A comparison of input types to a deep neural network-based forced aligner

Matthew C. Kelley & Benjamin V. Tucker

Introduction

- Forced aligners determine phone boundaries in audio
  - E.g., location of [k], [æ], [t] in recording of *cat*
- Most previous forced aligners hidden Markov model (HMM) based [4][6]
- Deep neural net (DNN) systems outperform HMM ones for general speech recognition [3]

Research question: DNNs → better forced alignment?

Prediction: DNN systems will outperform HMM ones

- Unclear if raw audio or engineered features better [5]

Data and Networks

- Trained on TIMIT speech corpus [1]
- One net uses raw audio, the other uses Mel-frequency cepstral coefficients (MFCCs)
  - Window length of 25 ms, taken at 1 ms intervals
  - For MFCCs, used 12 coefficients and energy term, plus delta and delta-delta coefficients
- Architecture kept same for both networks (Figure 1)
- All layers except output had ReLU activation

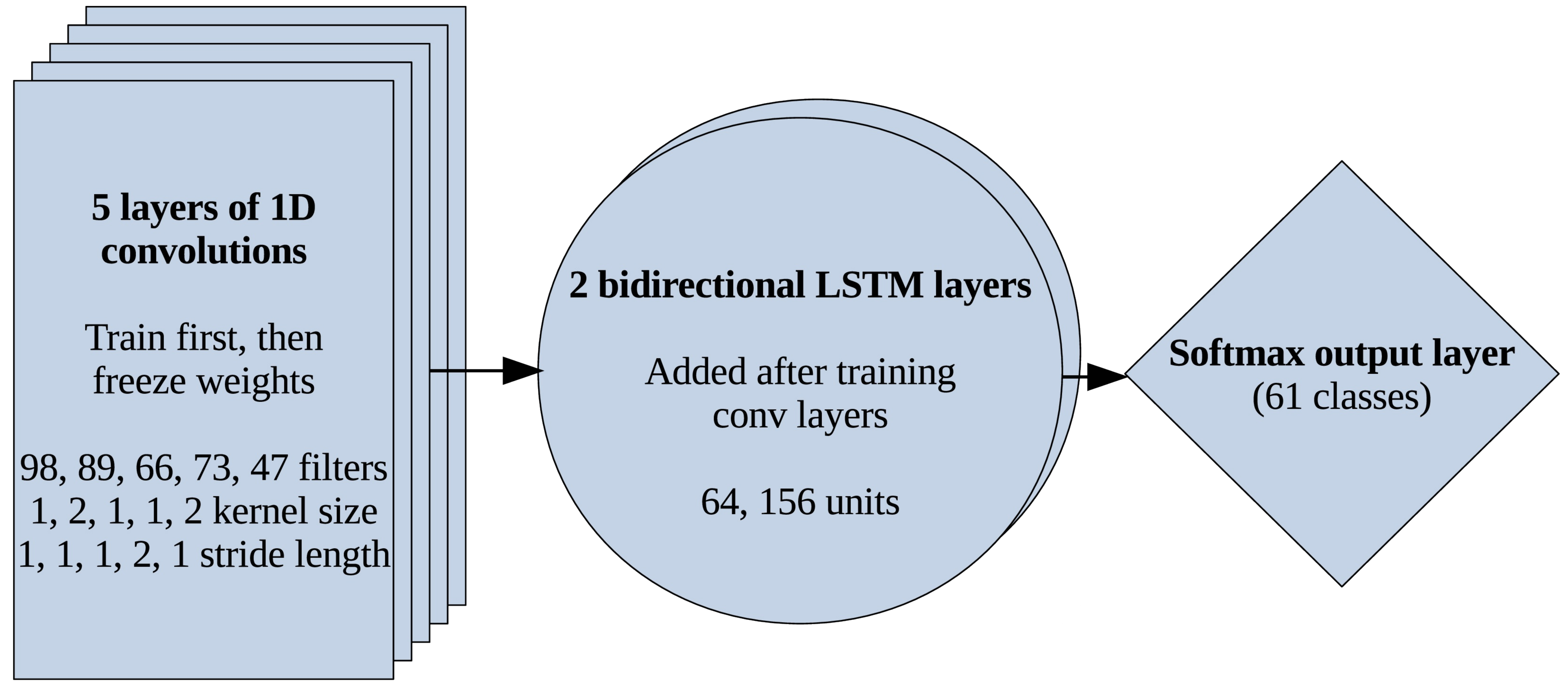

Figure 1. Network architecture and training procedure.

$$N = \begin{bmatrix} N_1^1 & N_1^2 & N_1^3 & N_1^4 \\ N_2^1 & N_2^2 & N_2^3 & N_2^4 \\ N_3^1 & N_3^2 & N_3^3 & N_3^4 \end{bmatrix}$$

$$O = \begin{bmatrix} N_1^1 & N_1^2 * O_1^1 & 0 & 0 \\ 0 & N_2^2 * O_1^1 & N_3^2 * max(O_1^2, O_2^2) & 0 \\ 0 & 0 & N_3^3 * O_2^2 & N_3^4 * max(O_2^3, O_3^3) \end{bmatrix}$$

Figure 2. Example of decoding process for a 3 label (rows), 4 time-step (columns) output. Neural network output is *N*, and output label matrix is *O*. Labels determined by backtracking and following the most probable previous steps through *O* (illustrated by the red arrows). Boundaries are taken as the point where the labels transition.

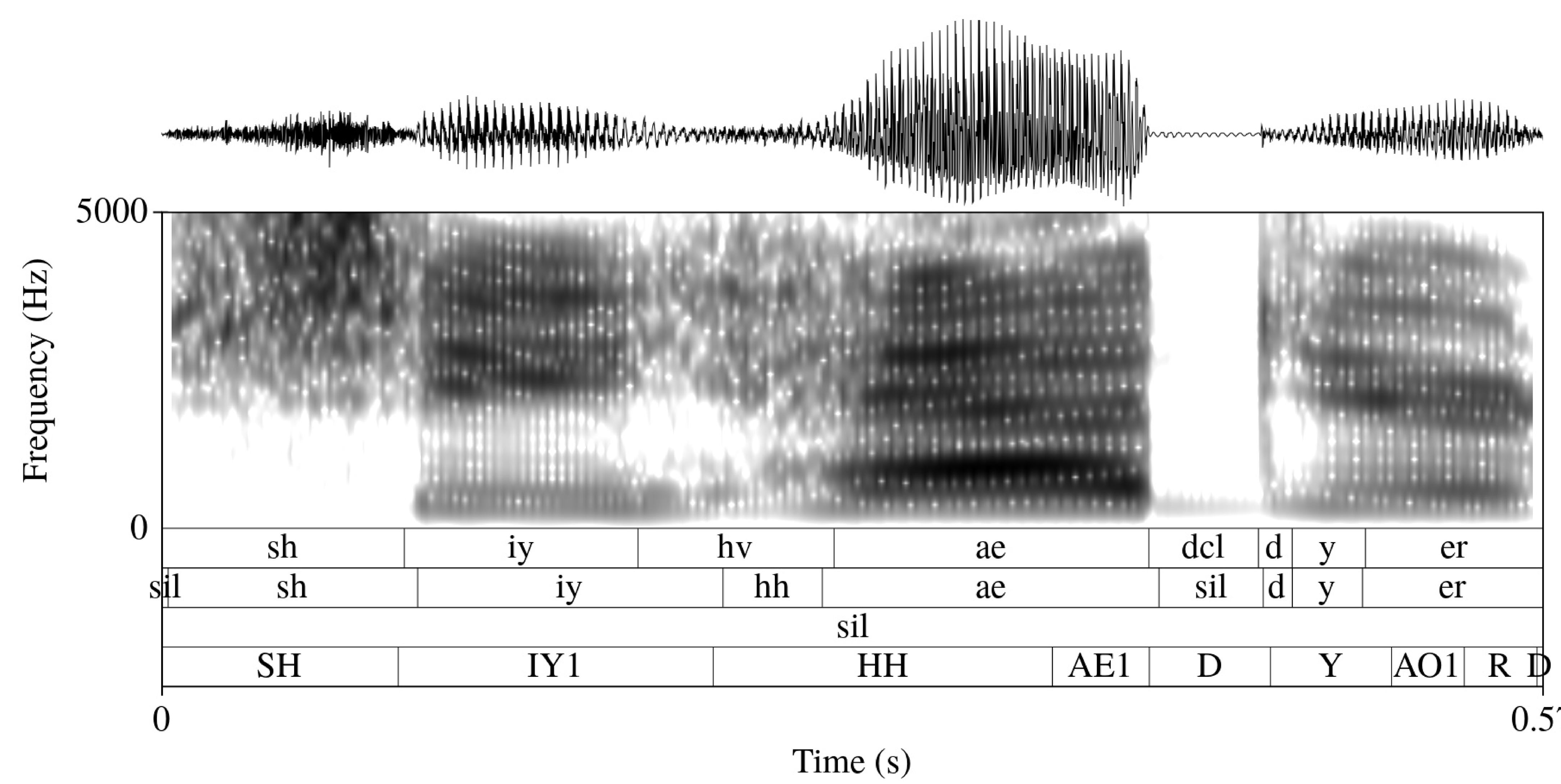

Figure 3. Sample phone alignment from aligners for "she had your." From top to bottom: ground truth, raw audio network, MFCC network, Montreal Forced Aligner. Closer to ground truth is better.

Results and Discussion

- Network output decoded for boundaries (Figure 2)
- Also evaluated Montreal Forced Aligner (MFA) for recent HMM system comparison

Aligner	Framewise acc.	MAE (s)
Raw audio	74.7%	0.008
MFCC	22.0%	1.55
MFA	72.1%	0.1

Table 1. Evaluation metrics for trained networks and MFA. Framewise alignment accuracy, framewise test accuracy, and median absolute error (MAE) of boundary timestamps. Test accuracy not available for MFA

- Raw audio system outperforms other tested systems
- Test accuracies not yet competitive with existing systems [2][5]
- Something wrong with MFCC network
- Raw audio alignment shows promise (Figure 3)
- Improving frame identification accuracy may improve alignment results
- Decoding algorithm may benefit from minimum durations

Acknowledgements: This research was funded in part by a SSHRC grant to the second author and the Kule Institute for Advanced Study through the Deep Learning for Sound Recognition group at the University of Alberta. We also acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research. Additionally, the authors would like to thank the participants in the course on deep learning for sound and behavior at the University of Alberta, and Vadim Bulitko and Terry Nearey in particular. Any errors or misunderstandings are, of course, our own.

References

[1] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., & Pallett, D. S. (1993). DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1. *NASA STI/Recon technical report n*, 93.

[2] Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks, 18*(5-6), 602-610.

[3] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal processing magazine, 29*(6), 82-97.

[4] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: trainable text-speech alignment using Kaldi. In *Proceedings of Interspeech* (pp. 498-502).

[5] Palaz, D., Collobert, R., & Doss, M. M. (2013). Estimating phoneme class conditional probabilities from raw speech signal using convolutional neural networks. *arXiv preprint arXiv:1304.1018*.

[6] Yuan, J., & Liberman, M. (2008). Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America, 123*(5), 3878.

Contact: {mckelley, bvtucker}@ualberta.ca