A Preliminary Speech Learning Tool for Improvement of African English Accents

Benedict Oyo
Department of Informatics
Tshwane University of Technology
Pretoria, South Africa
benoyo@gmail.com

Billy Mathias Kalema
Department of Informatics
Tshwane University of Technology
Pretoria, South Africa
KalemaBM@tut.ac.za

***Abstract*—**Speech recognition systems emphasise: accent recognition, recognition system performance through calculation of word error rate (WER), pronunciation modelling, speech-based interactions (tone, pitch, volume, background noise, speaker's gender and age, speaking speed and quality of recording equipment) and speech database solutions. However, research into the use of speech recognition systems for improvement accents is scarcely available. In this paper, we focus on development of an speech recognition system for recognizing African English accents and enabling the speakers improve their English accents. This is achieved by using a dual speech recognition engine: the first, a multiple accent recogniser receives African English speech input, classifies it and sends to the second recogniser that evaluates the speech against standard English pronunciations. Speech deviations from standard English pronunciations are captured and read by the system as a way of supporting the learner to improve his/her reading proficiency. Preliminary tests indicate that terminologies that are rarely used in ordinary conversations (e.g. enthusiasm, exuberant, vague, etc) are most poorly pronounced irrespective of the educational level of the reader.

Keywords—speech recognition; African English; speaker clustering; acoustic model

I. INTRODUCTION

There are many variations in pronunciation of English words across the different African English speaking countries in general and further differences within the respective countries. Researchers maintain that those who speak English, as a second language, in Africa demonstrate unique characteristics based on their previous language experience, i.e., certain races, ethnic groups or nations, speak a specific version of African English which gradually receives recognition [8][9]. In the Republic of South Africa for instance, five major different English accents have been identified in the literature: Black South African English, Afrikaans English, Cape Flats English, White South African English and Indian South African English [4]. For each category, learners' written English is highly influenced by the native pronunciations of words that end up impacting on their pass rates. Similarly, the East African English accents are more varied and structured around ethnicity groups comprising of the Bantu, Luo, Nilotics, Nilosaharan and Hermites [6][8]. Literature shows that the situation is also similar in other African countries as has been reported in West Africa [8]. Given the scale of these variations in African English accents, oral communication outside these accent groups is not effective.

In contrast, computer scientists and engineers have built robust systems with varying capabilities for automatic speech recognition, multilingual speech recognition, multi dialectal speech recognition, tone recognition, phonetic transcriptions, pronunciation dictionaries, speaker adaptations and accent identification. What is not common in these systems is computer aided speech instruction for improving English accents where English is a second language.

The rest of this paper is structured under five sections. Section II presents issues of spoken English in Africa by discussing real-life cases of pronunciation variations involving insertions, omissions and substitutions of letter sounds. Section III builds on the insights from section II to articulate the problem of speech deviations in African English Accents. Section IV introduces the speech technology solution to the problem by providing an experimental design of the speech recognition process for supporting learners improve their English accents. Section V presents the speech learning tool developed by following the design guidelines in section IV. The last section discusses study findings and a road map for future work.

II. ISSUES OF SPOKEN ENGLISH IN AFRICA

Linguistic studies of African English prioritize written English and to a lesser extent spoken English. These studies maintain that although reasons for different African English pronunciations are numerous, mother tongue effect is the main cause for the variations in African English accents [8][9]. As a result, much of the research effort and resources are devoted to discovering and promoting African Englishes rather than seeking convergence through standardising African English pronunciations.

A. Pronunciation Variations

Words are distinguishable from each other by correct pronunciations. African English accents deviate to varying degrees from standard English accent due to insertions of non-existent sounds in words, omissions of existing sounds in words and substitutions of existing sounds with another sound.

ISBN: 978-1-4799-6248-8/14/$31.00 ©2014 IEEE

The following examples have been extracted from a number of African English accents:

1) Insertions: The native Acholi language speakers in northern Uganda insert /h/ sound in words that begin with vowels, i.e, *air* is pronounced as *hair*, *ear* as *hear*, *is* as *his*, *our* as *hour* and *up* as *hup*. As depicted by these examples, variations in pronunciations can distort the meaning of pronounced words. Interestingly, the same native Acholi speakers also add /h/ sound in words that contain /sa/, /se/, /si/, /so/, /su/ sounds, e.g., *sake* is pronounced as *shake*, *sell* as *shell*, *sip* as *ship* and *soap* as *shop*. These pronunciation variations become more complex and interesting when other cases of omissions and substitutions are considered concurrently in real-life conversations.

2) Omissions: Following from the previous scenarios the same native Acholi speakers omit /h/ sound where it exists, e.g., *hat*e is pronounced as *ate*, *hear* as *ear*, *fish* as *fis*, *shop* as *sop/soap* and *shut* as *sat*. Comparing with insertion cases, some pronunciations are interchangeable making them very confusing to human listeners let alone to machines. For instance 'air' is pronounced as 'hair' and vice versa, similarly, *hear* and *ear*, *soap* and *shop*.

3) Substitutions: Focusing in other parts of Africa, the native Yoruba language speakers in South Western Nigeria tend to replace the /a/ sound with /o/ sound and again this can be very confusing when the deviated pronunciation corresponds with an established pronunciation such as: *bus* and *boss* or *lack* and *lock*, being pronounced the same way. Similar sound replacements are common among the South African countries like South Africa where /a/ sound is replaced by the /e/ sound, i.e., *land* and *lend* or *sand* and *send* are pronounced the same way. Broader pronunciation variations exist with most speakers of English as a second language in Africa which can be mitigated through speech recognition and speech correction systems.

B. Causes of Speech Variations

Studies on causes of speech variations show that an individual's original language or mother tongue affects how he/she speaks another language acquired later [8][9]. Other causes are biologically, i.e. speech variations occur due to ways in which the speaker uses different parts of the vocal tract, including lips, teeth, tongue (in terms of the tip, blade, body or root) or nasal cavity. Different languages may rely more heavily on a particular part of the vocal tract in comparison to other languages, hence affecting how a second language acquired later will be spoken. Mbogho and Katz [7] classify the Afrikaans language in South Africa as more of a guttural language (using the pharynx wall or the larynx) than English, which influences how an Afrikaans speaker pronounces English words. Other studies in phonology and phonetics reveal that speech constructions begin in the mind whereby the speaker constructs a phrase/sentence by choosing a collection of finite mutually exclusive sounds which are subsequently produced by the mouth [12]. Therefore by training the mind with correct pronunciations as is the case in this study, it seems possible to improve human accents.

C. Lexicon

The vocabulary of African Englishes comprises of the vocabulary of the standard English and specific African words that reflect the cultural and political heritage of Africa. The geographical range of African lexemes varies a lot and an in-depth analysis is beyond the scope of this study. Since the focus of this study is on improving African English accents, key words that have been incorporated into general English and codified in large English dictionaries such as the Oxford English dictionary, cannot be ignored.

The speech recognition engine developed in this study and discussed in section V has been trained to recognize standard English lexicon as well African English lexemes. More specifically, lexemes from East African English used in training data include: *askari, bwana, chai, safari, harambee, matooke, mwalimu, and ugali.* From South African English, a*masi, bakkie, braai, eina, gogo, khaya, and pap*, have been used. *Buckra, djembe, okra,* and *juju* have been used from West African English.

III. Problem

Linguistic researchers and speech system developers have shown that an individual's original language or mother tongue affects how he/she speaks another language acquired later [7][8][9]. This is particularly true for Africa where English is generally spoken as a second language. As a result, the current English accents across the entire English speaking African countries, deviates from the standard English accent, making African English pronunciations poor. For instance, listening to one of the West African English accent for the sentence "I left the rubber with our guard and will be coming to Sun city church by bus," which could sound as, "I left the *robber* with our *god* and will be *koming* to *Son* city *choch* by *boss,*" would be meaningless to listeners without experience in that accent. There is need therefore, to develop a speech recognition system that recognises and corrects poor English pronunciations thereby enabling better spoken English among African speakers.

IV. Speech technology solution

Dewey [2] highlights challenges in reading and speaking English because of its complex and unpredictable spelling system requiring some kind of dictionary to tell how to pronounce words whose spellings are known, i.e., English has 13.7 different spellings per sound, 3.5 sounds per letter and some letters having no sounds in words such as debt, budget and listen. Speech recognition research uses acoustic modelling (largely dominated by Hidden Markov Models) to deal with the fore mentioned challenge. An Hidden Markov Model (HMM) breaks speech sounds for each word into phonemes (smallest element of a sound) and using a software language model, compares the phonemes to words in the pronunciation dictionary [11].

Speech recognition systems for multiple accents have over the years been developed using two approaches: (i) by building different sets of acoustic models for each accent and (ii) by training a single accent independent acoustic model using pooled accent specific data from corresponding accents. Each of these approaches has its advantages. For instance, Humphries and Woodland [3] provide insights based on the former approach for building speech recognition systems for a wider range of speaker accents. They present a novel approach that uses pronunciation modelling for the synthesis of accent-specific pronuciation dictionaries directly from acoustic data.

Their research is motivated by the fact that electronically available pronunciation dictionaries for specific accents often do not exist and would be time consuming and expensive to build from scratch. On the other hand, speaker-independent speech recognition techniques such as the speaker clustering seem to be more convenient in dealing with multiple accent recognition compared to to accent-specific (speaker-dependent) systems.

In speaker clustering techniques, all speakers in the training data are clustered into speaker classes independent of the test speakers in the training step, while in the recognition step, the most appropriate speaker class model is selected utterance by utterance and used for recognition [5]. In the context of the present work, the speaker clustering method is more suitable as it reduces the complexity in building and integrating accent-specific regonisers.

Performance of the speech recogniser is one of the key concerns when building speech recognition systems. Performance levels vary depending on the functions of the recogniser. For instance, isolated words and command recognition systems have a higher performance due to finite word boundaries compared to continuous speech recognition systems without clear word boundaries [3][7]. Furthermore, accent-specific recognisers are known to perform better than accent-independent recognisers due to decreased speaker variation and speaking styles in the former [3][4]. In the context of the present work, a large data set of training data is used to reduce the effect of speaker variation on performance.

A. *Experimental Design*

Experiments in this study use two HTK-based [11] speech recognisers as shown in Fig. 1. The learner can provide his/her own words or sentences to read, or choose from available system words or sentences. Either way, the speech input is transferred to a speaker independent African English recogniser. If the speech is not recognized, i.e. because is not a known African English accent or is very poorly produced, then it can be discarded in any of the following ways: discard and try again, discard and quit, and, invoke system support and discard. These options are shown by the three arrows from "speech discard options" step in Fig. 1.

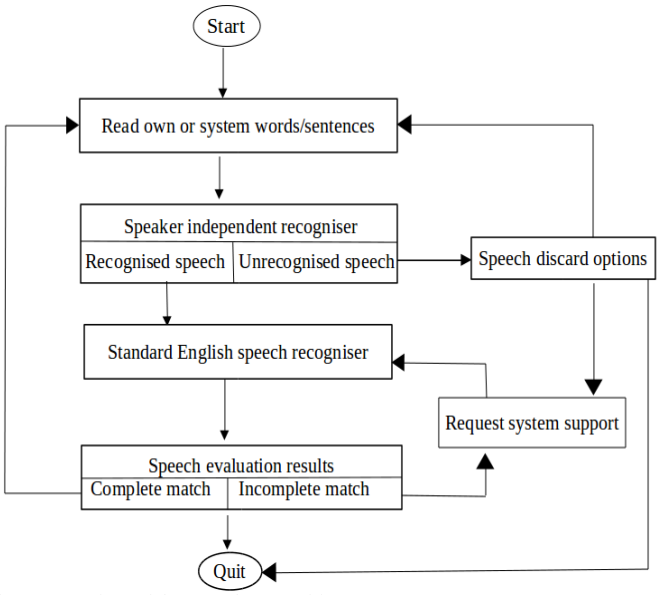

Fig. 1. Overview of the speech recognition process

On the other hand, the recognized speech is sent to the second recogniser that evaluates the speech against standard English pronunciations. At this stage, if the input speech is completely matched with the standard English pronunciations, then the learner can repeat the process by reading new words, otherwise, the unmatched speech is captured and read by the system through the "request system support" step in Fig. 1. This ensures that the system evaluates and supports learners' to improve their reading skills.

B. *Data Set*

Speech data for the two recognisers have been collected in tandem with the functions of the recognisers. As for the speaker independent recogniser, representative speech data has been collected from Uganda, Nigeria and South Africa. The data set for Uganda is so far the largest with 110400 utterances from two speaker groups (accent groups) of 60 speakers each. The Nigerian data set consists of 35 speakers of the same accent group with 32200 utterances. The South African data collection is still ongoing. The utterances were recorded at 48kHz sampling rate and 16 bits per sample, which is the ideal settings for the computer system used. Sound activation level was set at -60 dB which is the normal conversation setting. The recording software used is Audacity 2.0.5 for Linux [14].

The data set for the second recogniser consists of 28 speakers with a total of 28440 utterances. These speakers are carefully selected English language teachers in order to build a standard British English recogniser.

V. THE TOOL

The preliminary speech learning tool presented here is an implementation of the system design depicted in Fig. 1. The corresponding interface is developed using *Java Programming Language* and Java library for hidden markov models (Jahmm) [13], while the speech recognition engine is based on HTK. The name of the tool is *AAcents*, an acronym for African Accents with its logo fixed at the left side of the title bar as shown in Fig. 2. The interface has three menu options (*File, Accents and Admin)*. System words and sentences can be inserted/uploaded, edited or deleted from the *Admin* menu. In the current setup, 237387 words adopted from the British English Example Pronunciation (BEEP) dictionary have been uploaded. Sentences uploaded from the *Admin* section fall under six categories: historical information, scientific information, geographical information, jokes, biblical information and political information. This is meant to serve the different learners interests and ultimately maximize the tool for improving spoken English, moreover, other categories can be provided at the discretion of the learner. In addition, learners optionally register their names from the launch interface such that the new words or sentences provided by them are stored under their registered names. As shown in Fig. 2, a learner selects a sentence of choice from the preferred category and on clicking the read sentence button, a speech recorder is invoked. The temporally recorded speech is then transferred to the speaker independent recogniser for further processing as already discussed in the previous section.

ISBN: 978-1-4799-6248-8/14/$31.00 ©2014 IEEE

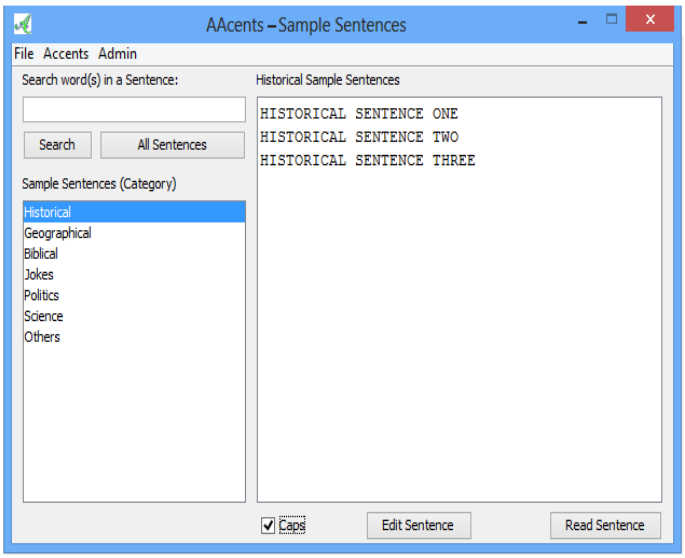

Fig. 2. A snapshot of user interface for reading system sentences

The *edit sentence* option in Fig. 2 allows the learner to modify an existing sentence into a new one and save or simply modify for the underlying reading session and discard. The sentence categories constitute a knowledge base that is not only meant for learners to quickly test their reading proficiencies, but also enable them to acquire new knowledge. As such, the content in this section is critical and yet to be completed. The rest of the functions can be explored when using the tool.

A. Benefits of the Tool

The tool is intended to provide the following benefits:

- Enable learners flexibly improve their English reading skills through self evaluation using system words/sentences or their own words/sentences.
- Enable teachers to electronically evaluate their learners reading skills by building a collection of desired sentences for testing the learners.
- Facilitate learners exposure to new words and how they are read thereby supporting learners to improve their English vocabulary.
- Support learners to improve their written English as a result of more exposure to both written and spoken English
- Boost the learner's confidence in public speaking through submitting written speeches and practice reading on the computer prior to actual event.
- Improve general knowledge while practicing speech using information from the knowledge base comprising of historical, geographical, biblical, political and scientific information.

VI. DISCUSSION AND CONCLUSION

We have presented a study on improvement of African English accents using a dual recogniser. The first one is a speaker independent recogniser for African English accents. In the current state, this recogniser has been trained on two Ugandan English accents, with each accent group having 60 speakers and 55200 utterances. Further experiments are ongoing for phone set extensions and augmentation of the pronunciation dictionary with accent specific units and their relevant probabilities for Nigerian and South African English accents.

The second is a standard British English recogniser trained with 28 speakers (English language teachers) and a total of 38440 utterances. Simple Gaussian models (mean and variance) were sufficient for the second recogniser since all speakers pronounced a given phoneme in very similar ways unlike the speaker independent recogniser which was trained with significant variations for phoneme pronunciations. Indeed unfamiliar words or rather rarely used English terminologies in the training data set were pronounced differently by most of the speakers. Furthermore, preliminary tests indicate that terminologies that are rarely used in ordinary conversations (e.g. enthusiasm, exuberant, vague, etc) are most poorly pronounced irrespective of the educational level of the reader. Fig. 3 gives a voiced speech linear frequency using hamming window function for pronunciations of the word "enthusiasm" by four randomly selected speakers in the training data. The frequency analysis in Fig. 3 should have been a close match if the pronunciations were similar. However, the significant deviations as can be seen in Fig. 3 attest to the fact that the four speakers read the same word differently.

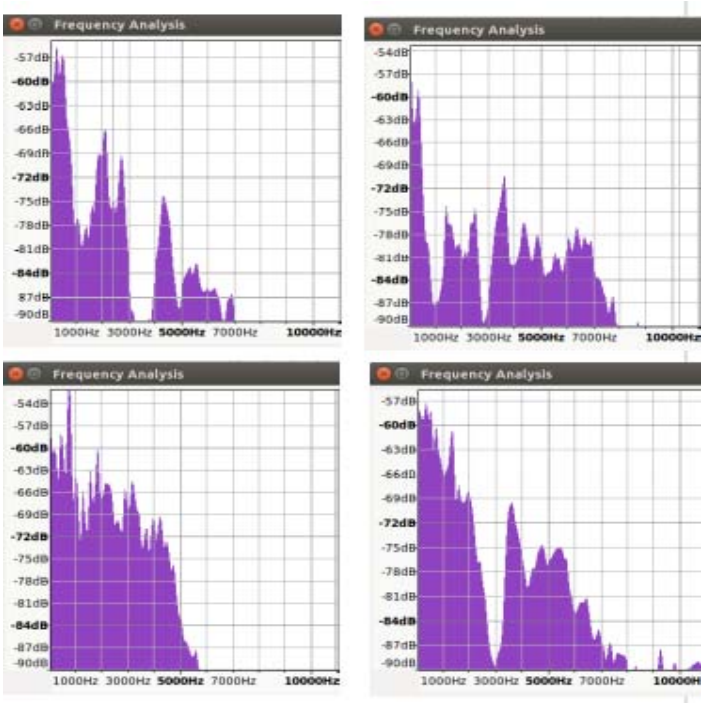

Fig. 3. A voiced speech linear frequency spectra using hamming window function

Interestingly, by listening to the actual speech corresponding with the spectral analysis in Fig. 3, it emerged that non of the speakers pronounced the word "enthusiasm" correctly. More specifically, the speakers reported here are fresh university graduates who by virtue of their educational background, i.e. having been instructed in English throughout their educational life time, are expected to be more fluent in

English. This further underscores the need for a speech learning tool for improvement of spoken English by native Africans irrespective of the educational background of the learners.

For the convenience of the learner, the tool presented in this paper allows for new sentences to be uploaded and used for speech practice. By grouping the sentences into six categories including: historical information, scientific information, geographical information, jokes, biblical information and political information, the tool impacts on the learner's knowledge and speech skills simultaneously. Furthermore, learners can add a new categories but cannot modify or delete the existing categories which are considered benchmarks standardised for the speech learning tool presented in this paper.

As more speech data for other African English accents are obtained, extensions will only be made on the speaker independent recogniser and not the standard British English recogniser. When this occurs, the final recogniser performance tests, e.g. word error rate, phone error rate and position independent word error rate, will be done. Fortunately, these tests will not require changes on the user interface and hence the functions of the tool as presented in this paper are binding.

ACKNOWLEDGMENT

The authors would like to thank M. Kaye of Gulu University for supporting the development of the user interface.

REFERENCES

[1] P. Fung and Y. Liu, "Effects and modeling of phonetic and acoustic confusions in accented speech recognition," Journal of the Acoustical Society of America, vol.118, issue 5, pp. 3279 -3293, 2005.

[2] G. Dewey, English Spelling: Roadblock to Reading. London: Teachers College Press, 1971.

[3] J.J. Humphries and P.C. Woodland, "The use of accent-specific pronunciation dictionaries in acoustic model training," Proceedings of ICASSP, Seattle, United States, pp. 317-320, 1998.

[4] H. Kamper and T.R. Niesler, "The impact of accent identification errors on speech recognition in South African English," South African Journal of Science, vol. 110, issue 1/2, 2014.

[5] K. Konno, M. Kato and T. Kosaka, "Speech recognition with large-scale speaker-class-based acoustic modeling," in Proc. of IEEE Signal and Information Processing Association Annual Summit and Conference, 2013.

[6] M.M. Michieka, "English in Kenya: a sociolinguistic profile," World Englishes, vol. 24, issue 2, pp.173–186, 2005.

[7] A. Mbogho and M. Katz, "The impact of accents on automatic recognition of South African English speech: a preliminary investigation," in Proc. of ACM, SAICSIT, pp. 187-192., 2010.

[8] M. Mutonya. "African Englishes: acoustic analysis of vowels," World Englishes, vol. 27, issue 3, pp. 434-449, 2008.

[9] J. Schmied. East African Englishes. In B.B. Kachru, Y. Kachru and C.L. Nelson (Eds.), *Handbook of World Englishes* (pp. 188-202). London: Blackwell Publishing, 2006.

[10] D. Vergyri, L. Lamel, J.L. Gauvain, "Automatic speech recognition of multiple accented English data," in INTERSPEECH, 2010, 1652-1655.

[11] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, The HTK Book, Cambridge University Engineering Department, 2006.

[12] C. Hulme and M.J. Snowling. "The interface between spoken and written language: developmental disorders," Phil. Trans. Roy. Soc. London, vol. B 369, 2013.

[13] URL:https://code.google.com/p/jahmm/

[14] URL:http://audacity.sourceforge.net/