AI 2022

A Text-Independent Forced Alignment Method for Automatic Phoneme Segmentation

Bryce Wohlan, Duc-Son Pham, Kit Yan Chan, and Roslyn Ward

Curtin University

Curtin University

Table of Contents

Curtin University

Introduction

- Speech sound disorders affect 1% to 4% pre-school children [1].
- Phoneme segmentation is important for the differential diagnosis of children with SSDs, however currently performed manually by speech language pathologists.
- Automatic speech recognition tools do not provide analysis at the phoneme level.
- We developed a tool that can segment phonemes automatically
  - No need for transcription;
  - Detect phoneme boundaries automatically with high accuracies.

`https://github.com/dsphamgithub/fatool`.

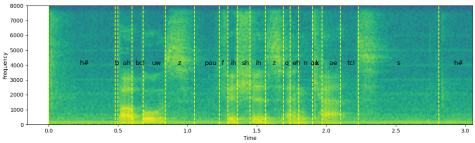
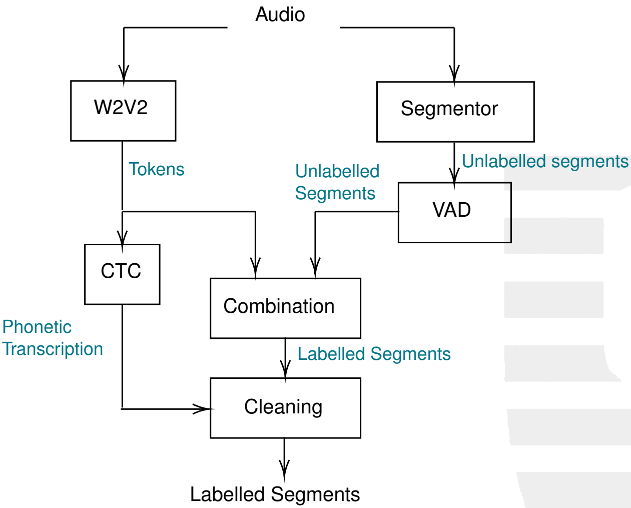
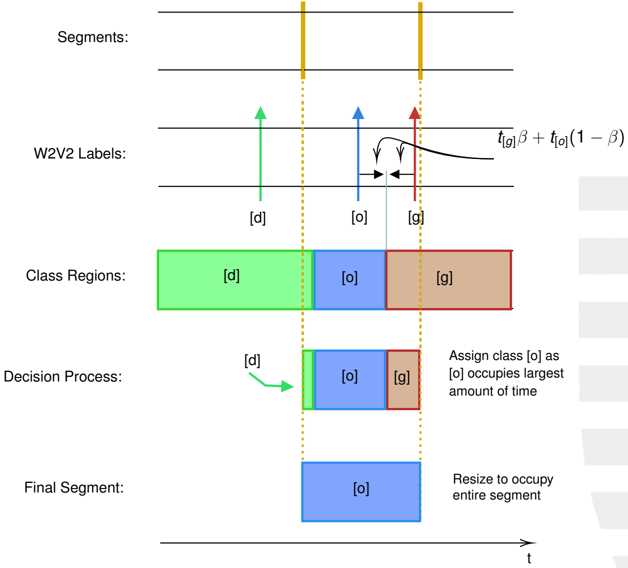


Fig. Utterance “Bubbles, fishes and cats”

Data Flow


Audio
W2V2
Segmentor
Tokens
Unlabelled Segments
Unlabelled segments
VAD
CTC
Combination
Phonetic Transcription
Labelled Segments
Cleaning
Labelled Segments

Main Ideas



Assign class [o] as [o] occupies largest amount of time

Resize to occupy entire segment

Algorithm - Boundary Calculation

Algorithm 1 Boundary Calculation

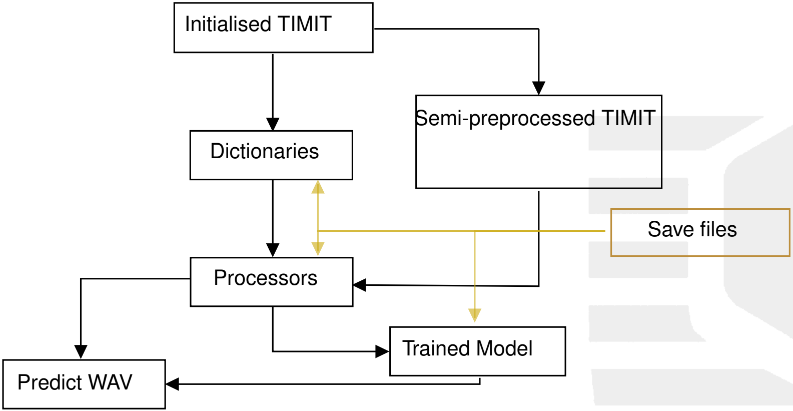
```
procedure DecisionBoundaryCalc(timedTokenList, seconds, bias)
    timedTokenList is a list of tuples of labels and their timings
    seconds is the duration of the speech sample in seconds
    bias is the bias factor which is positive and smaller than 1.
    DCB ← new List
    for ii in range of length timedTokenList do
        if ii equals length of timedTokenList - 1 then
            upper ← seconds
            lower ← timedTokenlist[ii - 1][time])*(1-bias) + timedTokenList[ii][time])*(bias)
        else if ii equals 1 then
            upper ← timedTokenlist[ii +1 1][time])*(bias) + timedTokenList[ii][time])*(1-bias)
            lower ← 0
        else
            upper ← timedTokenlist[ii +1 1][time])*(bias) + timedTokenList[ii][time])*(1-bias)
            lower ← timedTokenlist[ii - 1][time])*(1-bias) + timedTokenList[ii][time])*(bias)
        end if
        Append tuple (timedTokenList[ii][label], lower, upper)
    end for
    return DCB
end procedure
```

Algorithm - Forced Aligner

Algorithm 2 Forced Aligner

```
procedure LABELLED SEGMENTER(wavPath)
    signal, samplingFreq ← soundfile.read(wavPath)
    seconds ← length of signal / SamplingFreq
    wp ← Wav2Vec2PredictiorObject
    tokens ← wp.predictWavNoCollapse(wavPath)
    segPredictor ← UnsupervisedsegmenterPredictorObject
    segVect ← segPredictor.predict(wavPath, CheckpointPath)
    segVect ← VADFilterSegments(wavPath, SegVect)
    segVect ← toList(segVect)
    timedTokens ← tokensToTimedTokens(signal, samplingFreq, tokens)
    filteredTimedTokens ← new List
    for timedToken in timedTokens do
        if timedToken[label] is not ”[pad]” or ”[unk]” or ”—” then
            Append timedToken to filteredTimedTokens
        end if
    end for
    decisionBoundaries ← decisionBoundaryCalc(filteredTimedTokens, seconds, bias)
    strToUnicodeDict ← Read in from wav2vec2 object save
    MaxDCBinitdict ← dictionary fromkeys(strToUnicodeDict, 0)
    Insert 0 at index 0 to segVect
    Append seconds value to the end of segVect
    labelList ← MaxContribution(segVect, maxDCBInitDict, DCB)
    segList ← new List
    for ii in range of length of labelList do
        Append tuple (LabelList[ii], segVect[ii], segVect[ii+1]) to segList
    end for
    segList ← cleanSegs(segList)
    Convert list of tuples to list of dictionaries
    return segList
end procedure
```

Dependencies



Experiments

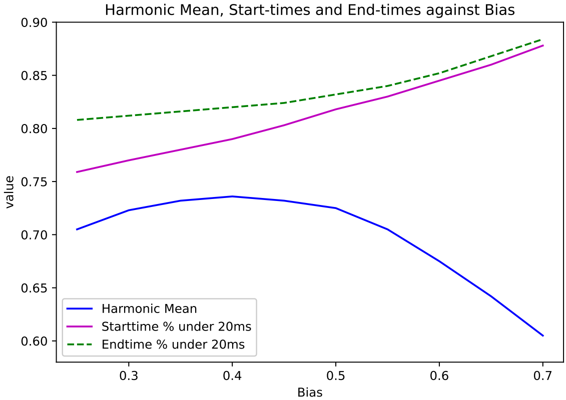


Fig. Performance with varied bias

- **wav2vec 2.0** [2]: phoneme error rate (PER)
  - Training: 10.6%; Validation set: 13.2%
- **UnsupSeg** [3]: R-val=0.83

Curtin University

Experiments

Dataset: TIMIT train/test = 4,620/1,680

Table I. Forced alignment results

Metric	With VAD	Without VAD
Accuracy of predictions	71.9%	82.35%
Proportion of manual labels correctly classified	71.4%	72.02%
Harmonic Mean	71.6%	76.88%

Curtin University

Experiments

Table II. Proportions of the segment boundary errors in milliseconds

	<20ms	<40ms	<60ms
Segment Start Time (w/ VAD)	81.73%	94.21%	97.22%
Segment End Time (w/ VAD)	87.53%	96.72%	98.70%
Segment Start Time (w/o VAD)	81.50%	94.07%	97.16%
Segment End Time (w/o VAD)	88.33%	96.87%	98.71%

Experiments

Table II. A comparison with other text-independent aligners

Model/Tool	P	R	F_1	Year
FAVE	0.57	0.59	0.58	2014
Gentle	0.49	0.46	0.48	2017
W2V2-CTC-20ms	0.31	0.30	0.31	2021
W2V2-FS-20ms	0.40	0.42	0.41	2021
W2V2-FC-20ms-Libris	0.57	0.59	0.58	2021
Ours	0.62	0.54	0.58	This work

Conclusion

- New text-independent FA tool
- Boundary adjustment via bias factor
- Competitive performance against many
- Improved with better pre-trained `wav2vec`

Thank you!

References I

[1] Sharynne McLeod et al. “Profile of Australian preschool children with speech sound disorders at risk for literacy difficulties”. In: **Australian Journal of Learning Difficulties** 22.1 (2017), pp. 15–33.

[2] Alexei Baevski et al. “wav2vec 2.0: A framework for self-supervised learning of speech representations”. In: **Proc. Advances in Neural Information Processing Systems**. Vol. 33. 2020, pp. 12449–12460.

[3] Felix Kreuk et al. “Phoneme Boundary Detection Using Learnable Segmental Features”. In: **Proc. ICASSP**. Barcelona, Spain: IEEE, May 2020, pp. 8089–8093.