WHAT DO FORCED ALIGNMENT LIKELIHOOD SCORES TELL US ABOUT THE ALIGNED SPEECH?

Ayushi Mrigen [*1], **Daniel Brenner** [†2] **& Benjamin V. Tucker** [‡2]
[1]Department of Mathematics, Indian Institute of Technology, Kharagpur
[2] Department of Linguistics, University of Alberta, Edmonton, Canada

1 Introduction

Forced alignment is an automatic speech recognition procedure frequently employed in the speech sciences. It is typically used to time-align a string of phones with a speech recording on the basis of some operationalization of acoustic similarities to averaged representations of the phones over a corpus. Ordinarily these similarity measures are not of interest to researchers, who simply want to know where in time the phones are. This study investigates the potential usefulness of these alignment scores for phonetic research, considering the phonetic properties which are and are not represented in the scores. The hypothesis is that these scores can be used to infer acoustic relationships along some dimensions between the aligned speech and the corpus on which the models were trained.

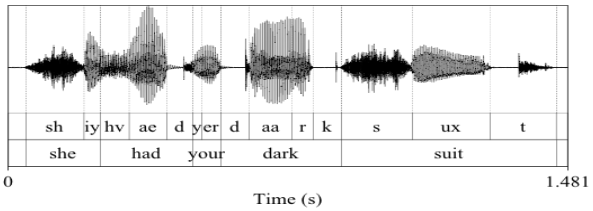

Figure 1: Sample waveform for a portion of a TIMIT sentence shown with the phone sequence aligned to it.

2 Method

2.1 What do alignment scores mean?

Forced alignment is essentially a system of optimized matching. The models of the system are trained on a corpus to provide an averaged acoustic representation of each phone in the corpus. During alignment, the system is given a string of phones and a sound recording, and it apportions continuous sections of the sound recording to the phones to maximize the similarity of the aligned portions to the averaged models. The typical forced aligner (like the Penn Aligner) is implemented using Hidden Markov Models [1]. This approach to symbol-sound mapping has a variety of practical applications.

For the problem of isolated word recognition, if each spoken word is represented by a sequence of speech vectors or observations O, defined as; $O = o_1, o_2, \ldots o_T$. The isolated word recognition problem can then be regarded as that of computing; arg max $\{P(w_i|O)\}$, where w_i is the i'th vocabulary word. However, since this probability is not computable directly we use the Bayes' rule which gives us: $P(w_i|O) = (P(O|w_i)P(w_i))/P(O)$

2.2 Modification in the typical forced aligner

A typical forced aligner, like the Penn Forced Aligner [4] takes two inputs, the audio file which contains the speech, and the text file with the transcription of the speech. The aligner produces a PRAAT [2] TextGrid file as output, which contains the alignment in typically two tiers: (1) the word level, and (2) the phone level (Figure 1). This output contains the name of the word or the phone, and the start and end time of occurrence of the same within the speech file. Within the alignment script, the Penn Aligner calls the function HVite. The original output of HVite is in the form of a MLF file, which contains, among other things the probability score which the Viterbi algorithm computes while aligning the file. This output, from the Hidden Markov model is of interest to us. In order to investigate more about the score, the code to convert the MLF file to the TextGrid was modified to simultaneously print these probability scores alongside the name of the phone/word.

2.3 Measuring distances between vowels

In order to test the hypothesis that the log likelihood scores obtained from forced aligners can be employed to tell us about phonetic distances, these scores need to be compared with an existing measure of distance. A straightforward perceptually relevant geometric distance can be computed for vowels in the F1 & F2 space, so we selected this metric for comparison. A set of eight vowel phones was selected and the distance between every two pairs was measured using the standard F1-F2 technique, and by the forced aligner. In order to measure the distance between each pair, every speech file was aligned with the text file for every vowel (tokens of [i] were aligned with the [ɪ] phone, the [e] phone, the [ɛ] phone, etc.). It was noticed that for each vowel ranking obtained in the descending order of scores matched approximately with the increasing order of F1-F2 distances.

2.4 Training on HTK

Different kinds of training have an effect on the pattern of alignment scores obtained. As a result, two sets of training models were created after training on different sets of speech data. The first model was trained on the TIMIT corpus [5], which is composed of read sentences. The other model was trained using the Buckeye corpus [3], which contains interview recordings. The two corpora were chosen

[*] ayushimrigen11@gmail.com
[†] brenner@ualberta.ca
[‡] benjamin.tucker@ualberta.ca

to compare the differences in the patterns of scores from the read speech and the conversational speech models.

2.5 Aligning larger corpora

For testing purposes a holdout section was selected from both TIMIT and Buckeye. While the TIMIT corpus already has a test section in the corpus, a bootstrapped holdout section was created for the Buckeye corpus. For both these sections, they were first aligned with their original transcription, using the models trained on data from the same corpus. From the TextGrid files obtained by this alignment, a list of files containing each vowel was created. The original vowel in the file was substituted one by one with every other vowel in the list ([i] was substituted with the [ɪ] phone, the [e] phone, the [ɛ] phone, etc. one by one). The scores of each such alignment was recorded. In this way we obtained the cross comparison scores between each vowel pair.

3 Results

A ranking of this cross comparison is shown in Table 1. The log likelihood scores were compared to the F1-F2 phonetic distances. The result of a regression for each vowel is shown in Table 1.

Table 1: LogLik ranking & correlations by TIMIT test vowel.

Vowel	Log Likelihood ranking	Correlation
aa	[aa, ow, ah, ae, ey, uw, ih, iy]	-0.847
ae	[ae, ey, ah, ow, aa, ih, uw, iy]	-0.61
ah	[ow, aa, ah, ae, ih, ey, uw, iy]	-0.761
ey	[ey, iy, ae, uw, ih, ow, ah, aa]	-0.800
ih	[ih, ae, iy, uw, ey, ow, ah, aa]	-0.761
iy	[iy, ey, uw, ih, ae, ow, ah, aa]	-0.846
ow	[ow, aa, ah, ae, uw, ey, ih, iy]	-0.884
uw	[ow, uw, iy, ih, ey, ah, ae, aa]	-0.429

These results show a negative correlation with the F1-F2 distances measured manually. However, when a similar analysis was done with Buckeye aligned on the two sets of models, and TIMIT aligned on Buckeye, the negative correlation appears weaker (Table 2).

Table 2: Correlation coefficients comparing data from the different models.

Vowel	Buck-Buck	Buck-TIMIT	TIMIT-TIMIT	TIMIT-Buck
aa	0.015	-0.102	-0.848	0.036
ae	-0.222	-0.312	-0.613	-0.428
ah	-0.021	-0.090	-0.761	0.100
ey	-0.576	0.522	-0.800	0.155
ih	0.056	0.480	-0.762	0.101
iy	-0.074	0.163	-0.846	0.164
ow	-0.109	-0.127	-0.884	-0.167
uw	0.054	0.108	-0.430	0.047

When comparing the Buckeye corpus model we find that the correlation coefficients between distances and scores for models trained on both corpus is low. With the TIMIT corpus the correlation coefficients are high in general when TIMIT is aligned with TIMIT models, except for 'uw' and the coefficient is low when it is aligned on Buckeye models. Figure 2 illustrates the distribution of scores for the diphthong [aɪ] for the two different holdout datasets when fit using the models from the two different corpora. This example illustrates the very different results that can be obtained by using different models and data.

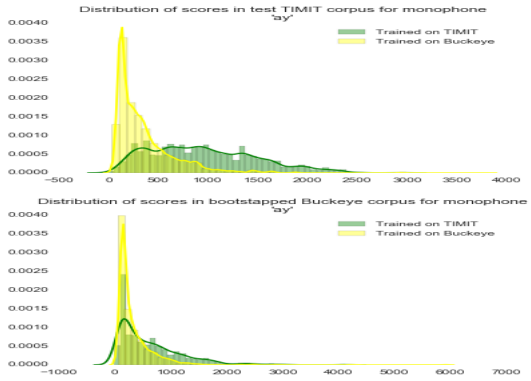

Figure 2: Distributions of alignment scores for the same phone on TIMIT and Buckeye, trained on the two corpora.

4 Conclusion

These probability measures seem to show a relationship with some acoustic characteristics of the segments. It is clear that these scores behave differently with different training, and this relationship may be exploitable to assess typicality of sounds within a given corpus context.

Acknowledgments

Funding for this project was provided by the University of Alberta Research Experience Program and Social Sciences and Humanities Research Council: #435-2014-0678.

References

[1] Baum, L. E.; Petrie, T. (1966). "Statistical Inference for Probabilistic Functions of Finite State Markov Chains". *The Annals of Mathematical Statistics* 37 (6): 1554–1563.

[2] Boersma, Paul & Weenink, David (2016). Praat: doing phonetics by computer [Computer program]. Version 6.0.19, retrieved 13 June 2016 from http://www.praat.org/

[3] Pitt, M. A., Johnson, K., Hume, E., Kiesling, S., & Raymond, W. 2005. The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication* 45, 89-95.

[4] Yuan, J. & Liberman, M. 2008. Speaker identification on the SCOTUS corpus. *Proceedings of Acoustics 2008*.

[5] Zue, V. & Seneff, S. Transcription and Alignment of the TIMIT Database. *Proceedings of the 2nd Meeting on Advanced Man -- Machine Interface through Spoken Language* 1988, 11.1-11.10.