Dialect-Specific Models for Automatic Speech Recognition of African American Vernacular English

Rachel Dorn
Virginia Commonwealth University
dornrm@mymail.vcu.edu

Abstract

African American Vernacular English (AAVE) is a widely-spoken dialect of English, yet it is under-represented in major speech corpora. As a result, speakers of this dialect are often misunderstood by NLP applications. This study explores the effect on transcription accuracy of an automatic voice recognition system when AAVE data is used. Models trained on AAVE data and on Standard American English data were compared to a baseline model trained on a combination of the two dialects. The accuracy for both dialect-specific models was significantly higher than the baseline model, with the AAVE model showing over 18% improvement. By isolating the effect of having AAVE speakers in the training data, this study highlights the importance of increasing diversity in the field of natural language processing.

1 Introduction

There have been tremendous improvements in recent years in automatic speech recognition (ASR). Models which approach or even surpass human performance on transcription tasks have been reported (Xiong et al., 2017). However, these increases in accuracy have not been evenly distributed across all speakers, performing worse for speakers of dialects other than Standard American English (SAE) (Tatman and Kasten, 2017). ASR systems are becoming more and more integrated into society, used in everything from call centers to medical transcriptions to asking for the weather. Poorer performance for individuals with accents leads to discrimination against significant percentages of the population, many of whom belong to already marginalized groups. This paper looks specifically at the dialect of African American Vernacular English (AAVE), which is spoken by an estimated 80% of African Americans in the United States (Lippi-Green, 1997) or about 35-40 million people.

A major cause of the bias against AAVE is the lack of examples of this dialect in major speech corpora. The TIMIT dataset is the most popular speech corpus offered by the Linguistic Data Consortium and is often used for training and benchmarking speech recognition systems. It does not specifically provide statistics for AAVE speakers, but of the 630 speakers, only 26 (4%) are black, compared to the 538 (85%) who are white (Garofolo et al., 1993). Switchboard 1, another of the most commonly-used speech datasets, does provide a list of included dialects, yet AAVE is not among them (Godfrey and Holliman, 1993). A model which is trained on biased data will result in uneven performance.

Despite this bias against AAVE in many ASR systems, there has been very little scholarship either identifying this issue or proposing solutions. A number of popular press articles have called attention to the challenges faced by speakers with accents attempting to use products such as Amazon Alexa (Paul, 2017) or Google Assistant (Harwell, 2018), but there is a gap in the academic literature on AAVE in ASR, which this paper aims to fill.

Many papers have been published related to ASR systems for under-resourced dialects in English and other languages, using a variety of techniques. (Cucu et al., 2012) apply the technique of statistical machine translation to Romanian. (Elmahdy et al., 2013) and (Lehr et al., 2014) use transfer learning to adapt ASR systems trained on news corpora to under-resourced dialects. Using a sequence-to-sequence model, (Li et al., 2018)

Proceedings of the Student Research Workshop associated with RANLP-2019, pages 16–20,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/issn.2603-2821.2019_003

Description	SAE Example	AAVE Example
consonant cluster reduction	first office	firs' office
devoicing of final consonants	bad	bat
variation of interdental fricatives	the	da
r-lessness	forever	fo'eva
-ing endings	walking	walkin'

Table 1: Common Phonetic Features of African American Vernacular English (Kendall et al., 2018)

combine language and pronunciation models into a single neural network for transcribing a variety of English dialects. Comparing all of these methods, and the many others which have shown success, for AAVE is beyond the scope of this paper, though these are promising avenues for future research. This preliminary study takes the approach of training a dialect-specific model for AAVE and SAE, drawing from the approach of (Soto et al., 2016).

1.1 African American Vernacular English

AAVE is a dialect of English commonly spoken some or all of the time by black persons in the United States, many of whom code-switch. It goes by other names including African American Language, Black English, and Ebonics (Kendall et al., 2018). While there are regional variations within the dialect, the majority of the phonetic and grammatical traits remain common across regions. Table 1 enumerates a few of the common phonetic features of AAVE; grammatical features were left out as this study's focus is on pronunciation.

1.2 Standard American English

SAE is the baseline to which other dialects are compared to; speakers of this dialect are perceived to be speaking without an accent. It is the language taught in classrooms and that spoken by newscasters and those in formal settings (Kretzchmar Jr., 2008). Labov observed that individuals exhibit more of the features of SAE as they increase attention to their language, indicating that the features of SAE are generally known and agreed upon, whether or not speakers use them in everyday speech (Labov, 2012).

2 Methodology

2.1 Data

Both the AAVE and the SAE speech data came from the Corpus of Regional African American Language (CORAAL) published by the University of Oregon (Kendall and Farrington, 2018). This dataset consists of audio files and time-aligned transcripts of interviews between SAE-speaking interviewers and AAVE-speaking interviewees. A subset of speakers located in Washington D.C. was used, to minimize the effects of regional accents within the dialect.

There were forty-two speakers–six interviewers and thirty-six interviewees–and 31,468 utterances across the training and testing sets. All speakers were adults, ranging in age from eighteen to seventy-seven. The gender-split of the AAVE speakers was ten female and twenty-six male; the genders of the interviewers were not provided.

2.2 Preprocessing

All audio files were split into utterances by the timecodes given in the transcripts. Each utterance was of a single speaker and was an average of four seconds long. These audio files were converted to 16-bit 16kHz mono WAVE files. The transcripts were sanitized to remove punctuation, lowercase all letters, and expand numerals into words. Regex matching was used to remove instances of sound effects, pauses, and other non-verbal entities present in the transcripts.

2.3 Model Training

Recent advances in neural networks have shifted the state of the art away from the hidden Markov models previously common in the field. Neural network models, rather than being passed lists of possible words, iteratively learn to match the patterns of words and phonemes to their written equivalents. The tool used for training in the study is Mozilla's DeepSpeech (Hannun et al., 2014), which is built on recurrent neural networks. This tool was chosen for ease of use as well as its demonstrated performance on noisy environments and with high speaker variation.

An AAVE model was trained using just the data from the AAVE-speaking interviewees. An SAE

Utterance Dialect	Model Dialect	Word Error Rate	Levenshtein Distance
AAVE	Combined	1.1509	18.7501
AAVE	AAVE	0.9363	15.2848
AAVE	SAE	1.1886	21.1567
SAE	Combined	1.1254	15.0589
SAE	SAE	1.0413	12.9967
SAE	AAVE	1.0492	15.5483

Table 2: Error Rates by Dialect

Gold Standard: you know take care of hisself

AAVE Model: you know tay car sef

Combined Model: you know tha co sa

Figure 1: Model Outputs for an Example AAVE Utterance

model was similarly trained on data from only the interviewers. To provide a baseline during evaluation, a model was trained using a combined set of all the data. Additionally, the dialect models were tested on the opposite dialect's utterances to examine whether there were major differences in the difficulty of the two datasets contributing to the error rates and to simulate how a large model trained only on SAE data would perform. A 70-20-10 split was used for the training, dev, and test sets. The splits were done to ensure that all speakers had proportional representation in each of the sets. For AAVE, the training set had 15,857 utterances, the dev set had 4,467, and the test set had 2,266. The SAE training set had 6,199 utterances, the dev set had 1,783, and the test set had 902. There were an average of 6.2 words per utterance for AAVE and 4.6 words per utterance for SAE.

3 Results

Transcripts were generated by passing each utterance in the test set through the baseline combined model, the dialect-specific model corresponding to the dialect of the utterance, and the dialect-specific model of the other dialect. These outputs were then compared to the gold standard transcripts provided by CORAAL.

Accuracy of the transcripts was measured using both word error rate (WER) and Levenshtein distance. The WER is a measure of the number of full word matches between the output and the gold standard, as a percent of the word count. It provides a good metric for judging the readability of the output, but the drawback of this measure is it does not allow for partial credit. The neural network based approach to building ASR models does not pass pre-defined word lists and often results in matches on certain phonemes within a word rather than the entire word. For instance, in the example output shown in Figure 1, both the AAVE-only and the combined model correctly matched ”you know” but missed the other three words, resulting in equal WER scores. However, the AAVE model picked up on the ”ar” sound in ”care” and the ”f” sound in ”hisself”, both of which the combined model incorrectly transcribed. In order to quantify this behavior, the Levenshtein distance was used, which is a measure of the number of character insertions, deletions, or substitutions needed to transform the output into the gold standard. For both WER and Levenshtein distance, a lower score indicates higher accuracy.

The hypothesis of the study was that the highest performance for each dialect's data would be from the same-dialect model and that the worst performance would result from running the opposite dialect's model. The results support this hypothesis, as Table 2 illustrates. In each section, the bottom row shows the performance when the utterance dialect is opposite that of the model. For both AAVE and SAE, this had the highest error rate as measured by the Levenshtein Distance. The middle row of each section, when the dialects were aligned between utterances and the model, the error rates were lowest for both WER and Levenshtein distance.

The WER rates are nearly all above 1 due to the outputs putting in extra word breaks. Though the overall accuracy is low, both the AAVE and the SAE models significantly improved accuracy over the baseline combined model for WER and Levenshtein Distance. Table 3 shows the percent improvement over the combined-dialect baseline for the two dialect models for both accuracy measures. The improvement for the AAVE model was over 18% when using the Levenshtein Distance to provide a character-level error rate. There was a larger increase for the AAVE utterances, likely due to the higher number of utterances for that dialect as well as the utterances being longer. Many of the SAE utterances were short interjections such as ”I see” and ”mm hm”.

Dialect	WER	Levenshtein Distance
AAVE	16.6%	18.5%
SAE	7.5%	13.7%

Table 3: Improvements in Error Rate Between Dialect-Specific Model and Combined Model

4 Conclusions and Future Work

Different dialects or accents within the same language can have conflicting patterns of phonemes. An automatic voice recognition tool which tries to handle all dialects with the same model is setting itself up for challenges. The ambiguities arising from having to learn opposing patterns can cause errors and lower overall accuracy. If the model does not handle these ambiguities, whatever pattern was seen more in the training data could win out, causing a bias against dialects with lower representation. Frequently, the majority dialect in training sets is Standard English. Those who do not speak Standard English, such as speakers of African American Vernacular English, are more often misunderstood. This study shows the potential of using dialect-specific models to remedy this situation.

Applications which handle diverse speech can benefit from using dialect-specific models for speech recognition. This is particularly useful in contexts where the speech data is gathered in advance, rather than in real-time. In this case, the increase in transcription accuracy would outweigh the small increase in processing time to classify the dialect of the speech.

Future directions for this research are growing the AAVE dataset used for training to improve the overall accuracy of the model and expanding to other dialects. A classifier could also be trained to automatically select the appropriate model for a given utterance to remove the preprocessing step of manually separating the utterances by dialect before passing them to the speech-to-text models. Additionally, transfer learning techniques could be explored as a comparison to the dialect-specific method explored in this study.

References

Horia Cucu, Laurent Besacier, Corneliu Burileanu, and Andi Buzo. 2012. Asr domain adaptation methods for low-resourced languages: Application to romanian language. In *20th European Signal Processing Conference*. pages 1648–1652.

Mohamed Elmahdy, Mark Hasegawa-Johnson, and Eiman Mustafawi. 2013. A transfer learning approach for under-resourced arabic dialects speech recognition. In *Workshop on Less Resourced Languages, new technologies, new challenges and opportunities (LTC 2013)*. pages 60–64.

John S. Garofolo, Lori F. Lamel, William M. Fisher, Jonathan G. Fiscus, David S. Pallett, Nancy L. Dahlgren, and Victor Zue. 1993. Timit acoustic-phonetic continuous speech corpus ldc93s1. *Linguistic Data Consortium* https://catalog.ldc.upenn.edu/docs/LDC93S1.

John Godfrey and Edward Holliman. 1993. Switchboard-1 release 2 ldc97s62. *Linguistic Data Consortium* https://catalog.ldc.upenn.edu/LDC97S62.

Awni Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng. 2014. Deep speech: Scaling up end-to-end speech recognition https://arxiv.org/abs/1412.5567.

Drew Harwell. 2018. The accent gap. *The Washington Post* .

Tyler Kendall and Charlie Farrington. 2018. The corpus of regional african american language. *The Online Resources for African American Language Project* (2018.10.06). http://oraal.uoregon.edu/coraal.

Tyler Kendall, Jason McLarty, and Brooke Josler. 2018. Oraal: Online resources for african american language: Aal facts. *The Online Resources for African American Language Project* https://oraal.uoregon.edu/facts.

William A. Kretzchmar Jr. 2008. Standard american english pronunciation. In *Varieties of English 2: The Americas and the Caribbean*, pages 37–51.

William Labov. 2012. *Dialect Diversity in America*, University of Virginia Press, chapter A Hidden Consensus, pages 9–10.

Maider Lehr, Kyle Gorman, and Izhak Shafran. 2014. Discriminative pronunciation modeling for dialectal speech recognition. *INTERSPEECH 2014* .

Bo Li, Tara N. Sainath, Khe Chai Sim, Michiel Bacchiani, Eugene Weinstein, Patrick Nguyen, Zhifeng Chen, Yonghui Wu, and Kanishka Rao. 2018. Multi-dialect speech recognition with a single sequence-to-sequence model. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. pages 4749–4753.

Rosina Lippi-Green. 1997. What we talk about when we talk about ebonics: Why definitions matter. *The Black Scholar* (27.2).

Sonia Paul. 2017. Voice is the next big platform, unless you have an accent. *Wired* .

Victor Soto, Olivier Siohan, Mohamed Elfeky, and Pedro Moreno. 2016. Selection and combination of hypotheses for dialectal speech recognition. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* .

Rachael Tatman and Conner Kasten. 2017. Gender and dialect bias in youtube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, pages 53–59. https://doi.org/10.18653/v1/W17-1606.

W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke. 2017. The microsoft 2017 conversational speech recognition system. *Microsoft AI and Research Technical Report MSR-TR-2017-39* https://arxiv.org/pdf/1708.06073.pdf.