Montreal Forced Aligner: an accurate and trainable aligner using Kaldi

Michael McAuliffe[1], Michaela Socolof[2], Sarah Mihuc[1], Michael Wagner[1], & Morgan Sonderegger[1]

michael.mcauliffe@mail.mcgill.ca, msocolof@umd.edu, sarah.mihuc@mail.mcgill.ca, chael@mcgill.ca, morgan.sonderegger@mcgill.ca

[1]Department of Linguistics, McGill University [2] Department of Linguistics, University of Maryland

91st LSA, Austin, Texas
Jan 5-8, 2017

BACKGROUND

Forced alignment

- Time align speech sounds, given
  - Sound file
  - Orthographic transcription
  - Pronunciation dictionary

Toolkits

- HMM Toolkit (HTK; Young et al, 2006)
  - State of the art in linguistics
  - Restrictive license
- Kaldi (Povey et al, 2011)
  - State of the art in automatic speech recognition
  - Actively maintained codebase
  - Permissive license

SYSTEM COMPARISON

System	Toolkit	Trainable	Acoustic model	Pretrained models	Supported platforms
MFA	Kaldi	Yes	Triphone GMM	English	Mac, Linux, Windows
Prosodylab-aligner[2]	HTK	Yes	Monophone GMM	English, French	Mac, Linux
FAVE-align/ P2FA[6,8]	HTK	No	Monophone GMM	English	Mac, Web, Windows
(Web) MAUS[7]	HTK	Non-trivial	Monophone GMM	English, French + 8 other languages	Linux, Web
EasyAlign[1]	HTK	No	Monophone GMM	English, French + 3 other languages	Windows
Gentle[3]	Kaldi	No	ANN	English	Mac, Web

EVALUATION

How do alignments from the Montreal Forced Aligner compare with a state-of-the-art system?

DATA

- Read speech from production experiment (48 minutes)
  - "Please say ___ again"
  - Target 1-2 syllable words with vowel + obstruent
- Vowel and obstruent of target word were hand annotated
  - Vowel begin, vowel end, and obstruent end
- Force aligned
  - Compared with Prosodylab-aligner
    - Also trainable
    - Uses similar acoustic models to other systems (Monophone GMM)
- Conditions:
  - Flat – Trained on limited data (48 minutes)
  - Pretrained on lab recordings (15 hours)
  - Pretrained on LibriSpeech (474 hours)

Montreal Forced Aligner

Features

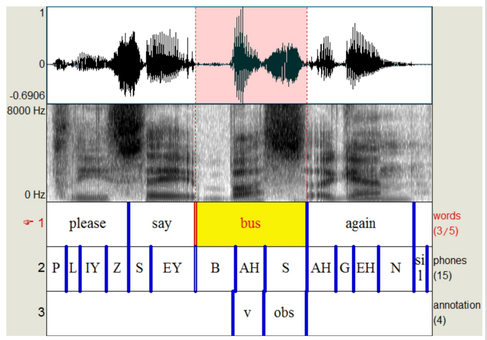
- Kaldi-based
- Trainable
  - Tested on 20+ languages
- Can model words not in the dictionary
  - Preserves alignments of other words
- Triphone acoustic models
  - Right and left context for phones (models coarticulation)
  - Acoustic features adapted by speaker
  - ⇒ more accurate alignment
  - Parallel processing helps scaling up
- Command line interface
  - Well-tested, easy-to-use
  - Actively maintained
  - Well-documented and open source
- Input
  - Orthographic TextGrid and label files
  - Wav files
- Output
  - Aligned TextGrids

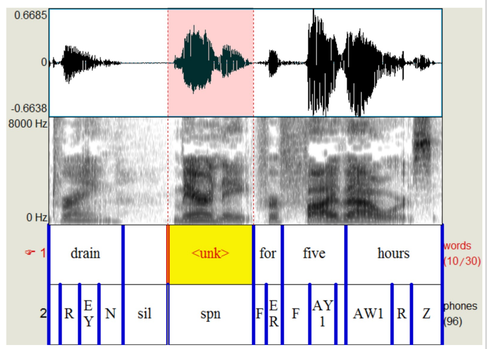
INPUT

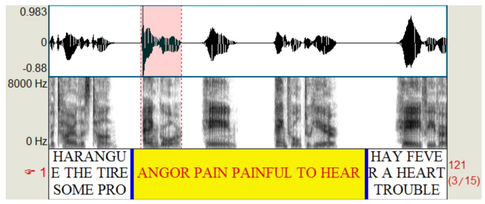
PLEASE SAY 'BUS' AGAIN

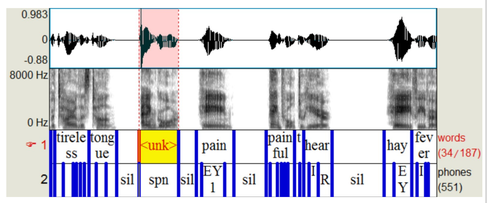
LET STAND FOR FIVE MINUTES AND DRAIN MARINATE FOR FIVE HOURS IN MELTED BUTTER LEMON JUICE AND VINEGAR SEASONING WITH SALT AND PEPPER DRAIN

OUTPUT

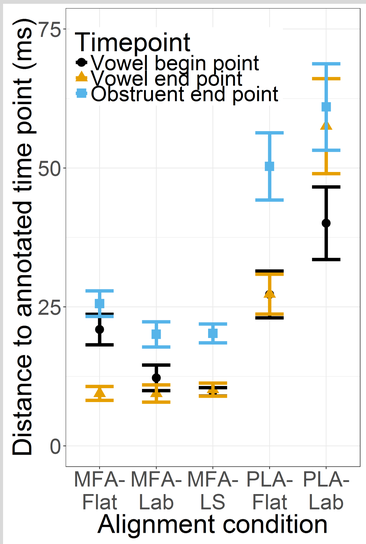


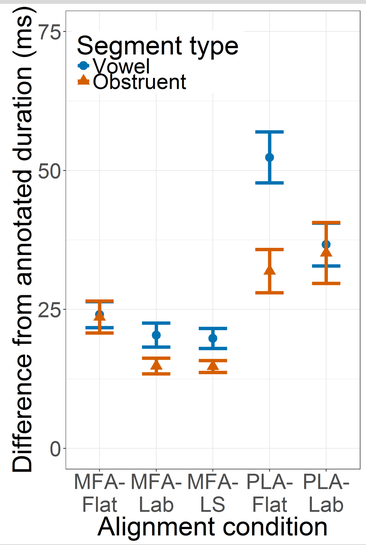






RESULTS





DISCUSSION

- Montreal Forced Aligner outperforms the Prosodylab-Aligner
- Pretrained models on larger datasets are generally preferable than only using the dataset to be aligned
- Larger data sets may be unnecessary if the style/recording conditions are the same

REFERENCES

[1] Goldman J. P. (2011). *EasyAlign: an automatic phonetic alignment tool under Praat*. Proceedings of InterSpeech, Firenze, Italy, September 2011. **[2]** Gorman, K. et al (2011). *Prosodylab-aligner: A tool for forced alignment of laboratory speech*. Proceedings of Acoustics Week in Canada. Canadian Acoustics, 39(3):192–193. **[3]** Ochshorn, R. and Hawkins, M. (2016). *Gentle [Computer Program]*. Version 0.9.1, retrieved May 27, 2016 from https://lowerquality.com/gentle/. **[4]** Panayotov, V. et al (2015). *Librispeech: an ASR corpus based on public domain audio books*. In 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 5206-5210). IEEE. **[5]** Povey, D. et al (2011). *The Kaldi speech recognition toolkit*. In IEEE 2011 workshop on automatic speech recognition and understanding (No. EPFL-CONF-192584). IEEE Signal Processing Society. **[6]** Rosenfelder, I. et al. (2011). *FAVE (Forced Alignment and Vowel Extraction) Program Suite*. http://fave.ling.upenn.edu. **[7]** Schiel F. (1999). *Automatic Phonetic Transcription of Non-Prompted Speech*. Proc. of the ICPhS 1999. San Francisco, August 1999. pp. 607-610. **[8]** Young, S. et al. (2006). *The HTK Book (Version 3.4)*. Cambridge University Engineering Department.