DEPARTMENT OF COMMUNICATIONS ENGINEERING
AHMADU BELLO UNIVERSITY, ZARIA - NIGERIA

Proceedings of the 1st National Communication Engineering Conference 2018

A Framework for Multiple Choice Multilingual Translation System Using Hidden Markov Model and Viterbi Algorithm

Peter Shadrach Anza
Department of Computer Science
Federal University of Technology
Minna, Nigeria.
shadytins@gmail.com

Muhammad Bashir Abdullahi
Department of Computer Science
Federal University of Technology
Minna, Nigeria.
el.bashir02@futminna.edu.ng

Abstract— **In the multilingual World, majority of languages are in parallel to each other, which make communication among different speakers difficult and burdensome. Most of the existing approaches to language translation focuses on either speech-to-text, text-to-speech, speech-to-speech or text-to-text, but do not consider user's preferences. In this paper, we present a framework for multiple choice multilingual translation system to convert the input English speech signals, text and printed text into Speech and/or text output for users in either Hausa, Igbo or Yoruba. Intuitively, the system consists of four modules, which include text extraction, speech recognition, text translation and speech synthesis modules. We used Mel Frequency Cepstral Coefficients (MFCC) to extract features from the speech signals of spoken words. Furthermore, we used Hidden Markov Model to train and test the audio files to get the recognized spoken word. The Viterbi Algorithm was used to get the most likely path and word combinations. For scanned images and printed documents, Optical Character Recognition was used for text extraction.**

Keywords—HMM, MFCC, OCR, Language translation, Speech recognition

I. INTRODUCTION

Language is one of the most important medium by which human communication can take place [1]. Albeit signs, is another medium the physically challenged (deaf) individuals employed. A language could be in written and spoken forms each with granularity of representation, which is referred to as smallest basic units or building blocks. These basic units for spoken and written forms are known as phonemes/syllables and graphemes/characters, respectively. Instinctively, there must be a consistent relationship between the spoken and written form. In a Multilingual world with diversity in geography and culture, the basic units of majority of languages are in parallel to each other and thus, make communication among different speakers difficult and burdensome. Consequently, as the world is becoming borderless day-by-day, the importance of a multilingual translation technology has been increasing, much more so with the advent and widespread use of communication tools, which has helped in the growth of the analysis and processing of speech signals. The goal of a Multilingual translation system is majorly to generate a speech signal or text in one (target) language that conveys the linguistic information contained in a given speech signal or text of another (source) language. Therefore, for effective communication using a translation system, languages are supposed to be in their most natural form (prosody) with emotions, expressions etc. not just in the machine voiced form. Although, different language pairs require different approach and technology due to the linguistic and tonal difference amongst the languages.

However, conventional approaches to language translation typically ignore key contextual information such as prosody, emphasis, and discourse state in the translation process. Capturing and exploiting the contextual information is especially important in machine-mediated translation system as it can serve as a complementary knowledge source that can potentially aid the end users in improved understanding and disambiguation [2]. Furthermore, majority of translation systems not only perform either Speech-to-Text [3], Text-to-Speech [4], Speech-to-Speech [2, 6, 7, 15] or Text-to-Text, but also do not consider user's preferences. Therefore, they prevent the user from many privileges such as the ability to obtain information in the target language of choice or users nearest language and not machine determined target Language.

From the aforementioned reasons, in this paper, a framework for multiple choice multilingual translation system (MMTS) was developed, which considers both speech and text as input. For translation, a user can enter the text manually or can choose an image of printed text as input. Similarly, a user can record speech sound or obtain an already recorded speech sound in the different available formats (wav, Mp3) as input. Furthermore, a user has

privilege to select a target language of interest. The target languages available to the user are Hausa, Igbo and Yoruba.

The contribution of this paper is two folds: (1) a framework for multiple choice multilingual translation system was designed. (2) A sample of translation into the three target languages were demonstrated, respectively.

The rest of the paper is organized as follows: In Section II, we discuss the related work. Section III describes the materials and methods used to design the MMTS framework. In Section IV, we demonstrate the results obtained by performing experiments in simulation. Lastly, we present concluding remarks in Section V.

II. Related Work

Debnath *et al.* [5] designed a Multilingual Speech Translator using MATLAB, which takes in manually written or image printed text in English and translate into speech in either of English, Hindi or Bengali. It is made up of stored sound library and translation is carried out based on iterative comparisons. The constraint or defect with the system is that the speech sounds are devoid of prosody and are not easily understood by the native speakers because they use machine sounds.

Nakamura *et al.* [6] developed an ATR multilingual speech-to-speech translation (S2ST) system, which mainly focused on translation between English and Asian languages (Japanese and Chinese). There are three main modules in the ATR S2ST system: large-vocabulary continuous speech recognition, machine text-to-text (T2T) translation, and text-to-speech synthesis. The main constraint of this system is that it is majorly speaker dependent.

Prasanthi *et al.* [7] designed a Speech to Text (STT) conversion system using Hidden Markov Model (HMM), the existed technique uses MFCC technique for feature extraction. The HMM uses forward-backward procedure and Baum-welch algorithm, the forward-backward procedure gives better prediction of states. Similarly, Hla and Tun [3] developed a Speech-To-Text Conversion (STT) System Using HMM. The system uses MFCC to extract speech signals at lower frequencies for greater energy. However, this system can only be used with minimal number of HMM states as much states affects the accuracy of the recognition.

Borgohain *et al.* [8] designed an English-Assamese Machine Translation System that translates English sentences (text) into Assamese sentences (text). The system finds the maximum probability of the target sentence using the probabilities of the language model and the translation model. It also adopts the use of a parallel corpus for the two languages. The main constraint of this system is that it depends on the probabilities of the text sequences rather than the accuracy of the translation. In addition, it works for only a limited amount of sentences, mainly simple sentences and does not take into account the word order of the sentence.

Badhe and Ghate [9] developed a Marathi Text to Speech Synthesis system. this system converts any given Marathi text as an input into sound as an output. The first step in the system is to break a given Marathi text into words. These words are then mapped with their respective English transliteration and the system now will check for the related audio file in the database. If the audio file is present in the database, it will concatenate the files and play them. A common constraint of this system is that it is devoid of the naturalness and the intelligibility of the Marathi language. In addition, the database contains limited words.

III. Materials and Methods

The MMTS framework is a domain specific translation system that can translate from a source English language to a target language of choice (i.e. either Hausa, Igbo or Yoruba). It takes in both Speech and/or text as input, as shown in Fig. 1. In addition, a user can enter the text manually or can choose to scan the text from an image of printed text. However, the Speech sounds can be recorded or obtain from already recorded Speech sound saved in the system in the different available formats (wav, Mp3). The texts alignment and correction was achieved with the aid of an editor.

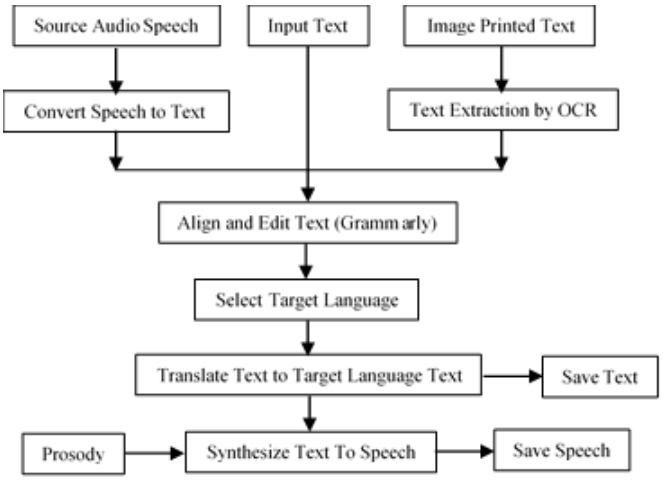

Fig. 1. The MMTS flow diagram

The MMTS framework consist of four modules namely: speech recognition module, language translation module, speech synthesis module and text recognition module.

A. Hidden Markov Model

A HMM is a Finite State Machine having a number of states. Therefore, since speech is a random process, the goal of using HMM is to model the feature extraction (phonemes) and find the parameters of the random variables in a well-defined manner. HMM process is observed through set of statistical processes that produce the sequence of observations. Thus, it presents a best way of quantifying speech patterns. A HMM is characterized by three matrices viz., a, b and π.

a - Transition Probability matrix $(N\times N)$

b - Observation Symbol Probability Distribution matrix $(N\times M)$

π - Initial State Distribution matrix $(N\times 1)$

where

N = Number of states in the HMM

M = Number of Observation symbols

$P(b|\text{Model})$ = Probability of the observation sequence b given the Model

$P(b|\text{Model})=P[b_1,b_2,...,b_n]=\prod_i^n P[b_i|b_{i-1}]$ gives the Markov Model.

We obtained the speech waveform (see Fig. 2) for a recorded speech for the sentence "the weather affected our crop yield" in 3 seconds. Then we build the model for the recorded speech for recognition as shown in Fig. 3.

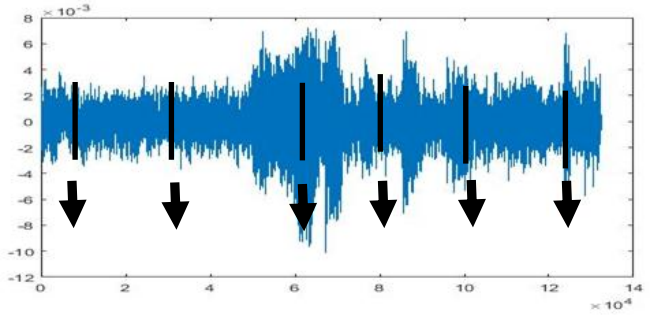

Fig. 2. the speech waveform

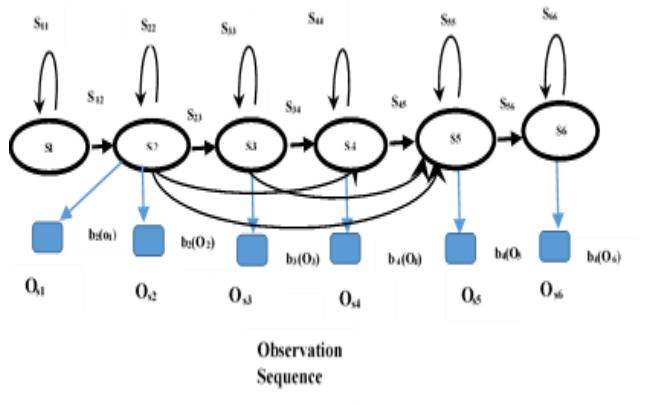

Fig. 3. The HMM for the speech waveform.

The phonemes are the smallest unit of an acoustic speech signal. The states S_1, S_2, S_3, S_4, S_5, S_6 represent the Phonemes from the acoustic speech signals. The Sequence $S_{12},\ldots,S_{56}$ as illustrated in Figure 3 represents the phonemes transition from the states. We use the vectors $O_{s1}, \ldots, O_{s6}$, which are the extracted features from the MFCC to predict the phonemes, which give us the Words and then the sentence. The observation symbol represents the probability distribution from one state to another, $b_2(O_1)$ represents the probability distribution of the feature vector O_{s1} from state 1 to state 2 etc. The basic idea is to find the most likely strings of phonemes, words and sentences given some acoustic speech input signals. We obtained the maximum probability of the word given the observation probability using Baye's rule as follows:

$$\arg\max_{w\in L} P(w|b)=\arg\max_{w\in L}\frac{P(b|w)P(w)}{P(b)} \qquad (1)$$

where w is a string of words, L is the language of interest, and b is the set of acoustic speech vectors gotten from the front end interface using MFCC.

For a single speech input (e.g. one sentence), the acoustic (b) will be constant and therefore, will be P(b). P(b|w) is called the likelihood model, P(w) is called the prior probability of the word string. In the case of word pronunciation P(b|w) is the probability of the acoustics given the word. Note that many acoustic sounds can make a word, so we choose an intermediate representation, the phones or phonemes

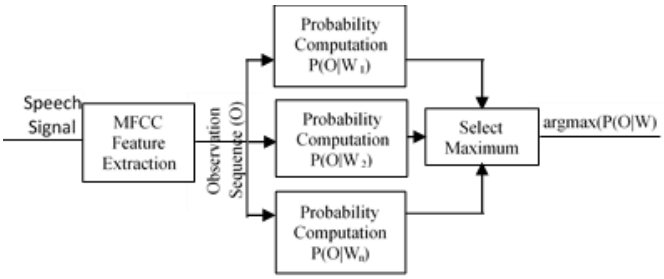

Fig. 4. HMM word model for recognition

MFCC feature extraction technique was carried on the Speech signals obtained from the front end interface, as shown in Fig. 4. The features obtained which form the vectors were used to calculate the maximum probability in the word model to obtain the appropriate word for recognition.

B. Speech Recognition

Speech Recognition (also known as Automatic Speech Recognition (ASR), or computer speech recognition is the process of converting a speech signal to a sequence of words, by means of an algorithm implemented as a computer program. The speech recognition tool takes in audio as Input and convert it into text (see Fig. 5), which serves as input for the next phase. It consists of a front end and back end. The front end processes the audio speech while the back end is a search engine that takes input from the front end and searches the database.

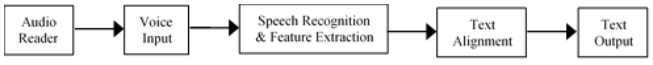

Fig. 5. Block diagram for speech recognition

C. *Feature Extraction*

Feature extraction is the process of extracting the information related to language or speech. The feature extraction of speech signal is based on the short term of the amplitude of the spectrum of the speech signal. During feature extraction the voice recording is cut into windows of equal length, these cutout samples are called frames, which are often 10 to 30*ms* long. The short section of speech signal is separated from the spectrum and then are processed. This processing is repeated for the entire duration of the waveform.

Therefore, to detect the required feature of speech signal, we adopted the MFCC technique. This is because of its wide usage and accuracy over the other techniques such as Linear Prediction Coding (LPC) and Auditory Spectrum-Based Speech Feature (ASSF). For this purpose, the first step to recognize the speech samples is to extract the features by identifying the components of the audio signal that are good for recognizing the linguistic content and discarding all the other stuff, which carries information like background noise, emotion etc. The aim of feature extraction is to reduce the data size of the speech signal before pattern classification or recognition. The steps involve in MFCCs calculation (see Fig. 6) are framing, windowing, Discrete Fourier Transform (DFT), Mel frequency filtering, logarithmic function and Discrete Cosine Transform (DCT).

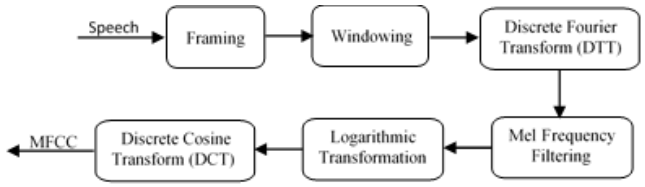

Fig. 6. Block diagram of MFCC for the feature extraction

After the feature pre-processing the speech-samples are extracted to features or coefficients by the use of MFCC. Finally, these MFCC coefficients are used as the input to HMM recognizer to classify the desired spoken word.

D. *Viterbi Algorithm*

The Viterbi Algorithm is a dynamic programming algorithm for finding the most likely sequence of hidden states called the Viterbi path, which results in a sequence of observed events especially in the context of Markov Chain and HMM. It has universal application in decoding convolution codes used in both CDMA and GSM digital cellular, dial–up modems [10] etc. For Speech recognition, the acoustic signal is treated as the observed sequence of events and a string of text is considered the 'hidden cause' of the acoustic signal. We use the Viterbi algorithm to find the most likely string of text given the acoustic signal or feature extractions. It generates a path: $X= \{X_1, X_2, X_3, \ldots X_n\}$ which is a sequence of states $X_n \in S= \{S_1, S_2, S_3, \ldots S_k\}$ that generates the observations $Y = (Y_1, Y_2, Y_3, \ldots Y_T)$ with $Y_n \in O =(O_1, O_2, O_3, \ldots O_N)$ (N being the count of observations)

INPUT

The observation space O=$\{O_1, O_2, \ldots, O_N\}$
The state space S= $\{S_1, S_2, \ldots, S_K\}$
An array of initial probabilities $\pi_i = (\pi_1, \pi_2, \ldots, \pi_k$ such that π_i stores the probability that $X_1==S_i$
a sequence of observations $Y= (Y_1, Y_2, \ldots, Y_T)$ such that $Y_t==i$ if the observation at time t is o_i
transition matrix A of size K x K such that A_{ij} stores the transition probability of transiting from state s_i to state s_j
emission matrix B of size K x N such that B_{ij} stores the probability of observing O_j from state S_i

OUTPUT

The most likely hidden state sequence $X= (x_1, x_2, \ldots, x_N)$

function *VITERBI*
(O,S, π ,Y,A,B):X
for each state
i ϵ {1,2, … ,K} **do**
T1[i,1] ← πi·Biy$_1$
T2[i,1] ← 0
end For
for each observation
i={2, 3,… ,T} **do**
for each state
j ϵ {1, 2,… ,K} **do**
T1[j,i] ← max k(T1[k,i-1]. Akj . Bjyi
T2[j,i] ← arg maxk (T1[k,i-1] . Akj . Bjyi)
end for
end for
zT ← argmaxk (T1[k,T])
$x_T \leftarrow s_{zT}$
For i ← T,T-1,...,2 **do**
$z_{i-1} \leftarrow$ T2[zi,i]
$x_{i-1} \leftarrow s_{zi-1}$
end for
return X
end function

E. *Text Extraction (Optical Character Recognition)*

We extracted text from images as well as printed text using Optical Character Recognition (OCR) as illustrated in Fig. 7. In our model, the basic 2D correlation coefficient is determined and classified according to the coefficient obtained. This pre-defined function returns the correlation coefficient between two matrices or vectors of the same size. The coefficient so obtained is compared with every character's position matrix in the library. Once the position matrix is matched with the coefficient, the character is detected and thus finally, output is displayed in the text format in the window of the Matlab Interface.

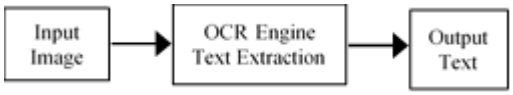

Fig. 7. Block diagram for Text extraction

F. Text Translation (Machine Translation)

In Text translation or Language Translation module, we take in Text from the source language as input and convert it into the target language (see Fig. 8). The text was matched and aligned with a text editor dictionary before conversion into the target language took place. In this system, we used Statistical Machine Translation, where we found the maximum probability of the target sentence by using the language and translation model. Now, we consider e to be the English sentence and h to be the possible Hausa sentence, the most likely Hausa translation is thus:

$$\hat{e} = \arg\max_e P(h|e) \quad (2)$$

Here, $P(h|e)$ is the probability that given an English sentence *e*, a Hausa sentence *h*, was produced from it. Now, By Bayes theorem, the probability, can be rewritten as,

$$P(h|e) = \frac{P(h)P(e|h)}{P(e)} \quad (3)$$

Therefore, equation (2) is

$$\hat{e} = \arg\max_e P(h|e) = \frac{P(h)P(e|h)}{P(e)} \quad (4)$$

Because, $P(e)$ is independent here, equation (4) can be simplified to obtain the equation,

$$\hat{e} = \arg\max_e P(h)P(e|h) \quad (5)$$

Therefore, for translation of the English sentence into Hausa sentence three components are used:

(1) Language Model: The kind of sentences that are likely to appear in the language 'h'. This is known as the language model $P(h)$.

(2) Translation Model: The way sentences in *e* are converted to sentences into *h*. This is called the translation model $P(e|h)$.

(3) Decoding: The decoding step find the best translation by picking up the translation sequence that gives the highest probability. The decoder generates the best output translation by finding the translation sequence with the highest probability among all the sequences

We computed the Probability of the occurrence of a word given another word $P(w_1|w_2)$ in the language model as follows: e.g. the weather.

$$P(w_1|w_2) = \frac{(count1, count2)}{count1} \quad (6)$$

Meaning that,

$$P(the|weather) = \frac{P(the|weather)}{P(the)}$$

Similarly,

$$P(w_1|w_2|w_3) = \frac{(count1, count2, count3)}{(count1, count2)} \quad (7)$$

Hence,

$$P(w_1|w_2|w_3|..|w_n) = \frac{(count1, count2, count3, ..., countn)}{(count1, count2, ..., countn-1)} \quad (8)$$

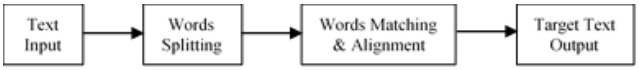

Fig..8. Block diagram for text translation

G. Speech Synthesis

Speech synthesis is the artificial production of human speech; a system used for this purpose is called a speech synthesizer. We created the synthesized speech by concatenating pieces of recorded speech that are stored in the database.

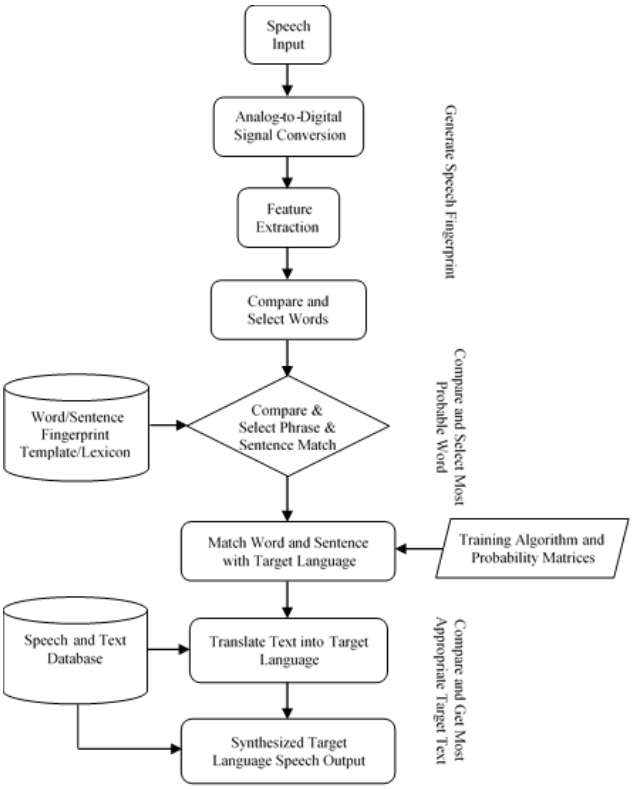

Fig. 9. Basic structure of the Proposed System.

The acoustic speech signal input is converted from analog to digital for feature extraction using the MFCC extraction technique for our word model to generate the source language text, which are then compare with the target language words from our corpus database and matched together (see Fig. 9). The target language text/sentence generated are then matched with the appropriate speech from the speech database.

IV. Results and Discussion

We experimented by pronouncing the word "crops" at a sample rate of 44100 with 16 bits per sample to obtain the waveform in Fig. 10 and generate our model thus:

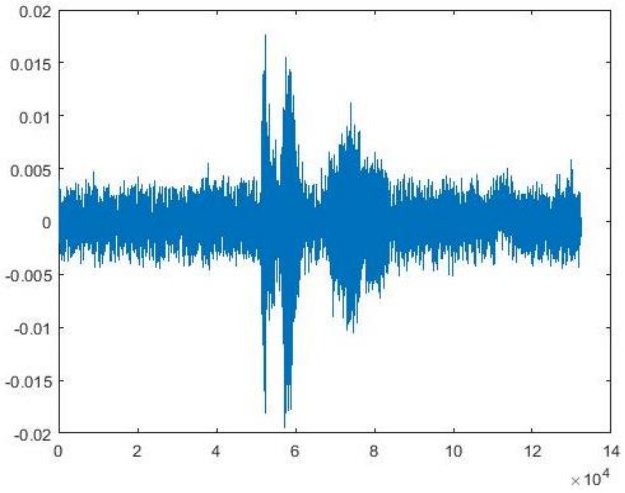

Fig. 10. Waveform for pronunciation of the word "crops"

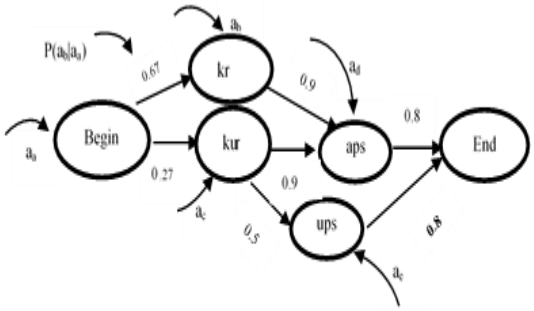

Fig. 11. Hidden markov word model for "crops"

After pronouncing the word "crops" 30 times, we arrived at the following probabilities. While pronouncing the word, the phonemes produce are 'kr_aps', 'kur_aps', and 'kur_ups' (see Fig. 11). We used probabilities from the word model to find the most appropriate pronunciation.

We represented the phone states as q_i s. P(b|w), which is the Probability of the word given the observation and it is broken into P(b|kur_aps), P(b|kr_aps and P(b|kur_ups).

$P(b|w_i)$ = P(b| Begin_kur_aps_ End) + P(b| Begin_kr_aps_ End) + P(b|kur_ups_End).

$P(b|w_{crops})$=0.27x0.9x0.8+0.67x0.9x0.8+0.27x0.5x0.8

$P(b|w_{crops})$ = 0.1944+0.4824+0.108

$P(b|w_{crops})$ =0.7848

TABLE I. PROBABILITIES

	Frequency of occurrence	Conditional Probabilities
No. of time 'kur_aps'	8(0.27 approx.)	$P(a_c\|a_a)$= 0.27
No. of time 'kr_aps'	20(0.67approx.)	$P(a_b\|a_a)$ = 0.67
No. of time 'kur_ups'	2(0.06 approx.)	$P(a_c\_a_e\|a_a)$ = 0.06
Probability of observed sound at 'b'		$P(b_0\|a_b)$ = 0.9
Probability of observed sound at 'c'		$P(b_0\|a_c)$ = 0.9
Probability of observed sound at 'd'		$P(b_1\|a_d)$ = 0.8
Probability of observed sound at 'e'		$P(b_2\|a_e)$ = 0.5

The individual $P(b_i|a_i)$ are estimated using a Gaussian likelihood estimator. The transitions $P(a_j|a_i)$ depend on the pronunciations of the words. For each word *n* we built a HMM model and estimated the model parameters that optimize the likelihood of the training set observation vectors from the nth word.

Below are our translation experimental results for language translation in the three languages:

We obtain our translation results at:

Sampling rate (Fs) = 44100

TABLE II. HAUSA

English	Hausa
The weather affected our crop yield.	Yanayin ya shafi yawan amfanin gonar mu.
The government is providing fertilizer for farmers this year.	Gwamnatin ta samar da taki ga manoma a wannan shekara.
The flood destroy our crops.	Ruwan tsufana ya rushe amfanin gona.

TABLE III. IGBO

English	Igbo
The weather affected our crop yield	Ihu igwe na-emtuta ihe ubi anyi
The government is providing fertilizer for farmers this year.	Goomenti na-enye ndi oro ugbu nri maka afo a
The flood destroy our crops.	Idei mmiri ahu na –ebibi ihe ubi anyi

TABLE IV. YORUBA

English	Yoruba
The weather affected our crop yield	Oju ojo naa pa ikore irugbin wa
The government is providing fertilizer for farmers this year.	Ijoba n pese awon ohun elo fun awon agbe ni odun yii
The flood destroy our crops.	Ikuno mi subu awon irugbin wa

For text extraction from our printed text image with the Matlab OCR we obtain the following for the sentence "the weather affected our crop yield" in image format.

Fig. 12. Printed text Image

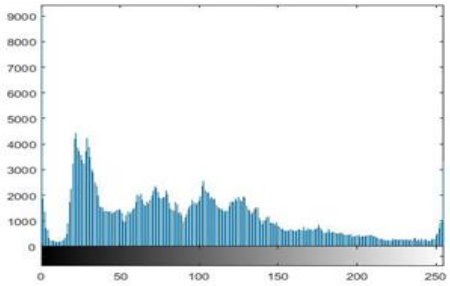

Fig. 13. Image histogram

TABLE V. IMAGE PIXELS EXTRACTS

255	254	254	254	254	253	252
254	255	255	255	255	255	254
252	255	254	255	255	255	254
248	253	253	254	254	254	254
248	251	253	253	253	253	254
246	250	252	248	252	252	254
242	247	248	243	249	249	248

TABLE VI. EXTRACTED TEXT PROPERTIES

Character Text Bounding	Character Confidence	Word bounding boxes	Word confidences
1	0.8050	74	0.4106
41	0.7369	143	0.7266
119	0.4106	229	0.8471
211	NaN	24	0.7728
211	0.8120	24	0.7865
74	0.8795	81	0.8836
86	0.9618	133	0.8207
101	0.7266	75	0.8677
120	0.9758	75	0.9499
129	0.8945	75	0.9499

Our character and word confidence values gives us an accurate confidence values on the text extracted which makes our translation successfully.

In carrying out the synthesis of the translated texts in the three languages we obtain the following results in our synthesis experiments with the following parameters:
Time = 3 seconds
Sampling rate=44100
Time resolution in millisecond = 500ms
Frequency resolution = 16Hz

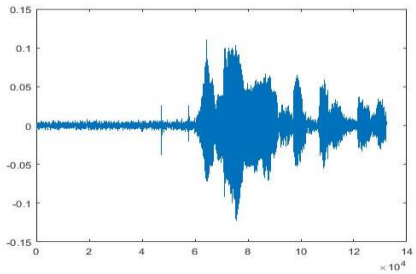

Fig.14. Synthesized Speech waveform for the Hausa sentence 'Yanyin ya shafe amfanin gonar mu'

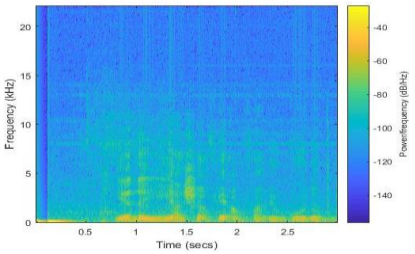

Fig.15. Synthesized Speech Spectrogram for the Hausa sentence 'Yanyin ya shafe amfanin gonar mu'

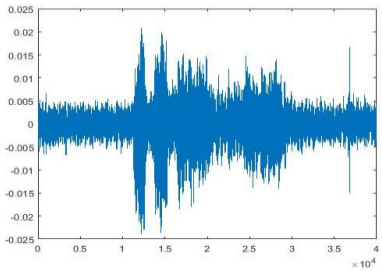

Fig.16. Synthesized Speech waveform for the Igbo sentence 'Ihu igwe na-emtuta ihe ubi anyi u'

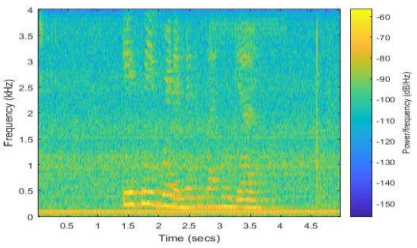

Fig.17. Synthesized Speech Spectrogram for the Igbo sentence 'Ihu igwe na-emtuta ihe ubi anyi u'

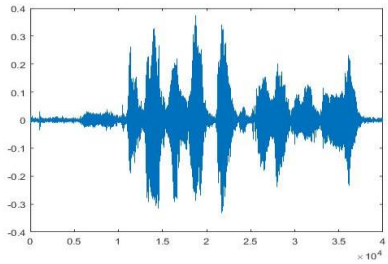

Fig.18. Synthesized Speech waveform for theYoruba sentence 'Oju ojo naa pa ikore irugbin wa'

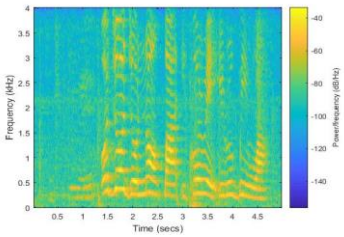

Fig.19. Synthesized Speech Spectrogram for the Yoruba sentence 'Oju ojo naa pa ikore irugbin wa'

V. CONCLUSION

This research work designs a framework and a system for Multiple – choice Multilingual Translation that ensures that information gets to individuals irrespective of their geographical location and their abilities/ literacy level. The proposed system translates information from a source language (English) to the target native language (Hausa, Igbo and Yoruba) depending on which language the local user can read or can understand. The proposed system will also be highly efficient for Agricultural Extension workers, Community health workers and also be of great help to individuals with different disabilities as it carries out both Speech and text translation.

In addition, the proposed system employs prosody in its translation so as to make the translation easily understandable with its naturalness and emotions. This will enable efficient and effective communication amongst users.

REFERENCES

[1] S. Bhardwaj, S. Pathania and R. Akela, "Speech Recognition using Hidden Markov Model and Viterbi Algorithm," in International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE), vol. 4, Issue 5, May 2015.

[2] V. Sridhar , S. Bangalore and S. Narayanan, " Enriching machine-mediated speech-to-speech translation using contextual information". www.sciencedirect.com doi:10.1016/j.csl.2011.08.00.

[3] S. Hla and M. Tun, "Speech-To-Text Conversion (STT) System Using Hidden Markov Model (HMM)" in International Journal of Scientific & Technology Research, vol 4, issue 06, June 2015.

[4] I. Isewon, J. Oyelade and O.Oladipupo, "Design and Implementation of Text To Speech Conversion for Visually Impaired People". International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA vol. 7, No. 2, April 2014 – www.ijais.org.

[5] R. Debnath, V. Hanumante, D. Bhattacharjee, D. Tripathi and S. Roy "Multilingual Speech Translator using MATLAB", in International Conference on Electrical, Electronics, Signals, Communication and Optimization (EESCO) – 2015.

[6] S. Nakamura, K. Markov, H. Nakaiwa and G. Kikui, "The ATR Multilingual Speech-to-Speech Translation System", IEEE Transactions on Audio, Speech, and Language processing, Vol. 14, No. 2, March 2006.

[7] G.P.S. Prasanthi , K. Sirisha , G. Ramya and B. Padma, "Speech to Text Conversion Using HMM" in International Journal of Advanced Research in Electronics and Communication Engineering (IJARECE) vol. 5, Issue 3, March 2016.

[8] R. Borgohain, M. Singh and S. Gohain, "An English-Assamese Machine Translation System". International Journal of Computer Applications (0975 8887), vol. 93 - No. 4, May 2014.

[9] Darshna Badhe and P. M. Ghate, " Marathi Text to Speech Synthesis – Using Matlab". International Journal of Computer Science and Network (IJCSN), vol. 4, Issue 4, August 2015.

[10] M. Ryan and G. Nudd, "The Viterbi Algorithm". Department of Computer Science, University of Warwick, Coventry, CV4 7AL, England.

[11] N. Dave, "Feature Extraction Methods LPC, PLP and MFCC In Speech Recognition" in International Journal for Advance Research in Engineering and Technology, vol 1, Issue VI, July 2013.

[12] M.Anusuyan and S.Katti, "Speech Recognition by Machine: A Review" in International Journal of Computer Science and Information Security (IJCSIS), vol. 6, No. 3, 2009.

[13] H. Krupakar, K. Rajvel, B. Bharathi, A.Deborah and V. Krishnamurthy, " A Survey of voice Translation Methodologies - acoustic dialect decoder", in International Conference On Information Communication & Embedded Systems (ICICES-2016).

[14] A. Supatkar and S. Mali, "Implementation of Feature Extraction Algorithm of Speech Signal in FPGA" in International Research Journal of Engineering and Technology (IRJET), vol. 3 Issue 8 | Aug-2016.

[15] K. Arora, S. Arora and M. Roy, "Speech to speech translation: a communication boon". CSIT, vol. 1, No. 3, pp 207–213, September 2013, DOI 10.1007/s40012-013-0014-4.