TMLAI TRANSACTIONS ON MACHINE LEARNING AND ARTIFICIAL INTELLIGENCE
SOCIETY FOR SCIENCE AND EDUCATION UNITED KINGDOM
VOLUME 2, ISSUE 3
ISSN: 2054 - 7390

Engineering Analysis and Recognition of Nigerian English: An Insight into Low Resource Languages

Sulyman A. Y. Amuda[1], Hynek Boril[2], Abhijeet Sangwan[2], John H. L. Hansen[2] and Tunji S. Ibiyemi[1]

[1]*Electrical & Electronics Engineering Department, University of Ilorin, Ilorin, Nigeria.*

[2]*Center for Robust Speech Systems, University of Texas, Dallas, USA.*

amudasulyman@gmail.com, hynek@utdallas.edu, sangwan@utdallas.edu, john.hansen@utdallas.edu, ibiyemits@yahoo.com

ABSTRACT

A comparative analysis between Nigerian English (NE) and American English (AE) is presented in this article. The study is aimed at highlighting differences in the speech parameters, and how they influence speech processing and automatic speech recognition (ASR). The UILSpeech corpus of Nigerian-Accented English isolated word recordings, read speech utterances, and video recordings are used as a reference for Nigerian English. The corpus captures the linguistic diversity of Nigeria with data collected from native speakers of Hausa, Igbo, and Yoruba languages. The UILSpeech corpus is intended to provide a unique opportunity for application and expansion of speech processing techniques to a limited resource language dialect. The acoustic-phonetic differences between American English (AE) and Nigerian English (NE) are studied in terms of pronunciation variations, vowel locations in the formant space, mean fundamental frequency, and phone model distances in the acoustic space, as well as through visual speech analysis of the speakers' articulators. A strong impact of the AE–NE acoustic mismatch on ASR is observed. A combination of model adaptation and extension of the AE lexicon for newly established NE pronunciation variants is shown to substantially improve performance of the AE-trained ASR system in the new NE task. This study is a part of the pioneering efforts towards incorporating speech technology in Nigerian English and is intended to provide a development basis for other low resource language dialects and languages.

Index Terms— Nigerian English, Limited Resource Language, Automatic Speech Recognition (ASR)

DOI: 10.14738/tmlai.24.334
Publication Date: 28th August 2014
URL: http://dx.doi.org/10.14738/tmlai.24.334

1 INTRODUCTION

English is spoken by about 130 million people in Nigeria as an official and also a colloquial language. Its unique linguistic characteristics constitute Nigerian English (NE) as a dialect of English. In spite of numerous experimental and instrumental studies of NE, so far a little attention has been paid to building viable speech processing technology for NE or even assessing the dialect-specific speech features from the system engineering perspective. This work presents the first of its kind audio-visual Nigerian English Corpus which consists of 45 hours of speech collected from approximately 530 speakers.

A comparative analysis of Nigerian English alongside with American English (AE) is presented. The two language dialects are closely related in terms of vocabulary but the apparent pronunciation and word choice differences prevent effective application of AE speech technology to NE environments. Hence, our focus is on identifying the major sources of mismatch between the two dialects from the perspective of acoustic signal modeling, evaluation of their impact on automatic speech recognition (ASR) performance, and proposal of an affordable strategy that will allow for a rapid adaptation of an existing ASR engine towards the target domain of the low resource dialect.

NE differs from its counterparts in terms of tones, prosody, phones, and unique lexical patterns that portray the influence of local Nigerian languages [1], [2], [3]. Analyses of speech rhythm and tonal and syllable structures of NE have revealed the tonal nature of the language. Particularly, the pitch employed by NE speakers is lexically significant, contractive, and relative. Additionally, NE is syllable timed and tends to suppress vowel contrast. The observed characteristics in Nigerian English are closely linked to the influence of the major local languages such as Yoruba, Igbo and Hausa. Besides prosodical differences, Nigerian English is also characterized by phonetic differences. The phonetic differences are more obvious when speakers encounter unfamiliar phones that are otherwise absent in their native language [3], often resulting in phone deletion, insertion, or omission. For example, Nigerian English speakers will introduce an unglided vowel structure and unnecessary nasalization of sounds when pronouncing unfamiliar phones while in other cases they may omit phones that are absent in their native language [2, 3].

This study analyzes acoustic-phonetic differences between AE and NE on the level of pronunciation variations, vowel locations in the formant space, mean utterance fundamental frequency, and distances between AE-trained acoustic models and models adapted to NE [4]. It is shown that the AE–NE acoustic mismatch has a strong impact on ASR. In the initial effort towards NE ASR, a combination of model adaptation and extension of an AE lexicon for the newly established NE pronunciation variants is proposed and shown to substantially improve performance of the AE-trained ASR system in the NE task. The results presented here highlight

the challenges brought forth by Nigerian English and are intended to motivate future development of speech systems for limited resource language dialects and languages.

1.1 The Challenges of Nigerian English

Most research works on NE proved that there exist some common phonological properties that can be used to identify NE despite its sub-varieties, hence this work adopts the principles of these properties [5,6,7]. This is based on the concessions of different research works that attempt to explain the accent variability as a direct result of sub-varieties of NE. Some researchers attributed the accent variability to be as a result of different ethnic groups while others believed this to be due to the education background and or the language function or purpose of usage of the language. Another perception for the variability is the influence of the first language on the second language [8] as English language is a second language to all Nigerians.

The common view of this baseline, which is more acceptable by these researchers, is that the standard NE is associated with the minimum of university education and that there is tendency of the speakers' local language to influence the NE in terms of speech rhythm, intonation and accentuation. The phonological aspect of the influence was critically examined by Ulrike Gut (2004) where the consequent effect of the three major languages phonemes on NE is well depicted with respect to British English. The submission is that, the local languages show some varying effects on NE or but there exists some common basis through which the NE clearly differs from the British English. Though Gut did not distinctively define this in terms of phonemes on general terms for NE, or show the common NE phonemic features across the major local languages, all the same it gives a lot of insight into the phonological challenge of the NE. Titi Ufomata (1995) gave an overview of the general Nigerian English phonology as compared with the British English (termed as Received Pronunciation), where pronunciation of some vowels and consonants in NE are undifferentiated and most times confusable or even in some instances totally different from received pronunciation. Some examples given include: "/u/" and "/ʊ/" are both pronounced as "[u]" such that *full* and *fool* are pronounced as [ful]: "/i/" and "/I/" are both pronounced as [i] such that *bead* and *bid* are pronounced as [bid]: "/a/" and "/ae/" are both pronounced as [a] such that *bard* and *bad* are pronounced as [bad]: While /ei/ and /au/ are monophthongized to [e] and [o] respectively, also the fricatives [θ] and [ɗ] are pronounced as [t] and [d] respectively, so as [ʃ] is often pronounced as [ʧ]. Based on these analyses, it was established that NE varied from other English languages in terms of stress, intonation and rhythm.

In spite of the depth of the research on NE, the establishment of NE phonemes is yet to be available. AE therefore, present a good new specimen to apply and expand the speech processing techniques as means of improving ASR systems for limited resources languages based on analysis of certain parameter of speech. The use of AE instead of British English is also justified with the recent steady improvement in relationship between Nigeria and USA. The

influence of the AE on NE is becoming more noticeable in Nigerian academic, social and political circles.

1.2 University of Ilorin Speech (UILSpeech) CORPUS

The UILSpeech corpus was collected as a pioneering effort to form a database for Nigerian English. The speech data in the UILSpeech corpus were exclusively collected at the University of Ilorin campus. Speakers were mostly undergraduate students with an average age of 20 years. The corpus consists of speech from about 300 males and females each. The speaker pool reflects the linguistic diversity of Nigerian English, as most speakers tended to be from 3 dominant linguistic backgrounds in Nigeria, namely, native speakers of Yoruba (South-Western Nigeria), Igbo (South-Eastern Nigeria), and Hausa (Northern Nigeria).

The UILSpeech corpus consists of isolated word recordings as well as continuous read speech data. The isolated word data were collected in a laboratory with the use of a hollow-shaped telephone mouth-piece. The mouth-piece was intended to help reduce speaker-induced variability while ensuring the posture of the speaker. The continuous sentences were recorded with a video camera with the image object distance set between 20 cm to 80 cm. The video data were recorded with a 6.0 mega pixel digital camera, with 640 x 480 resolution and frame rate of 30 frames/sec. Since most data in the corpus were collected in an office/laboratory environment, a low-level background noise is present in the speech utterances. The recorded speech data are

sampled at the rate of 8 KHz for the entire corpus. It is worth mentioning that the speakers were encouraged to speak in a natural manner, and sufficient breaks were given between recording sessions to ensure data quality. Furthermore, the speakers were also subjected to a listener quality evaluation where all speakers in the corpus scored a minimum of 80, 4, and 60 on the Diagnostic Rhyme Test (DTR), Mean Opinion Score (MOS) and Diagnostic Acceptable Measure (DAM), respectively [9, 10].

The isolated word recordings consist of 5 repetitions of 30 different words spoken by 30 different speakers of Nigerian English. A short pause is present between the word repetitions to ensure accurate end-point detection by human annotators and machines alike. The continuous read speech data consist of short utterances spoken by about 500 speakers. The utterances are about 5-15 words long with an average duration of 7.5 seconds. In this manner, the corpus consists of about 15,000 speech utterances in total. Additionally, the continuous speech recordings are also accompanied by a synchronous parallel video recording.

1.3 Dictionaries

Two different dictionaries were used to represent the pronunciations of an American English and Nigerian English. TIMITDICT is used for the AE while, the NE dictionary was developed based on the phonetic transcription of NE by phonetic specialists based on extension of AE

lexicon (this is later referred to as 'NE + NE' lexicon). Consistency and good representation were ensured over the wide range of the data by quantitative corroboration of intra and inter transcription results from the same sets of specialists. The developed NE dictionary only covers the words that were used in this research work whereas the TIMITDICT has over 300, 000 entries.

2 ACOUSTIC-PHONETIC AND SPECTRAL ANALYSIS

In this section, acoustic-phonetic and spectral differences between American English (AE) and Nigerian English (NE) were analyzed along with the impact of the AE–NE acoustic mismatch on ASR. In this study, all NE experiments were conducted on the isolated words portion of the UILSpeech corpus. In particular, 898 utterances from 20 females and 21 males capturing a total of 4490 words formed the NE experimental set. The AE data set were taken from the TIMIT database [11]. TIMIT consists of read speech utterances drawn from 630 speakers of AE (belonging to eight major dialects regions). The TIMIT subset used in the following experiments contains 136 female and 326 male sessions.

Table 1. Example of pronunciation differences in American (AE) and Nigerian (NE) English [4].

Ortographic Transcription	Phonetic Transcription	
	AE	NE
And	/and ae n d/	/ae n t/
Automation	/ao t ah m ey sh ah n/	/ao t ax m ey sh ix n/ /ao t ax m eh sh ix n/
Department	/d ah p aa r t m ah n t/	/d iy p ae t m eh n t/
Electrical	/ax l eh k t r ih k el/	/ax l eh k t r ih k ax l/ /ax l eh t r ih k ax l/ /ax l eh t r ih k ax/
Faculty	/f ae k el t iy/	/f ah k ax l t iy/
Laboratory	/l ae b r ix t ao r iy/	/l ah b ax r ix t r iy/
Numer	/n ah m b axr/	/n uh m b ax/
Zero	/z iy r ow/	/z eh r ax/

2.1 Fundamental Frequency Analysis

The fundamental frequency of speech (F_0) is known to be affected by stress [19, 20], emotions [19, 21], and talking styles [22]. Different languages may exhibit unique F_0 characteristics [23] and the same may be observed also for individual dialects of a language [24]. This motivates the comparative analysis of F_0 in the AE and NE recordings performed in this section. WaveSurfer [16] is used to extract F_0 tracks from the AE and NE utterances. Figure 1 summarizes the mean utterance F_0 values per each dialect ('*AE/NE - All*') followed by gender-specific values ('*AE/NE - Males/Females*'). The error bars represent 95% confidence intervals. It can be seen than in

overall, the NE speakers tend to produce higher-pitched speech compared to AE speakers - the trend being consistent for both genders.

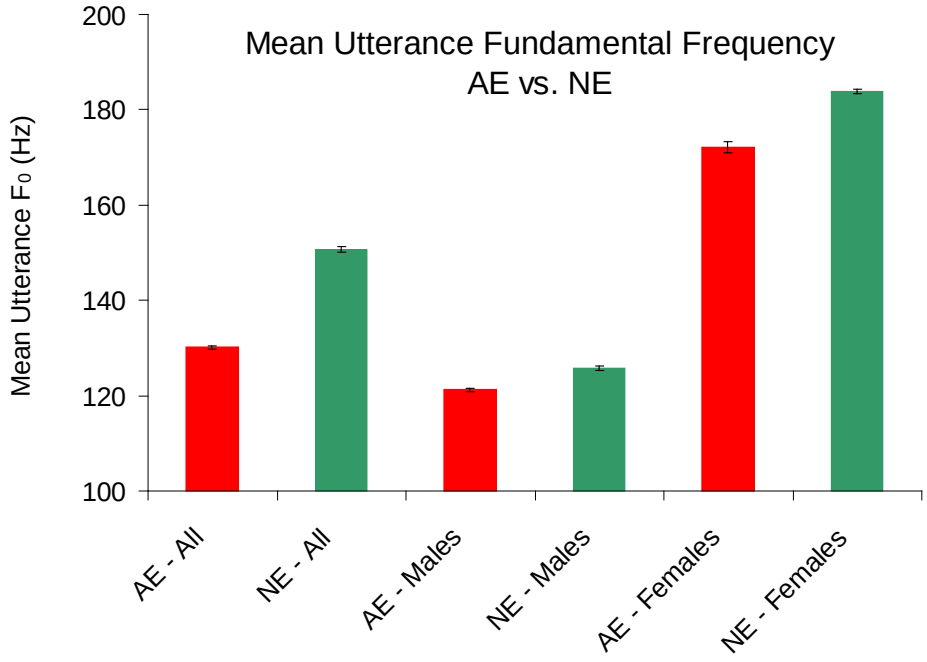
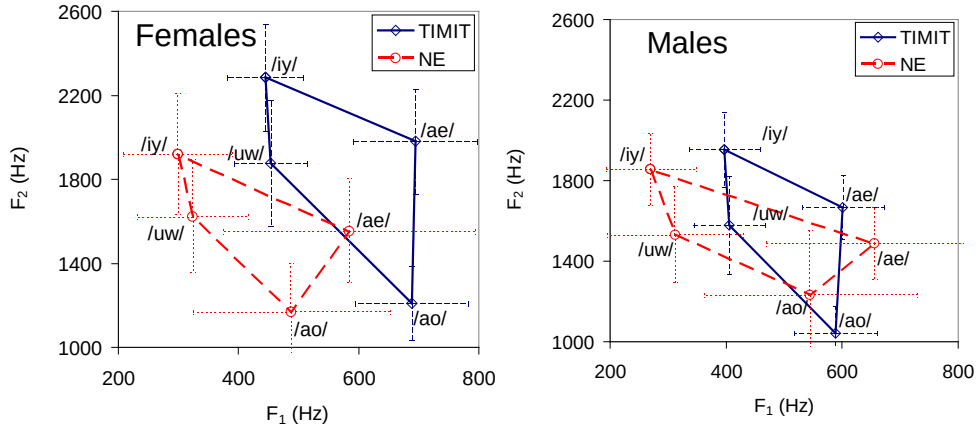


Fig. 1 Comparison of mean utterance fundamental frequency (F_0) in AE and NE recordings.

2.2 Formant Analysis

Past studies of the two languages suggest that there is a phonetic and acoustic mismatch in the AE and NE pronunciations of identical words and phonemes. To better understand the acoustic-phonetic mismatch in the AE and NE data, the locations of vowels in the F_1–F_2 (first and second formant) space are analyzed. Formant frequencies in individual phones are estimated by combining the output of formant tracking (WaveSurfer [16]) and the phone boundaries obtained from forced alignment. In the AE case, the AE lexicon was used in the forced alignment while the NE alignment utilized the 'AE+NE' lexicon. Gender dependent vowel analysis was conducted on the training data sets and the results are shown in Figure 2.

The error bars in the plots represent standard deviations of the F_1, F_2 sample distributions. Compared to native speakers of AE, both F_1 and F_2 vowel coordinates tend to be lower in the NE subjects. This suggests that the NE speakers produce vowels relatively further back and higher as F_1 varies inversely with tongue height and F_2 varies with the posterior-anterior dimension of the vowel articulation [17].



(a) (b)

Fig. 1 Comparison of the phone space of vowels /iy/, /uw/, /ao/, and /ae/ of American English (AE, TIMIT) and Nigerian English (NE) for both (a) female and (b) male [4]

2.3 Inter-HMM Distance Analysis

To further understand the ASR deterioration due to the dialect mismatch, and to get a more detailed insight about the similarities or confusions between the AE and NE phone sets, it is useful to compare the phone spaces of Nigerian and American English in terms of the learned HMM models. For this purpose, we utilize the KL-divergence measurement algorithm proposed in [18] to compute the distances between the baseline AE HMMs and adapted NE HMMs. Fig. 3 shows the KL-divergence between AE and NE vowel and consonant pairs. The articulation characteristics of /ax/ and /ix/ are found to be the closest to each other among AE and NE vowels. On the other hand, the vowels /aw/, /er/, /ay/, /ey/, and /oy/ seem to be the most unfamiliar vowels/diphthongs to NE speakers. The KL-Divergence between every AE and NE HMM pair is shown in Fig. 4. It can be seen that all adapted NE vowel HMMs tend to be closer to the AE /ax/ and /ix/ HMMs. This tendency could be a result of vowel substitutions employed by Nigerian speakers whenever a non-canonical vowel is encountered or when a canonical vowel is encountered in an unfamiliar syllabic position (here, non-canonical vowels refer to the vowels that are native to AE speakers but foreign to NE speakers). For example, the NE vowel /ah/ is close to AE /ah/, /ax/, /eh/, /ih/, /ix/, and /uh/. Here, it is possible that (i) /ah/ in NE is acoustically close to its AE counterpart, as well as (ii) NE speakers tend to substitute the usage of /ah/ with /eh/, /ax/, /ih/, /ix/, and /uh/ in some words. Similar observations can be made for other NE vowels as well, namely, /eh/, /ih/, /iy/, /uh/, and /uw/. For example, as seen in Fig. 4, /ay/ in NE seems to be substituted very frequently by phones /ih/ or /ix/.

Among the NE and AE consonants, /zh/ and /em/ show the largest mismatch, indicating an absence of these phones in NE or a large acoustic mismatch in the speech production. To a lesser degree, fricatives /s/ and /sh/ as well as affricatives /jh/ and /ch/ show a significant mismatch. In general, the acoustic space of the other NE and AE consonants seem to be well matched. However, significant substitutions are indicated among consonants based on the observed distance relationships.

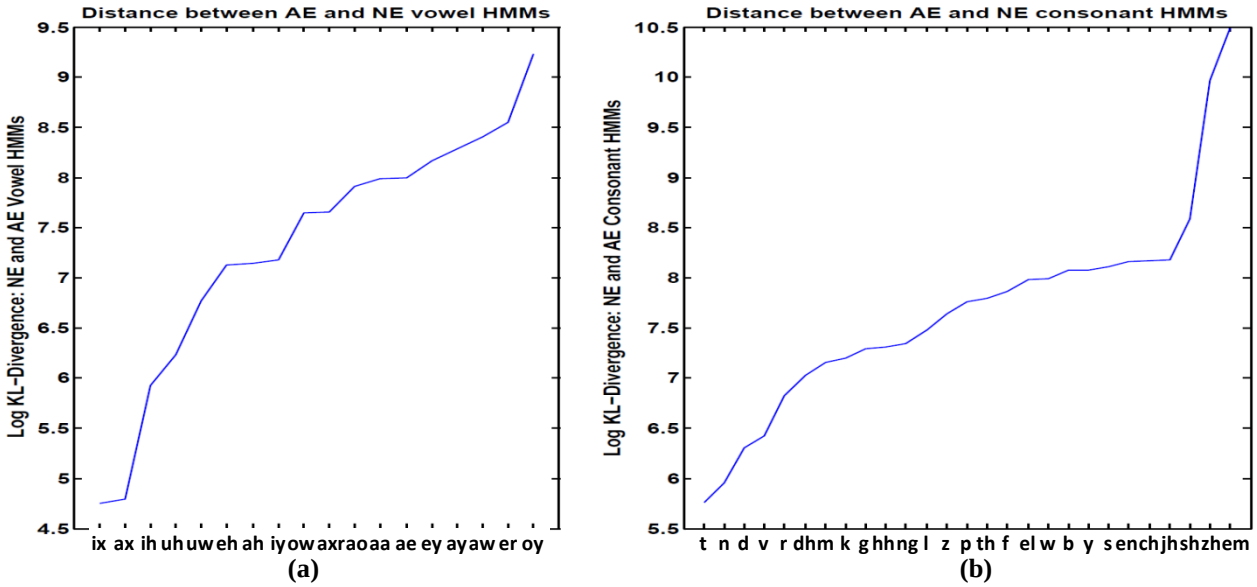


Fig. 3. Comparison of the KL-Divergence for vowels and consonants between (a) Corresponding Nigerian English (NE) and (b) American English (AE) HMMs [4]

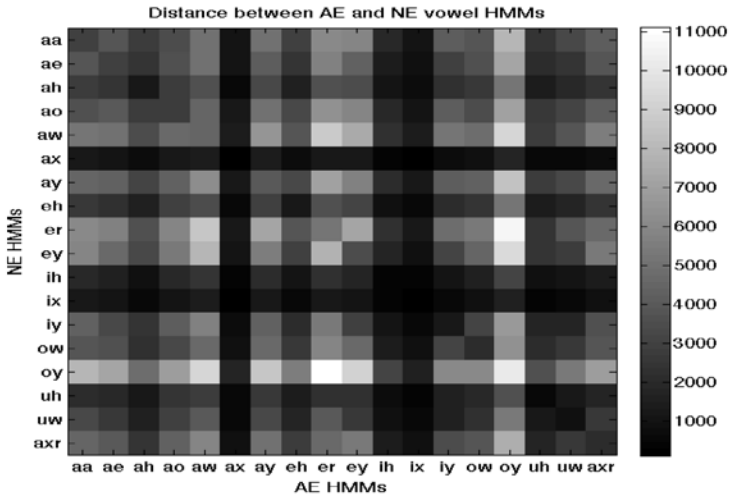


Fig. 4. Histogram-matching for vowels using KL-Divergence between corresponding Nigerian English (NE) and American English (AE) HMMs [4]

2.4 Focused Analysis

Other stimulus parameters of the two languages are studied through a focused analysis of selected word samples. Past research suggests that duration is an essential feature factor in the perception of different accents [11, 12]. Arslan and Hansen used the word final stop closure to carry out analysis of accent classification, with focus on the event before and after the stop consonant in a word. The approach proves to be effective in identifying accent-salient segments in the speech signal. This approach is adopted in our study for words that contain vowels preceding and succeeding a stop consonant, e.g., student, prudent, ardent, apart, etc. Our analysis suggests that in such words the NE speakers tend to spend longer time (put more emphasis) on the vowel after the consonant compared to AE speakers. Example spectrograms of the word 'student' from an NE speaker and an AE speaker are shown in Figure 5 (a) and (b), respectively.

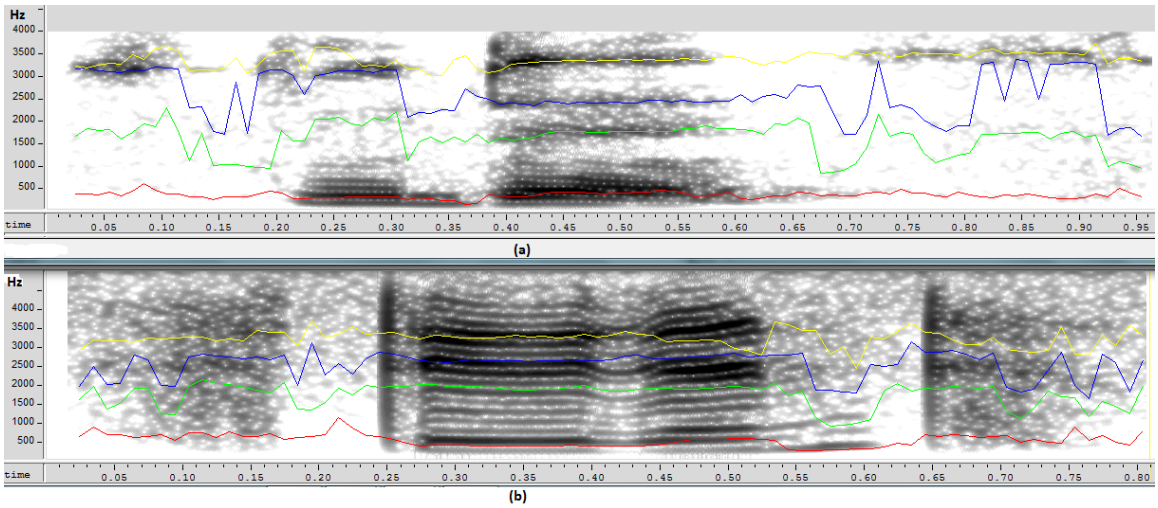


Figure 4. The spectrogram of the word "student" by (a) Nigerian English speaker and (b) American English speaker

3 ASR PERFORMANCE ANALYSIS

In spite of the above analysis, our focus is on rapid migration of an existing speech recognizer trained on American English (AE) to recognize Nigerian English (NE). This is a challenging task as AE and NE differ drastically along a number of critical speech parameters such as phonetic space, intonation patterns, and stress patterns. Considering this, the aim is to primarily mitigate the differences in the phonetic space by employing a two-pronged strategy: (i) developing a Nigerian English lexicon, and (ii) using a popular maximum-a-posteriori (MAP) model-adaptation technique to compensate for the acoustic phoneme pronunciation mismatch in the phone space [17].

3.1 ASR System Baseline

Detailed descriptions of the laboratory setup can be found in [4]. The performance of the baseline ASR system trained on the AE set and utilizing a TIMIT AE pronunciation lexicon is

shown in the second and third row of Table 2 for the complete AE set, denoted 'Devel+Test', and for the subset of AE comprising 1490 words from 9 speakers, denoted 'Test'. It can be seen that despite the simplicity of the small vocabulary task, the performance is very low, reaching approximately 50% word error rate (WER). It is believed that two major factors contribute to the poor performance: (i) the phonetic mismatch in the AE vs. NE pronunciations of the identical words, and (ii) the acoustic mismatch in the pronunciation of the identical phonemes in AE vs. NE.

In order to address the first factor, two trained phoneticians were asked to listen to a portion of the NE utterances and write down the most representative phonetic transcriptions of the 30 vocabulary words (see an example of AE–NE pronunciation differences in Table 1 as observed for TIMIT vs. UISpeech corpora). Subsequently, these transcriptions were used to extend the AE lexicon, yielding a lexicon denoted 'AE+NE'. As shown in rows 4 and 5 in Table 2, employing the extended lexicon helps to reduce WER by 2.5–3% absolute.

To address the phoneme pronunciation mismatch between AE trained acoustic models and NE test data, the acoustic models were adapted to the development ('Devel') set (1355 utterances from 32 female and male speakers who are distinct from the 'Test' set) using the MAP adaptation. First, forced alignment was performed on the development set given the known utterance transcriptions, yielding an estimation of the phone boundaries. Second, multiple MAP adaptation passes were performed. It was observed that 5 passes yielded reasonably adapted speaker-independent models (rows 6 and 7 in Table 2). Note that utilizing the combined 'AE+NE' lexicon in the adaptation process further reduces WER by 5.5% compared to using only the AE lexicon. Finally, an adaptation scheme where phone boundaries were re-estimated in every MAP adaptation iteration using the updated models was also evaluated (see the last row of Table 2). It can be seen that multiple re-alignments with the updated models do not significantly contribute to model refinement. When employing both lexicon extension and model adaptation, the overall absolute WER reduction over the baseline reaches 37%. Table 3 details the impact of the MAP adaptation on recognition performance when applied to a subset of phone models versus all models. It can be seen that adapting only consonant models (penultimate row of Table 3) has more substantial impact than adapting only vowel models - 20.3% absolute WER reduction versus 3.4% over the baseline unadapted models. However, adapting all models brings a further 14.3% WER reduction compared to adapting only consonants.

Table 2. ASR Performance of isolated word recognition part of UILSpeech Corpus. Test – re-alignments are performed every iteration [4].

Training	Lexicon	Set	WER (%)
No MAP	AE	Devel + Test	49.3
		Test	51.0
	AE + NE	Devel + Test	46.3
		Test	48.5
MAP	AE	Test	19.5
	AE + NE	Test	14.0
		Test*	13.9

It is noted that the improvements due to MAP adaptation may also be partly due to model adaptation to the acoustic environment of UILSpeech.

Table 3. ASR Performance of isolated word recognition on UILSpeech Corpus. The impact of MAP adaptation when applied to selected groups of phone models.

Adapted Phone Models	Open Test Set WER (%)
None	48.5
el, iy, ih, eh, ey, ae, aa, aw, ay, ah, ao, oy,ow, uh, uw, er, ax, ix, axr	45.1
b, d, g, p, t, k, m, n, ng, em, en, s, sh, z, zh, f, th, v, dh, jh, ch, l, r, w, y, hh	28.2
All	13.9

4 AUDIO-VISUAL ANALYSIS

Finally, we conduct an informal audio-visual analysis of the AE and NE speaker sessions. The goal is to relate the acoustic-phonetic properties of speech production with articulatory/facial movements. In this analysis, we study video recordings acquired while the subjects were reading three different sentences with three repetitions. The video samples are analyzed frame by frame using AVS Video Editor 4 [25] with a specific focus on the phonemes /dh/, /th/, /d/, /t/, /r/, /f/ and /v/ which are often confusable to Nigerian speakers.

Analysis of the the facial muscles and jaw positions in the video transcriptions reveals frequent substitution of the voiced flap in /dh/ for /d/ by most of the Nigerian speakers (see Figure 6). On the other hand, the expected substitution of /f/ for /v/ could not be ascertained. The two phonemes can be distinguished by a different mouth shape (lip height and width), however, the patterns here are strongly speaker dependent and the visual distinction is complicated by the fact that these phonemes are produced with most of the articulators covered by the lips. The

lip shape patterns for /t/ and /r/ were found strongly dependent on their position in the word (coarticulation effects) and unique to each speaker; that is, the lip shape pattern when producing the same utterance strongly varied across the speakers.

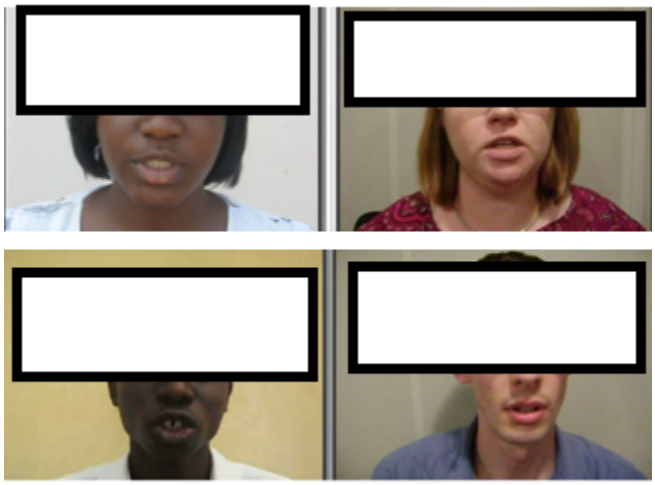


Fig. 6. Comparison of the lip shapes and jaw positions for the phoneme /dh/ production by both genders of Nigerian English (NE) and American English (AE) speakers

5 CONCLUSIONS

The data presented in this article consists of simultaneous speech audio and video tracks that capture isolated and read speech utterances. The corpus provides a unique opportunity for building a variety of speech systems such as speech/speaker recognition and dialect/accent identification for Nigerian English. Analysis of the American English and Nigerian English utterances on the lexical level and in terms of acoustic model distances, mean utterance fundamental frequency, and vowel location in F1–F2 space confirms substantial differences in American English and Nigerian English. Such differences cause a significant deterioration of American English-trained ASR when exposed to Nigerian English. A simple scheme that combines extended American English lexicon for Nigerian English pronunciation variants and multi-pass acoustic model adaptation showed a reduction of recognition errors by 37% absolute WER. These encouraging results suggest that such an approach may represent a viable ASR path also for other low resource dialects with limited availability of speech data and phonetic information. The results also show that while improved lexicon pronunciation is beneficial, corresponding advancement in acoustic modeling for the new language dialect domain is necessary to reach substantial performance gains. The audio-visual analysis of the lip shape patterns during speech production revealed strong speaker dependency for certain phonemes. Visual features extracted from lip shape patterns of these phonemes could be beneficial to speaker authentication applications.

ACKNOWLEDGMENTS

The authors appreciate the financial and moral support from the Fulbright foundation of IIE, U.S.A. and the Center for Robust Speech System (CRSS), Erik Jonsson School of Electrical and Computer Science, The University of Texas at Dallas. We also wish to thank all individuals that contributed to the data collections both in Nigeria and U.S. The portion of the study conducted in CRSS was funded partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

REFERENCES

[1] U. Gut and J.-T. Milde, "The prosody of Nigerian English," in *SP-2002*, 2002, pp. 367–370.

[2] C. T. Hodge, "Yoruba: Basic course," ED – 010 – 462 Report NDEA – VI – 375, US Foreign Service Institute, 1963.

[3] A. A. Fakoya, *Nigerian English: A Morpholecta Classification*, Ph.D. thesis, Lagos State University, 2007.

[4] S. Amuda, Boril, H., Sangwan, A. and Hansen, J. H. L. (2010). "Limited Resource Speech Recognition for Nigerian English." *Proc. of IEEE ICASSP'10*, 5090-5093.

[5] M. Jibril, "Phonological Variation in Nigerian English", Ph.D Thesis at University of Lancaster 1986

[6] T. T. Ajani "Is There Indeed A 'Nigerian English'?" *Journal of Humanities & Social Sciences*, 1(1), 2007.

[7] T. Ufomata "Setting Priorities in Teaching English Pronunciation in ESL Contexts", Seminar presentation as a British Academy Visiting Fellow at University College London, 1996.

[8] A. Bamgbose, "Language in Contact: Yoruba and English in Nigeria", *Education and Development*, 2(1), pp. 329-341, 1982.

[9] W. Voiers, I. Dynastat, and T. Austin, "Diagnostic Acceptability Measure for Speech Communication System," in *Proc. of IEEE ICASSP*, vol. 2, pp. 204–207, 1977.

[10] M. A. Koler, "A Comparison of the New 2400 bps MELP Federal Standard with other Standard Coders," in *Proc. of IEEE ICASSP*, 1997.

[11] L. M, Arslan and J. H. L. Hansen, "Language Accent Classification in American English", *Speech Communication*, vol. 18, pp. 353-367, ELSEVIER, 1996.

[12] L. M, Arslan and J. H. L. Hansen, "A Study of Temporal Features Frequency Characteristics in American English Foreign Accent", *Journal of Acoustical Society of America*, vol. 201(1), pp. 28-40, July, 1997.

[13] J. S. Garofolo, L. F. Lamel, J. G. Fisher,W.M. andFiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, LDC93S1, 1993.

[14] J.-L. Gauvain and Chin-Hui Lee, "Maximum a Posteriori Estimation for Multivariate Gaussian Mixture Observations of Markov Chains," *IEEE Transactions on Speech & Audio Processing*, 2(2), pp. 291–298, 1994.

[15] S. B. Davis and P. Mermelstein, "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4), pp. 357–366, 1980.

[16] K. Sjolander and J. Beskow, "WaveSurfer – An Open Source Speech Tool," in *Proc. of ICSLP'00*, Beijing, China, 2000, vol. 4, pp. 464–467.

[17] R. D. Kent and C. Read, *The Acoustic Analysis of Speech*, Whurr Publishers, San Diego, 1992.

[18] J. Silva and S. Narayanan, "Average Divergence Distance as a Statistical Discrimination Measure for Hidden Markov Models," *IEEE Transactions on Audio, Speech, and Language Processing*, 14(3), pp. 890–906, 2006.

[19] J. H. L. Hansen, "Analysis and Compensation of Speech Under Stress and Noise for Environmental Robustness in Speech Recognition," *Speech Communication*, 20(1-2), pp. 151–173, 1996.

[20] J. H. L. Hansen, E. Ruzanski, H. Boril, J. Meyerhoff, "TEO-Based Speaker Stress Assessment Using Hybrid Classification and Tracking Schemes," *International Journal of Speech Technology*, Springer, June 2012, DOI 10.1007/s10772-012-9165-1.

[21] T. Hasan, H. Boril, A. Sangwan, J. H. L. Hansen, "Multi-Modal Highlight Generation for Sports Videos Using an Information-Theoretic Excitability Measure," *EURASIP Journal on Advances in Signal Processing*, 2013:173, 2013.

[22] H. Boril, J. H. L. Hansen, "Unsupervised Equalization of Lombard Effect for Speech Recognition in Noisy Adverse Environments," *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6), 1379-1393, 2010.

[23] H. Boril, Q. Zhang, A. Ziaei, J. H. L. Hansen, D. Xu, J. Gilkerson, J. A. Richards, Y. Zhang, X. Xu, H. Mao, L. Xiao, F. Jiang, "Automatic Assessment of Language Background in Toddlers Through Phonotactic and Pitch Pattern Modeling of Short Vocalizations," *accepted to Workshop on Child Computer Interaction (WOCCI)*, September, Singapore, 2014.

[24] M. Mehrabani, H. Boril, J. H. L. Hansen, "Dialect Distance Assessment Method Based on Comparison of Pitch Pattern Statistical Models," *in Proc. of IEEE ICASSP'10*, 5158-5161, Dallas, TX, 2010.

[25] Link: http://www.avs4you.com (accessed on Aug 20, 2014).