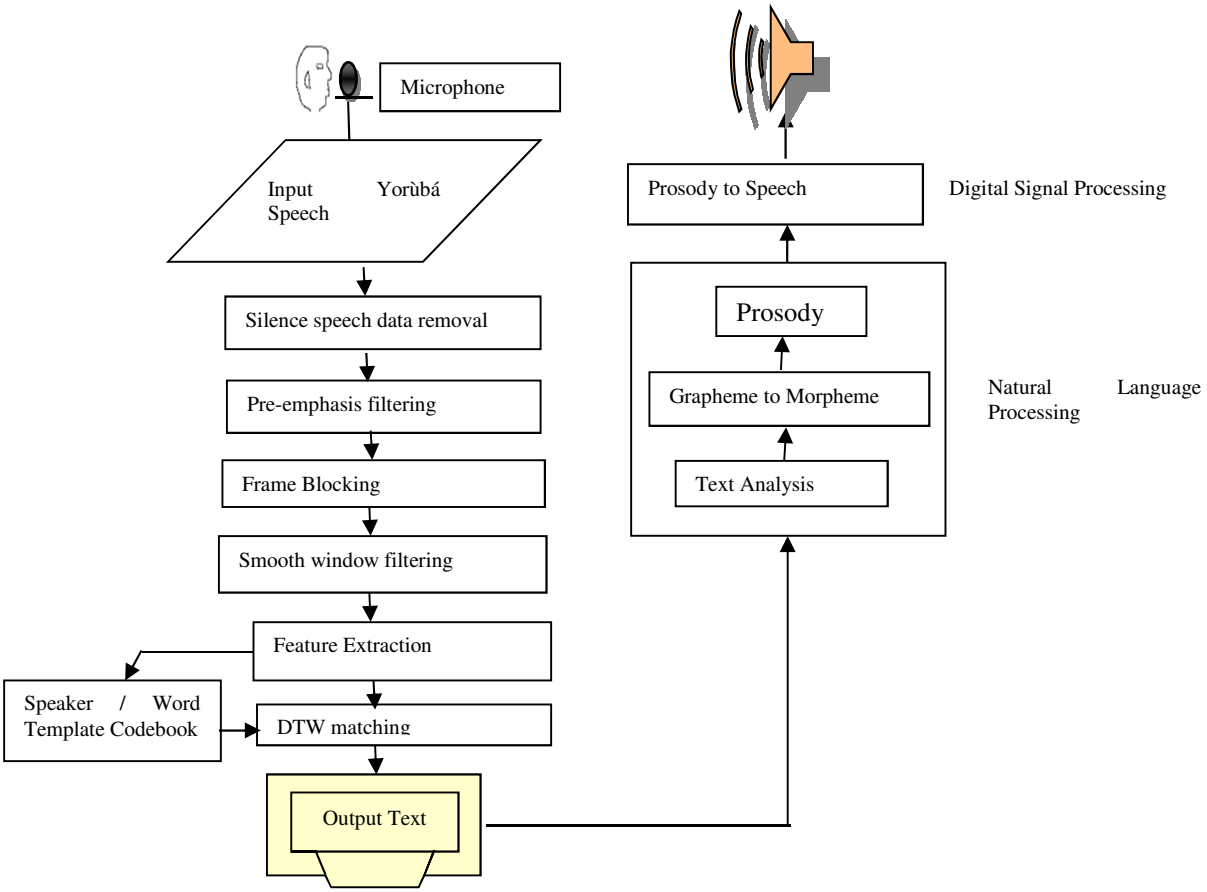

Figure 2: Proposed Machine-To-Man and Man-To-Machine Model

3.5 Pseudo Code for Text-To-Speech

{
Step 1: Text Analysis;
Step 2: Grapheme to Morpheme()
Step 3: Prosody generation()
Step 4: Prosody to Speech()
Step 5: End
}

4. CONCLUSION

Communication between man and man is by natural means such as sign language and gesture. Conversely, Man-To-Machine and Machine-To-Man communication is through electromechanical devices such as mouse, and keyboard. This communication involves application of Human Computer Interaction (HCI) which is not a natural means of communication for man. This research is aimed to design a robust architecture; for solving the non-natural means of communicating with machine in Yorùbá language using Speech Recognition and Text-To-Speech system. In order to extend HCI technology to the grass root, Machine-To-Man and Man-To-Machine communication in Yorùbá language is employed in this paper using Dynamic Time Warp, Vector Quantization, Guassian Mixture Model and Hidden Markov Model for a robust system architecture. Also, the pseudo codes for the proposed Speech Recognition and TTS system is presented.

REFERENCES

Akintola, A. (2011). Automatic Speaker Recognition-Based Telephone Voice Dialing in Yorùbá (Unpublished master's thesis). University of Ilorin, Ilorin.

Anusaga, M. & Katti, S. (2009). Speech recognition by machine: A review. *International Journal of Computer Science and Information Security*, 6, 181-205.

Bhusan, C. & Krishna, B. (2013). Nepali Text to Speech Synthesis System using ESNOLA Method of Concatenation *International Journal of Computer Application,* 62, 24-28.

Christogiannis, C., Varvarigou, T., Zappa, A., & Vamvakoulas, Y.(2000). Construction of the acoustic inventory

for Greek Text-To-Speech Synthesis system. In *Proceedings of International Conference on Spoken Language Processing*, (pp. 267-270).

Fallman, D. (2003). Design-oriented human—computer interaction. *CHI Letters*, 5, 225-232.

Ibiyemi, T. & Akintola, A. (2012). Automatic Speech Recognition for Telephone Voice Dialing in Yorùbá. *International journal of engineering Research and Technology*, 1, 1-6.

Karry, F. (2008). Human computer interaction: Overview on state of the art. *International Journal on Smart Sensing and Intelligent Systems*, 1, 137-159.

Mohammed, E., Sayed, M., Abdelnaiem, A., & Moselhy, A. (2013). LPC and MFCC performance evaluation with artificial neural network for controlling spoken language identification. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 6, 55-66.

Sak, H., Gungor, T., & Safkar, Y. (2006). A Corpus-Based concatenative speech synthesis system for Turkish. *Turkish Journal of Electrical Engineering and Computer Science,* 14, 209 - 223.

Sasirekha, D. & Chandra, E. (2012). Text-to-speech: a simple tutorial. *International Journal of Soft Computing and Engineering*, 2, 275-279.

Sher, Y., Chiu, Y., Hsu, M., & Chung, K. (2010). Development of Hmm Based Taiwanese Text- To-Speech system. In *International Conference on Software Techniques and Engineering*, (pp.330-333).

Sproat, R. & Olive, J. (1999). Text-to-Speech Synthesis In K. Madisetti & B. Douglas (Eds.) *Digital Signal Processing Handbook* () Boca Raton: CRC Press LLC.

Thurman, L. & Graham, W. (2000). *Body mind & voice: Foundations of voice education,* Collegeville: Pathenon.

Wikipedia. (2013). *Speech perception*. Retrieved July 12, 2013, retrieved from Wikipedia: http:// http://en.wikipedia.org/wiki/Speech_perception

Wikipedia. (2013). *Yorùbá Language*. Retrieved April 22, 2014, retrieved from Wikipedia: http://en.wikipedia.org/wiki/Yoruba_language

Wijoyo, S. and Thiang, (2011). Speech recognition using linear predictive coding and artificial neural network for controlling movement of mobile robot. In H. Shibasa & M. Oto (Ed.) *International Conference on Information and Electronics Engineering*, (pp. 179- 183). Singapore: IACSIT Press.

Zeki, M., Othman, O., Khalifa, A., & Naji, W. (2010) Development of an Arabic Text-To- Speech System. In *International Conference on Computer and Communication Engineering*, (pp. 11-14).

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

Academic conference: http://www.iiste.org/conference/upcoming-conferences-call-for-paper/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar