The Usage Guidelines for a Kaldi based Robust Korean Forced Aligner

Hyungwon Yang[1], Jaekoo Kang[2], Youngsun Cho[1], Yeonjung Hong[1], Seohyun Kim[1], Wiback Kim[1], Hyebin Yoon[1], Ingu Lee[1], Youngjun Kim[1], Hosung Nam[3]
[1]Korea University, [2]The Graduate Center, CUNY, [2,3]Haskins Laboratories
[1]{hyung8758, youngsunhere, yvonne_yj_hong, sh77, kwb425, hby1117, demiust, young35}@korea.ac.kr,
[2]jkang@gradcenter.cuny.edu, [3]hnam@korea.ac.kr

ABSTRACT

Korean Forced Aligner (Korean FA) is a tool that automatically aligns Korean speech signals and its corresponding orthographic transcription at the word and phone levels. This aligner uses Kaldi to attain its best-performed model after training and human-eye evaluating a number of models that we generated based on various algorithms. This paper mainly introduces (1) the process of building Korean language based alignment model, and (2) the direction of installing and applying Korean FA tool.

1. Introduction

Speech and text alignment information has been critical for many researchers in various fields [1,2], since it offers the information of how the graphical words are truly realized in human speech. However, aligning speech and text by human hands is not only tedious and time-consuming but also inaccurate because human's decision on setting phone boundaries could vary in case by case. Automatic forced alignment has been developed among researchers [3,4] to overcome this inconvenience and it was successful due to its fast application and consistency on alignment result.

2. Korean FA Model

2.1. Training data

Korean Read Speech corpus from National Institute of Korean Language (NIKL) was used for training Acoustic Model in Korean Forced Aligner. It consisted of approximately 17 GB (120 hours) of audio and text paired files. Audio files were recorded from various age groups among 20s, 30s, 40s, 50s, and 60s, and the participated speakers were born and raised in Seoul and Gyeonggi-do so they speak only standard Korean.

The speakers were given total 19 short Korean novels as reading materials and while they read them out loud in moderate speed with monotonous voice, their speeches were recorded through SM48 and SM57 SHURE microphones.

2.2. Training Procedure

Total 5 types of different alignment models were generated during a training experiment and went to a performance test by a few human researchers.

An experiment of training a Korean FA model was initialized by learning mono-phone features from a Korean corpus. Based on the mono-phone trained model, tri-phone training proceeded to enrich the alignment model by learning phone variation that happened in phone sequence. In tri-phone training, each phone learned its variation from possible combinations of phone sequences that located before and after of its own. After tri-phone training had finished, LDA, MLLT, and SAT algorithms were applied to the tri-phone model sequentially and then DNN training began on top of that model.

3. A Tool Usage

Korean FA's stability test was done on Mac OSX and Linux (Ubuntu) but not on Windows. Python 3.5 or above is required.

3.1 Installation

3.1.1 Kaldi installation

Korean FA is mainly dependant on Kaldi. Therefore, Kaldi should be installed before running this tool. Type below in command line.

```
$ git clone https://github.com/kaldi-asr/kaldi.git kaldi
  --origin upstream
$ cd kaldi
$ git pull
```

If you finished downloading git repository to your local directory, find README and follow the direction written there.

3.1.2 Package installation

Three packages are required and they need different command lines depending on your OS environment.

On Mac

```
$ brew install sox
$ brew install coreutils
$ pip3 install xlrd
```

On Ubuntu

```
$ apt-get install sox
$ apt-get install coreutils
$ pip3 install xlrd
```

3.2 Data Preparation

Prepare audio data for alignment and its transcribed text data as a pair set and locate them in a folder. Paired data set needs to share an identical nametag except their extension. (e.g., test01.wav, test01.txt)

3.2.1 Audio data

Provide audio data in WAV format at 16,000Hz sampling rate. If the given audio data's sampling rate does not match with 16,000Hz sampling rate, Korean FA tool will automatically resample the provided audio data to 16,000Hz and remove the original audio file.

3.2.2 Text data

Each text data should contain all the words mentioned in the paired audio data and meet all the following requirements as follows.

- Words in a text data should be written in Korean.
- Do not include symbols (e.g., period and comma), numbers, and English spelling in a text data.
- Remove white space or tab at the end of the line.

3.3 Direction

You are able to receive a current version of Korean FA tool or newly updated version by emailing to a developer.

If you downloaded the tool into the local directory, unzip the file and navigate to 'Korean_FA' directory. Open a forced_align.sh script with any convenient text editor and specify the Kaldi directory path. (Change Kaldi name variable. Initial setting: kaldi=/home/kaldi) After resetting the Kaldi directory path is finished, run a forced-align.sh script to initiate forced alignment process on your data.

```
$ bash forced_align.sh (options) (data directory)
$ bash forced_align.sh –nw ./example/readspeech
```

Option Instruction

1.	-h	--help	: Show tool instruction.
2.	-nj	--num-job	: Set the number of job threads to do the alignment process in parallel.
3.	-s	--skip	: Skip alignment for already aligned data.
4.	-nw	--no-word	: Delete word tier.
5.	-np	--no-phone	: Delete phone tier.

Once the alignment process is finished, TextGrid files will be saved into a data directory.

5. Correspondence

Hosung Nam
Dept. of English Language and Literature, Korea Univ.
Anam-dong 5-ga, Seongbuk-gu, Seoul 136-701, Korea
Email: hnam@korea.ac.kr

References

[1] Milne, P. M. (2011). Finding schwa: Comparing the results of an automatic aligner with human judgments when identifying schwa in a corpus of spoken French. *Canadian Acoustics*, 39(3), 190-191.

[2] Vosoughi, So., & Roy, D. (2012). A longitudinal study of prosodic exaggeration in child-directed speech. In *Speech Prosody* 2012.

[3] McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *interspeech* (pp. 498-502).

[4] Yun, J., Hwang, H. K., & Ko, S. (2012). Automatic Annotation for Korean Speech Corpus Analysis. In Poster presented at the The International Workshop on Corpus Linguistics and Endangered Dialects. *National Institute for Japanese Language and Linguistics*, Tokyo, Japan