College of Humanities and Social Sciences
LINGUISTICS PROGRAM
George Mason University

Confidence intervals for forced alignment with the Mason-Alberta Phonetic Segmenter

Matthew C. Kelley
George Mason University

Mason PhonLab

1. Introduction

- Forced aligners are error-prone, sometimes cascadingly
- Estimates of uncertainty might help flag bad alignments, but current systems do not estimate it
- The Mason-Alberta Phonetic Segmenter (MAPS) [1] now provides confidence intervals (CIs) using a neural network ensemble

2. Methods

- 10 different acoustic models trained during creation of MAPS
  - Forms an ensemble of models
- Assessed on TIMIT [2] and Buckeye [3] speech corpora
- Forced alignment run with each model in ensemble
- Center taken as median of ensemble's 10 different boundary estimates
- A 97.85% CI around the median is calculated with order statistics (2nd and 9th ordered values)
- Boundary accuracy compared using mean and median absolute error compared to “ground truth”

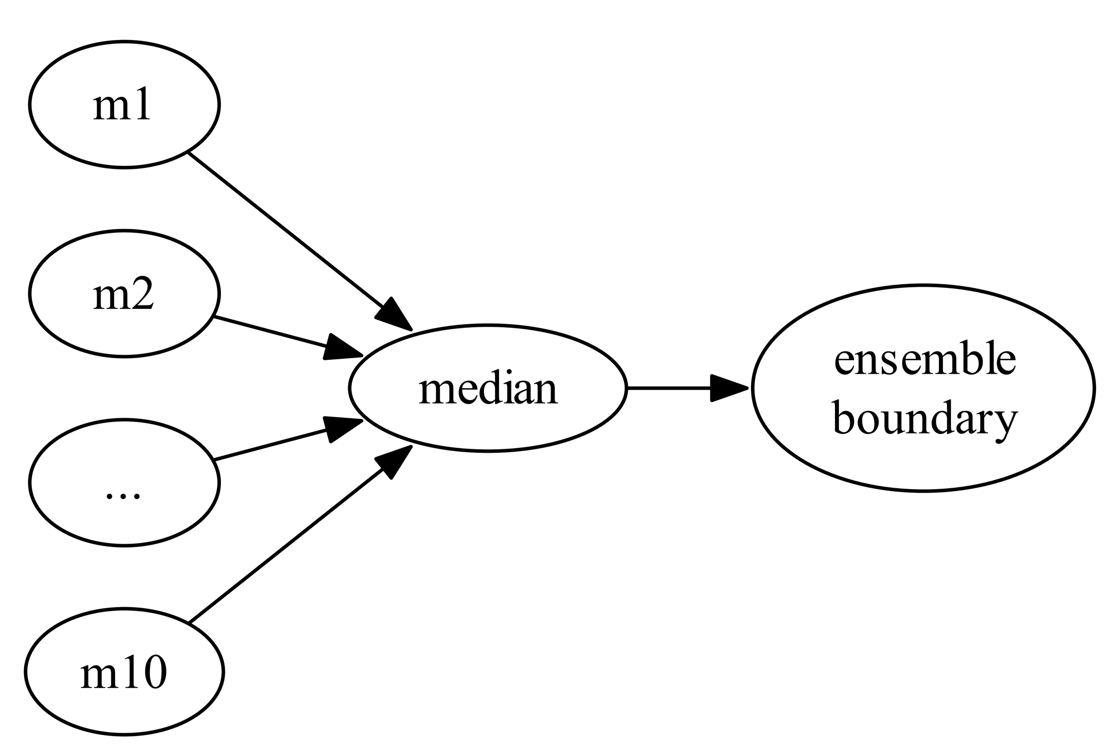

Figure 1. Median taken over each model's ("m1", "m2", etc.) forced alignment boundary. The result is the ensemble's estimate of the boundary. 97.85% CIs are also determined from this ensemble of boundaries, as the 2nd lowest and 9th highest items.

Table 1. Boundary error estimates for ensemble alignment when using manual transcriptions from the speech corpora. Singleton test is from Kelley et al. [1], using the 10 models individually; 95% CIs are provided for this set of results.

Data	Mean absolute error (ms)	Median absolute error (ms)
Train	14.13	6.48
Validation	14.18	6.85
Test	16.21	7.12
Singleton test [1]	16.75 ± 0.18	7.14 ± 0.05

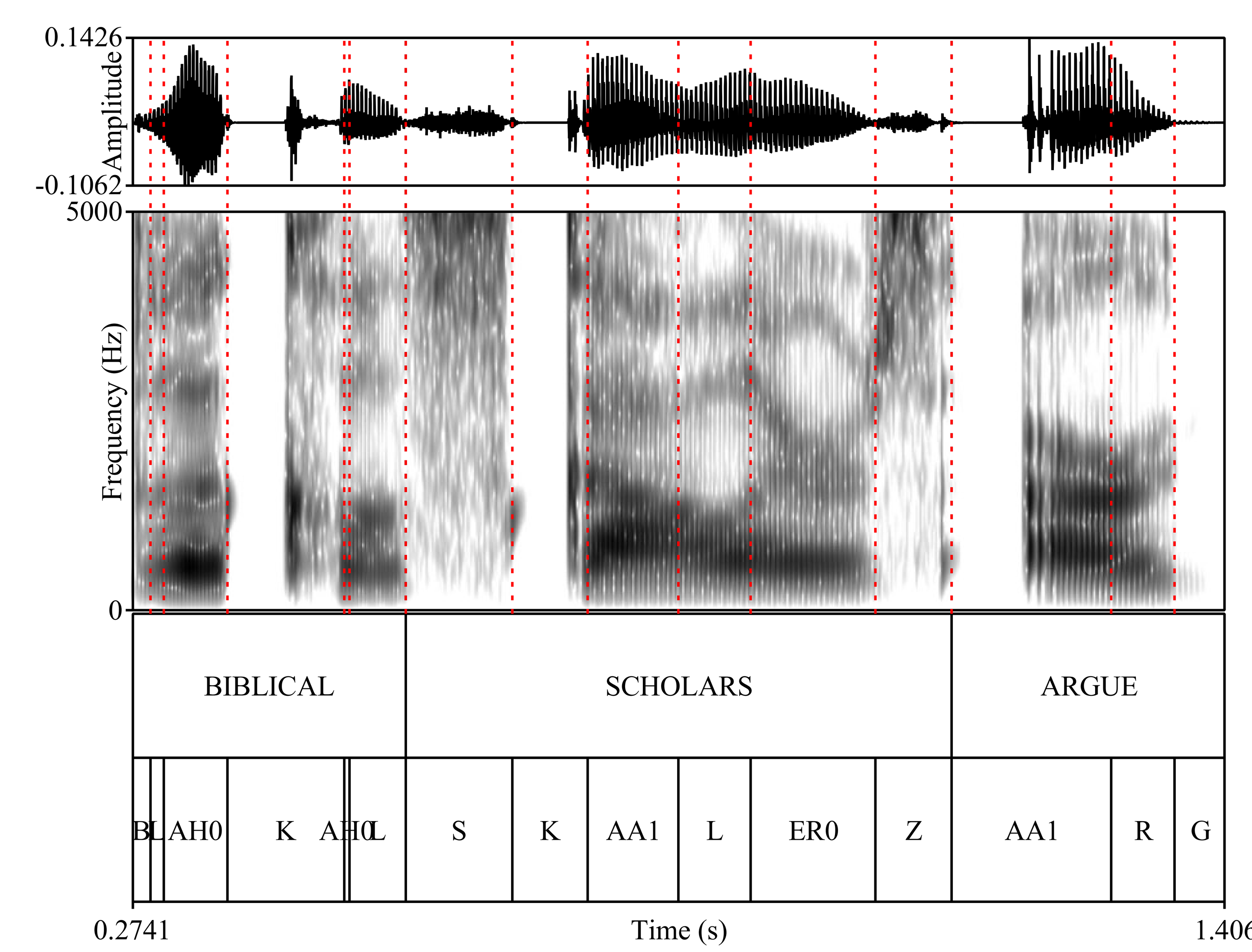

Figure 2. Sample of TextGrid format for segmentation of *...blical scholars arg...* from sentence "SX42" from speaker "FAEM0" in the TIMIT corpus. Segment boundaries are drawn with dashed red lines.

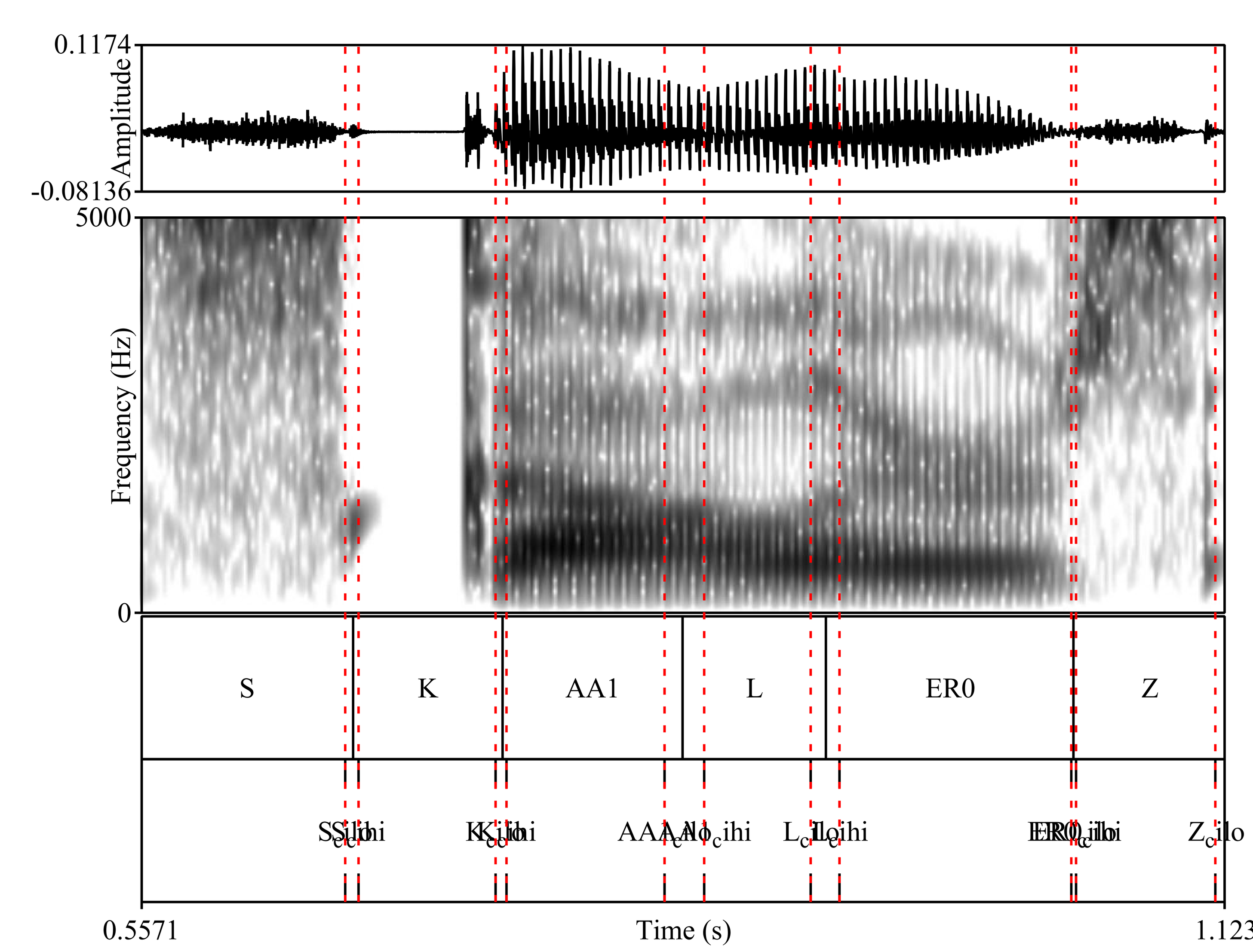

Figure 3. Sample CIs for *scholars* [ˈskɑlɚz] from *...blical scholars arg...* from "biblical scholars argue" in TIMIT corpus. High and low values for the CIs are drawn in dashed red lines. Note that the CIs are not symmetrical due to the order statistics method.

3. Boundary results

- Boundary error slightly improved using ensemble, compared with previous results [1]
- Quality of boundaries is reasonable (Figure 2, Table 1)
- Better boundaries are a perk but not a goal with CIs
- Ensemble helps prevent outliers and cascading alignment errors

4. Accessing the CIs

- Praat TextGrids don't permit boundary ranges
- Instead, a point tier is added to the TextGrid to indicate the low and high CI points (Figure 3)
- A table is also output from the alignment with the CIs for statistical or programmatic analysis

5. Discussion & conclusion

- Wide CIs can be used diagnostically to find alignment errors
- Uncertainty estimates of duration can be incorporated into Bayesian models
- Alignment takes 10 times as long, though this is not entirely escapable

Acknowledgements: I would like to thank the members of the Mason Phonetics and Phonology Lab, in addition to Pertti Palo, for their comments and discussion on previous versions of this project.

References:
[1] Kelley, M. C., Perry, S. J., & Tucker, B. V. (2024). The Mason-Alberta Phonetic Segmenter: A forced alignment system based on deep neural networks and interpolation. *Phonetica*, *81*(5), 451–508.
[2] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., Pallett, D. S., Dahlgren, N. L., & Zue, V. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1 (Technical Report No. 93; NASA/STI Recon). NASA/STI Recon.
[3] Pitt, M. A., Dilley, L., Johnson, K., Kiesling, S., Raymond, W., Hume, E., & Fosler-Lussier, E. (2007). Buckeye Corpus of Conversational Speech (Version 2nd release) [Computer software]. Department of Psychology, Ohio State University. www.buckeyecorpus.osu.edu

Contact: mkelle21@gmu.edu

Presented at the 188th Meeting of the Acoustical Society of America
New Orleans, LA, USA
1pSC12