2013 IEEE 3rd International Conference on System Engineering and Technology, 19 - 20 Aug. 2013, Shah Alam, Malaysia

# A Review of Yorùbá Automatic Speech Recognition

Shahrul Azmi Mohd Yusof
School of Computing
Universiti Utara Malaysia
Sintok, Kedah, Malaysia
shahrulazmi@uum.edu.my

Abdulwahab Funsho Atanda
School of Computing
Universiti Utara Malaysia
Sintok, Kedah, Malaysia
abdulwahabfu@gmail.com

M. Hariharan
School of Mechatronics
Universiti Malaysia Perlis
Perlis, Malaysia
hari@unimap.edu.my

***Abstract*— Automatic Speech Recognition (ASR) has recorded appreciable progress both in technology and application. Despite this progress, there still exist wide performance gap between human speech recognition (HSR) and ASR which has inhibited its full adoption in real life situation. A brief review of research progress on Yorùbá Automatic Speech Recognition (ASR) is presented in this paper focusing of variability as factor contributing to performance gap between HSR and ASR with a view of x-raying the advances recorded, major obstacles, and chart a way forward for development of ASR for Yorùbá that is comparable to those of other tone languages and of developed nations. This is done through extensive surveys of literatures on ASR with focus on Yorùbá. Though appreciable progress has been recorded in advancement of ASR in the developed world, reverse is the case for most of the developing nations especially those of Africa. Yorùbá like most of languages in Africa lacks both human and materials resources needed for the development of functional ASR system much less taking advantage of its potentials benefits. Results reveal that attaining an ultimate goal of ASR performance comparable to human level requires deep understanding of variability factors.***

**Keywords:** ***Automatic Speech Recognition, Robust ASR, Variability in ASR, Yorùbá speech processing.***

## I. INTRODUCTION

Speech is the most natural and a vital tool for human communications, thoughts and ideas are exchanged through speech [1][2][3]. Human auditory system (HSR) is endowed with robustness - ability to recognize speech with high accuracy regardless of speaker's characteristics and/or environmental conditions [4][5]. Speech is made up multi-layered temporal-spectral variation that embed words, intentions, style of speaking, accent, gender, intonation, expression, state of health and emotion of the speaker , speaker identity, sex, and age [6][7]. Technological advancement has made speech technology an indispensible tool for socioeconomic development [8]. A spoken language system must have both speech recognition and speech synthesis capabilities – understanding and dialogue component plus domain knowledge.

Despite appreciable progress ASR has recorded over the past six decades, there still exist a wide performance gap between ASR and HSR with HSR performance far higher than ASR [5][9][10]. This degradable performance of ASR is attributed to non-cognizance of variability in real-world situation during ASR design [5]. The fundamental problem of speech recognition like any other pattern recognition problem is variability which result is low recognition rate (higher word error rate) and its consequent degradable performance. Sources of speech variability include duration, spectral, emotion, accent, contextual, and noise. However the most challenging of this variability includes accent, co-articulation, and background noise [7]. While [11] remarks that ASR systems are highly susceptible to speaker variability and that aside gender, the next source of variability is speech is accent, and went ahead to suggest that ASR should be designed considering variation in accents rather than base on native speakers alone.

Reference [6] while given further explanation on the performance gap between ASR and HSR remarked that the deficiencies probably come from a combination of factors such as the feature representations used for ASR may not contain all the useful information for recognition, the modeling assumptions may not be appropriate, and the applied features extraction and the modeling approaches may be too sensitive to intrinsic speech variability, amongst which are: speaker, gender, age, dialect, accent, health condition, speaking rate, prosody, emotional state, spontaneity, speaking effort, articulation effort. In the work of [12], WER of 30% was observed due to noise and 45% to non-native speakers (accent). This thus calls for serious attention to noise and accent as viable means of achieving ASR robustness.

Like any other technology, research in ASR for African language is still at infancy stage [8]. Yorùbá being one of African languages is considered under-resourced for speech processing [13][14]. Reference [15], remarks that developing ASR capability for new languages requires considerable amount of time, money and expertise that constitute a major impediment in developing nations bedeviled with scarcity of both human and financial resources.

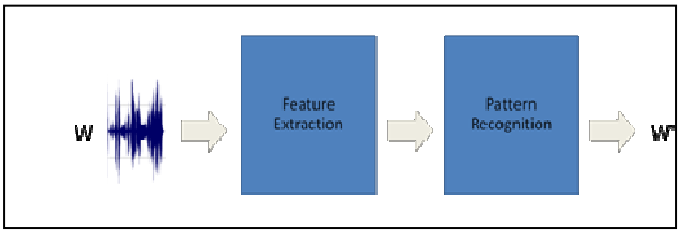
Among wide range of ASR systems applications includes the relatively simple isolated-word recognition systems for name-dialing, automated customer service and voice-control of cars and machines, continuous speech recognition as in auto-dictation or broadcast-news transcription [7]. ASR technologies has become part of daily lives activities of developed nations ranging from voice commands to control electronics, voice dialing, video games interface to voice dictation and dialogue systems [16]. This adoption of ASR technologies has lead to more conveniences for citizens that have ASR in their languages [17]. Remarks by [17] that speech technology has the potential to bridge the digital divide of the developing nations, should be a motivating factor spurring research in Yorùbá ASR.

978-1-4799-1030-4/13/$31.00 ©2013 IEEE

## II. AUTOMATIC SPEECH RECOGNITION (ASR)

Automatic speech recognition (ASR) involves the process of converting recorded acoustic speech signal into equivalents sequence of words. Its involves generation of words from recorded speech (STT) [18][19]. ASR is a technique aimed at converting speech signal into spoken word equivalents in text or action [20] in an accurate and efficient manner [21] as shown in figure 1. However, the conversion process should not depend on the speaker, acoustic signal or channel medium. In otherwords, ASR should not only be accurate and efficient, but must also be robust to variation either by the speaker or due to environmental interference [5]. ASR performs two basic operations of speech signal modeling and pattern matching. ASR can be classified based on process and approaches.

Fig. 1. Block diagram of ASR




### A. *ASR Process*

ASR system like any other patter recognition system, consist of two major processes of signal modeling and pattern matching [16][22]. Signal modeling involves the conversion of speech signal into a set of parameter. The operations involves in Signal modeling includes: spectral shaping – conversion of speech signal from analogue to digital; feature extraction – extraction of speech features such as energy, pitch, formants etc from speech signal; parametisation - transforming extracted speech features into signal parameters by differentiation and concatenation; and statistical modeling: converting signal parameters into observation vectors. While Pattern matching/classification involves finding parameters set from memory that match the parameters obtained from feature extraction process. Basic operations of pattern matching are: training and testing. Of all the process involves in ASR, feature extraction is of significant role as the accuracy of pattern classification depends on successful feature extraction.

i. Feature Extraction (FE): FE is the process of extracting useful and meaningful features from the recorded speech signal [23]. Accuracy of ASR system depends to a large extent an effective and efficient FE [24]. FE is therefore the most important component of ASR as the success of pattern classification depends on an accurate FE process [25]. An efficient ASR tasks requires FE to be designed for normalizing number of effects irrelevant for the decoding of speech signal. These effects include: noise, channels, microphone characteristics, and speaker-dependent characteristics such as: vocal track length, accent, emotion, and illness [24]. This works argues that successful ASR systems to a large extent depend on an effective and efficient FE process. Approaches to FE can be classified into: temporal analysis (TA) and spectral analysis (SA). In TA, the speech waveform is used directly for processing. Though this approach is simple and involves less computation, it's however limited to few and simple speech features such as power and periodicity. SA uses spectral representation of speech signal for feature analysis. SA is the most used of the two analysis approaches. There are several techniques of SA which include: critical band filter, cepstral analysis, Mel cepstrum analysis, LPC, perceptually linear predictive analysis (PLP), and MFCC. In SY ASR, MFCC and LPC are the most widely used techniques.

ii. Pattern classification (PC): Approaches to pattern classification is ASR can be divided into: Rule based and data driven. Though rule based approach is generalisable, easy to use and understand, it however requires multi-disciplinary experts which SY lacks. Data driven approach (machine learning) such as classification and regression tree (CART), Hidden Markov Model (HMM), Bayesian model, and ANN involves modelling and training using speech data for recognition purposes. In SY ASR research, commonly used approaches are ANN and HMM with hybrid of both [26][27]. Figure 2 below shows the ASR process.

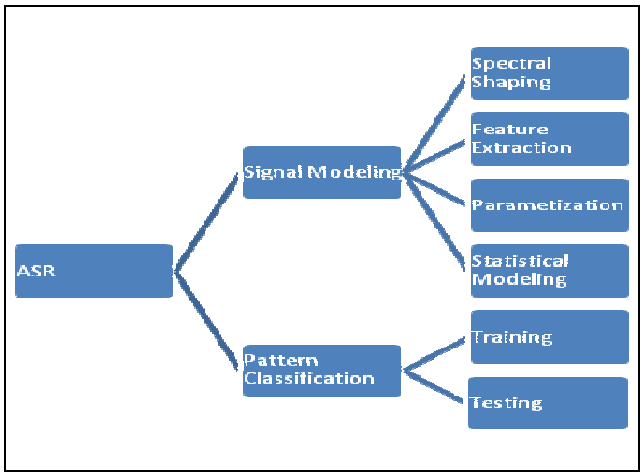



Fig. 2. ASR process

### B. *ASR Approaches*

Speech recognition approaches can be classified into three broad categories of:

i. Acoustic-Phonetic Approach: This approach exploits the theory of acoustic-phonetics of existence of finite phonetic units characterized by set of properties in the speech signal (Rabiner & Juang, 1993). It attempts to decode speech signal based on the known relationship between acoustic features of the signal and phonetic unit. This approach consists of three process of: spectral analysis and features extraction, segmentation and

labeling, and valid word/ string identification. Despite being one of the earliest approaches acoustic-phonetic has not recorded wider use in commercial applications [18].

ii. Pattern Recognition Approach: In pattern recognition approach, only speech pattern are used directly in speech recognition based on mathematical formulation. It consists of two processes of pattern training and recognition. Methods used include MFCC, HMM and LPC. Pattern recognition approach is the most widely used approach due to its simplicity of use, robustness, and high performance [18][28].

iii. Artificial Intelligence Approach: Is a combination of both acoustic-phonetic and pattern recognition approach resulting into a hybrid approach. The ideas and concepts of acoustic-phonetic and pattern recognition approach is being exploited [18]. The aim of artificial intelligence approach is to mechanize the recognition process similar to the manner in which humans applies intelligence in visualizing, analyzing, and characterizing speech based on a set of measured acoustic features. Techniques commonly include expert system (e.g., a neural network) that integrates phonemic, lexical, syntactic, semantic, and even pragmatic knowledge for segmentation and labelling, and uses tools such as artificial neural networks for learning the relationships among these phonetic events [28].

## III. YORÙBÁ LANGUAGE SYSTEMS

Yorùbá (Standard Yorùbá - SY ) like most of Africa languages is a tonal language and one of the twelve languages of the Edekiri sub-branch from the great family of the West Benue-Congo of the Niger-Congo phylum of African languages. It is natively spoken in south western part of Nigeria (the second largest ethnic group in number). Nigeria has about 30 million speakers of SY. It is also spoken in Togo, Republic of Benin, Ghana, Sudan, Sierra-Leone and Côte D'Ivoire [27]. SY is also been spoken beyond the Africa to countries such as Brazil, Cuba, including Trinidad and Tobago where a large number of speakers of the language can be found [29]. SY being a tonal language like Cantonese and Thai, unlike non-tonal languages such as French and Malay in which word meaning can be inferred from spelling; the tone of pronunciation which is associated with each syllable of a SY word determines the meaning of that word. The homographic nature of SY also makes it a complex language. In homographic language, a single word can have several meaning based on the tone pronunciation [27].

Although Yorùbá language has many dialects, the most generally used is Standard Yorùbá (SY). SY alphabets consists of seven vowels (a, e, ẹ, i, o, ọ , u ), eighteen consonants ( b, d, f, g, gb, h, j , k, l, m, n, p, r, s, □, t, w, y ), five nasalized vowels (an, ẹn, in, ọn, un), and 2 syllabic nasals (m, n). SY has a diagraph gb which is a consonants with two letters. SY has three tone levels: high tone, mid tone, and low tone represented by acute accent symbol ( ´ ), macron ( □ ), and grave accent symbol (`) respectively. Two contrastive tones of rising (R) and falling (L), are also part of SY. Tones are realised on vowels and sometimes on nasal consonants [30]. A SY syllable can be formed with a combination of vowels (V), consonants (C), and/or nasal vowels (n) to give syllable combinations: CV, $CV_n$, V, N and $V_n$ [32].

## IV. SOURCES OF SPEECH VARIATION

Successful improving ASR systems performance is directly linked with successfully identifying sources of variability before counteracting the effects can be designed [6]. Sources of speech variability can be broadly categorized into two: speaker's intrinsic characteristics and environmental sources [5].

### *A. Intrinsic variability in speech*

Intrinsic variation in speech is due to factors that are directly related to speaker's characteristics. Such factors includes: gender, age, rate of speech (ROS), accent, dialect, and emotion [6][7]. Of all these factors, accent constitutes a major source of variation in ASR [33][34][35].

Accent variation

Reference [36] defines accent as the way a speaker produces the sounds of a language. An accent can indicate the speaker's first language (native or non-native), where they were born (regional accent), religious affiliation, ethnic group, or socio-economic class. Accents affect both acoustic (e.g. formants) and prosodic (e.g. intonation, duration, and rate) aspects of speech. Accent constitutes a major factor that impede on the performance of ASR [7]. Accent variation is as a result of variation in pronunciation between native and non-native speakers and therefore pronunciation modeling plays a critical part is performance improvement of ASR due to accent variation [6]. Reference [37] works on the variability between speakers using statistical analysis methods reveled that the two sources of variation are gender and accent. This assertion is corroborated by the findings of both [38][39] that performance degrades when recognizing accented speech and non-native speech. This is as a result of replacement of unfamiliar phoneme absent in the native language phoneme inventory of the speaker, with the sound considered as the nearest in their native language phoneme inventory [40]. Reference [7], remarks that test data of different accent from the training data can result in degradable performance of ASR, and hence emphasized the need for accent identification and modeling. The result from their experiment of accents of British, American, and Australia corroborates this assertion.

### *B. Environmental variability*

Before now, the main source of variability is ASR is attributed to environmental - environmental noise. This is evident in the number of publications related to environmental/noise robustness. Lately, distortion as a result of transmission channels and reverberation has been identified as environmental factor of speech variability.

Several methods are adopted in mitigating the environmental effect of ASR performance. Speech enhancement is aimed at generating clean speech input signal devoid of environmental contaminations. Speech enhancement approaches includes using noise-cancelling microphone or microphone array [41][42]. Approaches in counteracting reverberation effects are proposed by [43][44][45]. Voice activity detection (VAD) aimed at eliminating both noise and reverberation effect through detection and elimination of speech boundaries or noise portion of speech signal. VAD methods includes energy thresholds and zero crossing rate[46][47][48][49]. Others approaches are: feature extraction [50][51][52][53][54], acoustic modelling [28][55][56][57], and pronunciation modelling [8][58].

## V. Research Progress in SY ASR

The fundamental problem of speech recognition like any other pattern recognition problem is variability which result is low recognition rate (higher word error rate). Sources of speech variability include duration, spectral, speaker, accent, contextual, and noise. However the most challenging of this variability includes accent, co-articulation, and background noise [7] While [11] remarks that ASR systems are highly susceptible to speaker variability and that aside gender, the next source of variability is speech is accent, and went ahead to suggest that ASR should be designed considering variation in accents rather than base on native speakers alone. The main acoustic cue for tone is pitch (fundamental frequency F0) whiles the first and second formants (F1 and F2) constitute acoustic cues for phoneme [13][20]. Most ASR performs poorly on intelligibility and naturalness due to non consideration of tonal cue in their design due to complexity in modeling tonal cue [13]. Also [20] observed that ASR for tone language involves complex task of simultaneously identifying tone and phoneme in speech signal.

In ASR research, tones have becomes an interesting attraction due to its: syllabic associativity - tones are associated with syllables which are building units of speech; unique fundamental frequency (F0) - in isolated utterance, each tone is characterized by a unique F0 curve [20][27],and also the fact that in SY, tones are realized on vowels [20][30]. F0 is used to distinguish between lexical tones words that are otherwise phonemically identical [31]. Though [27] opined that considering the syllabic associativity and uniqueness of F0 tones, the complexities of designing speech recognition for tone languages can be reduced considerably, taking into cognizance the tonal realisation on syllable can lead to further reduction in ASR design complexities for tonal language. The importance of given cognizance to the fact that tone recognition is a significant step in the recognition of speech in tone languages emphasized by [20][59]. Most syllables in SY words end with a vowel or a nasal sound, and there are no consonant clusters. It is common in some dialects of SY to combine the pronunciation of two syllables if one ends in a vowel and the next begins with one.

In ASR systems, premium should be on FE that is robust (invariant) to both speaker and environmental variations [60]. Despite appreciable progress ASR has recorded over the past six decades, challenges still remains prominent of which is speech variability due to speaker and environmental factors [8]. SY being a scarce resource and tonal language has recorded only a handful number of researches mostly focusing on Text-to-Speech (TTS) [8][13][14][26][32][61], automatic voice dialing (ATD), [62], tone realization and prosodic modeling [63][64], and ASR [20][32]. Automatic Voice Dialing (AVD) is an application of ASR to human-machine interface by speech in mobile phone dialing [62]. Reference [62] developed and implements an algorithm in C-language for speaker authentication and speech recognition for mobile phone voice dialing in SY. The speech corpus was made up of 2,600 isolated words recorded from 20 subjects each reading SY numeral 0-9 and words pè, gbè, and fònú 10 times. Speech signal features were extracted using Mel frequency cepstral coefficient (MFCC) and pattern matched by Euclidean distance measure. Recognition rate of 94% and 82% is achieved for speaker identification and speech recognition respectively.

Reference [27] experimented with Artificial Neural Network (ANN) to recognition of tones for SY. The experiment employs F0 features extracted from syllable recordings of four speakers using PRAAT. Multilayered Perceptron (MLP) and Recurrent Neural Network (RNN) were modelled for the classification and recognition of extracted features of the tones in the recorded syllable. Results from of both MLP and RNN reveal that both perform on training data than on test data with overall performance of RNN greater than MLP. Lack of speech technology for SY motivated the work by [20]. The paper asserted that recognition of vowels is a necessary prerequisite for utterances identification of tonal languages. Fundamental frequencies (F0) and first two formant (F1 and F2) features were extracted from vowels recording of ten subjects. Fuzzy logic (FL) and ANN models were developed for vowels and phoneme recognition. Performance evaluation of the two models reveals that while ANN performs better with training data and FL is better with test data. Though the work of both [20] and [27] are similar in their approaches, their results however show little dissimilarities. While the results of [27] shows that both MLP and RNN perform on training data than on test data with overall performance of RNN greater than MLP. That of [20] reveals that while ANN performs better with training data and FL is better with test data with ANN models performs better than FL on the overall. Reference [26] experimented on the effect of Voice Activity Detection (VAD) on SY ASR. Hybrid of MFCC and Linear Predictive Coding (LPC) was used for feature extraction while ANN was used in recognition stage. Recognition accuracy of about 61.2% was achieved for the hybrid feature extraction which is higher than 60% for LPC and 58% for MFCC. Results from the experiment reveal an interesting relationship between VAD and speech intelligibility, the higher the VAD value determine by frame size, the lesser the intelligibility of the speech.

The problem of scarce resources for the development of ASR for resource-scarce languages such as SY ASR motivated researchers to developed small-vocabulary ASR for resource-scarce languages. Notable among these was the development of Small-Vocabulary Speech Recognition for Resource-Scarce Languages of Yoruba and Hebrew using cross-language phonemic mapping [15]. Though their system achieved more than 90% accuracy, due to its small vocabulary it can only applied in limited applications. Absence of speech corpus for Nigerian English (NE) motivated the development of Nigerian English Corpus (UISpeech corpus [65]. Their work demonstrates the possibility of developing ASR for resource-scarce language from an existing and established ASR. The result from UISpeech reveals that there is pronunciation variation between American English (AE) and NE due to accent dissimilarity. Similar researches include The Salaam method, PMSR.

The fact that vowels are the main tone bearing element in tone language speech, makes their recognition in tone languages to be crucial and central. Recognizing vowels first will make the task of recognizing bigger units, such as syllables and words to be accomplished much more easily. F0 is the acoustic cue for the tone while the first and second formants (F1 and F2) frequencies constitute the acoustic cues for the phoneme [20].

## VI. Conclusion and feature work

Ability to recognize tone is a necessary step towards accurate utterances recognition in SY. The extent of the consideration given to tonal and phonemic content in the recognition process to a large determines the accuracy of ASR for tonal languages. Considering the uniqueness of F0 and syllabic associativity of tones can greatly reduce the complexities is the development of SY ASR. Though ASR technologies has recorded considerable progress and improved comfort to advanced worlds, likewise the tonal languages of Asia such as Thai, Cantonese, and Mandarin. African tonal languages of which SY belongs is still at infancy. SY being a scarce-resource language is under research. Most of ASR in SY are on speech synthesis, voice dialing and based on isolated word. There is lack of research work on continuous speech recognition, independent speaker or on robustness of ASR. Achieving the major goal of speech recognition not only in terms of accuracy but to the level comparable to human capability requires robust ASR. Therefore, future works will be concentrated on how to develop a robust ASR for SY.

## References


[1] Furui, S. (2004). Fifty years of progress in speech and speaker recognition. The Journal of the Acoustical Society of America, 116, 2497.

[2] Hariharan, M., Chee, L. S., Ai, O. C., & Yaacob, S. (2012). Classification of speech dysfluencies using LPC based parameterization techniques. Journal of medical systems, 36(3), 1821-1830.

[3] Huang, C.-L., & Wu, C.-H. (2007). Spoken document retrieval using multilevel knowledge and semantic verification. IEEE Transactions on Audio, Speech, and Language Processing, , 15(8), 2551-2560.

[4] Scharenborg, O., & Cooke, M. (2008). Comparing human and machine recognition performance on a VCV corpus. Paper presented at the Proc. Workshop on Speech Analysis and Processing for Knowledge Discovery.

[5] You, H., & Adviser-Alwan, A. (2009). Robust automatic speech recognition algorithms for dealing with noise and accent: University of California at Los Angeles.

[6] Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., et al. (2007). Automatic speech recognition and speech variability: A review. Speech Communication, 49(10), 763-786.

[7] Yan, Q., Vaseghi, S., Rentzos, D., & Ho, C.-H. (2007). Analysis and synthesis of formant spaces of British, Australian, and American accents. IEEE Transactions on Audio, Speech, and Language Processing, 15(2), 676-689.

[8] Adedjouma Sèmiyou, A., Aoga, J. O., & Igue, M. A. (2012). Part-of-speech tagging of Yoruba standard, language of Niger-Congo family.

[9] Lippmann, R. P. (1997). Speech recognition by machines and humans. Speech Communication, 22(1), 1-15.

[10] Stern, R. M., & Morgan, N. (2012). Hearing is believing: biologically-inspired feature extraction for robust automatic speech recognition.

[11] Faria, A. (2006). Accent classification for speech recognition *Machine Learning for Multimodal Interaction* (pp. 285-293): Springer.

[12] Kirchhoff, K. (1999). Robust speech recognition using articulatory information.

[13] van Niekerk, D. R., & Barnard, E. (2013). Predicting utterance pitch targets in Yorùbá for tone realisation in speech synthesis. Speech Communication.

[14] Wong, S. H. S., & Beaumont, A. J. (2007). A fuzzy decision tree-based duration model for Standard Yorùbá text-to-speech synthesis. Computer Speech & Language, 21(2), 325-349.

[15] Qiao, F., Sherwani, J., & Rosenfeld, R. (2010). Small-vocabulary speech recognition for resource-scarce languages.

[16] Azmi, S. (2010). Feature extraction and classification of malay speech vowels.

[17] Qiao, F., Sherwani, J., & Rosenfeld, R. (2010). Small-vocabulary speech recognition for resource-scarce languages. Paper presented at the Proceedings of the First ACM Symposium on Computing for Development.

[18] Anusuya, M., & Katti, S. (2010). Speech recognition by machine, A review. arXiv preprint arXiv:1001.2267.

[19] Paulraj, M., Yaacob, S., & Yusof, S. M. (2008). Vowel recognition based on frequency ranges determined by bandwidth approach. Paper presented at the International Conference on Audio, Language and Image Processing, 2008. ICALIP 2008.

[20] Àkànbí, L. A., & Odéjobí, O. À. (2011). Automatic recognition of oral vowels in tone language: Experiments with fuzzy logic and neural network models. Applied Soft Computing, 11(1), 1467-1480.

[21] Holmes, J., & Holmes, W. (2002). Speech recognition and synthesis: Taylor & Francis, London, UK.

[22] Duda, R. O., Hart, P. E., & Stork, D. G. (2001). Pattern classification. *New York: John Wiley, Section, 10*, l.

[23] Deller, J. R., Proakis, J. G., & Hansen, J. H. (2000). Discrete-time processing of speech signals: Ieee New York, NY, USA:.

[24] Sun, Z., Yuan, X., Bebis, G., & Louis, S. J. (2002). Neural-network-based gender classification using genetic search for eigen-feature selection. Proceedings of the 2002 International Joint Conference on Neural Networks, 2002. IJCNN'02.

[25] Hamdy, A., Hefny, H., Salama, M. A., Hassanien, A. E., & Kim, T.-h. (2012). The importance of handling multivariate attributes in the identification of heart valve diseases using heart signals. Paper

presented at the 2012 Federated Conference on Computer Science and Information Systems (FedCSIS).

[26] Aibinu, A. M., Salami, M. J. E., Najeeb, A. R., Azeez, J., & Rajin, S. (2011). Evaluating the effect of voice activity detection in isolated Yoruba word recognition system. Paper presented at the 2011 4th International Conference on Mechatronics (ICOM).

[27] ỌDélọBí, Ọ. À. (2008). Recognition of Tones in YorÙbÁ Speech: Experiments With Artificial Neural Networks Speech, Audio, Image and Biomedical Signal Processing using Neural Networks (pp. 23-47): Springer.

[28] Rabiner, L., & Juang, B.-H. (1993). Fundamentals of speech recognition.

[29] Williamson, K., & Blench, R. (2000). Niger-Congo. African languages: An introduction, 1, 42.

[30] Dopamu, P. (2004). Understanding Yoruba life and Culture. Trenton NJ: Africa.

[31] Xu, Y. (2004). Transmitting tone and intonation simultaneously-the parallel encoding and target approximation (PENTA) model. Paper presented at the International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages.

[32] Ọdẹ́ọbí, O., Beaumont, A., & Wong, S. (2004). Experiments on stylisation of standard Yorùbá language tones: Technical Report CS-001, Aston University, Birmingham, United Kingdom.

[33] Hanani, A., Russell, M. J., & Carey, M. J. (2013). Human and computer recognition of regional accents and ethnic groups from British English speech. Computer Speech & Language, 27(1), 59-74.

[34] Humphries, J., & Woodland, P. (1998). The use of accent-specific pronunciation dictionaries in acoustic model training. Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, 1998.

[35] Tjalve, M., & Huckvale, M. (2005). Pronunciation variation modelling using accent features. Proceedings of INTERSPEECH 2005.

[36] Felps, D. (2013). Articulatory-based Speech Processing Methods for Foreign Accent Conversion.

[37] Huang, X., Acero, A., & Hon, H.-W. (2001). *Spoken language processing* (Vol. 15): Prentice Hall PTR New Jersey.

[38] Kubala, F., Anastasakos, A., Makhoul, J., Nguyen, L., Schwartz, R., & Zavaliagkos, E. (1994). Comparative experiments on large vocabulary speech recognition. 1994 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP94).

[39] Lawson, A. D., Harris, D. M., & Grieco, J. J. (2003). Effect of foreign accent on speech recognition in the NATO N-4 corpus. Paper presented at the Eighth European Conference on Speech Communication and Technology.

[40] Hao, Y.-C. (2012). Second language acquisition of Mandarin Chinese tones by tonal and non-tonal language speakers. Journal of phonetics, 40(2), 269-279.

[41] Martin, R. (2005). Statistical methods for the enhancement of noisy speech Speech Enhancement (pp. 43-65): Springer.

[42] Wu, M., & Wang, D. (2006). A two-stage algorithm for one-microphone reverberant speech enhancement. IEEE Transactions on Audio, Speech, and Language Processing, 14(3), 774-784.

[43] Nakatani, T., Juang, B.-H., Kinoshita, K., & Miyoshi, M. (2005). Harmonicity based dereverberation with maximum a posteriori estimation. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005.

[44] Neely, S. T., & Allen, J. B. (1979). Invertibility of a room impulse response. The Journal of the Acoustical Society of America, 66, 165.

[45] Yegnanarayana, B., & Murthy, P. S. (2000). Enhancement of reverberant speech using LP residual signal. IEEE Transactions on Speech and Audio Processing, 8(3), 267-281.

[46] Agarwal, A., Jain, A., Prakash, N., & Agrawal, S. (2010). Word boundary detection in continuous speech based on suprasegmental features for hindi language. 2010 2nd International Conference on Signal Processing Systems (ICSPS).

[47] Laskowski, K., Jin, Q., & Schultz, T. (2004). Crosscorrelation-based multispeaker speech activity detection. Paper presented at the subm. Proc. ICSLP-2004.

[48] Tu, C.-C., & Juang, C.-F. (2012). Recurrent type-2 fuzzy neural network using Haar wavelet energy and entropy features for speech detection in noisy environments. Expert Systems with Applications, 39(3), 2479-2488.

[49] Wrigley, S. N., Brown, G. J., Wan, V., & Renals, S. (2005). Speech and crosstalk detection in multichannel audio. IEEE Transactions on Speech and Audio Processing, 13(1), 84-91.

[50] Deng, L., Droppo, J., & Acero, A. (2004). Enhancement of log mel power spectra of speech using a phase-sensitive model of the acoustic environment and sequential estimation of the corrupting noise. IEEE Transactions on Speech and Audio Processing, 12(2), 133-143.

[51] Dimitriadis, D., Maragos, P., Pitsikalis, V., & Potamianos, A. (2002). Modulation and chaotic acoustic features for speech recognition. Control and intelligent systems, 30(1), 19-26.

[52] Hui, Y., Climent, N., & Volker, H. (2010). Pitch-and formant-based order adaptation of the fractional Fourier transform and its application to speech recognition. EURASIP Journal on Audio, Speech, and Music Processing, 2009.

[53] Pelecanos, J., & Sridharan, S. (2001). Feature warping for robust speaker verification

[54] Pitsikalis, V., & Maragos, P. (2006). Filtered dynamics and fractal dimensions for noisy speech recognition. Signal Processing Letters, IEEE, 13(11), 711-714.

[55] Saraclar, M., Nock, H., & Khudanpur, S. (2000). Pronunciation modeling by sharing Gaussian densities across phonetic models. Computer Speech & Language, 14(2), 137-160.

[56] Acero, A., Deng, L., Kristjansson, T., & Zhang, J. (2000). HMM adaptation using vector Taylor series for noisy speech recognition. Paper presented at the Proc. ICSLP.

[57] Daoudi, K., Fohr, D., & Antoine, C. (2003). Dynamic Bayesian networks for multi-band automatic speech recognition. Computer Speech & Language, 17(2), 263-285.

[58] Kenny, P., Boulianne, G., & Dumouchel, P. (2001). Inter-speaker correlations, intra-speaker correlations and Bayesian adaptation. Paper presented at the ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition.

[59] Connell, B. A., Hogan, J. T., & Rozsypal, A. J. (1983). Experimental evidence of interaction between tone and intonation in Mandarin Chinese. Journal of phonetics, 11(4), 337-351.

[60] Kesarkar, M. (2003). Feature extraction for speech recognition. Electronic Systems, EE. Dept., IIT Bombay.

[61] Wong, S. H. S., & Beaumont, A. J. (2008). A modular holistic approach to prosody modelling for Standard Yorùbá speech synthesis. Computer Speech & Language, 22(1), 39-68

[62] Ibiyemi, T., & Akintola, A. (2012). Automatic Speech Recognition for Telephone Voice Dialling in Yorùbá. International Journal of Engineering, 1(4).

[63] Nagano-Madsen, Y. (1993). The grouping function of F0 and duration in two prosodically diverse languages-Eskimo and Yoruba. Paper presented at the ESCA workshop on prosody.

[64] Van Niekerk, D., & Barnard, E. (2012). Tone realisation in a Yoruba speech recognition corpus.

[65] Amuda, S., Boril, H., Sangwan, A., & Hansen, J. H. (2010). Limited resource speech recognition for Nigerian English. 2010 IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP2010).