APhL Aligner: A Neural Network Forced-Alignment System

Matthew C. Kelley, Scott James Perry, & Benjamin V. Tucker

181st Meeting of the Acoustical Society of America
Seattle, WA, USA
December 3, 2021

UNIVERSITY OF ALBERTA
DEPARTMENT OF LINGUISTICS

APhL
Alberta Phonetics Laboratory

Introduction

- Forced alignment commonly used in phonetics and sometimes in speech recognition
- Automatic calculation of temporal boundaries of segments in speech
- Some notable challenges to overcome

Alberta Phonetics Laboratory

Challenge 1: Segment separability

- A forced aligner's acoustic model is designed to separate speech segments from each other
- Ladefoged & Broadbent (1957) found that a given stimulus can be assigned to different categories based on surrounding context
- Acoustic context can't resolve confusable (Miller & Nicely, 1955) pairs like [f] and [θ] in *fin* and *thin*
- Segment separation may not be learnable

Challenge 2: Time sampling and boundaries

- Forced aligners often classify 25 ms windows of speech every 10 ms
- Maximum precision of 10 ms
- More precision requires
  - Faster sampling (e.g., 1 ms in Kelley & Tucker, 2018)
  - Error correction models (e.g., Stolcke et al., 2014)
  - And/or some sort of interpolation

Proposed solutions

- To resolve segment separability: Relax crisp separability requirement
  - Train network to treat segment categories as tags (e.g., “this sound has features of [f] and [v]”)
  - He & Xia (2018) call this a “joint binary network”
- To resolve boundary precision: Use linear interpolation during alignment
  - Treat Viterbi/dynamic time warping path as finite discrete approximation of smooth function

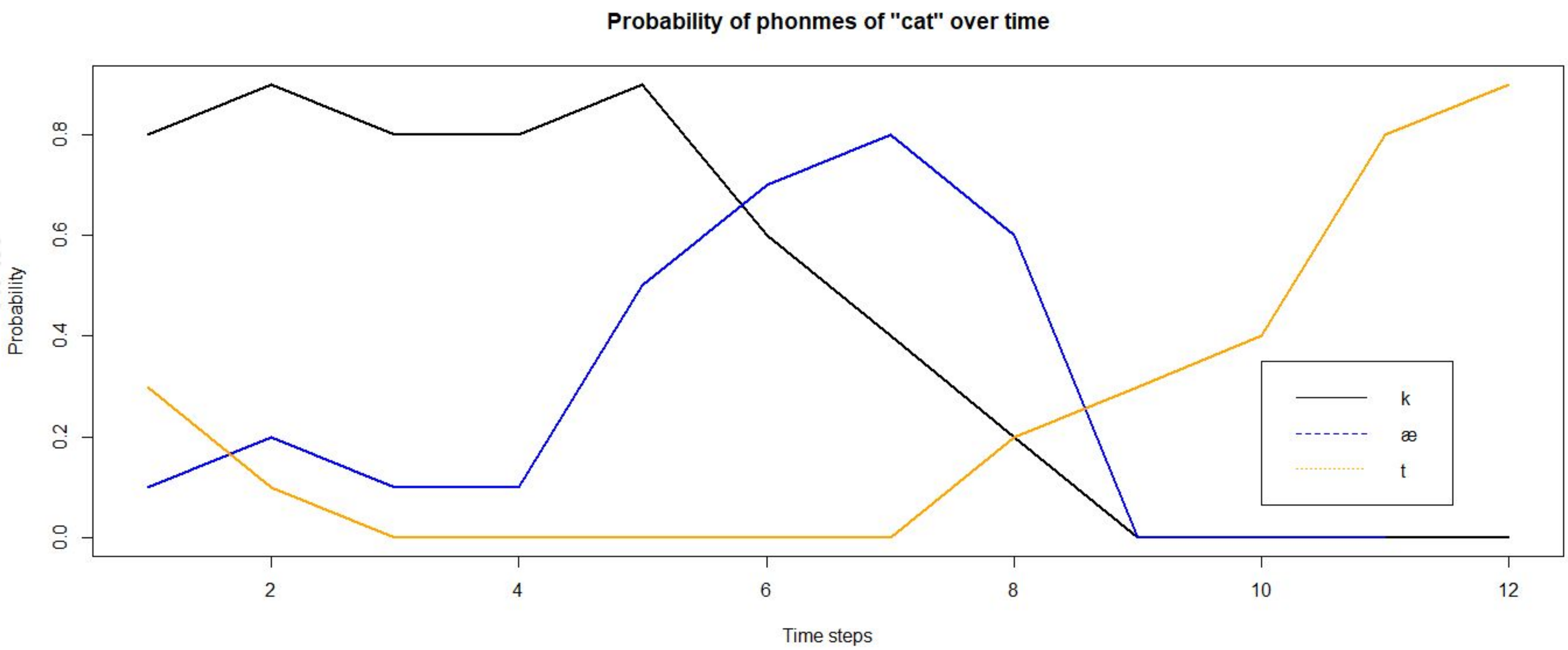
Alberta Phonetics Laboratory

On crispness relaxation

- Unclear how to determine what segment categories should be assigned to a given sound
  - Besides its original label in training data
- Using empirical approach
  - Train the network as a normal segment classifier first, similar to Graves & Schmidhuber (2005)
  - For each input, reassign targets as original segment category plus all categories that received more activation
- Result is a network with sparse instead of crisp output

Alberta Phonetics Laboratory

Interpolation schematic



Data

- Mix of TIMIT (Garofalo, 1993) and Buckeye (Pitt et al., 2005)
- Buckeye extracted as phrases based on silence periods
- Validation data held out as 5% from training data
- Some speakers held out from Buckeye for test set
- Standard TIMIT test set used

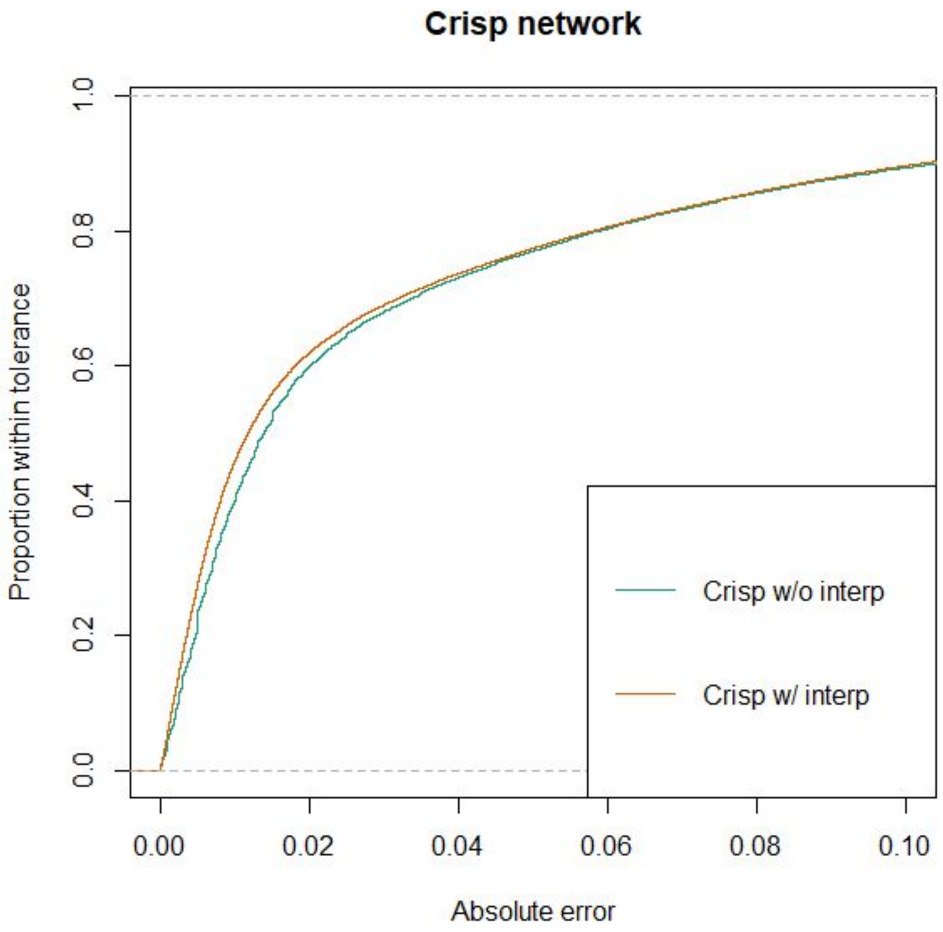
Alberta Phonetics Laboratory

Network

- 3 bidirectional LSTM layers with 128 units each
- Dropout of 0.5 between layers during training
- Output 40 classes
- Batch size of 64
- Trained for 50 epochs and used model with best validation accuracy

Alignment results

- Relaxing crispness of predictions had little to no improvement
- Interpolation had a bigger effect
  - Most notable on boundaries within 20 ms of target
- Best performance was crisp with interpolation



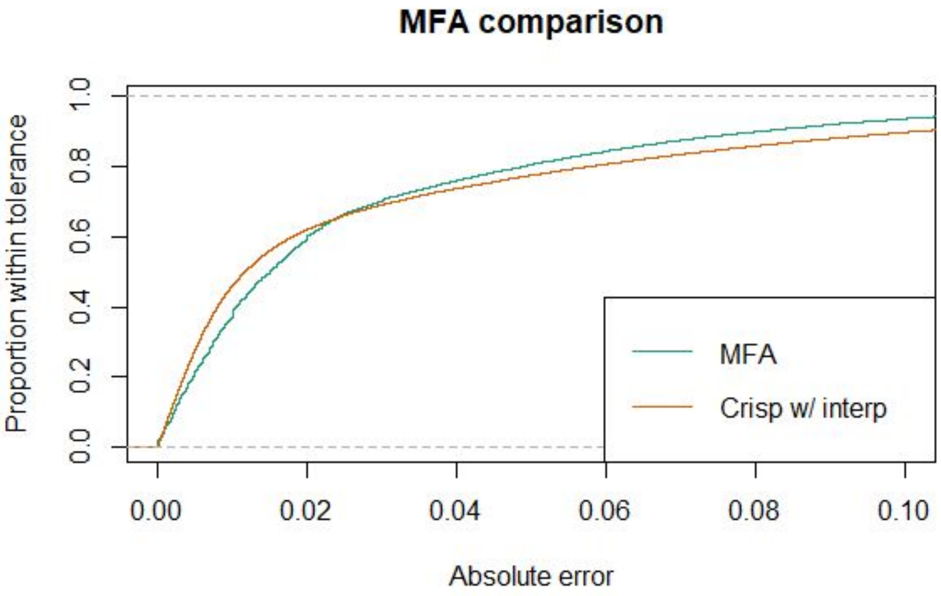
Alberta Phonetics Laboratory

Comparison with off-the-shelf model

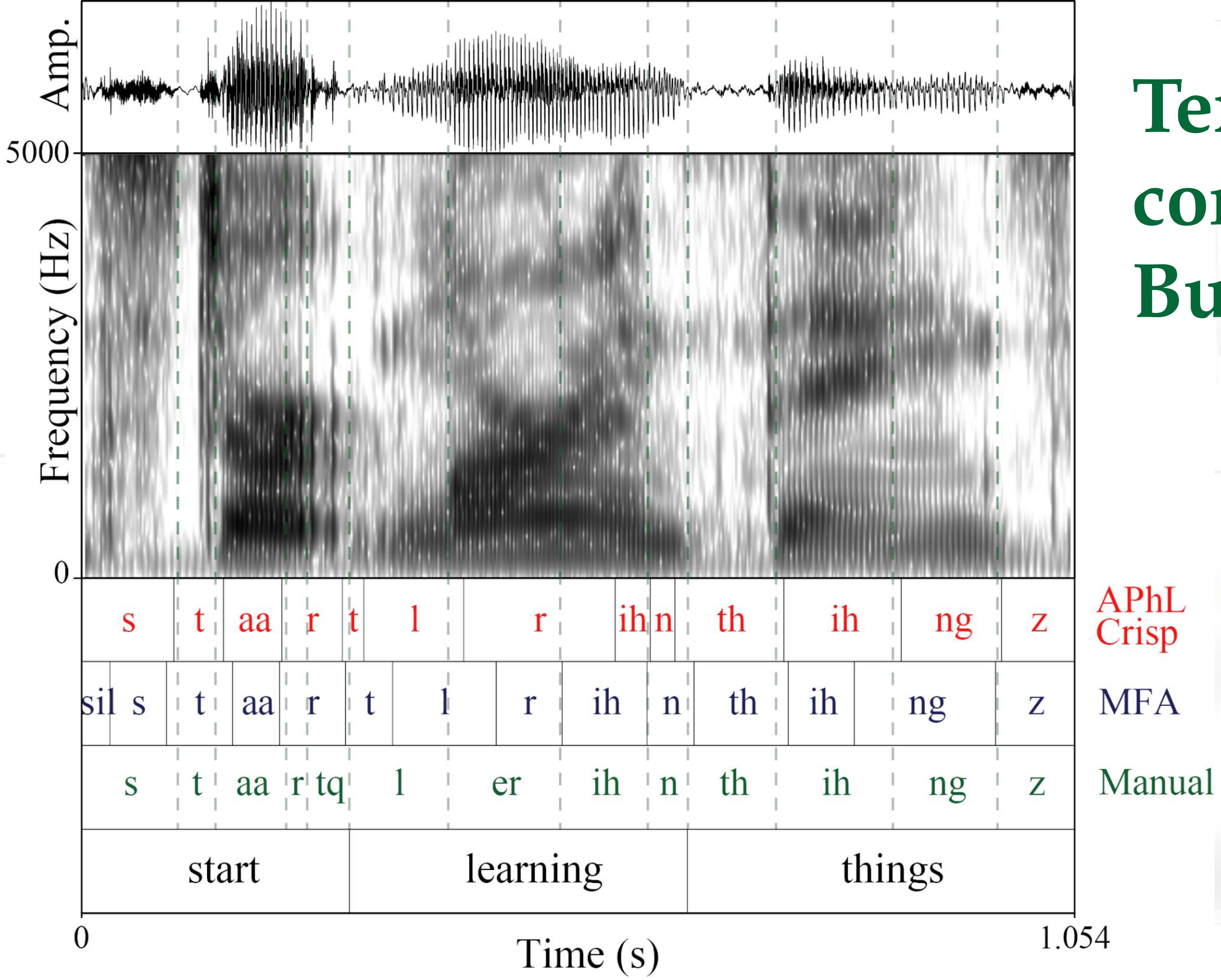
- Montreal Forced Aligner (MFA) is current state-of-the-art trainable aligner (McAuliffe et al., 2017)
- Trained MFA v1.0.1 on same data as neural network model
  - Used `train_and_align` function
  - Was able to align most but not all data in the training set

Comparison with Montreal Forced Aligner

- Crisp model better at lower end
- MFA has fewer large errors
- Some discrepancy could be due to some programming differences
  - E.g., collapsing repeated phones like [d d] in *red dog*



TextGrid comparison: Buckeye



Discussion

- Numerical comparisons not very useful
- Interpolation yields qualitatively better boundaries
- Improving the acoustic model can only do so much
- Aligner's performance depends on the quality of aligned transcriptions it is trained on

Alberta Phonetics Laboratory

Future directions

- Complete validation and testing
- Train on more data
- Explore more sophisticated interpolation with splines or polynomials
- Evaluate aligner with behavioral tasks using trained phoneticians
- Give consideration to the feasibility of forced alignment as a task

References

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.

Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5), 602–610. https://doi.org/10.1016/j.neunet.2005.06.042

He, H., & Xia, R. (2018). Joint binary neural network for multi-label learning with applications to emotion classification. CCF International Conference on Natural Language Processing and Chinese Computing, 250–259.

Kelley, M. C., & Tucker, B. V. (2018). A Comparison of Input Types to a Deep Neural Network-based Forced Aligner. Interspeech 2018, 1205–1209. https://doi.org/10.21437/Interspeech.2018-1115

Ladefoged, P., & Broadbent, D. E. (1957). Information Conveyed by Vowels. The Journal of the Acoustical Society of America, 29(1), 98–104. https://doi.org/10.1121/1.1908694

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. Interspeech 2017, 498–502. https://doi.org/10.21437/Interspeech.2017-1386

Miller, G. A., & Nicely, P. E. (1955). An Analysis of Perceptual Confusions Among Some English Consonants. The Journal of the Acoustical Society of America, 27(2), 338–352. https://doi.org/10.1121/1.1907526

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(1), 89–95. https://doi.org/10.1016/j.specom.2004.09.001

Stolcke, A., Ryant, N., Mitra, V., Yuan, J., Wang, W., & Liberman, M. (2014, May). Highly accurate phonetic segmentation using boundary correction models and system fusion. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5552-5556). IEEE.

Alberta Phonetics Laboratory