# A Text-Independent Forced Alignment Method for Automatic Phoneme Segmentation

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

Curtin University

## Contributions

This work provides a novel text-independent FA tool based on two models: `wav2vec 2.0` and an UnsupSeg. To provide labels to the segments, the class regions that are obtained by 1-$NN$ classification with `wav2vec 2.0` labels pre-CTC collapse as the reference points. Maximal overlap between the class regions and the segments determines class label. Additional post-processing steps, such as overfitting cleaning and application of voice activity detection, are also performed to further improve the segmentation performance. All the models used to create the tool are self-supervised, and thus can leverage great amounts of unlabelled data to reduce the need for labelled data. When evaluated on the TIMIT dataset, our implementation achieved a harmonic mean score of 76.88%, competitive against other alternatives.

## Phoneme Segmentation

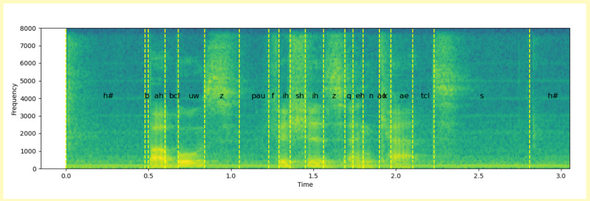



Fig. Utterance "Bubbles, fishes and cats"

- A key task in the diagnosis of children with speech sound disorders (SSD)
- Currently, a labour intensive process for speech language pathologists (SLPs)
- Tools are limited ⇒ new tools are needed

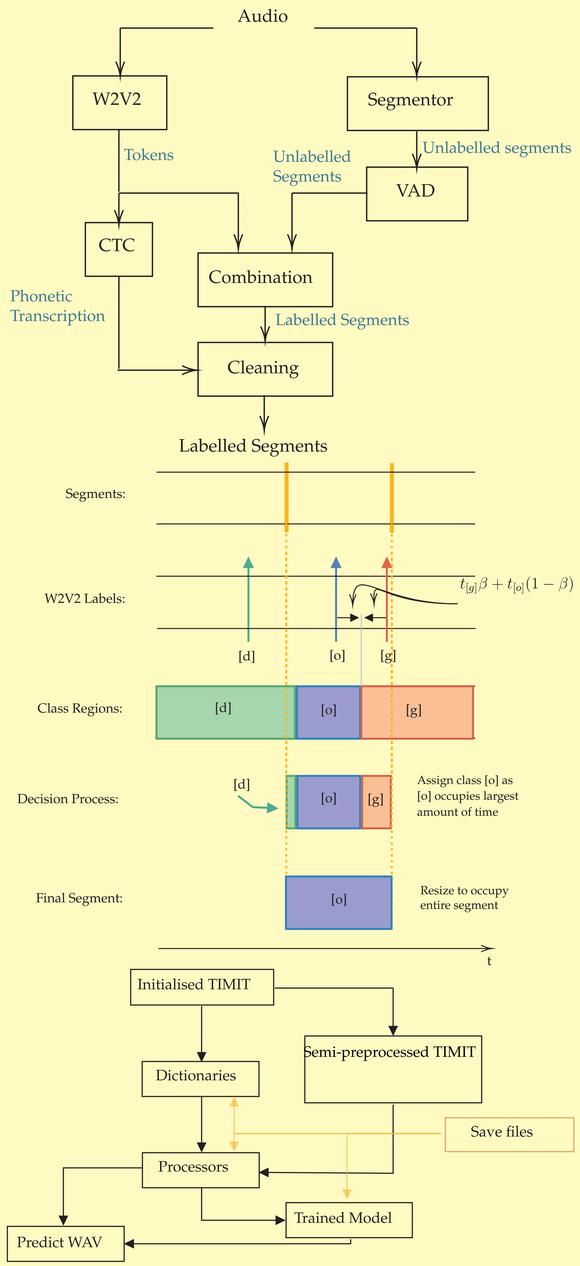
We consider the most challenging setting

- Transcript is found automatically from input speech
- Forced alignment: align phoneme boundaries with the computed transcript

Our FA tool is based on previous work

- `wav2vec 2.0`[1]: an engine to recognise the phonemes
- UnsupSeg [2]: initial boundaries

## Proposed Method




**Preprocessing**

- Encoding of phone labels

$$ARPABET \leftrightarrow Unicode \leftrightarrow NumericID$$

- `wav2vec2` feature extractor
- Fine tuning: `wav2vec2-large-xlsr-53`

## Results

**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% |

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% |

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 |

## Performance

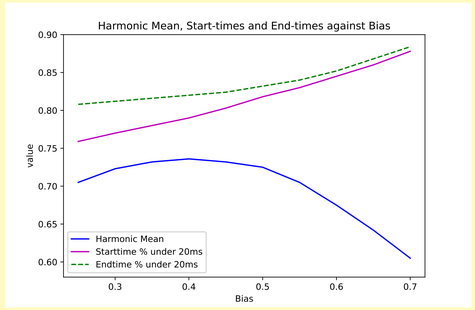



Fig. Performance with varied bias

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

## Key Takeaways

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

## References & Github

[1] Baevski, A. and Zhou, Y. and Mohamed, A. and Auli, M. `wav2vec 2.0`: A framework for self-supervised learning of speech representations. In *Proc. NeurIPs*, 2020.

[2] Kreuk, F., Keshet, J., Adi, Y. Self-Supervised Contrastive Learning for Unsupervised Phoneme Segmentation. arXiv:2007.13465, 2020.

**Github repository**

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