A Translation Framework for Multimodal Spoken Units Discovery

Liming Wang, Mark Hasegawa-Johnson

University of Illinois at Urbana-Champaign

Asilomar 2021

Overview

Motivation

Multimodal word discovery (MWD)

What is MWD?

A Translation Model for MWD

From MWD to Multimodal Phoneme Discovery (MPD)

A Translation + Compression Model for MPD

Machine vs Human in Learning Speech

- *Machine*:
  - Needs large amount of transcribed speech more than 99% of world's languages have
  - Does not transfer well across different domains
  - Learns from only speech and text
- *Human*:
  - Needs only noisy, untranscribed speech for training
  - Generalizes well
  - Learns from a wide range of information sources besides speech

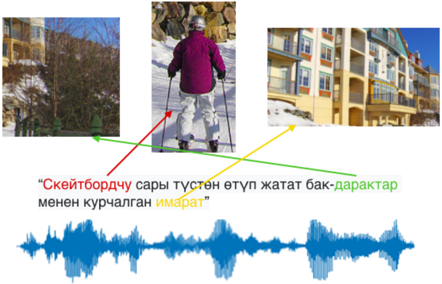
Overview

Motivation

Multimodal word discovery (MWD)
- What is MWD?
- A Translation Model for MWD

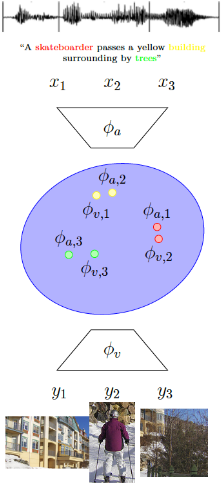
From MWD to Multimodal Phoneme Discovery (MPD)
- A Translation + Compression Model for MPD

Multimodal Word Discovery (MWD): Learn to listen by looking



- Discover word-like units by associating the visual objects with visual words in the speech

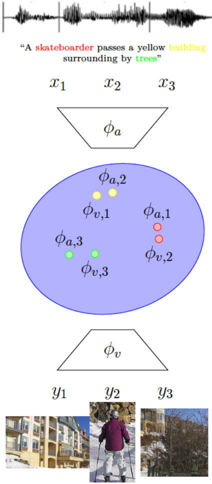
Association mechanism 1: retrieval-based model



---

[a]recall at 10 for speech-to-image retrieval
[b]recall at 10 for image-to-speech retrieval

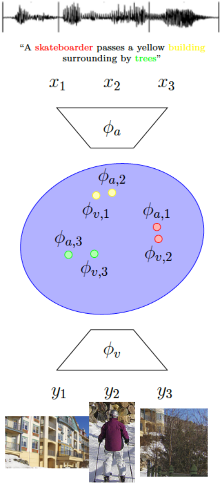
Association mechanism 1: retrieval-based model



- **Mismatch of objective**: Perform well in retrieval, but badly in word discovery

[a] recall at 10 for speech-to-image retrieval
[b] recall at 10 for image-to-speech retrieval

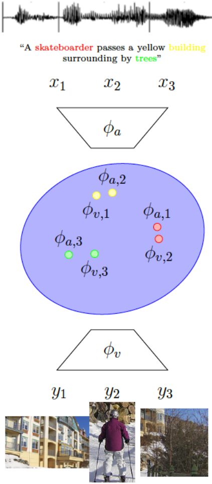
Association mechanism 1: retrieval-based model



- **Mismatch of objective**: Perform well in retrieval, but badly in word discovery
- **Under-constrained**: Learning good sentence embedding $\neq$ learning good word embedding

[a] recall at 10 for speech-to-image retrieval
[b] recall at 10 for image-to-speech retrieval

Association mechanism 1: retrieval-based model



- **Mismatch of objective**: Perform well in retrieval, but badly in word discovery
- **Under-constrained**: Learning good sentence embedding $\neq$ learning good word embedding
- **Results on SpeechCOCO (Havard et al. 2017)**:

	S2I@10[a]	I2S@10[b]	Alignment F1
(Harwath et al. 2018)	57	59	37
Random	1	1	20

[a] recall at 10 for speech-to-image retrieval
[b] recall at 10 for image-to-speech retrieval

Association mechanism 2: Probabilistic Translation Model

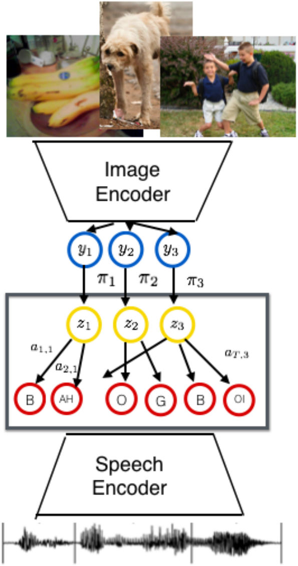


Figure: MWD Translator

- *Image Encoder*: maps ROIs to visual concept probabilities

Association mechanism 2: Probabilistic Translation Model

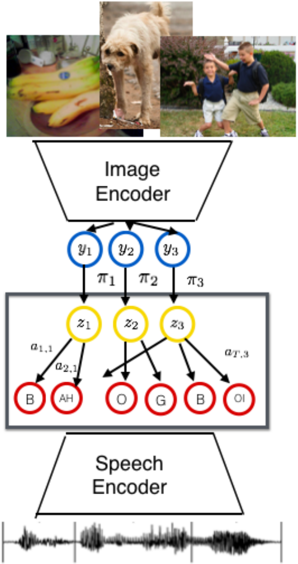


Figure: MWD Translator

- *Image Encoder*: maps ROIs to visual concept probabilities
- *Speech Encoder*: maps spoken segments to phone probabilities

Association mechanism 2: Probabilistic Translation Model

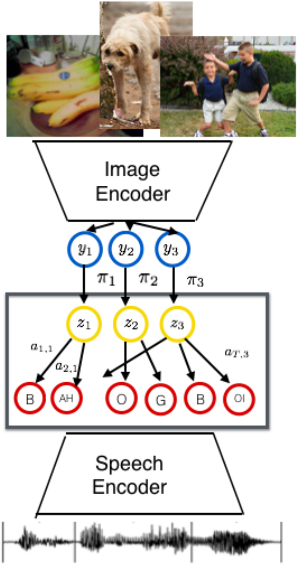


Figure: MWD Translator

- *Image Encoder*: maps ROIs to visual concept probabilities
- *Speech Encoder*: maps spoken segments to phone probabilities
- *Hidden Markov Model Aligner*: Learn the alignment from the phone and concept probability vectors

Association mechanism 2: Probabilistic Translation Model

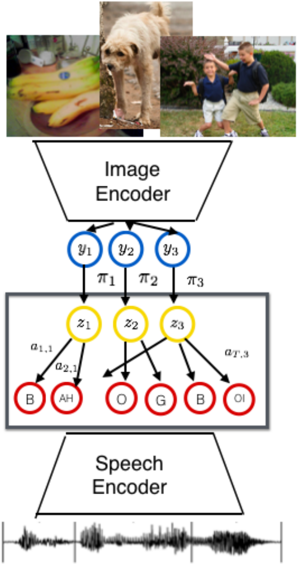


Figure: MWD Translator

- *Image Encoder*: maps ROIs to visual concept probabilities
- *Speech Encoder*: maps spoken segments to phone probabilities
- *Hidden Markov Model Aligner*: Learn the alignment from the phone and concept probability vectors
- *Training objective*: maximum likelihood with expectation maximization algorithm

Evaluation Metrics

- **Alignment F1**: Harmonic mean between the alignment recall and precision:
  - **Alignment recall**: the average probability that a word is aligned correctly over each true position
  - **Alignment precision**: the average probability that a word is aligned correctly given each predicted position
- **Retrieval Recall@1, 5, 10**: The empirical probability that the model retrieves a matching image/caption after $1, 5, 10$ trials respectively

Experimental Results

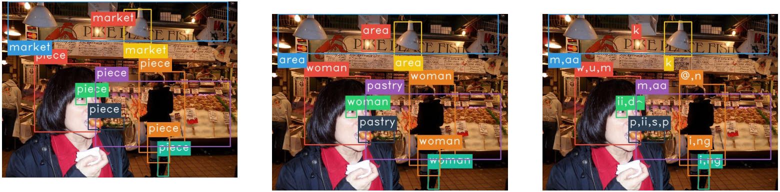
	S2I @1	@5	@10	I2S @1	@5	@10
Cosine+TDNN (Harwath et al. 2018)	**12**	**38**	**57**	**12**	**41**	**59**
SMT	3	13	20	0.1	0.5	1
SMT (phones)	7	24	36	4	16	28

Table: Speech-to-image (S2I) and image-to-speech (I2S) retrieval performance of various systems on SpeechCOCO

	Alignment Recall	Alignment Precision	Alignment F1
Cosine+TDNN	54.9	27.8	36.9
SMT	**60**	**30**	**40**

Table: Word discovery performance of various systems on SpeechCOCO; Results are evaluated only with words that describe one of the 80 concepts

Visualization of Discovered Words

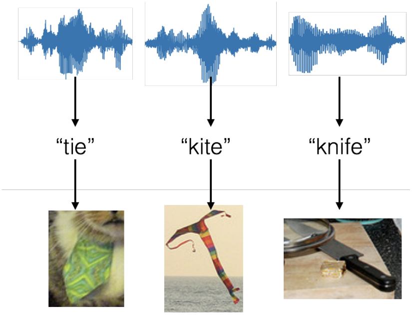


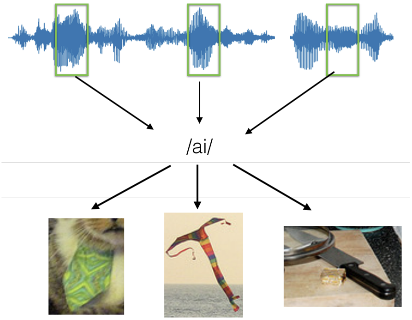
(a) audio-level cosine+TDNN (b) audio-level SMT (c) phone-level SMT

Figure: Word discovery results of different systems on the image-caption pair "a woman eating a piece of pastry in a market area." The texts are not available in the first two figures during training and are shown for ease of understanding.

Overview

From MWD to Multimodal Phoneme Discovery (MPD)





Word:

- Unit most directly related to meaning
- Large vocabulary size, large sample complexity
- Unreliable for understanding unseen words, not universal across languages

Phoneme:

- Smallest meaning-preserving unit
- Low vocabulary size, relatively low sample complexity
- Shared among words, more universal across languages

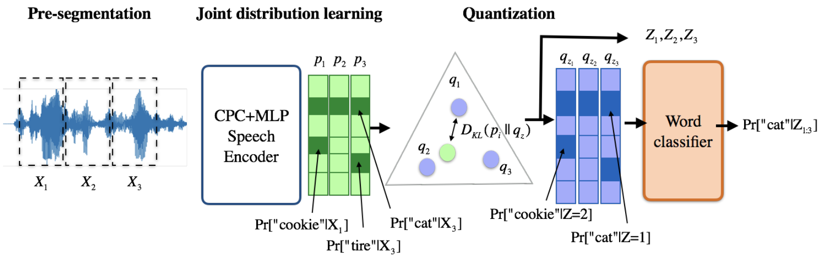
Acoustic Units (AU) as Information Bottleneck (IB)

- **The information bottleneck objective** (Tishby et. al., 1999): For Markov chain $Z - X - Y$, Z is an information bottleneck of (X, Y) if $(P^*_{Z|X}, P^*_{Y|Z})$ is the optimal solution of

$$\max_{P_{Z|X}, P_{Y|Z}} \quad I(Z;Y)$$
$$s.t. \quad I(Z;X) \leq I_0.$$

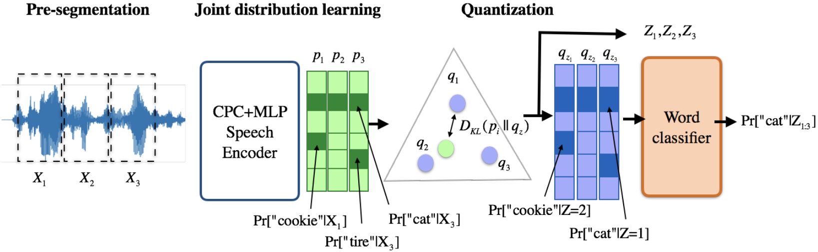
- **MAUD as special cases of IB**:
  - $X = [X_1, \cdots, X_T]$ is the sequence of spoken segments, $Y \in \mathcal{Y}$ is the visual word and $Z = [Z_1, \cdots, Z_T] \in \{1, \cdots, K\}^T$ is the AU sequence represented by X.
  - MWD: T is the number of words, $I_0 \approx H(\text{Word}) \times T$
  - MPD: T is the number of phonemes, $I_0 \approx H(\text{Phoneme}) \times T$

Information Quantizer (IQ): A Translation + Compression model for MPD



- **Pre-segmentation**: Either use an algorithm based contrastive predictive coding (CPC) representation (Kreuk et al. 2020), or simply use framewise representation from a convolutional neural net (CNN)

A Translation + Compression Model for MPD

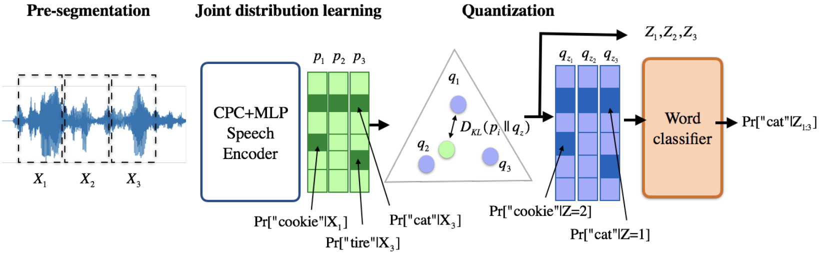


- **Joint distribution learning objective**: $P^{\theta}_{Y=y|X=x} := \Pr[Y = y|X = x]$ is learned by a multilayer perceptron (MLP); $q(\cdot) : \Delta^{|\mathcal{Y}|} \to \{q_1, \cdots, q_K\} \subset \Delta^{|\mathcal{Y}|}$ is some quantizer on the probability simplex

$$\min_{\theta, q(\cdot)} \sum_{i=1}^{n} \log q_{y_i}(P^{\theta}_{Y|X=x_i}) \quad \text{(with ST[1])} \quad \text{or} \quad \sum_{i=1}^{n} \log P^{\theta}_{Y|X}(y_i|x_i) \quad \text{(w/o ST)}$$

[1] Straight-through gradient

A Translation + Compression Model for MPD



- **Quantization (IB) learning objective**:

$$\min_{\theta, q(\cdot)} \sum_{i=1}^{n} D_{KL}(\mathrm{sg}[P^{\theta}_{Y|X=x_i}] || q(P^{\theta}_{Y|X=x_i})) + D_{KL}(P^{\theta}_{Y|X=x_i} || \mathrm{sg}[q(P^{\theta}_{Y|X=x_i})])$$

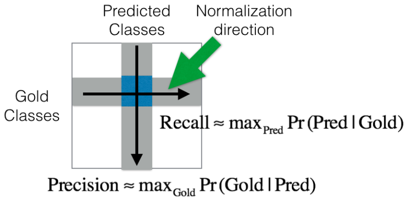
[1] sg[·]: Stop-gradient operator

Datasets

- **Visual-word only datasets**: Created by cutting out visually salient noun segments from the utterances using forced alignments
  - **Flickr audio [Harwath & Glass 2015]**:
    - Visual words extracted from Flickr30kEntities with frequency at least 50 ($|\mathcal{Y}| = 258$) over the whole dataset
    - Training: 23741 words
    - Test: 2491 words
  - **LibriSpeech**:
    - Same set of visual words as Flickr audio
    - Training: 42015 words from train-clean-100 and train-clean-360
    - Test: 595 words from dev-clean
- **Whole-sentence dataset**:
  - Training: LibriSpeech with three subsets of words:
    - Visual words: same set as Flickr, $|\mathcal{Y}| = 224$
    - Visual words + top-300 words: $|\mathcal{Y}| = 524$
    - Visual words + top-600 most frequent words: $|\mathcal{Y}| = 824$
  - **TIMIT**: the whole dataset excluding `SA` utterances, 5040 utterances

Evaluation Metrics

- **Token F1**: Harmonic mean between token recall and precision
  - **Token recall**: the average probability of the most likely cluster over each phoneme
  - **Token precision**: the average probability that the most likely phoneme over each cluster



- **Normalized Mutual Information (NMI)**: Computed using the empirical joint distribution between the predicted (clusters) and gold classes (phonemes) as

$$NMI := \frac{I(\text{Pred}, \text{Gold})}{\text{avg}(H(Pred, H(Gold)))}$$

- **Boundary F1**: between each predicted phoneme boundary times and the gold boundary times with a tolerance of 20ms

Phoneme Discovery Results: Visual Word-only Datasets

Flickr Audio Word	Token Precision	Recall	F1
Continuous representation			
CPC+k-means (Nguyen et al. 2020)	31.3	39.8	35.1
k-means	31.6	43.5	36.6
Discrete representation			
Gumbel VIB (Alemi et al. 2017)	34.2	**51.6**	41.1
DIB (Strouse et al. 2016)	51.1	42.9	46.6
IQ (Ours), K=44	55.4	50.5	**52.9**
IQ (Ours), K=100	**61.2**	42.3	50.0
IQ (Ours), K=256	60.8	40.0	48.3

Table: Phoneme discovery results on isolated visual words from Flickr Audio. The baseline results are obtained with $K = 44$. All results use gold segmentation.

LibriSpeech Word	Token Precision	Recall	F1
Continuous representation			
CPC+k-means (Nguyen et al. 2020)	41.1	55.5	47.2
k-means	57.5	49.4	53.1
Discrete representation			
Gumbel VIB (Alemi et al. 2017)	39.9	65.1	49.5
DIB (Strouse et al. 2016)	61.8	61.2	61.6
IQ (Ours), K=39	**62.2**	**63.1**	**62.6**

Table: Phoneme discovery results on LibriSpeech visual words with ground-truth segment boundary. The baseline results are obtained with $K = 39$. All results use gold segmentation.

Phoneme Discovery Results: Whole-sentence Dataset

TIMIT	Token F1	NMI	Boundary F1
(Harwath et. al. 2020)	-	35.9	54.2
(Yusuf et. al. 2020)	-	40.1±0.1	76.6 ±0.5
(Feng et. al. 2021, GP only, K=50)	-	36.8	70.5
+ gold segmentation	-	51.2	97.8
+ gold segmentation, K=39	-	50.4	97.1
(Ours) IQ, $\lvert\mathcal{Y}\rvert$=224, K=39	37.9±1.2	38.6±0.7	77.1±0.1
+ training on TIMIT	39.3	39.2	77.2
+ gold segmentation	51.8	59.8	98.0
(Ours) IQ, $\lvert\mathcal{Y}\rvert$=524, K=39	42.4±0.1	43±0.5	**79.4**±0.1
+ training on TIMIT	45.7	44.3	79.1
+ gold segmentation	55.7	61.6	98.0
(Ours) IQ, $\lvert\mathcal{Y}\rvert$=824, K=39	43.9±0.1	44.3±0.2	79.2±0.0
+ training on TIMIT	**46.0**	**45.2**	79.1
+ gold segmentation	55.3	**63.4**	98.0

Table: Phoneme discovery results on TIMIT

- More vocab helps
- Training on TIMIT helps
- Large (19%) gap between using or not using gold segmentation

Visualization of Discovered Phonemes

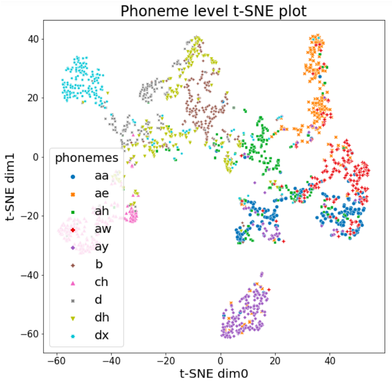


Figure: t-SNE plots of phoneme clusters discovered by IQ with gold segmentation on TIMIT

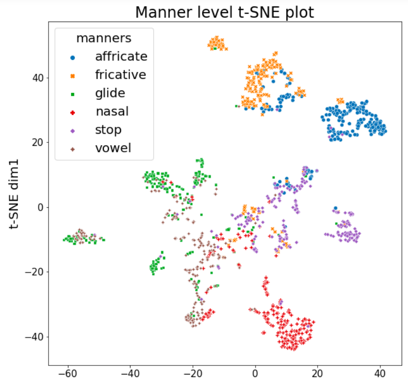


Figure: Manner-level t-SNE plots of phoneme clusters discovered by IQ with gold segmentation on TIMIT

Codeword Distribution of Predicted Phonemes

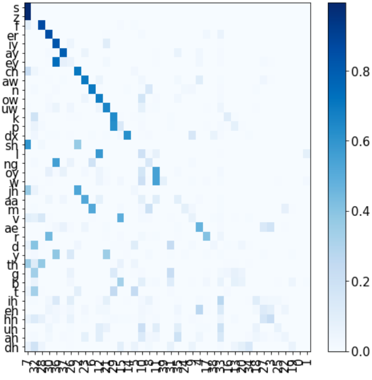


Figure: Codeword distribution of phoneme clusters discovered by IQ with gold segmentation on TIMIT

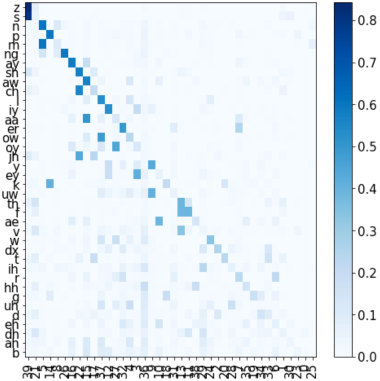


Figure: Codeword distrbution of phoneme clusters discovered by IQ with predicted segmentation on TIMIT

Confusion between Phonemes: Gold Segmentation Case

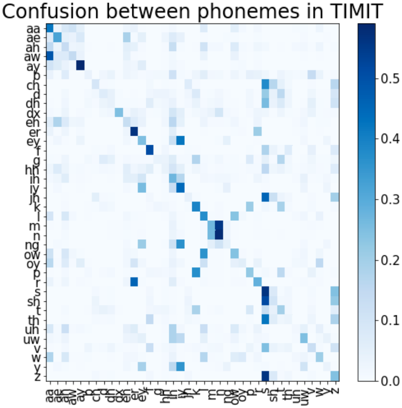


Figure: Confusion matrix of phonemes by IQ with gold segmentation on TIMIT

Phoneme Pair	Error Prob.
ae, aa	1.00
ch, ah	0.85
sh, s	0.82
ah, aa	0.82
aw, aa	0.77
z, s	0.75
n, m	0.73
p, k	0.70
r, er	0.67
iy, ey	0.60

Table: Top-10 most confusing phoneme pairs by IQ with gold segmentation on TIMIT

Confusion between Phonemes: Predicted Segmentation Case

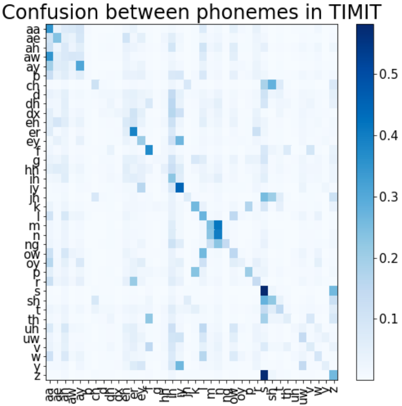


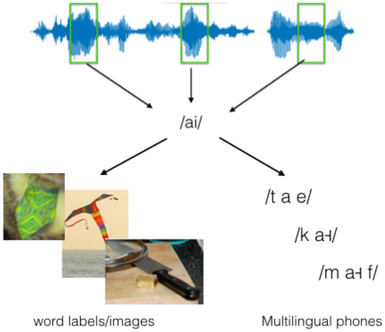
Figure: Confusion matrix of phonemes by IQ with predicted segmentation on TIMIT

Phoneme Pair	Error Prob.
ae, aa	1.00
ah, aa	0.81
z, s	0.78
aw, aa	0.72
ay, aa	0.54
n, m	0.49
sh, s	0.48
iy, ey	0.45
dh, ah	0.42
ch, ah	0.41

Table: Top-10 most confusing phoneme pairs by IQ with predicted segmentation on TIMIT

Conclusion and Current Work

- Translation and compression are useful metaphors for exploiting multi-modal information in speech technology
- Current work: incorporate multilingual information into the IB framework; apply the model to a low-resource language called Mboshi



Further Reading

- [Wang et al, 2021] Align or Attend? Toward More Efficient and Accurate Spoken Word Discovery using Speech-to-image Retrieval.
  Liming Wang, Xinsheng Wang, Mark Hasegawa-Johnson, Odette Scharenborg, Najim Dehak.
  ICASSP 2021.
- [Wang and Hasegawa-Johnson, 2020] A DNN-HMM-DNN Hybrid Model for Discovering Word-like Units from Spoken Captions and Image Regions.
  Liming Wang, Mark Hasegawa-Johnson.
  Interspeech 2020.