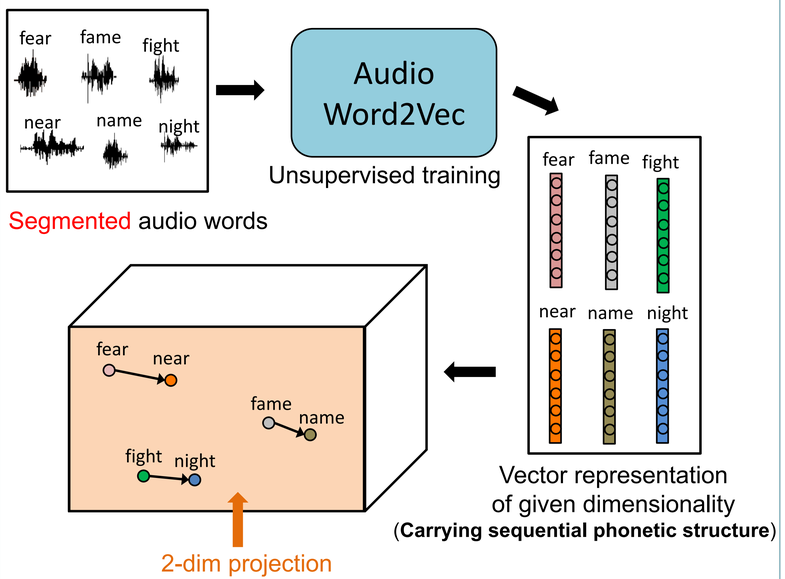
Segmental Audio Word2Vec: Representing Utterances as Sequences of Vectors with Applications in Spoken Term Detection

Yu-Hsuan Wang, Hung-yi Lee, Lin-shan Lee
National Taiwan University

1. Introduction

Previous Work: Audio Word2Vec



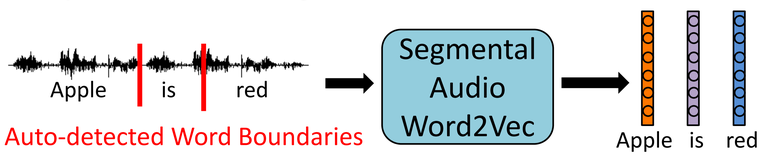
Similar phonetic difference ➡ similar movement

✻Problem: segmented audio words are needed

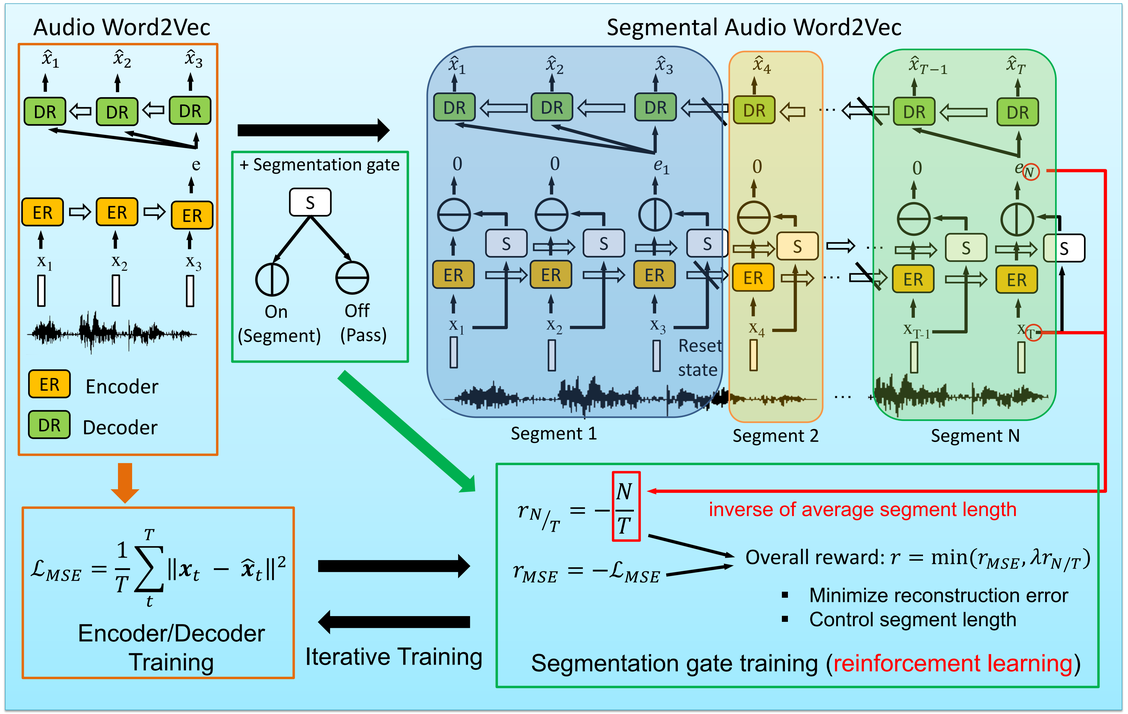
Ref: Audio Word2Vec: Unsupervised Learning of Audio Segment Representations using Sequence-to-sequence Autoencoder, Chung et al. INTERSPEECH 2016

This paper: Segmental Audio Word2Vec

Jointly learn acoustic word segmentation and audio Word2Vec

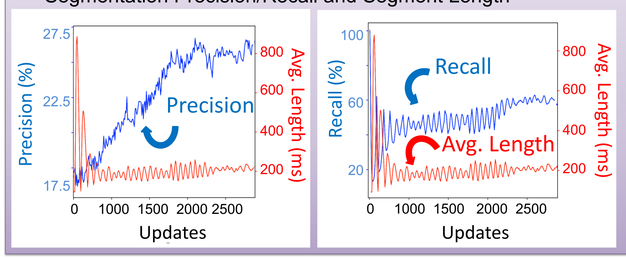


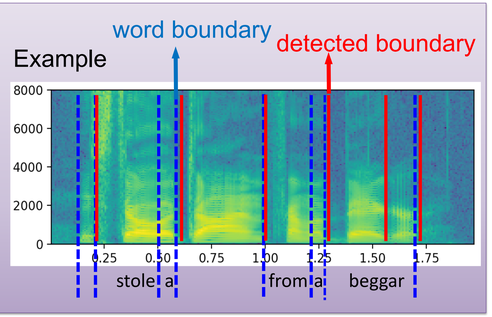
2. Proposed Approach



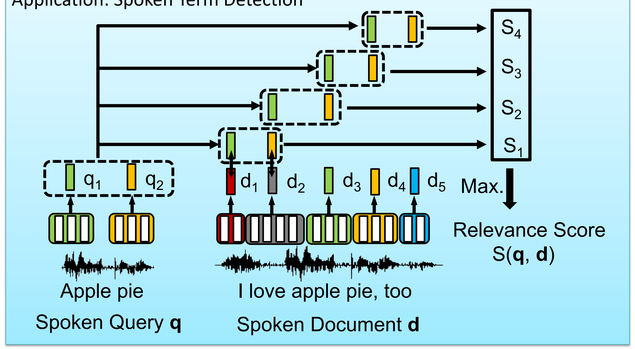
Acoustic Word Segmentation Learning

Segmentation Precision/Recall and Segment Length





Application: Spoken Term Detection



3. Experiments

- 4 languages
  English (TIMIT), Czech, French, German (GlobalPhone)
- Encoder and decoder: 1-layered 100 LSTM Cells
- Segmentation gate: 2-layered 256 LSTM Cells
- Spoken term detection setup
  1. Evaluation measure: mean average precision (MAP)
  2. Queries: words cropped from training set utterances
  3. Document archive: testing set utterances

Spoken Term Detection Results

MAP		Audio Word2Vec (Different Segmentation)			
Language	DTW	GAS	HAC	Proposed	Oracle
English	12.02	8.29	0.91	23.27	30.28
Czech	16.59	0.68	1.13	19.41	22.56
French	11.72	0.40	0.92	21.70	29.66
German	6.07	0.27	0.26	13.82	21.52

- GAS, HAC: other segmentation methods
- Oracle: using ground truth word boundary
- DTW: frame-level dynamic time warping
- Proposed approach significantly better than DTW
- **Obtained audio word representations did carry sequential phonetic Information**