Greek Forced Alignment: Assessing the Accuracy of the Montreal Forced Aligner

Anonymous ACL submission

Abstract

Forced alignment has allowed for the rapid creation and annotation of corpora. In this study we examine the Montreal Foreced Aligner and its accuracy of aligning Greek data. Using a conversational Greek corpus we train a small grapheme-to-phoneme model and use this model to align the entire corpus. We compare our results to various previous studies of the MFA and other forced alignment software and conclude that forced alignment greatly increases the ability to create new corpora for low-resource and understudied languages.

1 Introduction

Forced alignment software has seen an increased use in the fields of sociophonetics and corpus research. The manual alignment of word and phone boundaries can easily take hundreds of hours even for smaller corpora. Forced aligners such as the Munich Automatic Segmentation System (Kisler et al., 2017), the Forced Alignment & Vowel Extraction suite (Rosenfelder et al., 2014), and the Montreal Forced Aligner (MFA, McAuliffe et al., 2017) allow for the automatic segmentation and alignment of words and phones. Forced alignment has allowed researchers to speed up the alignment process giving them more time to spend on the actual research question at hand.

Various recent studies have shown the usefulness and accuracy of Forced Alignment Software used in the study of English (Gonzalez et al., 2020, MacKenzie and Turton, 2020, Meer, 2020). These results are to be expected, however, given that the software used contains various pre-trained acoustic models for English. With the development of the Prosodylab-Aligner (PL-A, Gorman et al., 2011) and the newer MFA the use of forced alignment has been extended to understudied and low-resource languages. Johnson et al. (2018) showed that the PL-A provides a substantial increase in efficiency for the aligning of understudied languages.

In the current study we will investigate the use of the MFA on a spoken Greek dataset. We will investigate the ease of use of the MFA alignment pipeline and compare the results with a human annotated segment. The paper is divided into 4 main sections. In section 2 we discuss what forced alignment entails and how it works. In section 3 we outline the pipeline that we created for aligning Greek data. In section 4 we discuss our results and how the forced alignments compare with human aligned data. In section 5 we will give a summary of our research and some suggestions for further study.

2 Background

Forced alignment is the automatic segmentation of a transcript and aligning the phonemic boundaries to the speech signal. The process of forced alignment starts with creating a grapheme-to-phoneme (G2P) dictionary which is used to map the orthographic transcription to the phonemic representation. The data is then segmented and the phonemic representation is aligned with the continuous speech signal creating a time-aligned transcript of the data. Forced Aligners like the MFA often come with pre-trained models for languages like English, German, and Spanish. However, minority languages and sparsely documented languages often do not have pre-trained models due to the lack of data available. The MFA has the ability to train a custom model for new languages not included in the MFA. The MFA uses the Kaldi toolkit (Povey et al., 2011), which is based on a Hidden Markov Model, to generate models for aligning the data.

For our alignment task we are working with a corpus of interviews conducted in Greek (Anastasiadis et al., 2017). Each audio file contains an interview conducted between two people. The type of speech in this corpus is colloquial and informal and does not contain much overlap in speaking due to the interview register. Since the interviews were

File	Phone Count
004	73
013	53
120	77
128	105
400	90
403	45

Table 1: Number of phones analyzed per File

conducted one-on-one there is often only one person speaking at a time and additionally there is not much background noise. This means that forced alignment is a viable option for speeding up the corpus alignment. In this study we only aligned the speech of the person being interviewed. We will do a quantitative analysis comparing the alignments from the MFA to those corrected by a human. In table 1 we show the number of phones that were analyzed.

3 Pipeline

Our alignment pipeline is composed of three steps: (1) *pre-processing*, both of the transcripts and the audio files, (2) *alignment*, and (3) *consolidation* of the aligned data.

3.1 Pre-Processing

We are only interested in aligning the speech of the person being interviewed which means we need to extract the tier corresponding to that speaker. When selecting a tier to be used as the base for the alignment it is important to pick a tier of medium length. Intervals shorter than 100ms will not be aligned by the MFA (McAuliffe et al.) and if long intervals are used the alignment may deteriorate towards the end of the interval. For the alignment discussed in this paper we used the "PrWord" tier and extracted it from the TextGrid files. This tier contains prosodic words and is a good fit for forced alignment.

After the tier is extracted we ran a script which fixes various punctuation issues. Some issues we fixed were removing extra spaces and adding spaces after various punctuation symbols.

After processing the TextGrid files we processed the audio files. The MFA is based on the Kaldi toolkit which requires .wav files by default. We used the command line tool "ffmpeg" to convert the audio files to the .wav format. After obtaining the .wav audio files we used a praat script to resample the audio files to 16kHz as required by the MFA.

3.2 Alignment

Because the MFA does not come with a pretrained model for Greek we have setup a system to bootstrap the alignment process. First we trained a small Grapheme to Phoneme (G2P) dictionary using 1 hour and 40 minutes of audio. This G2P model is then used in our pipeline to construct a pronunciation dictionary of the entire corpus that is being aligned. After constructing this dictionary we align the corpus using the generated pronunciation dictionary.

3.3 Consolidation

After all alignments are generated we extract the manually generated tiers such as the sentence-level and prosodic word-level tiers and merge these tiers with the MFA generated tiers into a single TextGrid file. At the end of this step we have an annotated corpus where each audio file contains sentence, prosodic word, word, and phone alignments.

4 Results

In this section we will first present the quantitative results that were obtained by comparing the human aligned data with that of the MFA data. We will then compare our results to those of previous studies where the accuracy of forced alignment software was reported on. We will focus only on phone alignments.

4.1 Quantitative Results

Our quantitative results are calculated based only on the phone onset boundaries. We do not calculate the phone offset since the onset of one phone is the offset of the previous phone, with exception of the very first phone.

As can be seen in table 2, the mean onset boundary displacement ranges from 18.6 ms to 25.0 ms with an average of 65.8% of boundaries, over all files sampled, having a displacement within 20 ms.

4.2 Performance Assessment

First we will make some comparisons to studies looking at the alignment of languages for which a pre-trained model is provided. (McAuliffe et al., 2017) reports an accuracy of 77% within a tolerance of 25 ms for the English Buckeye corpus (Pitt et al., 2007) which consists of 20.7 hours of conversational speech. While this study uses a different

File	Mean Displacement in ms	% within 20 ms
004	19.8	69.9%
013	25.0	66.0%
120	20.3	67.5%
128	23.7	57.1%
400	18.6	70.0%
403	19.6	64.4%

Table 2: Summary Statistics

cutoff we can see that our accuracies are not far off from this baseline study.

(Johnson et al., 2018) tested the PL-A on Tongan data to see whether forced alignment is a viable option for speeding up alignment of understudied and low-resource languages. They reported accuracies of 73% and 75% when comparing forced alignments generated by the PL-A to those of a human annotator within a 20 ms tolerance. (Tang and Bennett, 2019) used the MFA to train alignment models for two Mayan languages and reported an accuracy of 71.7% within 20 ms.

5 Conclusion

We have shown that the MFA can be used effectively to speed up the alignment process of corpora for languages that do not have a pre-trained model included with the MFA. With minimal effort a pipeline can be setup for any language that has at least a few hours worth of audio data that can be used to create a small G2P model.

While a forced aligned corpus may not be accurate enough for further quantitative inquiries without human correction, for many languages it will be beneficial to have a search-able corpus that contains roughly aligned phones and word boundaries as this could be a stepping stone towards deeper insights and studies.

References

Tassos Anastassiadis, Angela Ralli, Sakis Gekas, Panayiotis Pappas, Alexandra Siotou, Charis Tsimpouris, and Symeon Tsolakidis. 2017. Immigration and language in canada: Greeks and greek-canadians. [electronic database] retrieved from https://db.immigrec.com/cas.

Simon Gonzalez, James Grama, and Catherine E. Travis. 2020. Comparing the performance of forced aligners used in sociophonetic research. *Linguistics Vanguard*, 6(1).

Kyle Gorman, Jonathan Howell, and Michael Wagner. 2011. Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, 39(3):192–193.

Lisa M. Johnson, Marianna Di Paolo, and Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing prosodylab-aligner with tongan data. *Language Documentation Conservation*, 12:194–203.

Thomas Kisler, Uwe Reichel, and Florian Schiel. 2017. Multilingual processing of speech via web services. *Computer Speech Languages*, 45:326–347.

Laurel MacKenzie and Danielle Turton. 2020. Assessing the accuracy of existing forced alignment software on varieties of british english. *Linguistics Vanguard*, 6(1).

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. *Montreal Forced Aligner [Computer program] version 2.0.0a24*.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. 2017. Montreal forced aligner: Trainable text-speech alignment using kaldi. volume 2017, pages 498–502.

Philipp Meer. 2020. Automatic alignment for new englishes: Applying state-of-the-art aligners to trinidadian english. *The Journal of the Acoustical Society of America*, 147(4):2283–2294.

Mark A. Pitt, Laura Dilley, Keith Johnson, Scott Kiesling, William Raymond, Elizabeth Hume, and Eric Fosler-Lussier. 2007. Buckeye corpus of conversational speech (2nd release).

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society.

Ingrid Rosenfelder, Josef Fruehwald, Keelan Evanini, Scott Seyfarth, Kyle Gorman, Hilary Prichard, and Jiahong Yuan. 2014. Fave (forced alignment and vowel extraction) program suite v1.2.2.

Kevin Tang and Ryan Bennett. 2019. Unite and conquer: Bootstrapping forced alignment tools for closely-related minority languages (mayan). In *Proceedings of the 19th International Congress of Phonetic Sciences*, pages 1719–1723.