PROSODYLAB-ALIGNER: A TOOL FOR FORCED ALIGNMENT OF LABORATORY SPEECH

Kyle Gorman,[1] Jonathan Howell,[2] and Michael Wagner[2]

[1]Department of Linguistics and Institute for Research in Cognitive Science, University of Pennsylvania, 619 Williams Hall,
255 S. 36th St., Philadelphia, PA, U.S.A., 19104-6305
[2]Department of Linguistics, McGill University, 1085 Dr. Penfield, Montreal, QC, Canada, H3A 1A7

1. INTRODUCTION

Recordings of speech, whether spontaneously generated or elicited in the lab, are an important data source for linguists, but the time required to produce by hand the time indices necessary to perform acoustic feature extraction is often prohibitive. The Penn Forced Aligner (Yuan & Liberman 2008) automates the alignment process using the Hidden Markov Model Toolkit (HTK),[1] a speech recognition software package. However, it is limited to certain sample rates, and can only perform alignments according to an included North American English acoustic model; it does not support training of new acoustic models for different languages or domains.

2. A NEW TOOL

We have constructed a new software package, Prosodylab-Aligner, which, like the Penn Forced Aligner, uses HTK for forced alignment. However, it also permits the experimenter to use transcribed (but not necessarily aligned) audio to train new acoustic models. Prosodylab-Aligner is open-source and available for free online.[2]

The core of Prosodylab-Aligner is `align.py`, a script which performs acoustic model training and alignment. This script automates calls to HTK and SoX,[3] an open-source command-line tool which is capable of resampling audio. The included `README` file provides instructions for installing HTK and SoX on Linux and Mac OS X, and can also be run on Windows.

2.1 Preprocessing

The `align.py` script checks for missing data files, and terminates if an audio or transcript file (indicated by the extensions `.wav` and `.lab`, respectively) is lacking its companion label or audio file, after printing a list of "orphan" data files. It also checks for words in the transcripts which are not found in the pronunciation dictionary, if any out-of-dictionary words are found, terminates after printing this list. Both of these steps permit the experimenter to correct for missing data before proceeding with alignment or training.

2.2 Acoustic Model

The Prosodylab-Aligner acoustic models are monophone Gaussian mixtures consisting of 39 Mel frequency cepstral coefficients (Mermelstein 1976).

2.3 Training Routine

By default, the input data is aligned using a precomputed acoustic model trained from North American English laboratory speech (see Section 4), but with the `-t` flag, the experimenter can provide training data for estimating a new acoustic model. In many cases, it is desirable for the training set will be the same as the test set.

During training, the model is initialized with flat-start monophones, which are then submitted to a single round (by default, four iterations) of model estimation. Then, a tied-state "small pause" model is inserted and used in a second round of estimation. The data is then aligned once to choose the most likely pronunciation of all homonyms (i.e., dictionary entries with the same orthographic form), and a final round of estimation is performed. The optimal alignments are computed and the resulting word and phone alignments are written to Praat[4] TextGrid files.

A separate flag (`-T`) can be used to trigger a final round of speaker-dependent training and speaker-dependent alignment.

2.4 Helper Scripts

Several other scripts are included for related tasks: aligning a single audio/label pair (`align_ex.sh`), fixing errors in label files (`fix_lab.py`), or downloading the CMU English pronunciation dictionary (`get_dict.sh`).

3. EXPERIMENTAL PIPELINE

Prosodylab-Aligner plays an integral part in an experimental pipeline for speech data, both elicited in the laboratory with controlled experiments, and harvested from the web (Howell & Rooth 2009; Rooth et al. 2011).

3.1 Laboratory Data

Laboratory data is elicited using a suite of MATLAB scripts designed for production experiments. The experimenter enters the stimuli into a spreadsheet, indicating "words of interest" found in target stimuli, and organizes stimuli into items and conditions. These are them

[1] http://htk.eng.cam.ac.uk/

[2] http://prosodylab.org/tools/aligner/

[3] http://sox.sourceforge.net/

[4] http://www.fon.hum.uva.nl/praat/

randomized and presented as the elicitation “script” for the subject. The experimenter then verifies that the resulting utterance produced matches the text presented to the subject.

3.2 Web Data

Web audio (in MP3 format) is downloaded from Ramp,[5] a company which indexes radio and television programming, including NBC, PBS, Fox and CBS Radio, and processed using standard UNIX tools. The transcriptions produced by Ramp are sufficient for identifying a targeted linguistic construction (with approximately 50% accuracy), but the experimenter must still listen to the audio files to eliminate false positives.

3.3 Data Comparison

It is desirable to be able to compare lab and web data, and thus to use the acoustic model for alignment. While WAV files and a uniform sample rate is used in the laboratory, web audio recording quality is inconsistent (bit rate between 32-256 kbit/s and sample rate between 11025-44100Hz). For this reason, when training is performed by `align.py`, the experimenter may also specify sample rate (with the `-s` flag): any training or test audio data that does not conform to this sample rate is automatically resampled using SoX.

3.4 Post-alignment Processing

A script marks “words of interest” in the aligned TextGrids, and a Praat script extracts acoustic measurements from each word of interest for later statistical analysis.

4. EVALUATION

The North American English acoustic models provided with Prosodylab-Aligner produce alignments of impressionistically high quality. To quantify alignment quality, we use a single annotator's hand alignment from a prior study as a gold standard: Howell & Rooth (2009) and Howell (2011) gathered tokens of the phrase “than I did” embedded in longer utterances harvested from the web, and 127 tokens of this phrase hand-aligned at the phone level.

For each file and for each phone boundary in this phrase, we compute the magnitude of the distance from the gold standard for Prosodylab-Aligner, as well as for the Penn Forced Aligner. The results in Table 1 suggest that the two aligners produce alignments of comparable quality. Figure 1 shows that both aligners have comparable magnitude difference compared to the gold standard.

5. CONCLUSIONS

The Penn Forced Aligner is a useful tool for linguists interested in acoustic properties of speech. Prosodylab-Aligner further empowers the experimenter by automating the difficult task of constructing domain-appropriate acoustic models to use for forced alignment.

Table 1. Magnitude difference compared to hand-aligned gold standard

	Mean $|\Delta|$ (s)	Median $|\Delta|$ (s)
Prosodylab-Aligner	0.3060	0.0119
Penn Forced Aligner	0.2061	0.0124

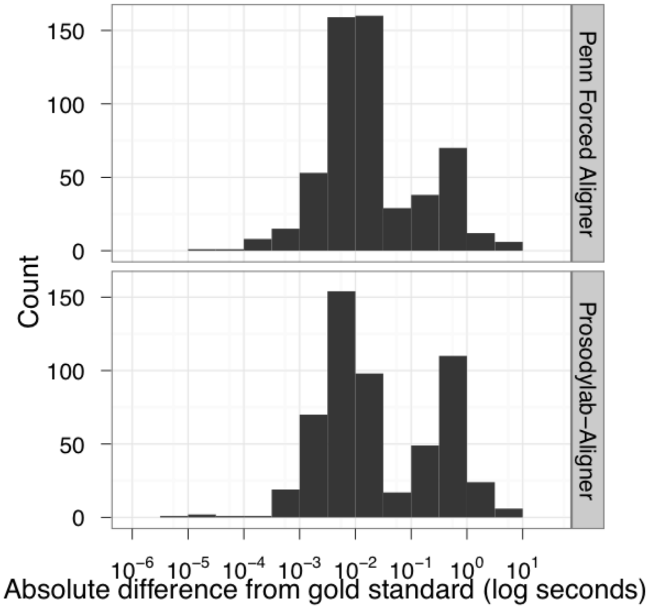

Figure 1. Magnitude difference compared to hand-aligned gold standard

REFERENCES

Howell, J. And Rooth, M. (2009). Web harvest of minimal intonational pairs. In I. Alegria, I. Leturia, and S. Sharoff (ed.), *Proceedings of the 5th Web as Corpus Workshop,* 45-52. San Sebastian: Elhuyar Fundazioa.

Howell, Jonathan. (2011). Meaning and prosody: On the web, in the lab and from the theorist's armchair. Doctoral dissertation, Cornell University.

Mermelstein, P. (1976). Distance measures for speech recognition, psychological and instrumental. In C-H Chen (ed.), *Pattern recognition and artificial intelligence*, 374-388. New York: Academic Press.

Rooth, M., Howell, J., and Wagner, M. (2011). Harvesting speech datasets for linguistic research on the web. Paper presented at Digging into Data Challenge Conference.

Yuan, J. And Liberman, M. (2008). Speaker identification on the SCOTUS corpus. In *Proceedings of Acoustics 2008,* 5687-5690.

ACKNOWLEDGEMENTS

This research was supported by grants to the third author from FQRSC (Nouvelle NP-132516) and SSHRC (Digging into Data Challenge Grant 869-2009-0004; Canada Research Chair 218503).

[5] `http://www.ramp.com/`