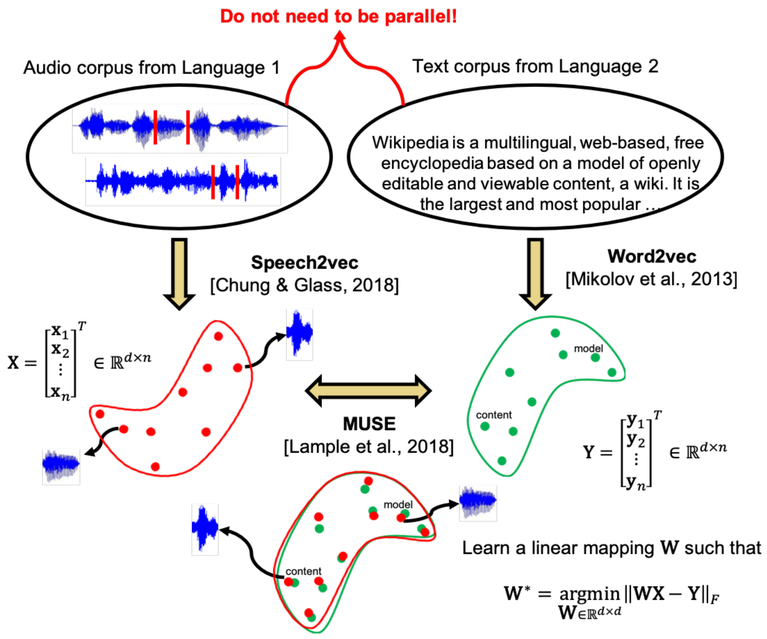
Unsupervised Cross-Modal Alignment of Speech and Text Embedding Spaces

Yu-An Chung, Wei-Hung Weng, Schrasing Tong, James Glass

Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA

1. Overview

- **Goal: To learn a linear mapping between speech & text embedding spaces**



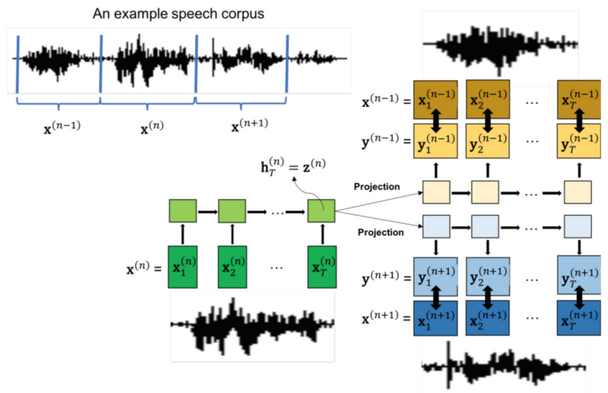
2. Learning Embeddings

Text Embedding Space

- Train Word2Vec [Mikolov et al., 2013] on the text corpus
- Unsupervised learning of distributed word representations that model word semantics

Speech Embedding Space

- Train Speech2Vec [Chung and Glass, 2018] on the speech corpus
  - The corpus is pre-processed by an off-the-shelf speech segmentation algorithm such that utterances are segmented into audio segments corresponding to spoken words
  - Speech version of Word2Vec: unsupervised semantic audio segment representations



3. Embedding Spaces Alignment

- Both embedding spaces are learned from corpora based on distributional hypothesis (e.g., skip-grams) → approximately isomorphic
- Construct the synthetic mapping dictionary to learn a linear mapping matrix between the two embedding spaces

$$W^* = \underset{W \in \mathbb{R}^{d_2 \times d_1}}{\operatorname{argmin}} \|WX - Y\|^2$$

Adversarial Training

- Make the aligned embeddings indistinguishable

$$\mathcal{L}_D(\theta_D|W) = -\frac{1}{m}\sum_{i=1}^{m} \log P_{\theta_D}(\text{speech} = 1|Ws_i) - \frac{1}{n}\sum_{j=1}^{n} \log P_{\theta_D}(\text{speech} = 0|t_j)$$

$$\mathcal{L}_W(W|\theta_D) = -\frac{1}{m}\sum_{i=1}^{m} \log P_{\theta_D}(\text{speech} = 0|Ws_i) - \frac{1}{n}\sum_{j=1}^{n} \log P_{\theta_D}(\text{speech} = 1|t_j)$$

Refinement (Orthogonal Procrustes Problem)

- Use the W learned from the adversarial training step as an initial proxy and build a synthetic parallel dictionary
- Consider the most frequent words

4. Experimental Settings

Datasets

Corpus	Train	Test	Words	Segments
English LibriSpeech	420 hr	50 hr	37K	468K
French LibriSpeech	200 hr	30 hr	26K	260K
English SWC	355 hr	40 hr	25K	284K
German SWC	346 hr	40 hr	31K	223K

Details of Training

- Speech2Vec with Skip-grams, window = 3
  - Encoder: single-layer bidirectional LSTM
  - Decoder: single-layer unidirectional LSTM
  - SGD with fixed learning rate of 0.001
- Word2Vec fastText implementation
- Both dimensions = 50
- Discriminator in adversarial training
  - 2 layers, 512 neurons, ReLU

Method Comparison

Configuration	Speech2Vec training		Unsupervised
	How word segments were obtained	How embeddings were grouped together	
A & *A**	Forced alignment	Use word identity	✗
B	Forced alignment	k-means	✗
C	BES-GMM	k-means	✓
D	ES-KMeans	k-means	✓
E	SylSeg	k-means	✓
F	Equally sized chunks	k-means	✓

References

- Chung and Glass. Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. INTERSPEECH 2018.
- Lample et al. Word translation without parallel data. ICLR 2018.
- Mikolov et al. Distributed representations of words and phrases and their compositionality. NIPS 2013.

5. Results

Task I | Spoken Word Recognition

- Accuracy decreases as the level of supervision decreases
- Unsupervised alignment approach is almost as effective as it supervised counterpart (A vs. A^*)
- Word segmentation is a critical step
- Apply to different corpora settings

Corpora	$EN_{ls} - en_{ls}$	$FR_{ls} - fr_{ls}$	$EN_{swc} - en_{swc}$	$DE_{swc} - de_{swc}$	$EN_{ls} - en_{swc}$	$EN_{swc} - en_{ls}$
Nonalignment-based approach						
Word Classifier	89.3	83.6	86.9	80.4	–	–
Alignment-based approach with cross-modal supervision (parallel dictionary)						
A^*	25.4	27.1	29.1	26.9	21.8	23.9
Alignment-based approaches without cross-modal supervision (our approach)						
A	23.7	24.9	25.3	25.8	18.3	21.6
B	19.4	20.7	22.6	21.5	15.9	17.4
C	10.9	12.6	14.4	13.1	6.9	8.0
D	11.5	12.3	14.2	12.4	7.5	8.3
E	6.5	7.2	8.9	7.4	4.5	5.9
F	0.8	1.4	2.8	1.2	0.2	0.5

Task II | Spoken Word Synonyms Retrieval

- The output actually contain both synonyms and different lexical forms of the audio segment
- Also consider synonyms as valid results

Rank	Input audio segments			
	beautiful	clever	destroy	suitcase
1	lovely	cunning	destroyed	bags
2	pretty	smart	**destroy**	suitcases
3	gorgeous	**clever**	annihilate	luggage
4	**beautiful**	crafty	destroying	briefcase
5	nice	wisely	destruct	**suitcase**

Corpora	$EN_{ls} - en_{ls}$		$FR_{ls} - fr_{ls}$		$EN_{swc} - en_{swc}$		$DE_{swc} - de_{swc}$		$EN_{ls} - en_{swc}$		$EN_{swc} - en_{ls}$	
Average P@k	P@1	P@5	P@1	P@5	P@1	P@5	P@1	P@5	P@1	P@5	P@1	P@5
Alignment-based approach with cross-modal supervision (parallel dictionary)												
A^*	52.6	66.9	46.6	69.4	47.4	62.5	49.2	63.7	41.3	54.2	39.0	49.4
Alignment-based approaches without cross-modal supervision (our approach)												
A	43.2	57.0	42.4	58.0	36.3	50.4	32.6	48.8	33.9	47.5	33.4	45.7
B	35.0	48.2	35.4	50.4	33.8	44.6	29.3	45.4	30.0	42.9	31.1	40.7
C	27.7	37.3	26.4	35.7	21.1	30.3	26.2	34.5	22.4	28.9	17.1	26.3
D	26.7	35.2	27.2	36.3	21.1	28.2	25.3	33.2	21.2	29.3	18.7	25.1
E	17.7	24.2	20.8	28.4	17.3	21.8	18.3	23.0	15.2	21.1	11.2	17.8
F	3.5	5.7	5.2	6.9	3.8	5.8	2.7	4.9	3.2	5.7	2.9	4.4

Task III | Spoken Word Translation

- More supervision yields better performance
- Translation using the same corpus outperforms those using different corpora

Corpora	$EN_{ls} - fr_{ls}$		$FR_{ls} - en_{ls}$		$EN_{swc} - de_{swc}$		$DE_{swc} - en_{swc}$		$EN_{ls} - de_{swc}$		$FR_{ls} - de_{swc}$	
Average P@k	P@1	P@5	P@1	P@5	P@1	P@5	P@1	P@5	P@1	P@5	P@1	P@5
Alignment-based approach with cross-modal supervision (parallel dictionary)												
A^*	47.9	56.4	49.1	60.1	40.2	51.9	43.3	55.8	34.9	46.3	33.8	44.9
Alignment-based approaches without cross-modal supervision (our approach)												
A	40.5	50.3	39.9	50.9	32.8	43.8	33.1	43.4	31.9	42.2	30.1	42.1
B	36.0	44.9	35.5	44.5	27.9	38.3	30.9	40.9	26.6	35.3	25.4	38.2
C	24.7	35.4	23.9	37.3	22.0	30.3	20.5	29.1	19.2	26.1	14.8	23.1
D	25.4	33.1	24.4	34.6	23.5	29.1	20.7	31.3	20.8	25.9	14.5	22.4
E	15.4	20.6	16.7	19.9	14.1	15.9	16.6	17.0	14.8	16.7	9.7	11.8
F	4.3	5.6	6.9	7.5	4.9	6.5	5.3	6.6	4.2	5.9	1.8	2.6