EKAB-OSOWO TAWO

Undergraduate Thesis (Chapter 1&2)

OCTOBER 10, 2018

Chapter 1: Introduction

Today, the increase in the patronage of speech recognition systems have become quite evident as giants in the technology sector such as Apple, Google, Amazon, Microsoft, IBM to name a few have inculcated forms of speech recognition in their devices or have speech recognition systems of their own. The growth and patronage of these brands have mostly increased in the world as the use of speech recognition systems exist on almost every computer system.

In Africa, with numerous languages and accents, the use of these systems has proven to be intriguing as specific, at the same time common errors are present. Such errors include mispronunciation of words, misinterpretation of words and so much more. The root of this problem is the fact that speech recognizers do not understand specific accents spoken to it. As such, the mismatch in accents can increase the error rate of a speech recognition system by more than 100% (Yan & Vaseghi, 2002) as it may not understand the adequate pronunciation of words. Though advancements in technology have subjected the significant causes of errors in speech recognition systems to noise and speaker variations, accents also form a blocking stone in the progression of these systems. Now, a problem that speech recognition systems have that may provide a reason as to why accent affects the credibility and reliability of these systems is that the training of speech recognizers occurs in the UK or US English accents (Yan & Vaseghi, 2002). Though England or other European countries colonized a large proportion of Africa, the accents of Africa are unique and relatively different thus proving to be a hindrance in speech recognition systems recognizing African speech regardless of it being in English.

Speaker variability is a factor that affects the performance of automatic speech recognition (ASR). The attributes of this speaker variability involve gender and accents to name a few. Unlike gender which has been relatively resolved, little research has been factored into accents. The accent

issues have two related research, and these are accented adaptation through pronunciation modeling and accent identification. People with heavy accents are prone to make more pronunciation errors regarding standardized errors (Benzeghiba et al.). The African accent is a heavy accent as the pronunciation of English words varies from location to location. Even in similar geographical areas, the difference in mother tongue also affect the accents and thus change the pronunciation of words.

The motivation for this paper comes from the fact that using various speech recognition technologies, be it Siri, Google Voice or Cortana, errors are prone to occur. As earlier stated in the paper, accents have a substantial effect on the reaction from this recognition technologies and as such mistakes are likely to happen every day when accessing any of this speech recognition technologies. The impact of accents on speech recognition systems are crucial in undergoing research because of firstly, the technological advancement that occurs every day around the world. With regards to this point, the prevalence of this systems would mainly increase because the world is moving towards the Internet of Things and Artificial Intelligence and more and more speech recognition instruments to get about daily activities. Also, with the globality of the world, Africa would compete at almost the same level with other technological powerhouses and these advancements and the use of technologies such as speech recognition systems to name a few would be a norm in the continent.

Having looked at speech recognition systems and the various factors of it, as well as the importance of accents in understanding, acquiring credibility and reliability of these systems, **the objective of this research is to understand and access to a considerable extent the impact of the African accent on English speech recognition systems**. Moreover, there has been researching into the effects of differences in accents in speech recognition, the comparison of the

UK and US accents in speech recognition, as well as the impact of the South African accent with regards to speech recognition. This paper seeks to look at the African continent to work with a more diverse experimental scope.

For this project, the use of human subjects would be used for the experiment. The participants are most likely going to be college students precisely freshman from a diversified college in West Africa. These participants would be either randomly chosen or would be given the opportunity to volunteer to the success of this research. Though the selection of the participants may be random, a more stratified approach would be considered, as the freshmen would first be grouped based on their nationalities before the randomization occurs. The focus of the various African nationalities for this research include; Nigerians, Ghanaians, Zimbabweans, Kenyans, and Rwandans. The exact number of participants from each group, ranges from 2-5 participants, with ages 17-21. Also, all forms of gender would be accepted for the research. Furthermore, though all gender forms are applicable, the experiment which would be undertaken for this research would strive to have gender balance. The need for the gender balance is to strive for consistency during the project as gender is a possible factor for the occurrence in error rates by speech recognition systems.

Aside from the participants, various speech recognizers would be made available and these recognizers may be trained before the experiment begins or no training and just straight to the experiment. The considered speech recognizers include; IBM speech recognition system, Microsoft's Cortana, Google's Hello Google, and CMU to name a few.

1.1 Research Question

- How does the African Accent affect English speaking speech recognition systems?

Chapter 2: Literature Review/Related Work

There is first a need to understand what accents are and if at all they affect speech recognition systems in order to aid in the progress of this research. By doing this, as previously stated, it would assist in giving or guiding the purpose of this paper to a fruitful conclusion and the objectives of the article can be met.

Accents are patterns of pronunciation used by a speaker for whom English is the native language or, it involves this speaker patterns being distinctive to a community or social grouping. Furthermore, an accent is something which every speaker has, and it is something unique to that individual. Also, accents to a considerable extent are characteristics of people or a group of persons belonging to some geographical region, and belonging to a particular social class, and, it may even be representative to a speaker's sex, level of education or level of education. Around the world, the native languages of some continents, countries, and regions are not English, and as such the accents with which these nationals would speak it may be referred to as Foreign Accent. Foreign accents are pronunciation patterns seen as typical of speech of those who English is not their native language. These patterns would expect reflections from many phonological and phonetic characteristics of their mother tongue. An individual's accent is a very powerful tool as it helps as indicators of various aspects of that individual. Based on a person's pronunciation, the knowledge of where that individual comes from, where he/she grew up and, in some cases, where that individual lives may be observed or known. Accents also to some extent reveal a person sex and ethnicity. In the case of the sex of an individual, this isn't limited to the standard biological differences, but also how the individual sounds. For ethnicity, there exists an infusion of phonetic characteristics which are specific to their mother tongue (Wells, 1982).

Thus, since to a reasonable extent, understanding of accents have been established, the need to gain knowledge on what speech recognition systems are, how it works as well as examples would be useful. Speech is a sequence of words encoded by a speaker into a continuous acoustic signal. The idea of speech recognition and the speech recognition problem involves a system, device or computer adequately identifying the specific words which are used by the speaker or more correctly, the words which were intended by the speaker. A basic overview of speech recognition is speaking to a device and having that device answer you or actively perform the task you want it to act or to interpret what has been said to it effectively. This form of technology has become increasingly needed, for example, an idea of speech recognition involves aiding deaf people with communication. This form of speech technology is about translating speech to text (Houde, 1979), however, as important as this may be, it doesn't connote what the paper wants to achieve or what this paper is trying to address. As earlier stated, this paper is trying to understand accents, African accents to be precise against English speaking speech recognition systems. That is, speaking to the speech recognition systems for them to trigger an effect or result into action.

Theories of acoustic-phonetics influenced the early attempts to develop speech recognition systems or automatic speech recognition systems (ASR). Acoustic phonetics describe phonetic elements of speech as well as trying to explain how the basic sounds of languages are acoustically realized in the spoken utterance. Practically, this meant that for example in creating a vowel sound, the vocal cords need to have some form of vibration. The year 1952 was when Balashek, Davis, and Biddulph of Bell laboratories constructed a system for isolated digit recognition for a speaker. This system was only able to recognize spoken numbers from 1-9. Other systems built during the 1950's involved Olson and Belar of RCA Laboratories system which could identify ten syllables of a single talker. At MIT Lincoln Lab, Forgie made a speaker-independent ten vowel recognizers.

In the 1960's, the Japanese laboratories showcased their skills in building some form of unique hardware to carry out speech recognition tasks. The vowel recognizer of Suzuki and Nakata at the Radio Research Lab Tokyo was a notable recognizer built by the Japanese. Another early recognition system was constructed by Fry and Denes of University College in England. Their recognizer could recognize four vowels and nine consonants; however, by teaching statistical information about allowable phoneme sequences in English, they were able to increase the overall phoneme recognition accuracy for words consisting of two or more phonemes.

In the 1970's, recognition of common vocabularies (order of 100-1000 words) using simple template-based, pattern recognition methods were developed. During this decade, the critical technology created was the pattern recognition models, the pattern clustering methods for speaker-independent recognizers to name a few. The 1980's brought about the tackling of speech recognition problems based on statistical methods that involve a wide range of networks handling language structures. The critical technologies during this period that were introduced involved the hidden Markov model (HMM) and the stochastic language model. These models together enabled powerful new methods for virtually handling any continuous speech recognition problems effectively and with high performances. The 1990's came with the ability to build large vocabulary systems with unconstrained language models as well as constrained task syntax models for continuous speech recognition and understanding. The last few years have brought about the introduction of extensive vocabulary systems with full semantic models, integrated with text-to-speech (TTS) synthesis systems and multi-modal inputs. During this period, the use of machine learning to improve speech understanding and speech dialogs as well as the introduction of mixed-initiative dialog systems to enable user control (Juang & Rabiner, 2004).

In no time the existence of speech recognition technology was being paved and now large technology companies today have their speech recognition systems. Amazon's Alexa, Googles Hello Google, Apple's Siri, Microsoft Cortona to name a few are examples of popular and mostly patronized speech recognition systems. Speech recognition technologies have entered the marketplace benefiting these large technological powerhouses with revenue as well as innovative advancements.

The first step in beginning the work with speech recognition systems and accents is the accent identification. Accent identification is an adequate tool that can help in training the recognizer to be aware of what kind of accent that is being made available to it. In their research (Teixeira, Trancoso, Serralheiro) worked on trying to identify foreign accents using an automatic identification accent system. Though the reasons for their study was based on identifying accents from 6 European countries, the idea is still valid. From their work, it showed that it is preferable to train a recognizer with a mixture of accents as it would aid in accurate identification from speech recognizers. Their research also noted the fact that automatically identifying non-native accents is a difficult task as well as the fact that their system or analysis takes into consideration only pronunciation networks per word. This is a restriction in accent identification as it doesn't take account the multiple ways which an accent can be pronounced. For their current work as well as future work, (Teixeira et al.) are working on deriving various pronunciation networks and for the future, they intend to explore rhythmic cues to make a more effective accent identification.

After the identification of accents, the next research (Yan & Vaseghi), looked at the comparison between the UK and US English and how it affects the results of speech recognition systems. The experiments conducted in the research were performed doing a detailed study of the acoustic correlates of accents using intonation pattern and pitch characteristics. The paper further

goes to explain the differences in accents and the elements of the differences between the UK and US accents. The early stages of their experiment were to do a cross accent recognition quantifying the accent effects between the British accent (BrA) and American accent (GenAm) on speech recognition. After training the specific speech recognizers based on the national accents, the results from the first experiment is below.

Accent	British model	American model
British input	12.8	29.3
American input	30.6	8.8
Average	21.7	19.1

Per their beginning research, it was recorded that the American English achieves 31% less error than the British English in a matched accent condition. However, an inconsistent accent of the recognition systems deteriorates the performance as the results got worse with 139% for recognizing British English with American models and 232% for identifying American English with British models. The next step in their experiments was to an acoustic feature of the two accents in question. This feature was done using duration, pitch characteristics, and prosody. These features helped determine the understanding that apart from phonetics, the difference in the slope of rising and fall exist in bringing about contrasting results when there is a mix of accents in speech recognition.

With the breakthrough and innovative advancements and development of speech recognition, speaker variability still affects the performance of these recognition systems. Part of the factors that affect speaker variability, accents is among the most important. The purpose of their research was to show the accent issues that occur in large vocabulary continuous speech

recognition. Based on cross-accents experiments, it shows that accent problems are very dominant in speech recognition. The research areas that are related to accent issues in speech recognition involve accent adaptation through pronunciation and accent identification. Speakers with heavy accents tend to make more pronunciation errors regarding standard pronunciation. For their research, the looked at the impact of accents on speech in two views; the cross-accent speech recognition experiments were carried out and secondly, multivariate analysis tools, PCA and ICA, were applied to confirm the importance of accent in speaker variability qualitatively.

The research carried out extensive research to show or to investigate the impact of accent on state-of-the-art automatic recognition systems. Furthermore, they examined some critical factors of speaker variability and how they correlate with each other. The whole experiment consisted of 980 speakers with 200 utterances per speaker, and they were from 2 accent areas, Beijing (BJ) and Shanghai (SH).

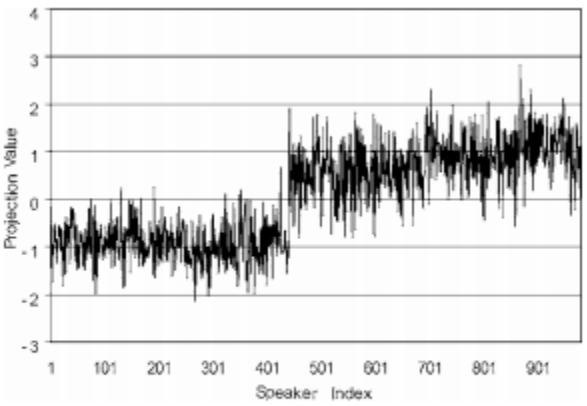
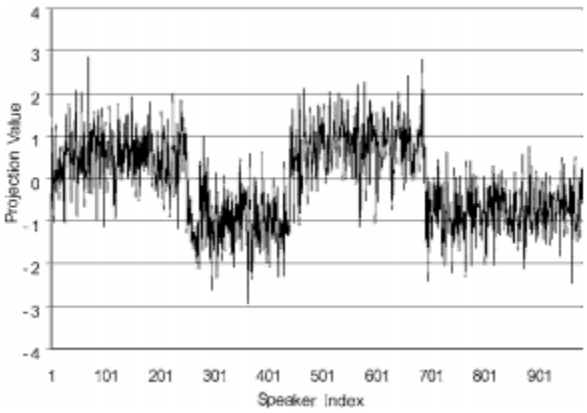


Figure 1. Projection of all the speakers onto the first independent component (The first block corresponds to the speaker sets BJ-F and SH-F, and the second block corresponds to the sets BJ-M and SH-M).

Table 4. Speaker distribution for speaker variability analysis.

	Beijing	Shanghai
Female	250 (BJ-F)	190 (SH-F)
Male	250 (BJ-M)	290 (SH-M)



The images above show the gender and accent distribution, and from the pictures, it can be concluded that the first independent component corresponds to gender characteristics of a speaker as projections from this component almost separate all speakers into gender categories. In the next figures, four subsets occupy four blocks, and from that based on their findings, it is evident that the components have strong correlations with accents. Thus, from their research, they were able to establish the fact that both cross-accent experiments and speaker variability analysis show that accent is one of the most critical factors leading to fluctuating performance of ASR systems. Conclusively in their research, they were able to come up with the fact that yes speaker variability does affect speech recognition performance significantly as well as the fact that accent is one of the main factors that cause variability and should impact the recognition. The paper showed the fact that there was a 40-50% error increase from cross-accent speech recognition. Also, based on their experiment with PCA/ICA, they could qualitatively confirm the fact that accent is another dominant factor, adding to gender in speaker variability (Huang, Chen, & Chang, 2004). The second aspect of their research involved accent adaptation and identification. This aspect of the study is heavily linked to (Teixeira et al.). For this speech adaptation technique, the researchers

used data available to sort of train the recognizers for them to see a substantial difference between the accents and for them to be able to classify these accents.

Furthermore, (Huang, Chen, Li, Chang & Zhou), analyzed speaker variability. Speaker variability is an essential study in speech recognition as it highlights the various factors that affect the reliability and the interpretation which the speech recognition systems would provide based on the speech spoken to them. The paper highlights that gender and accent are significant factors in the analysis of speaker variability. For the conduction of their research, two powerful statistical multivariate analysis method was used. The plans include the principal component analysis (PCA) and the independent component analysis (ICA). The research yielded that, using ICA representation, they achieved about 6.1% and 13.3 error rates in gender and accent classification.

(Kat & Fung, 1999) Speaks about how the performance of a speech recognition would be degraded when the accent being presented to it is different from the training set. Furthermore, the purpose of this paper involved using a much faster approach to accent classification using phoneme-class model. The article also shows how the transformation of native accent pronunciation dictionary, can be changed to that for accented speech. The results from the research conducted show that the accent-adapted dictionary reduces recognition error rate by 13.5%. Kat and Fung used a mixture of feature-based and model-based discrimination, and since it was a small amount of data, they used phoneme-class HMMs. This phenome consists of six classes which include; stops, affricates, fricatives, nasals, semivowels & glides and finally vowels. The paper further uses various features to show the differences between speaking styles and structures of two languages. The first feature used was energy. The figure below shows the differences between the main accents used in the research based on the energy feature.

Figure 1: Average mean energy of various phone classes

Phone classes	American	Cantonese
vowels	1035.1	506.65
nasals	601.70	252.28
stops	147.74	55.63
affricates	370.66	89.87
fricatives	224.79	117.99
semi-vowels & glides	1282.49	522.93

Another feature that was considered in the paper was Fundamental frequency. Beneath this feature, it is found that human perception indicates that listeners base their accent classifications partly on prosodic features such as pitch movements, rhythms and pausing. It was another feature that was considered in the research, and the results between the measured accents are below.

Figure 4: The pitch contour of the same utterance spoke by Cantonese speaker(top) and English speaker(bottom)

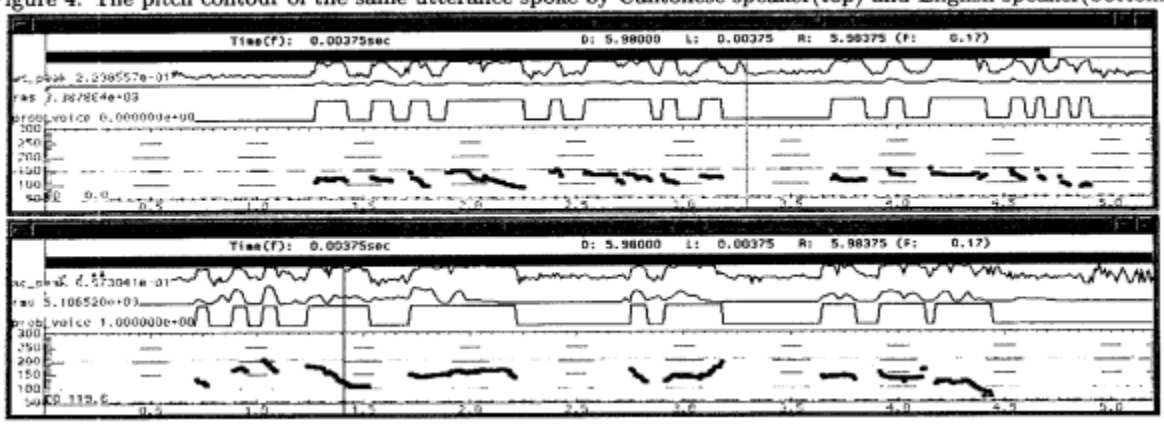


Figure 5: F0 information reduces classification error

parameter set	error rate
full set	14.52%
no dd(F0)	14.58%
no F0	14.65%
no d(F0)	14.67%
no F0 info.	15.33 %

The above images, the first shows the differences in pitch between the two accents used in the paper. The analysis done on the pitch per the research conducted shows that the reason for the higher pitch is that people tend to add their native or first language speaking tongue into foreign tongues. The second image shows the increase in error rate by 5.6% if the first derivatives are ignored and the second derivatives are masked.

(Benzeghiba et al., 2007) Research gives some review or briefing to the understanding of automatic speech recognition and speech variability. The paper speaks about the specific barriers to flexible solutions and user satisfaction under some circumstances. Several factors are related to this issue. From the article, it states that such factors include environment or weak representation of grammatical and semantic knowledge. Also, in the review, it indicates that there are other deficiencies in the dealings of variations that are naturally present in speech. These deficiencies are part of the many factors that affect speech realization to name a few. The factors such as the speaker, gender, speaking rate, vocal effort, regional accent, speaking style to name a few may are variations that may not be modeled well.

In their research, (H, F.j, & T, 2012) looked at accessing and investigating the best possible way to combine speech data from five South African accents to improve the performance of speech recognition. For this paper, since there is difficulty in dealing with multiple accents in under-resourced environments, there are complications in working with automatic speech recognizers. In the article, three acoustic modeling approaches are considered, they include; separate accent-specific models, accent-independent models obtained by pooling training data across accents, and multi-accent models. From the conclusion of their research, it shows that multi-accent models help in offering a mechanism whereby speech recognition can have its performance optimized

automatically as well as aid in hard decisions involving which data to pool and which separate to be avoided.

Thus, based on the research conducted by other scholars, it can be said that accent is a factor of speech variability which in turn affects the way or the means by which speech recognition systems can effectively interpret.

References

Teixeira, C., Trancoso, I., & Serralheiro, A. (1996). Accent identification. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP '96* (Vol. 3, pp. 1784–1787 vol.3). https://doi.org/10.1109/ICSLP.1996.607975

Huang, C., Chen, T., & Chang, E. (2004). Accent Issues in Large Vocabulary Continuous Speech Recognition. *International Journal of Speech Technology*, *7*(2), 141–153. https://doi.org/10.1023/B:IJST.0000017014.52972.1d

Wells, J. C. (1982). *Accents of English:* Cambridge University Press.

Yan, Q., & Vaseghi, S. (2002). A comparative analysis of UK and US English accents in recognition and synthesis (Vol. 1, p. I-413). https://doi.org/10.1109/ICASSP.2002.5743742

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., … Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication*, *49*(10), 763–786. https://doi.org/10.1016/j.specom.2007.02.006

Kat, L. W., & Fung, P. (1999). Fast accent identification and accented speech recognition. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)* (Vol. 1, pp. 221–224 vol.1). https://doi.org/10.1109/ICASSP.1999.758102

H, K., F.j, M. M., & T, N. (2012). Multi-accent acoustic modelling of South African English. https://doi.org/10.1016/j.specom.2012.01.008

Houde, R. (1979). Prospects for Automatic Recognition of Speech. *American Annals of the Deaf*, *124*(5), 568–572.

Automatic Speech Recognition – A Brief History of the Technology Development - Semantic Scholar. (n.d.). Retrieved November 7, 2018, from

https://www.semanticscholar.org/paper/Automatic-Speech-Recognition-%E2%80%93-A-Brief-History-of-Juang-Rabiner/1d199099a2f4f8749c7e10480b29f5adaecad4a1

Huang, Chao / Chen, Tao / Li, Stan / Chang, Eric / Zhou, Jianlai (2001): "Analysis of speaker variability", In *EUROSPEECH-2001, 1377-1380.*