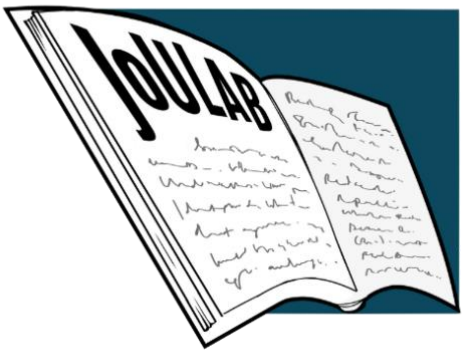


THE JOURNAL OF THE UNDERGRADUATE LINGUISTICS ASSOCIATION OF BRITAIN

ISSN 2754-0820

COMPARING AND CONTRASTING FORCED ALIGNED ACOUSTIC MODELS FOR VERNACULAR SPEECH

Amy Cox

University of Alabama
amyecox02@gmail.com

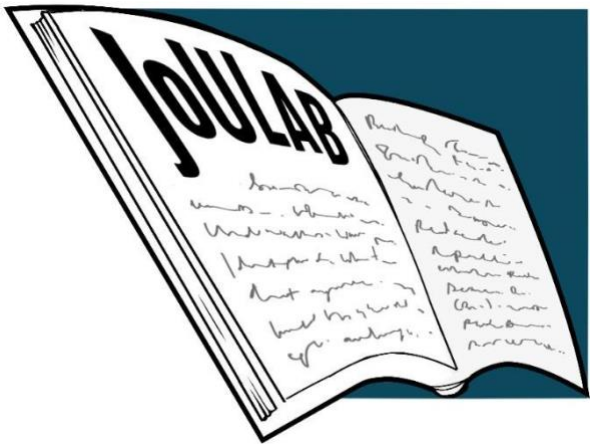
Keywords: Montreal forced aligner; forced alignment; sociolinguistics; Appalachian English

JoULAB URL: https://www.ulab.org.uk/journal/volumes/3/issues/1/articles/1

DOI: 10.5281/zenodo.17408121

Citation: Cox, A. (2025). Comparing and Contrasting Forced Aligned Acoustic Models for Vernacular Speech. *Journal of the Undergraduate Linguistics Association of Britain*, 3(1).

Received: 04/04/2023
Revisions: 16/05/2024
Accepted: 20/05/2025



The *Journal of the Undergraduate Linguistics Association of Britain* (ISSN 2754-0820) was founded in July 2020 as a Subcommittee of the Undergraduate Linguistics Association of Britain. It is the only academic journal in the world taking submissions solely from undergraduates in any area of linguistics. The Journal exists to provide a forum for the publication of exceptional undergraduate research in linguistics for students across the globe and from any background. In so doing, it seeks to offer undergraduate students an introduction to academic publishing and standard practices in academia, including the opportunity for undergraduates to have their research reviewed. Every manuscript is peer-reviewed over multiple rounds by two members of the Board of Reviewers, which consists of doctoral students in linguistics from a plethora of countries and institutions.

Copyright Disclosure

Accessibility Declaration

The Journal is committed to the maximum accessibility of its output. As such, provisions have been sought wherever possible to mitigate accessibility issues in the production of this Issue: for example, figures and tables have been designed to ensure accessibility for those with colour vision deficiencies. However, should you encounter access difficulties with this publication, please contact the Journal's Editorial Committee via email (below) and we will be happy to assist.

Disclaimer

The views expressed in the articles published herein do not necessarily reflect those of the Journal nor of the Undergraduate Linguistics Association of Britain; they are the authors' own. Although maximum effort has been expended to ensure no factual errors or inaccuracies are contained herein, the Journal apologises for any remaining.

Useful Links

Email: ulabjournal@gmail.com
Instagram: @ULAB_Journal
Bluesky: @ulabjournal.bsky.social
LinkedIn: JoULAB
Website: www.ulab.org.uk/journal

Comparing and Contrasting Forced Aligned Acoustic Models for Vernacular Speech

Amy Cox
University of Alabama

Abstract. Research into dialect variation has been streamlined with the introduction of forced alignment software. The purpose of this study is to evaluate the accuracy of the Montreal Forced Aligner (MFA) when analysing non-mainstream dialects. To achieve this, the alignment from a pretrained speech model was compared to hand aligned data. The pretrained model uses English speakers from the LibriSpeech database which is an open-source corpus of 1000 hours of speech. Four interviews of speakers from the Appalachian region of the United States were run through the MFA with the pretrained model. The speakers were chosen due to variation in demographic information like age and gender. These four files were compared to hand-aligned variations to identify the model's accuracy on a non-mainstream dialect. No systematic issues were noted based on analysis of the original alignment of the four files. Altogether, there was 0.54–1.3% of errors that occurred across speakers. Errors primarily occurred when a word was out of the vocabulary or with transcription errors. While most of the errors were minor, there were instances, specifically with boundaries, where the alignment was off by multiple seconds. However, the aligner performed better than expected with overlapping speech and with productions that differed significantly from the training data. Therefore, a researcher could be confident that the MFA will perform at a usable level for Appalachian English with high quality recordings.

Plain English Abstract. Studying dialectal variation is important when considering sociological analyses, language documentation, and differentiating dialects and disorders. To study the differences in dialects, researchers identify the individual sounds within words and phrases and analyse key aspects within these sounds. The process of identifying sounds in an audio recording is called time alignment, and while time aligning is essential, it is a tedious and time-consuming process. Forced aligning was created to use computers programs to speed up the process of identifying these speech sounds. However, most aligners are trained on standard American or British English which may not generalise with different variations. For my research project, I am looking at one of these forced aligning programs, the Montreal Forced Aligner (MFA), and seeing how accurate it performs compared to when this process is done by hand. Specifically, my research focuses on dialects and how the computer's output changes when presented with speech that varies significantly from the data it was trained on. We used interviews from the Appalachian region of the United States because there are noted differences in speech sounds and grammar from mainstream American English. We analysed four recordings from this region and our analysis revealed predictable and very low percentages of error per speech sound even with the sounds that were known to vary greatly from the training data. Ultimately, a researcher can be confident in the performance of MFA even with non-mainstream dialects.

Keywords: Montreal forced aligner; forced alignment; sociolinguistics; Appalachian English

---

1 Introduction

Linguistics research often requires analysis and the identification of phonemes in audio recordings. In recent years, forced aligning software such as the Montreal Forced Aligner (MFA), the Forced Alignment & Vowel Extraction (FAVE) suite, and the Dartmouth Linguistic Automation (DARLA) suite have all become popular ways to accelerate the process of time aligning audio recordings. Time alignment is an essential step in research involved in phonetics, sociolinguistics, psycholinguistics, and language documentation. Despite how crucial time aligning is for linguistics research, it is generally a long and tedious portion of the work. Forced alignment was introduced to speed up the process of aligning speech using computer

programs. While forced alignment software has increased the speed of aligning speech, the accuracy can be affected by factors such as the quality of the recording (Reddy & Stanford, 2015) or the accuracy of the transcript (McAuliffe et al., 2017). Additionally, past research has shown that alignment software performs well with certain dialectal variations; however, when more variability from the training data was present, more major errors occurred in alignment (Mackenzie & Turton, 2019). Currently, research is focused on comparing existing alignment software (Gonzalez et al., 2019) and analysing the accuracy of the software on mainstream speech varieties (McAuliffe et al., 2017). Our research goal is to ascertain the accuracy of pretrained forced alignment software on nonmainstream varieties of American English, specifically, Appalachian English.

1.1 Appalachian English

Appalachia is a mountainous region in the Eastern United States which is officially comprised of 420 counties across 13 states, encompassing 26.1 million people (Pollard et al., 2022). Colloquially, however, the term Appalachia is more commonly used to refer to a much smaller region (Figure 1) in the Southeastern United States encompassing a small portion of six states (Ulack & Raitz, 1981; Reed, 2018). For our study, we used speakers from Hancock County, Tennessee. Hancock County is located in the northeast portion of Tennessee. The population of Hancock County is estimated to be 6846 people with 97% of the population being white and 79% of the population being over the age of 18 (United States Census Bureau, V2022).

Figure 1: *Image of Appalachia* (*adapted from Reed, 2018, and Ulack and Raitz, 1981*).

The Appalachian region is known to have significant variations in speech production, particularly in vowel production (Wolfram & Christian, 1976; Montgomery et al., 2020) and prosody (Reed, 2020). Grammatical differences also distinguish the varieties, with differences such as the pluralisation of nouns or *a*-prefixing on present participles of verbs (Montgomery & Hall, 2004). Pronunciation differences are primarily

documented in vowel changes such as the Southern Vowel Shift, which is characterised by the rotation of the nuclei of the tense and lax mid and high front vowels (Labov et al., 1972). Moreover, Appalachian English is characterised by monophthongisation of /a͡ɪ/ and vowel breaking (Thomas, 2003). Monophthongisation of /a͡ɪ/ is a hallmark of Southern speech but appears differently depending on individual and regional dialectal differences (Thomas, 2003). For Appalachian English, there are noted monophthongisation of /a͡ɪ/ in both pre-voiced and pre-voiceless contexts, which is different from other Southern US English varieties (Reed, 2016).

1.2 Montreal Forced Aligner

Of the many alignment software packages available, we chose to use MFA because it has unique features that make it better suited for analysing non-mainstream speech. For example, MFA uses triphone acoustic models and accounts for unknown words in the corpus. The triphone acoustic models had previously shown improved accuracy compared to other aligners. The MFA models are trained first using a monophone model, which means that the phonemes are looked at regardless of the surrounding phonological context. After the first pass with the monophone model, the data is run through a triphone model where the surrounding phonological context is considered when creating the model. A third pass focuses on key features that make phonemes different using LDA+MLLT. Finally, speaker differences are considered and the MFCC is calculated in the fourth pass (McAuliffe et al., 2017). MFA accounts for unknown words in the corpus, which allowed us to analyse how accurate the aligner is with no changes to either the transcript or the model. MFA treats unknown words as if they are their own unique phone, which allows for surrounding words to be aligned without major changes needing to be made to the dictionary (McAuliffe et al., 2017). Both factors make MFA a suitable candidate for aligning speech when there are minimal changes to the original data utilised by the computer program.

In addition, prior research showed MFA as having predictable error patterns (Gonzalez et al., 2020) and highly accurate transcriptions even with dialectal variations (Mackenzie & Turton, 2019). Mackenzie and Turton found that DARLA, which is a graphical user interface for MFA, was accurate with British English, which differs from the mainstream American English training data. However, the aligner performed worse when there were major differences from mainstream British English, such as with the Westray variety of Scots. In addition, the aligner produced more errors when speech was more rapid or had reduced speech patterns such as those found in the Sunderland variety of British English. The Sunderland variety features faster speech rates (Mackenzie & Turton, 2019) in addition to elisions such as *h*-dropping and some glottalisation of /p/, /t/, and /k/ (Burbano-Elizondo, 2008). For example, a reduction from "over and" to [ovn̩] in the Mackenzie and Turton study led to an extreme misalignment of the boundaries. They also reported many significant errors occurring due to laughter in the data sets. Whereas Mackenzie and Turton focused on British English, there has not been a comparison with divergent varieties of American English. Thus, our research focuses on how accurately MFA will align Appalachian English, a non-mainstream variety of American English.

1.3 LibriSpeech Model

MFA provides a model that was trained following the steps provided in Section 1.1 on data from the LibriSpeech study (Panayotov et al., 2015), a collection of opensource audiobooks. The LibriSpeech study chose audiobook data because it is closer to conversational speech than other sources. While creating the

dataset, many of the phonological processes such as deletions, substitutions, and transpositions had to be removed. Additionally, any words that did not align directly with the text were removed from the dataset as well. The LibriSpeech dataset was balanced for male and female speakers. There were 900,000 words isolated from the cleaned data that was collected, and of those the most frequent 200,000 words were included in the dataset. The corpus was created in varied sizes to account for differing needs of individuals; MFA used the full 1000 hour when creating their model. When Panayotov et al. (2015) tested the LibriSpeech model on the Wall Street Journal (WSJ) test sets, it was shown to be more accurate than the model trained on the WSJ directly. The dataset created by Panayotov et al. (2015) is open source and was released with Kaldi scripts, allowing for easy integration into the MFA, which is also based in Kaldi.

Because of the differences in dialects between the LibriSpeech training data and our data, sociolinguistic interviews from Appalachia, we hypothesised that there would be inaccuracies in the alignment specifically for the features of Appalachian English that vary from mainstream varieties. By studying one of the non-mainstream dialects of Appalachian English, we can offer a better understanding of the limits of a pretrained forced aligner when the test data varies greatly from the training data. By using MFA on our data, we anticipate that we will obtain pertinent information on reliability of forced aligning software in addition to patterns in errors with differences from training data. Thus, our study aims to inform future researchers of the accuracy of forced aligners when using pretrained data on non-mainstream dialects.

2 Methods

To assess the performance of MFA, we aligned four recordings obtained from the Appalachian region. The demographics of the recordings, illustrated in Table 1, include three male participants and one female participant of varying ages, including one young adult, two middle aged adults, and one older adult. The recordings were semi-structured sociolinguistic interviews obtained by Reed (2016) from Hancock County, TN. The recording ranged from about 45 to 90 minutes long depending on the speaker. Both recordings and orthographical transcription were obtained from Reed (2016) for our analysis.

Table 1: *Participant Demographics*

Participant	**Demographic**
Nathan	40-year-old Male
Misty	37-year-old Female
Joey	29-year-old Male
Hugh	84-year-old Male

After obtaining the raw data, we first converted the files into a single (mono) channel using a script in Praat. From there, we obtained the acoustic model trained on the LibriSpeech dataset and the LibriSpeech dictionary. We then force-aligned our recordings following the guidelines put in place by McAuliffe et al. (2017).

After aligning the files, we hand-corrected the entire length of each recording. We identified the locations where the aligner was incorrect and separated them into four categories: boundary errors (Section 2.1), phoneme errors and gap errors (Section 2.2), and other errors (Section 2.3). The errors were recorded onto a spreadsheet with the timestamp of the error, error type, and the phoneme the error occurred on. After identifying errors in the alignments, we compared the total number of errors that occurred to the total

number of phonemes in each recording, in addition to charting the error type and the phoneme the error occurred on.

2.1 Boundary Errors

Boundary errors were determined where the beginning or end of the boundary placement by MFA was more than 100 ms different from the hand-corrected version. McAuliffe et al. (2017) found an average difference in boundary placements to range from 17 ms to 26 ms, which is comparable to the intertranscriber variations by human annotators. By focusing on larger differences than those reported in previous research, we are ensuring any errors noted were serious deviations from that produced by human annotators. Therefore, boundary errors in the present work represent major differences in the alignment. For example, Figure 2 depicts the differences in the realisation of the word <was> with the top two rows depicting the MFA alignment and the bottom two rows the hand-corrected version. The phonemes are depicted using the ARPAbet, as the MFA alignment is provided with the ARPAbet.

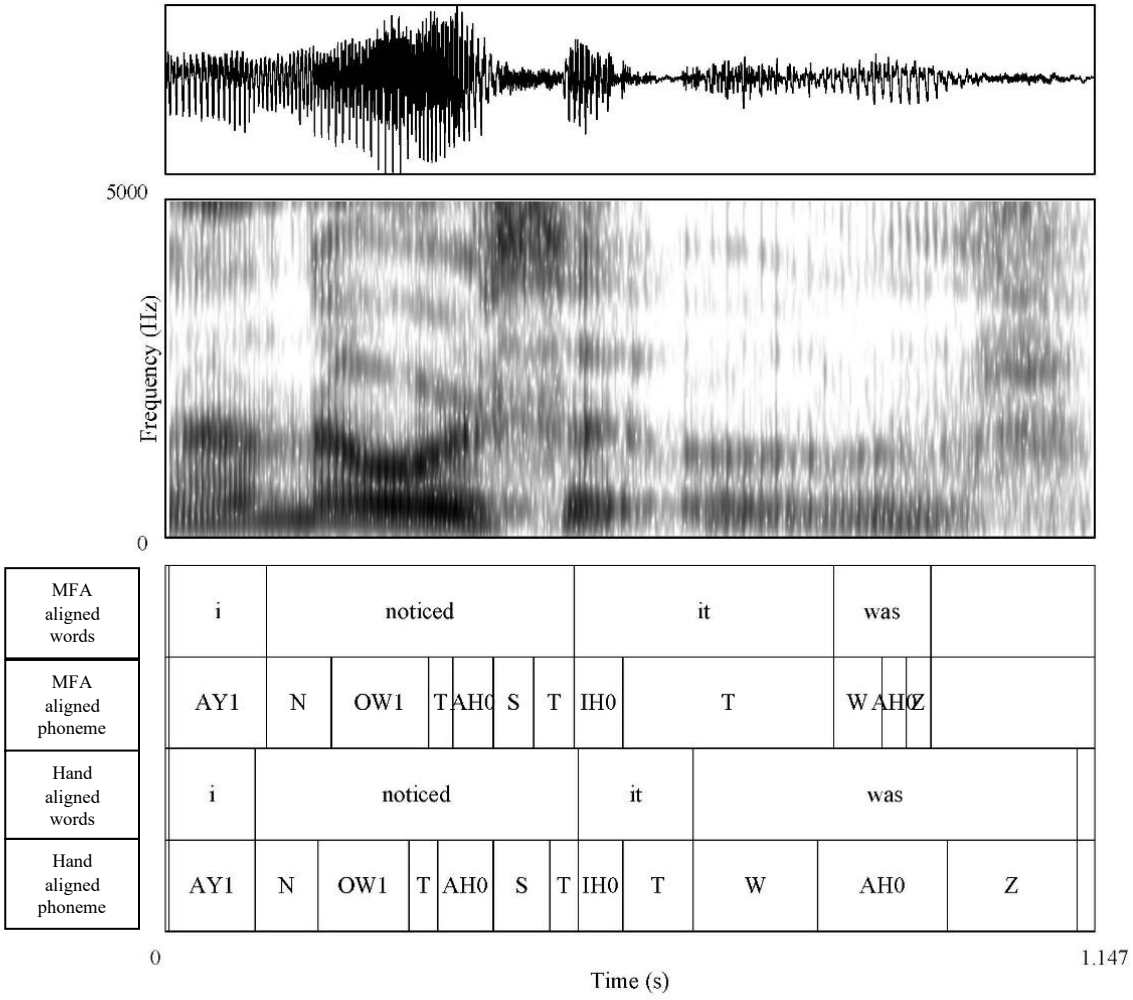


Figure 2: *Example boundary error.*

2.2 Phoneme and Gap Errors

Phoneme errors are differences in phonemes from the hand-corrected version to the MFA aligned version. For example, a common error occurred with the indefinite article <a> which has two variant productions. Often /e͡ɪ/ was used in place of /ə/ or vice versa; in these instances, the speaker's production and MFA's identification were inconsistent for this word. It is important to note that while this paper depicts phonemes using the International Phonetic Alphabet, the English version of MFA utilises the ARPAbet to transcribe phonemes. Some of the dialectal variation could not be fully captured because the results are output using the ARPAbet which does not have characters to capture some of the nuances of the monophthongisations and vowel breaking typical to this region.

Gap errors occurred when a period of silence was placed within a phoneme in the MFA alignment. Typically, they were depicted as having a phoneme followed by a gap and then the same phoneme with no gap present in the recording. For example, the sibilant /s/ would appear to have been produced twice in a row with a gap between productions, despite being from the same phoneme.

2.3 Other Errors

The other category was split into errors in the transcription, errors in the dictionary, errors due to background noise, and errors due to overlapping speech. Errors in the transcription and dictionary were attributed to words absent from the LibriSpeech dataset or words misspelled in the original orthographic transcript. Errors due to overlapping speech or noise tended to be errors in boundaries that occurred due to competing signals in the recording. For example, Figure 3 depicts the differences due to overlapping speech between the words <little bit> with the top four rows depicting the MFA alignment and the bottom four showing the hand-corrected version.

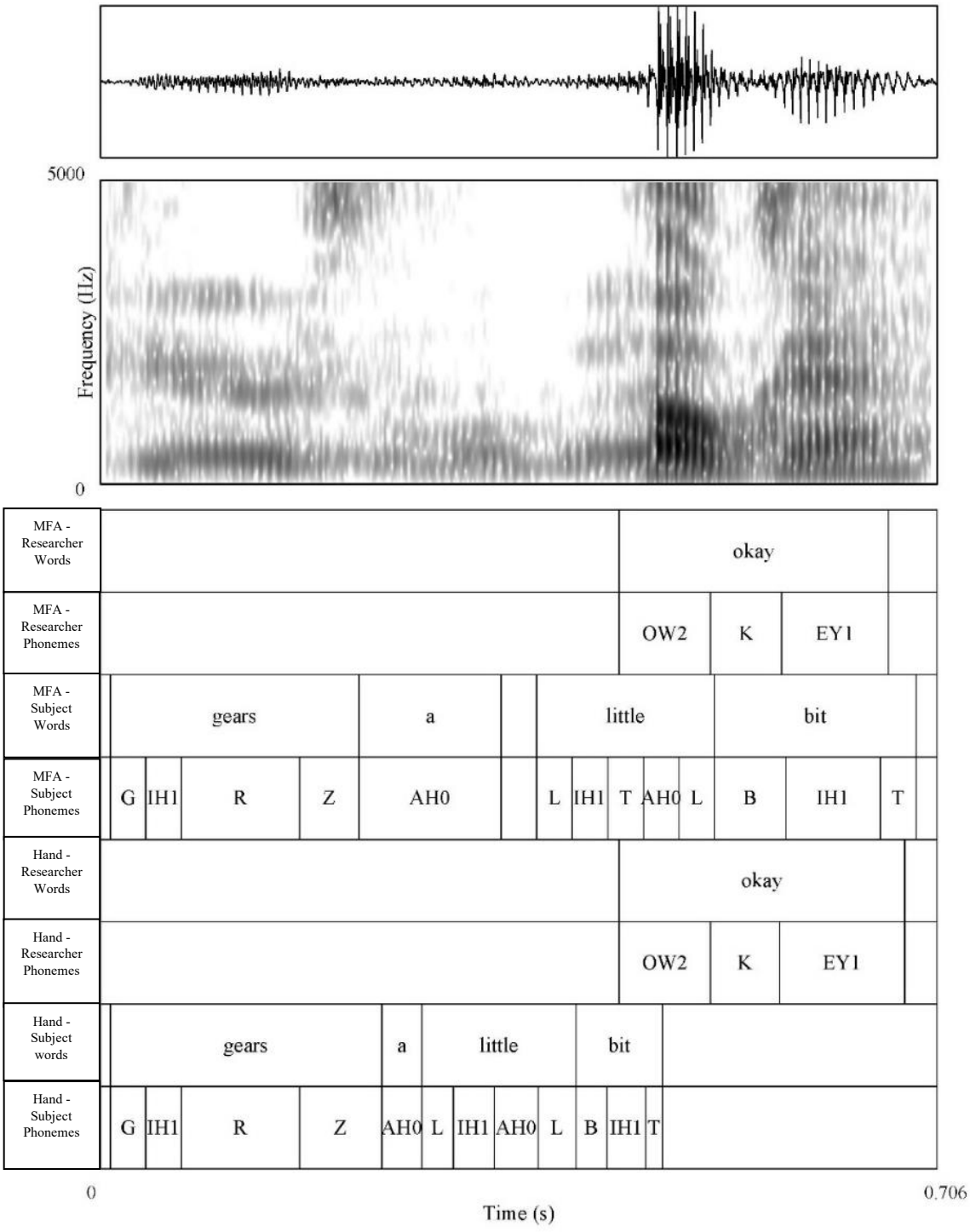


Figure 3: *Example of overlapping speech error.*

3 Results

After comparing the hand aligned version to those aligned by MFA, we found a low error rate with typically less than 1% of phonemes having errors (Table 2).

Participant	Percent of Error
Nathan	0.75%
Misty	0.87%
Joey	0.54%
Hugh	1.3%

Table 2: *Percentage of error for each participant*

Of the errors that occurred, 45.5% were due to errors in the other category such as from dictionary, transcript or overlapping speech problems. The second most common error was errors in boundary placement, which made up 38.4% of the errors. The final 16.1% of errors came from phoneme and gap error types.

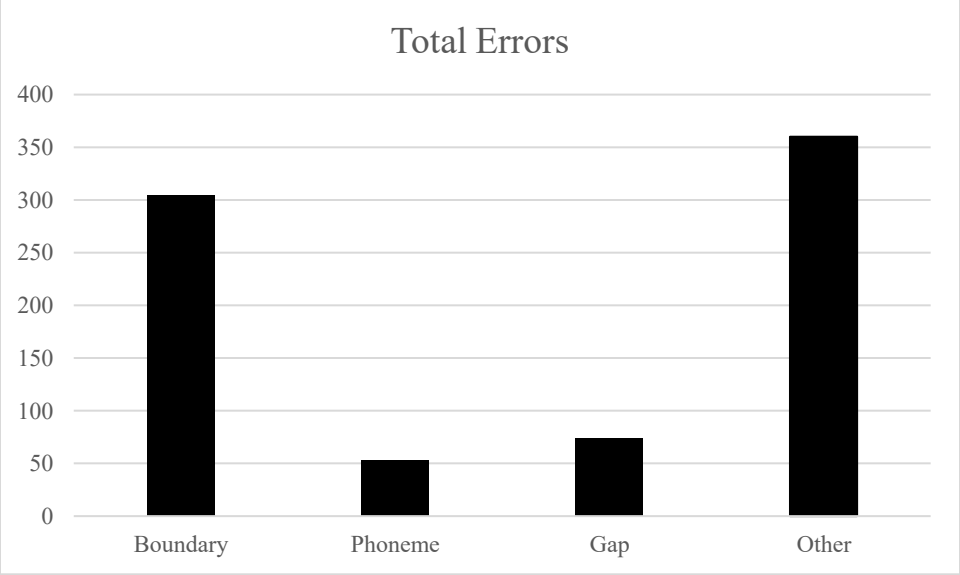


Figure 4: *Total number of errors by type.*

The other errors are primarily characteristic of the use of MFA without changes to the dictionary or acoustic model. This category is composed of four subcategories and the total number of errors per subcategory can be found in Figure 5. Many of the errors that occurred were due to out-of-vocabulary (OOV) words or words not in the dictionary. The majority of the words labelled OOV were local terminology, such as place names, which were not within the dictionary. Incorrectly spelled words in the transcription made up the next largest section of OOV words. In addition, a small portion of the OOV words came when the transcription used numerals instead of spelling out numbers.

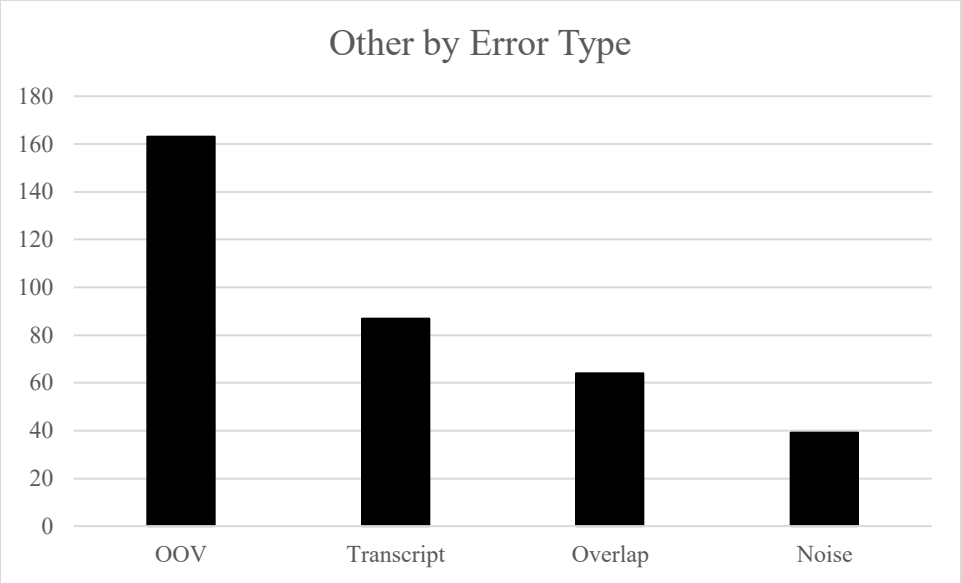


Figure 5: *Number of errors by type in the other category.*

Boundary errors are the second most common error with roughly 1/3 of the errors being from this category. The errors occurred primarily with vowels. Specifically, /ʌ/ was the most common phoneme for errors, though errors also occurred for /æ/, /a͡ɪ/, and /i/. Boundary errors were also common with fricatives and affricates. Of the fricatives, the phoneme /ð/ had the highest rate of error, followed by /s/ and /z/. Boundary errors were the highest for Hugh, the older adult speaker, and Misty, the female speaker. There were very few boundary errors for both the young and middle-aged male speakers.

Though phoneme and gap errors made up the smallest portion of errors seen, there are a few notable occurrences in this category. In the gap category, errors primarily occurred on the sibilant /s/ and were again more common for the older adult and female speakers. The were few phoneme errors though those that did occur primarily affected vowels and approximates, specifically /ɝ/ and /ɹ/. Errors occurred for the phoneme /ɝ/ for all three male speakers and the issues with /ɹ/ primarily occurred with the female speaker.

4 Discussion

Ultimately, analysis of the performance of MFA revealed low percentages of errors even with respect to the features unique to Appalachian English varieties. We expected higher error rates, particularly with regional specific variations and background noise. However, the aligner performed surprisingly well in these conditions. It should be noted that we had a small sample size, so some of the issues that occurred may not generalise in a larger population of Appalachian English speakers. Given our results and those of the past research noted in the introduction, researchers can expect high quality alignments for Appalachian varieties of English even when the baseline model varies.

There was one error that should be noted which occurred on sibilants in the gap category and more specifically the /s/ phoneme. Though gap errors account for a small percentage of the total error, these errors occurred for both our female and older male participants. This error stood out because the /s/ phoneme is not one that has noted acoustic differences for the Appalachian English variation as compared to other

American English varieties. Additionally, Gonzalez et al. (2020) found that MFA performed better with fricatives like the /s/ phoneme when compared to most other speech sounds. Further research should delve in to whether this issue occurs with other speakers focusing on a larger sample or different dialects.

When considering variations in dialect, the use of the ARPAbet for the English version of MFA lead to some of the errors. The ARPAbet is used to represent the phonemes and allophones of mainstream American English (Klautau, 2001), but it is not all encompassing especially for changes in vowels such as monophthongisations. There were errors in the phonemes that had no better alternative within the ARPAbet. Therefore, when studying a language with known monophthongisations or diphthongisations, researchers should use a dictionary that utilises IPA to better depict the variations.

Many of the errors occurred due to the issues in the transcript, especially in relation to out-of-vocabulary (OOV) words. The LibriSpeech dictionary is comprised of the 200,000 most frequent words within the LibriSpeech dataset. Therefore, while it accounts for frequently used words and phrases, it neglects more regionally specific terms or productions. For example, local place names frequently occurred within our recordings. Because the place names are not common terms for all English speakers, they were marked as OOV. In addition, the production of place names can vary based on context or speaker and would therefore require multiple variations of the word to be added to the dictionary. Errors with overlapping speech and noise were both low for our dataset. However, a consideration that must be recognised is the high-quality recordings including minimal background noise and overlapping speech. When either of these instances occurred, errors commonly accompanied them. In recordings, where noise and overlapping speech may be an issue, there may be a higher rate of error. Thus, further research needs to be completed into the accuracy of MFA with increased noise or overlapping speech.

5 Conclusion

Regarding the future, there are two main ways to expand on the present study. Firstly, the research in this study focuses on the use of MFA with a pretrained model. MFA offers a functionality to create a language specific model that can be applied to Appalachian English. Because the aligner performed beyond expectations with the baseline model, it would be interesting to analyse the change in accuracy with a dialect specific language model. In addition, analysing the accuracy of aligners on archival data or data with more background noise could be illuminating. The recordings used in our study were taken in relatively quiet environments with efforts made to reduce noise when possible. Also, when noise was present in the recording, there were typically errors noted. Therefore, future works should look at the accuracy of MFA on recordings with more background noise. Furthermore, we had a small sample size, so some of the issues we found may not be generalised or may be a larger issue. Thus, subsequent research should look at a larger sample of speakers from the region. Finally, further research could focus on the standard performance of MFA on other non-mainstream dialects.

Overall, MFA performed well with Appalachian English, as seen by the low error rate and predictable errors. Many of the errors that occurred were due to the transcription and dictionary, so adding novel words and double-checking transcription will eliminate many of these errors. Being aware of boundary errors is important but with the low percentage of errors, spot checking is going to be more important than analysing the entire recording. Regardless, researchers can have confidence that MFA can perform well with Appalachian English when using high quality recordings.

6 References

Burbano-Elizondo, L. (2008). *Language variation and identity in Sunderland*. [Doctoral dissertation]. University of Sheffield.

Gonzalez, S., Grama, J. & Travis, C. E. (2020). Comparing the performance of forced aligners used in sociophonetic research. *Linguistics Vanguard*, 6(1). 10.1515/lingvan-2019-0058.

Klautau, A. (2001). ARPABET and the TIMIT alphabet [Archived file]. Retrieved 12/03/20, from <https://web.archive.org/web/20160603180727/http://www.laps.ufpa.br/aldebaro/papers/ak_arpabet01.pdf>.

Labov, W., Yaeger, M., and Steiner, R. (1972). *A quantitative study of sound change in progress: Volumes 1 and 2*. Report on National Science Foundation Contract NSF-GS-3287. University of Pennsylvania, Philadelphia, PA.

Mackenzie, L., & Turton, D. (2020). Assessing the accuracy of existing forced alignment software on varieties of British English. *Linguistics Vanguard*, 6(1), 20180061.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Proc. Interspeech* (pp. 498-502). 10.21437/Interspeech.2017-1386.

Montgomery, M., Heinmiller, J. K. N., & Hall, J. H. (2021). *Dictionary of Southern Appalachian English*. The University of North Carolina Press.

Montgomery, M. B., & Hall, J. S. (2004). *Dictionary of Smoky Mountain English*. University of Tennessee Press.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015, April). LibriSpeech: an ASR corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech, and signal processing (ICASSP)* (pp. 5206-5210). IEEE.

Pollard, K., & Jacobsen, L. (2022). The Appalachian Region: A Data Overview from the 2016-2020 American Community Survey. *Appalachian Regional Commission*. Retrieved from <https://www.arc.gov/report/the-appalachian-region-a-data-overview-from-the-2016-2020-american-community-survey/>.

Reddy, S., & Stanford, J. (2015). A Web Application for Automated Dialect Analysis. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (pp. 71-75). Association for Computational Linguistics.

Reed, P. E. (2016). *Sounding Appalachian: /ai/ monophthongization, rising pitch accents, and rootedness* (Doctoral dissertation, University of South Carolina).

Reed, P. E. (2018). The importance of Appalachian identity: A case study in rootedness. *American Speech: A Quarterly of Linguistic Usage*, 93(3-4), 409-424.

Reed, P. E. (2020). The Importance of Rootedness in the Study of Appalachian English Case Study Evidence for a Proposed Rootedness Metric. *American Speech*, 95(2), 203-226.

Thomas, E. R. (2003). Secrets revealed by Southern vowel shifting. *American Speech*, 78(2), 150-170.

Ulack, R., & Raitz, K. (1981). Appalachia: A Comparison of the Cognitive and Appalachian Regional Commission Regions. *Southeastern Geographer*, 21(1), 40-53.

United States Census Bureau. (V2022). Hancock County, Tennessee. Retrieved 10/05/2024, from <https://www.census.gov/quickfacts/fact/table/hancockcountytennessee/PST045223/>.

Wolfram, W., & Christian, D. (1976). *Appalachian speech*. Virginia Center for Applied Linguistics.

About the Author

Amy Cox is a former student at the University of Alabama in the United States. Her research looked at the integration of computer based alignment tools and the Appalachian English dialect. She currently is in graduate school pursuing a Doctorate of Audiology and Master of Public Health.

Acknowledgements

I would like to thank Dr. Paul Reed for all of his support through the years while I was working on this project. I would also like to thank all of the amazing professors in the Randall Research Scholar Program, Dr. Gray, Mrs Batson, Darren and Dr. Sharpe. I learned so many strong skills in computer programming and public speaking through my time in the Randall Research program.