A COMPARATIVE SURVEY OF DTW AND HMM USING HAUSA ISOLATED DIGITS RECOGNITION IN HUMAN COMPUTER INTERACTION SYSTEM

Yakubu A. Ibrahim [1], Tunji S. Ibiyemi [2]

[1] **Department of Computer Science, Bingham University, Karu, Nasarawa State, Nigeria**
[2] **Department of Electrical Engineering, University of Ilorin, Ilorin, Nigeria**

Corresponding Author: Yakubu A. Ibrahim, talktoibro80@gmail.com

ABSTRACT: Speech Recognition is a vital part of different computer-based applications in communication and security systems. However, there has been very little research in the aspect of speech Human Computer Interaction system for African languages such as Hausa, hence, the need to extend the research in order to bring in, the different systems based on speech recognition. Also, Hausa is an important ethnic tribe of lingua franca in both west and central Africa countries. Isolated word recognition is an easy speech type because it demands the user to pause between each word. In this study, the two algorithms that were used to implement a system of Recognition of Hausa isolated digits are Dynamic Time Warping and Hidden Markov Model. To perform the recognition efficiently, speech endpoint, framing blocking, speech normalization, vector quantization and Mel Frequency Cepstral Coefficient techniques were used to process the speech. The accuracy of about 94% was obtained for recognition with HMM-based system. In a very noisy environment, the performance of the two techniques is bad but the pattern matching using HMM is better than the pattern matching using DTW.

KEYWORDS: DTW, Hausa Language, HMM, Speech Recognition, MFCC.

I. INTRODUCTION

Nowadays, people most a times needed to reduce the rate at which efforts are made by humans to get work done. A suitable and effective interface for Human Computer Interaction is a vital technology issue. The commonly used human-computer interface is keyboard or a mouse device for input and a visual display unit, a speaker or a printer for output. Human communication is dominated by speech and hence people expect human computer interaction to be via speech interface ([Hol88]). Automatic speech recognition technology is adopted for several computer-based technologies by many individuals such as scientists, doctors, lawyers, teachers and students. ASR technology therefore, allows human speech to be used through human-computer interaction to carry out certain computer based activities. Speech recognition system can only perform a predefined task if and only if the system notices and recognizes a sound or string of sounds ([BN96]). Speech produces continuous sound pressure waves of diverse frequencies and amplitudes. Speech recognition can only occur when a corresponding sequence of discrete units of words or sentences are derived from sound waves ([Moo94]). The purpose of most computer-based speech recognition systems is to model perfect human speech recognition. However, computer speech based systems applications do not have the capability and flexibility of understanding speech in communication as humans do.

II. SPEECH RECOGNITION

Speech based systems essentially entails extraction of features and classification of a audio speech. The measurements that could be performed on the speech waveform include energy, zero crossings, extrema count, formants, LPC cepstrum ([SR75], [Toh86]) and the Mel Frequency Cepstrum Coefficient (MFCC) ([R+79]). The LPC method provides a dependable and accurate method for estimating the parameters that characterize the linear, time-varying, system which is recently used to approximate the nonlinear, time-varying system of the speech waveform. The MFCC method uses the bank of filters scaled according to the Mel scale to smooth the spectrum, performing a processing that is similar to that executed by the human ear. As to classification of an input speech, the most successful speech recognition methods are the pattern matching using Dynamic Time Warping (DTW) ([Sha60]), Vector Quantization (VQ) ([J+82]) and Hidden Markov Model (HMM) ([B+70], [Lee90]). Since the same word uttered by the same person may have different duration of the same phoneme, the DTW process nonlinearly expands or contracts the time axis to match the same phoneme or landmark positions between the input speech signal and reference templates signal. Vector quantization is a process of mapping vectors from a large vector space to a finite number of regions in that space. Each region in the space is called a cluster and can be represented by its center called a codeword. The collection of all code words is called a codebook. The VQ is a data compression principle introduced by Shannon ([J+82]). When VQ is applied to speech compression, a training sequence is used to produce a set of reproduction vectors (codeword), called the codebook of the speech. Moreover, the selection of a perceptually meaningful distortion measure in clustering and the construction of an optimal codebook are difficult. It is also not always easy to apply the VQ to a large vocabulary because it has high computational cost in clustering ([MJA06]).

ASR can also be seen as successive transformations of the acoustic micro-structure of the speech signal into its implicit phonetic macro-structure. The major purpose of any ASR system is to realize the mapping between the two given structures. HMM is most used approach to the ASR systems. Many types of HMMs have been developed and applied to the ASR systems. In this vain, HMM is pleasing in terms of algorithmic complexity that is why it is has been examined in different studies ([AFH02]). As a matter of fact, HMM method has significantly minimized the computational cost and has been adopted for large vocabulary of isolated and continuous speech recognition applications ([E+03]). The existing speech recognition system techniques were widely adopted for certain international languages such as English or Japanese. For African languages such as Hausa language, research efforts remain limited.

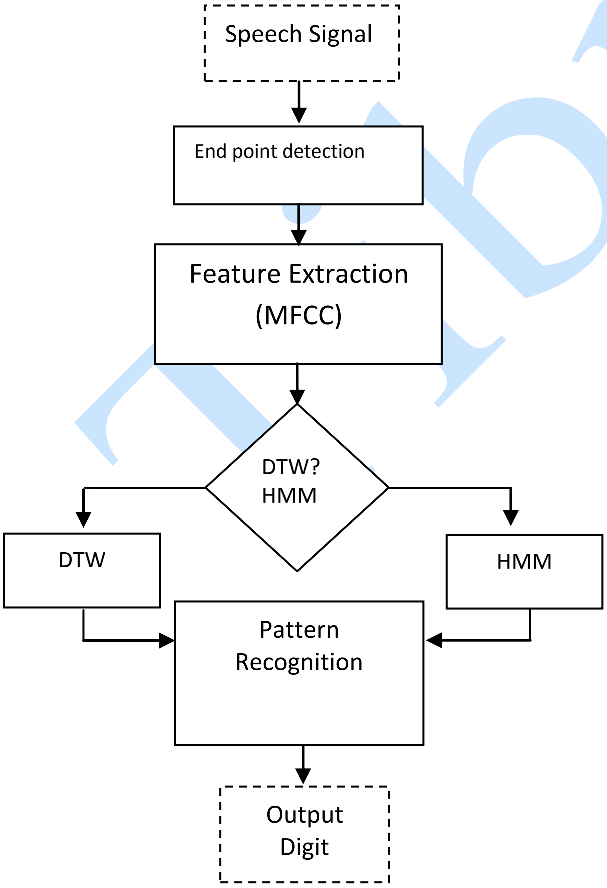

Figure 1: Block diagram of hausa digit speech recognition using dtw and hmm

This paper shows development of an experimental isolated word recognizer for isolated digits recognition system in Hausa Language. The study is further extended to benchmark of speech recognition system for small vocabulary of speaker dependent isolated spoken words using the HMM and DTW method (Figure 1). The presented work in this study focuses on template-based recognizer approach using MFCC with dynamic programming computation and vector quantization with Hidden Markov Model based recognizers in isolated word recognition tasks, which also significantly reduces the computational costs. The analysis, design and development of the two automation systems are done in MATLAB 8.5, using ([TA14], [SC78])

III. HAUSA LANGUAGE

Hausa is known to belong to a member of the Chadic language family, which puts it with the Semitic and Cushitic languages in the Afro-asiatic language stock. In population, over 25 million speakers, Hausa is widely spoken in West Africa ([Bur92]). It is the most frequently used language for communication in the Northern part of Nigeria and as well as many parts of Western Africa. However, Hausa is spoken by over 23 million people as a first language and by over 5.5 million as a second language. It is the major language spoken in Northern Nigerian states such as Kano, Kaduna, Katsina, Zaria, Katsina, Bauchi, Jigawa, Zamfara, Kebbi, Gombe, and Sokoto, to name only a few. Frankly speaking, Hausa is also spoken in Benin, Burkina Faso, Congo, Cameroon, Central African Republic, Chad, Eritrea, Ghana, Niger, Sudan, and Togo ([Bur92]).

Table 1: Hausa Digits with English Digits Equivalent

Digits	Hausa Word of Digits	English Meaning of the Digits
0	Sifiri	Zero
1	Ɗaya	One
2	Biyu	Two
3	Uku	Three
4	Hudu	Four
5	Biyar	Five
6	Shida	Six
7	Bakwai	Seven
8	Takwas	Eight
9	Tara	Nine

In other words, Hausa is a tonal language. This means that word meanings can change according to pitch differences in syllables. Tone is indicated, in the written script, with accent marks placed on vowels. The cities of northern region - Kano, Sokoto, Zaria, and Katsina, to name only a few, are among the largest commercial centers of sub-Saharan Africa. Hausa people also live in other countries of West Africa like Cameroon, Togo, Chad, Benin, Burkina Faso, and Ghana. About one-fourth of Hausa words come from Arabic words ([Els90]). Hausa's modern official orthography is a Latin-based alphabet called boko, which was imposed in the 1930s by the British colonial administration. Hausa consists of 22 characters of the English alphabet *(A/a, B/b, C/c, D/d, E/e, F/f, G/g, H/h, I/i, J/j, K/k, L/l, M/m, N/n, O/o, R/r, S/s, T/t, U/u, W/w, Y/y, Z/z) plus Ɓ/ɓ, Ɗ/ɗ, Ƙ/ƙ, Ƴ/ŷ called "hook letters", " ' "called "a glottal stop " and three digraphs which are kw, sh, and ts* . Hausa has five vowel alphabets: *a, e, i, o, u*.

There are three basic tones in Hausa, namely: low tone, high tone and mid/falling tone. Each of the five vowels /a/, /e/, /i/, /o/, and /u/ may have low tone, high tone, or mid/falling tone ([Bur92]). Additionally, it is distinguished between short and long vowels which can also affect word meaning ([Ipa99]). Neither the vowel lengths nor the tones are marked in standard written Hausa.

IV. DYNAMIC TIME WARPING

To deal with the alignment work from an algorithmic point of view, we need algorithms that are capable of quantifying similarity or dissimilarity in an optimal way, by taking into account the fact that parts of one sequence may need to be locally extended or compressed, so that the optimal alignment is achieved ([RS75]). The operations of extending and compressing speech parts are collectively known as *time warping*. Due to the fact that time warping is determined dynamically during the joint inspection of two feature sequences, the term *Dynamic Time Warping* is often used to denote the alignment of two feature sequences that represent speech signals ([TA14]). To move further, certain definitions must first be given. Let **X** = {*x(i)*; *i* = 1, . . . , T_x} and {**R** = *r(j)*; *j* = 1, . . . , T_r }, be the two feature sequences to be aligned. In general, T_x is not equal to T_r and *x(i)* and *r(j)* are *K*-dimensional feature vectors. Also, let *d(a, b)* be a dissimilarity function, e.g. the Euclidean distance function, which is defined as Dynamic:

$$d(a,b) = \sqrt{\sum_{l=1}^{K}(a_l - b_l)} \qquad (1)$$

The DTW algorithms begin with the building of a *cost grid*, *C*. Assuming that sequences **R** and **X** are positioned on the horizontal and vertical axis, respectively, the coordinates of node *(i, j)* are the *i*th feature vector of **X** and the *j*th feature vector of **R**. Therefore, a dynamic programming algorithm method is employed. The algorithm begins from the first node and the nodes of the grid are visited row-wise from left to right (an equivalent column-wise visiting scheme can also be used). At every node, the accumulated cost to reach the node is computed. At the very end of this processing stage, the cost at node *(*T_x*,* T_r *)* stands for the total matching cost between the two feature sequences. A zero cost shows a best match. The higher the matching cost, the more different (less similar) the two sequences.Time Warping algorithm is adopted to recognize an isolated word sample by comparing it against a number of stored word templates to determine the one that perfectly matches it. This aim is complicated by a number of decisions. One, different samples of a given word will have somewhat different durations. The issue can be eliminated by simply normalizing the templates and the unknown speech so that they all have an equal duration. Dynamic Time Warping (DTW) is a good method for finding optimal nonlinear alignment between a template and the speech sample. The major issue of systems based on DTW is the small amount of learning words, high computation rate and large memory requirements ([B+07]).

V. VQ/HMM SYSTEM

A Hidden Markov Model is a Model in which the system being modelled is assumed to be a Markov process with unidentified or hidden states ([HC95]). The issue is to find all the appropriate hidden parameters from the observable states. In speech recognition, the use of HMM is subject to the following constraint: (1) must be based on a first order Markov chain; (2) must have stationary states transitions; (3) observations-independence and (4) probability constraints. In other words, HMM is a collection of states connected by transitions. Each transition carries two sets of probabilities ([RG03]):

i. A transition probability which provides the probability for taking a transition from one state to the next, and
ii. An output probability density function (*pdf*), which defines the conditional probability of emitting each output symbol from a finite alphabet given that that transition is taken.

HMMs have become a widely-adopted approach for speech recognition due to the existence of maximum likelihood methods to estimate the parameters of the models and algorithms that perfectly find the most likely state sequence ([S+94]).

Also, *HMM is* a stochastic automaton that undergoes distinct transitions among states based on a set of probabilities. When HMM gets to a state, it emits an observation, can be a multidimensional element (feature vector of discrete or continuous elements. The emission of observations is always governed by probabilistic rules. During the design phase of an HMM it is of necessary importance to decide on the number of states and assign a meaningful stochastic interpretation to every state, so that the nature of the issue is reflected successfully on the structure of the HMM.

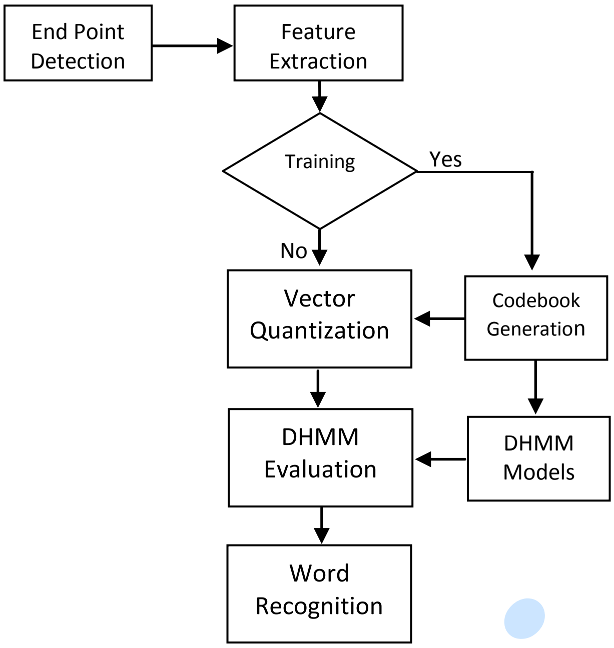

Figure 2: Building Blocks for Teaching Module Speech Recognition

Definitions

Let S_i, $i = 1, . . . ,N$ be the set of HMM states and b_j, $j = 1, . . . ,M$ the discrete symbols of the alphabet of observations. The following three matrices stand for the HMM:

i. The vector of *initial state probabilities*, π. Element, $\pi(i)$, $1, . . . ,N$ is the probability that the HMM starts its operation at the i^{th} state.

ii. The *state transition matrix*, A. define that $A(i, j) = p(j \mid i)$, i.e. $A(i, j)$ is the probability that the model jumps from state i to state j. Obviously, A is a $N \times N$ matrix and

$$\sum_{j=1}^{N} A\,(i,j) = 1, \forall i \qquad (2)$$

iii. The matrix of *emission probabilities*, B. We define that $B(m, i) = p(b_m \mid i)$. This means that $B(m, i)$ is the probability that the m^{th} symbol of the alphabet is emitted from the i^{th} state. It follows that B is a $M \times N$ matrix and that

$$\sum_{m=1}^{M} B\,(m,j) = 1, \forall i \qquad (3)$$

Vector Quantization ([HC95]) is always applied to ASR. VQ is useful for speech coders, that is, efficient and perfect data reduction. Since transmission rate in ASR is not a major issue, the utility of VQ now depends on the efficiency of compact codebooks for reference models and codebook searcher in place of more costly evaluation methods. To demonstrate an application of HMMs for speech recognition, the implementation for isolated word digits recognition system on Hidden Markov Models is showed in figure 2. If have a vocabulary of L words to be recognized, and each word is to be modelled by a distinct HMM. The training sets will consist of K utterances of each word. To obtain a word recognizer, the following steps are performed.

A. VQ Codebook

In building a good HMM system, the continuous feature vector space is subdivided by a vector quantized into non overlapping subsets and each subset is represented with a codeword. The set of available code words is called the codebook. The VQ codebook is built by an unsupervised cluster algorithm.

B. Re–Estimation of HMM

For every word of the vocabulary, HMM is constructed, that is, the model parameters is approximated to optimize the likelihood for the training set of observation sequences. There are different criteria that can be adopted for this problem. For this issue the Baum-Welch algorithm ([TA14]) developed by Baum which is one of the most successful optimization methods was used.

C. Recognition

For every unknown word to be recognized by the system, the likelihood models is calculated for all achievable models, and selected the models with the maximum likelihood. The probability calculation was performed using Viterbi algorithm, accurately the logarithm of the highest likelihood ([Moo94]).

VI. EXPERIMENTAL SETUP

A. Database preparation

A corpus was created from all ten (10) Hausa digits (0-9). Five (5) Hausa speakers (three (3) males and two (2) females) were asked to utter all digits 10 times each. At the recording session, every utterance was played back to ascertain that the entire digit was included in the recorded speech signal. Speech recognition performance depends on accurate endpoint detection, for all experiments; each speech file from the database was analyzed by an endpoint detection program in order to locate more accurate endpoints ([MF09]).

B. End Point Detection

All speech recognition system contains a speech or non-speech detection segments. This issue is often referred to as the endpoint location problem. The method adopted in this work uses two methods of the speech signal: the Zero Crossing Rate (ZCR) and the Energy. This algorithm was performed in MATLAB 8.5 ([TA14]), and applied to the corpus.

C. Feature Extraction

Speech signal, was sample at 10 KHz, is pre-emphasized by a first-order digital filter in order to spectrally flatten the speech signal.

$$\hat{S} = S(n) - \mu S(n-1) \qquad (4)$$

With µ=0.96. The signal is segmented into frames by using a 25.6ms Hamming window with 10 ms shifting. For every frame, the MFCCs, their corresponding first and second directives, are computed ([MF09]). Each frame is thus represented by an acoustic vector Xt as follows:

$$Xt = \{MFCC\ (m), \Delta MFCC\ (m), \Delta MFCC\ (m)\} \qquad (5)$$

The first and second order derivatives of cepstral coefficient were estimated respectively by equations (2) and (3) given as follows:

$$\Delta MFCC\, l(m) = \sum_{k=-k}^{k} k\big(MFCC\, l - k(m)\big) \qquad (6)$$

$$\Delta\Delta MFCC\, l(m) = [\Delta MFCC\, l + 1(m) - \Delta MFCC\, l - 1(m)] \qquad (7)$$

Where *k* and *l* frame indexes, *m* the MFCC component

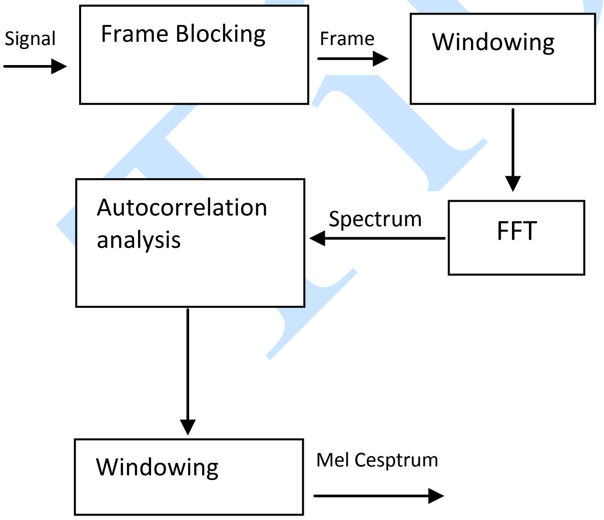

Figure 3: Block Diagram of the MFFC Processor

The modelling of the glottal flow is a difficult issue and few studies attempt to accurately decouple the source tract components of the speech signal. Standard feature extraction method MFCC simply ignores the pitch component and roughly compensates the spectral tilt by applying a pre-emphasis filter prior to spectral analysis ([H+93]). The speech samples (8 kHz sampling rate) are windowed into overlapping 25 ms frames with a frame shift of 10 ms. A 256 point FFT was used to compute the power spectrum that is used in an emulated filter-bank composed of 24 triangular weighting functions on a Mel scale. The natural logarithm is then applied to the 24 filter-bank energies. Mel spectrum coefficients are also real numbers; they can be converted to time domain using the Discrete Cosine Transform (DCT). In this paper, a HMM-based speech recognizer is used for the recognition of isolated digits. Also, HMM for every word have five traversable states. Transitions between states are permitted only in left-to-right direction with no jumping of states.

VII. RESULTS

A. DTW recognition

To search for efficient signal characteristics, first, add the log Mel cepstrum energy coefficient to each frame. Thereafter, obtain 13 coefficients per frame. To give correct account for the temporal properties, these 13 coefficients are derived twice to obtain respectively the vectors ΔMFCC and ΔΔMFCC. The recognition percentage is then calculated for each case in figure 4.

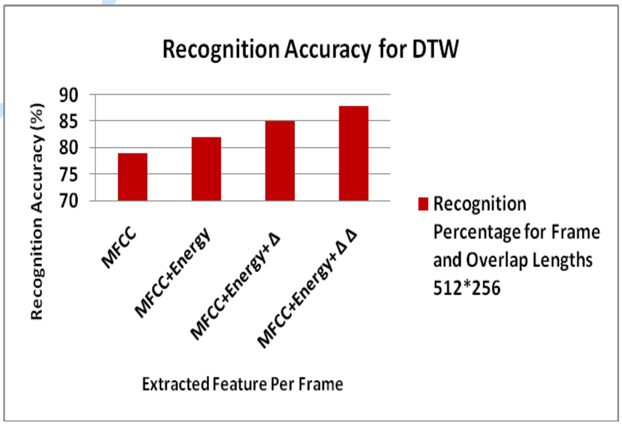

Figure 4: Recognition percentage for different extracted features using DTW

The obtained results in figure 4 present that the extracted features "MFCC + Energy + Δ + ΔΔ" per frame offer the best system performances. In this case, the recognition percentage gets to 88%.

B. Effects of Vector Quantization Codebook Size (HMM)

Size of speaker codebook differs from 8 to 128. The results are shown in figure 5. The calculated Word Error Rate (WER) decreases when the codebook size increases. In other words, this can be explained as the codebook size increases, the distortion (quantization) error decreases.

The size of the codebook of more than 256 can perform better than the size of the codebook 128, but the average time to recognize one speaker is higher compared to codebook size of 128. This can affect the response time of identification system. Hence, codebook size of 128 was used throughout the experiment.

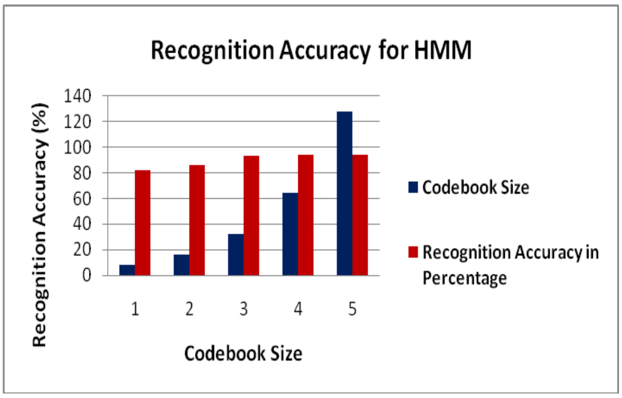

Figure 5: Recognition accuracy for varying codebook size

C. Effects of additive noise (HMM and DTW)

To study the effect of noise on the two ASR systems developed; Gaussian noise was added to the original speech signals.

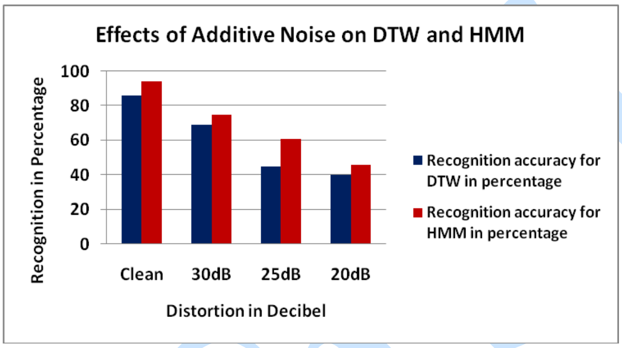

Figure 6: Effect of additive noise distortion on the recognition performance

Figure 6 shows the comparison of Hausa digits recognition accuracy after the noise with various signals to noise ratios (SNRs). SNR is the ratio of the power of the correct signal to the noise and it is measured in decibel (dB). The results of the study show that the recognition performances for the two ASR are worse with the noise but the pattern recognition using HMM is better than the pattern using DTW in noisy conditions.

VIII. CONCLUSION

Speech utterance is the effective and the most convenient means of communication between humans. For reasons starting from curiosity in research to build machines that mimic humans to the desire for automation of simples tasks is not a modern phenomenon, but one that went back more than one hundred years in history. ASR has attracted scientists as an important discipline and has created a technological impact on society and is expected to flourish further in this area of human computer interaction. It is hope this paper brings about understanding and inspiration amongst the research communities of ASR. In this study, DTW and HMM based isolated Hausa digits systems recognition was designed and evaluated.

DTW system recognition that adopted the MFCC coefficients as basic characteristics vector leads to recognition accuracy of 79%. This recognition accuracy can as well achieve 88% by using additional characteristics as power information and differential information (Δ and ΔΔ). HMM system recognition with codebook size 128 shows interesting recognition accuracy measurement of about 94%. In a very noisy environment, the recognition performances for the two ASR systems are worse but the pattern recognition using HMM is better than the pattern using DTW. The performance of the two methods DTW and HMM based isolated Hausa digits for automatic systems recognition can be increased with further research by using a larger corpus of Hausa language or using any of the other related African languages.

REFERENCES

[AFH02] **W. Alkhaldi, W. Fakhr, N. Hamdy** - *Multi-Band Based Recognition of Spoken Arabic Numerals Using Wavelet Transform*, The 19th National Radio Science Conference Alexandria, March 19-21. 2002.

[Bur92] **D. A. Burquest** - *An Introduction to the Use of Aspect in Hausa Narrative*, Language in context: Essays for Robert E. Longacre, Shin Ja J. Hwang and William R. Merrifield (eds.), 1992.

[BN96] **C. Barber, J. Noyes** - *Automatic speech recognition in adverse environments. Human Factors*, 38, 142-156, 1996.

[B+07] **M. Benzeghiba, R. de Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, C. Wellekens** - *Automatic speech recognition and speech variability: A review*, Speech Communication 49 763–786, 2007.

[B+70] **L. E. Baum, T. Petrie, G. R. Soules, N. Weiss** - *A maximization technique* *occurring in the statistical analysis of probabilistic functions of Markov chains*, Ann. Math. Statist. 41 (1) 164–171, 1970.

[Els90] **M. Elshafei** - *Toward an Arabic Text-to-Speech System*, The Arabian Journal for Science and Engineering, Vol. No. 16, Issue No. 4B, pp. 565-83, 1990.

[E+03] **F. A. Elmisery, A. H. Khalil, A. E. Salama, H. F. Hammed** - *A FPGA Based HMM for a Discrete Arabic Speech Recognition System*, ICM, Cairo, Egypt, 2003.

[Hol88] **J. N. Holmes** - *Speech Synthesis and Recognition*, Van Nostrand Reinfold (UK) Co, Ltd, pp.102-113, 1988.

[HC95] **Q. Huo, C. Chan** - *Contextual Vector Quantization for Speech Recognition with Discrete Hidden Markov Model*." Pattern Recognition, vol. 28, no. 4, pp. 513-517, 1995.

[H+93] **X. D. Huang, H. Hon, M. Hwang, K. Lee** - *A comparative study of discrete, semi Continuous and Continuous Hidden Markov Models*, Computer Speech and Language, voil 7 pp. 359-368, 1993.

[J+82] **B. H. Juang, D. Y. Wong, A. H. Gray** - *Distortion performance of vector quantization for LPC voice coding*, IEEE Trans. Acoust. Speech Signal Process. 30 (2) (1982) 294–303, 1982.

[Lee90] **K. F. Lee** - *Context-dependent phonetic hidden Markov models for speaker independent continuous speech recognition*, IEEE Trans. Acoust. Speech Signal Process. ASSP-38 (4) 599–609, 1990.

[Moo94a] **D. W. Moore** - *Automatic speech recognition for electronic warfare verbal reports"*, Unpublished master's thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA, 1994.

[Moo94b] **R. K. Moore** - *Twenty things we still don't know about speech*, Proc.CRIM/ FORWISS Workshop on Progress and Prospects of speech Research and Technology, 1994.

[MF09] **M. P. Meysam, F. Fardad** - *An Advanced Method for Speech Recognition.* World Academy of Science and Technology, pp. 995-1000, 2009.

[MJA06] **D. Mohamed, P. H. Jean, H. Amrane** - *A Vector Quantization Approach for Discrete Speech Recognition System*, International Journal of Computing. Vol. 5, Issue 1, pp 72-78, 2006.

[RG03] **E. Robert, A. Granat** - *A Method of Hidden Markov Model Optimization for Use with Geophysical Data Sets*, International Conference on Computational Science, pp.892-901, 2003.

[RS75] **L. R. Rabiner, M. R. Sambur** - *An Algorithm for Determining the Endpoints of Isolated Utterances*, The Bell System Technical Journal, Vol. 54, No. 2, February 1975, pp. 297-315, 1975.

[R+79] **L. R. Rabiner, S. E. Levinson, A. E. Rosenberg, J. G. Wilson** - *Speaker independent recognition of isolated words using clustering techniques*, IEEE Trans. Acoust. Speech Signal Process. 27 (1979) 336–349, 1979.

[Sha60] **C. E. Shannon** - *Coding theorems for a discrete source with a fidelity criterion*, R. E. Machol (Ed.), Information and Decision Processes, McGraw-Hill, New York, 1960, pp. 93–126, 1960.

[SC78] **H. Sakoe, S. Chiba** - *Dynamic programming algorithm optimization for spoken word recognition*, IEEE, Trans. Acoustics, Speech and Signal Proc., Vol. ASSP-26, 1978.

[SR75] **M. R. Sambur, L. R. Rabiner** - *A speaker-independent digit recognition system,"* B.S.T.J. 54 (1) 84–102, 1975.

[S+94] **J. C. Segura, A. J. Rubio, A. M. Peinado, P. Garcia, R. Roman** - *Multiple VQ Hidden Markov Modeling for Speech recognition*, Speech Communication, vol. 14, pp 163-170, 1994.

[Toh86] **Y. Tohkura** - *A weighted cepstral distance measure for speech recognition*, IEEE ICASSP 86, Tokyo, 1986, pp. 761–764, 1986.

[TA14] **G. Theodoros, P. Aggelos** - *Introduction to Audio Analysis: A MATLAB Approach*, Elsevier Academic Press USA, pp 185-210, 2014.