LIMITED RESOURCE SPEECH RECOGNITION FOR NIGERIAN ENGLISH

Sulyman Amuda[+]

Department of Electrical Engineering
University of Ilorin, Nigeria.

Hynek Bořil, Abhijeet Sangwan, John H. L. Hansen[*]

Center for Robust Speech Systems (CRSS),
Erik Jonsson School of Eng. and Comp. Science,
The University of Texas at Dallas, U.S.A.

ABSTRACT

In this study, we introduce the UISpeech corpus which consists of Nigerian-Accented English audio-visual data. The corpus captures the linguistic diversity of Nigeria with data collected from native-speakers of Yoruba, Hausa, Igbo, Tiv, Funali and others. The UISpeech corpus comprises isolated word recordings and read speech utterances. The new corpus is intended to provide a unique opportunity to apply and expand speech processing techniques to a limited resource language. Acoustic-phonetic differences between American English (AE) and Nigerian English (NE) are studied in terms of pronunciation variations, vowel locations in the formant space, and distances between AE-trained acoustic models and models adapted to NE. A strong impact of the AE–NE acoustic mismatch on automatic speech recognition (ASR) is observed. A combination of model adaptation and extension of 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 represents the first step towards incorporating speech technology in Nigerian English.

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

1. INTRODUCTION

Nigerian English differs from its counterparts in terms of tones, prosody, and phones, coupled with some unique lexical patterns that portray the influence of local Nigerian languages. An analysis of the speech rhythm, tonal, and syllable structures of Nigerian English has revealed the tonal nature of the language. Particularly, the pitch employed by Nigerian English speakers is lexically significant, contractive, and relative [1]. Additionally, the language is syllable-timed and there is no vowel contrast. In fact, these characteristics observed in Nigerian English are closely linked to the influence of local languages (such as in Yoruba and Hausa languages, [2]). Finally, it has also been observed that Nigerian English speech rhythm has higher vowel intervals when compared to British English [1]. Other than prosodical differences, Nigerian English is also characterized by phonetic differences. These phonetic differences are more obvious when speakers encounter unfamiliar phones that are otherwise absent in their native language [3]. For example, Nigerian English speakers will introduce an unglided vowel structure and unnecessary nasalization of sounds when pronouncing unfamiliar phones. Additionally, some speakers will tend to omit the phones that are absent in their native language [3, 2].

While English is spoken by about 130 million people in Nigeria, little attention has been paid towards building viable speech processing technology for Nigerian English. The biggest hurdle in this direction has been the lack of a quality speech corpus. This paper represents a pioneering effort towards establishing the first of its kind audio-visual Nigerian English Corpus. The corpus consists of approximately 30 hours of speech collected from approximately 600 speakers.

This project was supported by [+] Fullbright Foundation, and by [*]USAF under a subcontract to RADC, Inc. under FA8750-05-C-0029. Approved for public release; distribution unlimited.

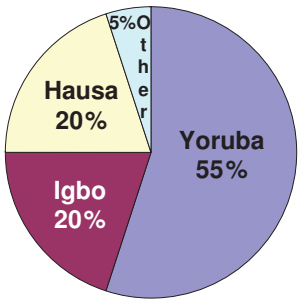

Fig. 1. Representation of different linguistic backgrounds in UISpeech corpus.

A study of acoustic-phonetic differences between American English (AE) and Nigerian English (NE) is conducted on the level of pronunciation variations, vowel locations in the formant space, and distances between AE-trained acoustic models and models adapted to NE. It is shown that the AE–NE acoustic mismatch has a strong impact on automatic speech recognition (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 as well as motivate development of future speech systems for limited resource languages.

2. UISpeech CORPUS

The UISpeech (University of Ilorin Speech) corpus was collected as a pioneering effort to form a database for Nigerian English. The speech data in the UISpeech corpus was exclusively collected on the University of Ilorin campus. Speakers were mostly undergraduate students with an average age of 20 years. In terms of gender, the corpus consist of speech from about 300 males and females each. The speaker pool reflects the linguistic diversity of Nigerian English. Most speakers tended to be from 3 dominant linguistic backgrounds in Nigeria, namely, natives speakers of Yoruba (South-Western Nigeria), Igbo (South-Eastern Nigeria), and Hausa (Northern Nigeria). Fig. 1 shows the composition of the UISpeech corpus in terms of the major Nigerian languages. In this manner, the corpus is a good

Table 1. ASR Performance of isolated word recognition part of UISpeech Corpus. Test* – re-alignments are performed every iteration.

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

reflection of the true proportions of speaker diversity in Nigeria.

The UISpeech corpus consists of isolated word recordings as well as continuous read speech data. The isolated word data was collected in a laboratory environment with the use of a hollow-shaped telephone mouth-piece. The mouth-piece was intended to help reduce speaker induced variabilities 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 was 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 was collected in an office/laboratory environment, a low-level ambient yet realistic noise is present in the speech utterances. The recorded speech data is sampled at the rate of 8kHz for the entire corpus. Here, it is also useful to note 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 [4, 5].

The isolated word recordings consists 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 consists 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.

3. ACOUSTIC-PHONETIC ANALYSIS AND SPEECH RECOGNITION EXPERIMENTS

In this section, acoustic-phonetic differences between American English (AE) and Nigerian English (NE) are analyzed along with the impact of the AE–NE acoustic mismatch on ASR. In this study, all NE experiments are conducted on the isolated words portion of the UISpeech 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 was taken from the TIMIT database [6]. 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 2. Example of pronunciation differences in American (AE) and Nigerian (NE) English.

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/

3.1. Baseline Speech Recognition System

Our focus is on rapidly migrating 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 [7] to compensate for the acoustic phoneme pronunciation mismatch in the phone space.

The baseline HMM-based (Hidden Markov Model) ASR system consists of 45 context-independent monophone models and two silence models. Each HMM comprises 3 emitting-states modeled by 64-mixture GMM's (Gaussian Mixture Models). The MFCC's (Mel-Frequency Cepstral Coefficients) [8] are used as the front-end acoustic feature vectors. Particularly, 13 static cepstral coefficients, (i.e., c_0–c_{12}) along with their first and second order time derivatives are extracted. The MFCC extraction incorporates 26 mel filters spread over the band of 0–8 kHz. Using the above described scheme, gender-independent phoneme models are trained on the subset of TIMIT in 58 iterations. The task is to recognize utterances containing sequences of 5 isolated words coming from the vocabulary of 30 words.

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 1 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.

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 2 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 1, employing

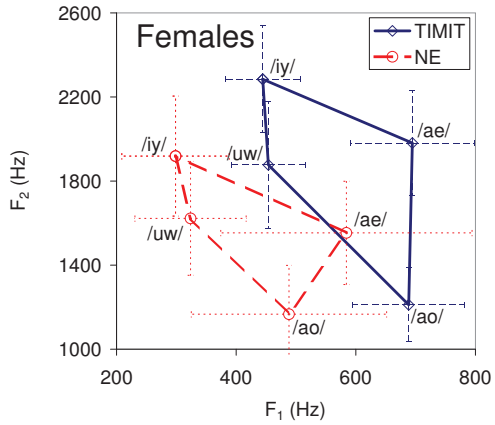

Fig. 2. Comparing the phone space of vowels /iy/, /uw/, /ao/, and /ae/ for female speakers of American English (AE, TIMIT) and Nigerian English (NE).

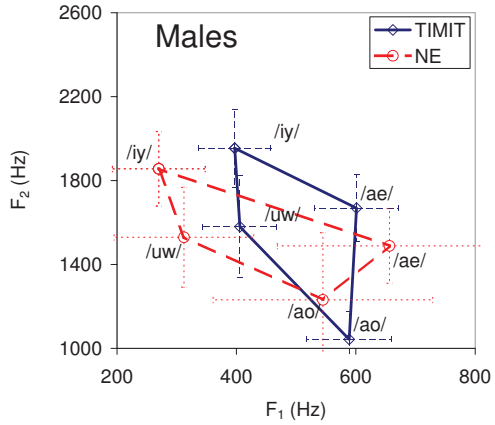

Fig. 3. Comparing the phone space of vowels /iy/, /uw/, /ao/, and /ae/ for male speakers of American English (AE, TIMIT) and Nigerian English (NE).

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 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 1). Note that utilizing the combined 'AE+NE' lexicon in the adaptation process further reduces WER by 5.5 % compared to simply 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 1). It can be seen that multiple re-alignments with the updated models do not significantly contribute to model refinement. When employing both the lexicon extension and model adaptation, the overall absolute WER reduction over the baseline reaches 37 % absolute. It is noted that the improvements due to MAP adaptation may also be partly due to model adaptation to the acoustic environment of UISpeech.

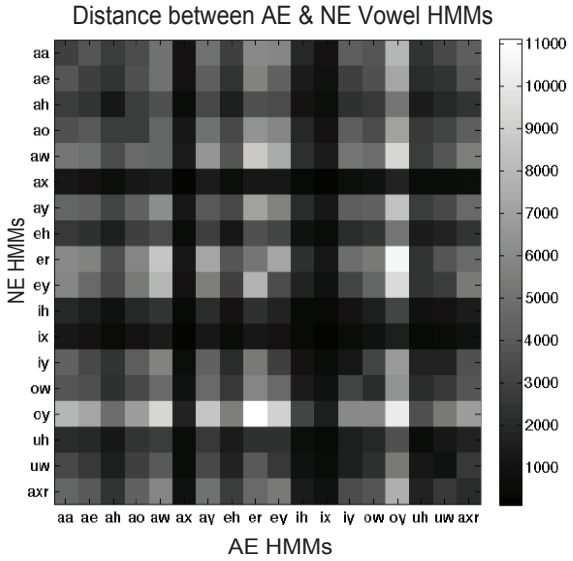

Fig. 4. Comparing the KL-Divergence between Nigerian English (NE) and American English (AE) HMMs for vowels and consonants.

3.2. Formant Analysis

To better understand the acoustic-phonetic mismatch in AE and NE data, the locations of vowels in the F_1–F_2 (first and second formant) space are analyzed. Formant frequencies for individual phones are estimated by combining the output of formant tracking (WaveSurfer [9]) and the phone boundaries obtained from forced alignment. In the AE case, the AE lexicon was used in the forced alignment while in the NE alignment utilized the 'AE+NE' lexicon. The gender-dependent vowel analysis was conducted on the 'Devel+Test' set for NE. The results are shown in Figs. 2 and 3. The error bars in the plots represent standard deviations of the F_1, F_2 sample distributions. As compared to native speakers of AE, both F_1 and F_2 vowel coordinates tend to be lower in NE subjects. This suggests that the NE speakers produce vowels relatively further back and higher in the vocal tract as F_1 varies inversely with tongue height, and F_2 varies with the posterior-anterior dimension of the vowel articulation [10].

3.3. Inter-HMM Distance Analysis

It is of interest 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 [11] to compute the distances between the baseline AE HMMs and adapted NE HMMs. Fig. 5 shows the KL-divergence between AE and NE vowel/consonant-pairs. From this figure, it can be observed that the articulation characteristics of /ax/ and /ix/ are 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. From this figure, it can also be observed that all the 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 if 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

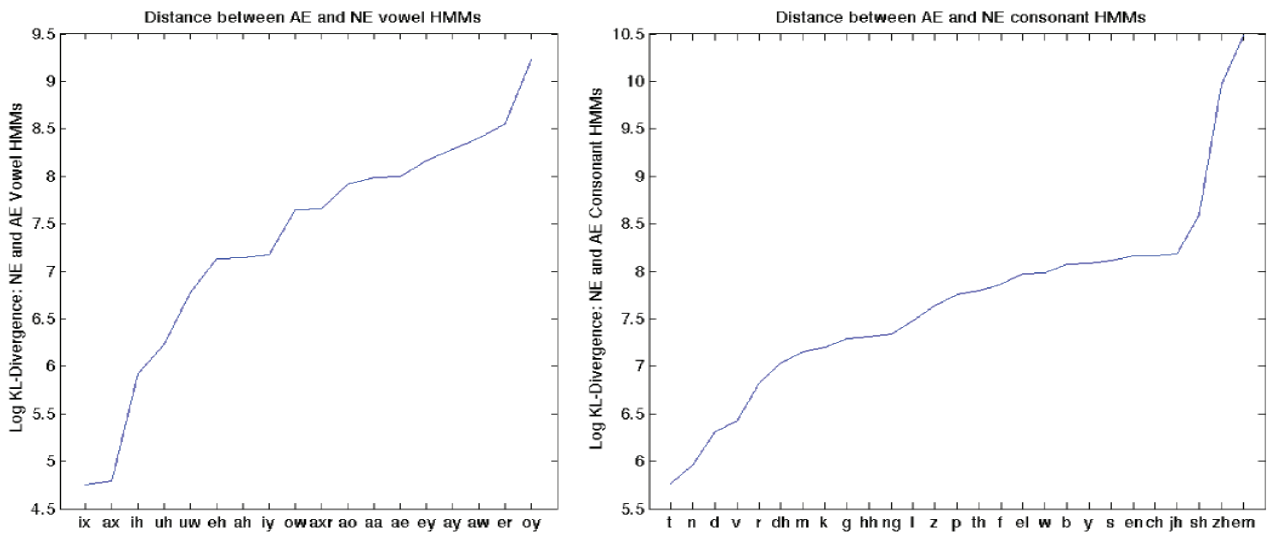

Fig. 5. Comparing the KL-Divergence between corresponding Nigerian English (NE) and American English (AE) HMMs for vowels and consonants.

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 production (see Fig. 5). 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.

4. CONCLUSION

This study has presented a newly acquired audio-visual corpus on Nigerian English (UISpeech). UISpeech consists of 30 hours of speech collected from 600 speakers that reflect the linguisitic diversity of Nigeria. The data consists of simultaneous speech as well as video recordings of 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. An analysis of American English and Nigerian English utterances on a lexical level and on the level of acoustic model comparison and vowel location in F_1–F_2 space confirmed substantial differences in AE and NE. Such differences caused a significant deterioration of AE-trained ASR when exposed to NE. A simple scheme combining an extension of AE for NE pronunciation variants with acoustic model adaptation showed a reduction in ASR WER by 37 % absolute, which suggests that such an approach may be a reasonable step towards NE ASR in the case of limited availability of NE data. The results show also that while improved lexicon pronunciation can help, without the advancement in acoustic modeling for the new language domain, the lexicon impact is small.

5. ACKNOWLEDGMENTS

The authors would like to thank Jan Volín and Radek Skarnitzl of Institute of Phonetics, Charles University in Prague, Czech Republic for helping with the phonetic transcriptions of Nigerian English.

6. REFERENCES

[1] Ulrike Gut and Jan-Torsen Milde, "The prosody of Nigerian English," in *SP-2002*, 2002, pp. 367–370.

[2] Carleton 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] W. Voiers, I. Dynastat, and T. Austin, "Diagnostic acceptability measure for speech communication system," in *ICASSP*, 1977, vol. 2, pp. 204–207.

[5] M. A. Koler, "A comparism of the new 2400 bps MELP Federal standard with other standard coders," in *ICASSP*, 1997.

[6] 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.

[7] 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*, vol. 2, no. 2, pp. 291–298, 1994.

[8] 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*, vol. 28, no. 4, pp. 357–366, 1980.

[9] K. Sjolander and J. Beskow, "WaveSurfer – An open source speech tool," in *Proc. of ICSLP'00*, Beijing, China, 2000, vol. 4, pp. 464–467.

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

[11] 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*, vol. 14, no. 3, pp. 890–906, 2006.