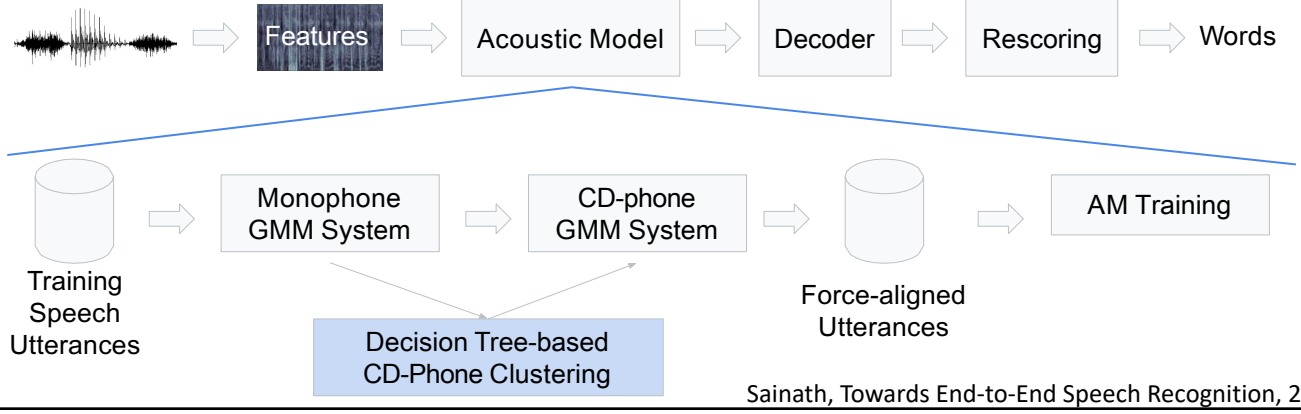
Towards Unsupervised Speech Processing

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

1

A Circa 2012 Perspective on Speech Recognition

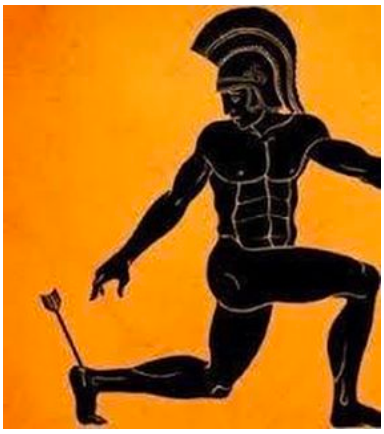
- Dramatic speech recognition progress in English & other major languages
- The (re)rise of neural-network acoustic models; hybrid HMM models
- Critical dependence on annotated speech data, pronunciation dictionaries
- Relatively modest learning paradigm



Sainath, Towards End-to-End Speech Recognition, 2019

2

The "Achilles Heel" of Modern AI



- Many successful machine learning tasks rely on large quantities of *annotated* training data
  - Annotated data comes in {Input, Output} pairs

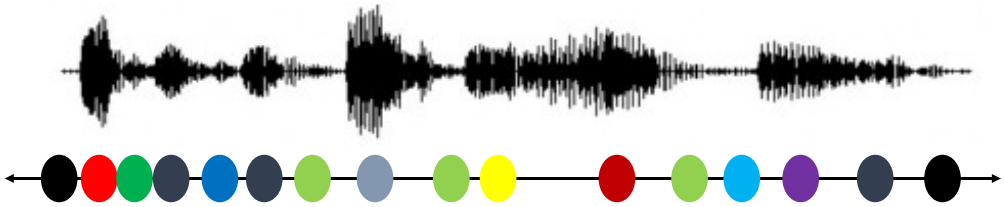
Input ⇨ **Model** ⇦ **Output**

- Issues:
  - Models learn the biases in the training data
  - Training data should match the "testing" conditions
  - Annotating large corpora is time-consuming and expensive
- Challenge:
  - There is far more raw data in the world than annotated data
  - Can we build models that learn with much less supervision?

3

The Speech Recognition Learning Paradigm

- The training paradigm for speech recognition is >40 years old
  - {Speech, words} pairs enable alignment at phone/character level
  - Training becomes an exercise in aligning "beads on a string"

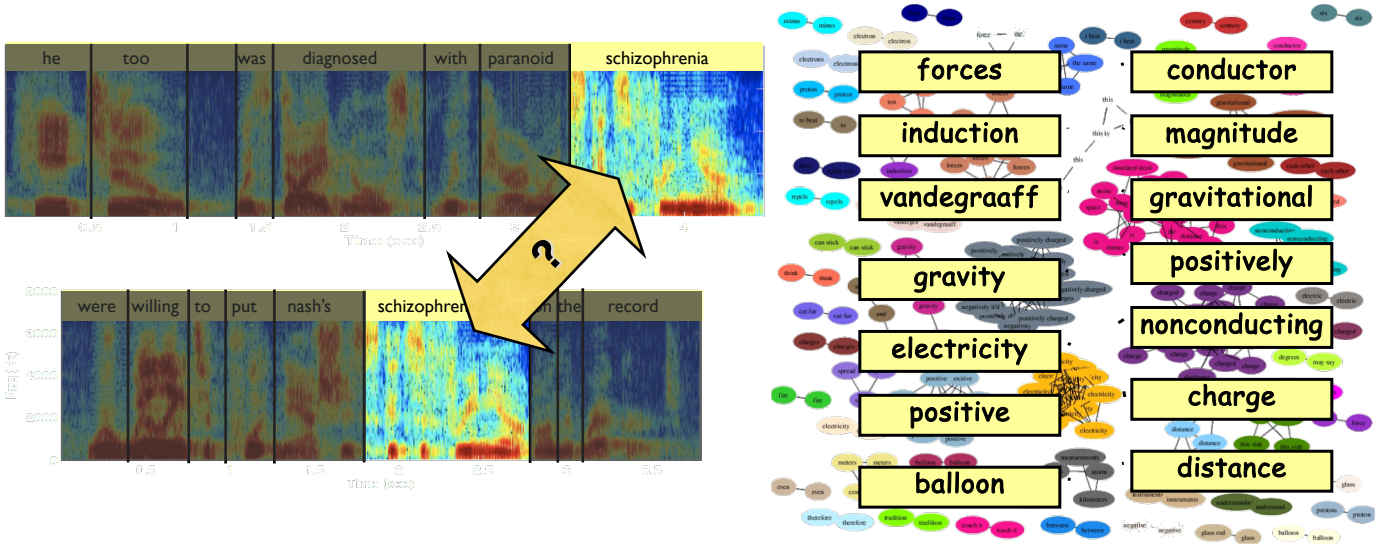


- Cost of annotations limits ASR to major languages of the world
- An ability to learn 1) with weakly constrained inputs from 2) freely available data, will be a major paradigm shift for low-resource speech tasks
- P.S. *This is not how humans learn speech!*

4

Unsupervised Pattern Discovery in Speech

- Finding reoccurring speech sequences in raw audio corpora

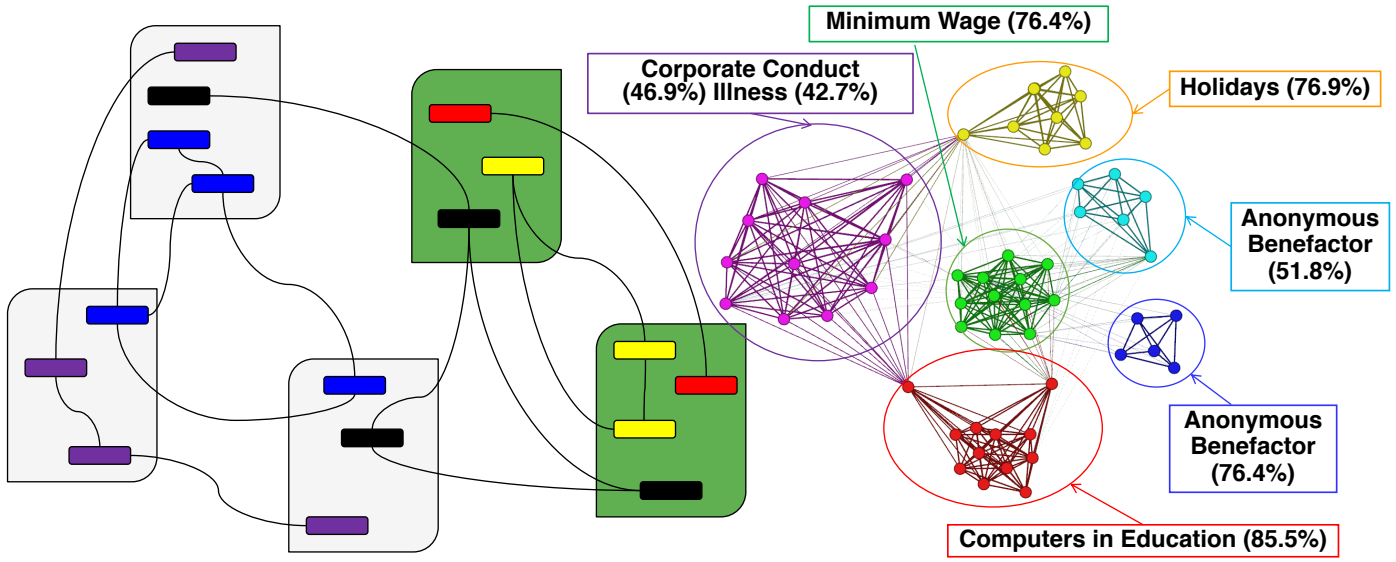


Park & Glass, Unsupervised Pattern Discovery in Speech, ASLP 2008

5

Unsupervised Speech Topic Discovery

- Clustering unannotated speech conversations based on latent topics

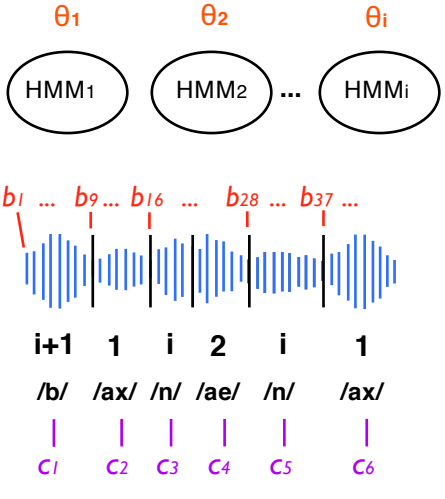


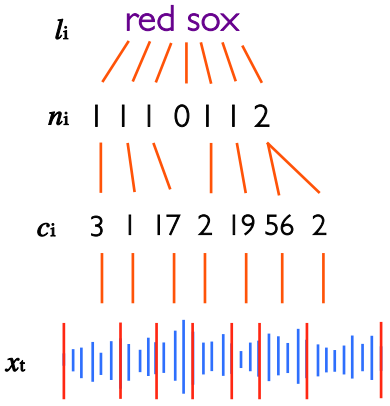
Harwath et al., Zero Resource Spoken Audio Corpus Analysis, ICASSP 2013

6

Unsupervised Acoustic Unit Discovery

- Bayesian approaches for learning speech units & pronunciations





mountain	Monterrey
m aw n tcl t ax n	m aa n tcl t axr ey
m ow n tcl t ax n	m aa n tcl t r ey

Borneo	Marseille
bcl b ao r n iy ow	m aa r s ey
bcl b ao r n y ax l	m aa r s ey ax

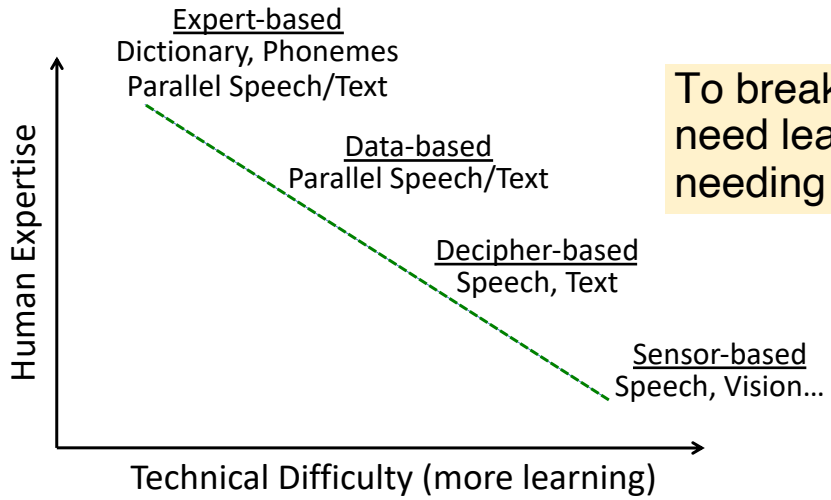
Jakarta	freeze
dcl jh ax kcl k aa r tcl t ax	f r iy z
dcl jh ae kcl k aa r tcl t ax	f r iy z

Lee & Glass, A Nonparametric Bayesian Approach to Acoustic Model Discovery, ACL 2012

7

Speech Processing Challenges

Most (~98%) of the worlds languages have not been addressed by resource and expert intensive supervised speech recognition training methods



To break the speech barrier we need learning techniques needing less human annotations!

Glass, Towards Unsupervised Speech Processing, ISSPA 2012

8

Spoken Language Processing Challenges

	Speech	Text	Vocabulary	Alignments	Phonemes	Dictionary	Modalities
Expert	✓	✓	✓	✓	✓	✓	✗
Data	✓	✓	✓	✓	✗	✗	✗
Decipher	✓	✓	✓	✗	✗	✗	✗
Sensor	✓	✗	✗	✓	✗	✗	✓
…	✓	?	?	?	?	?	?
Unsupervised	✓	✗	✗	✗	✗	✗	✗

- There are many reasonable low-resource speech learning scenarios
  - Large amounts of unannotated & small amounts of annotated data
  - Can have phoneme inventory & dictionary, but no alignments

9

2012-2022: A Remarkable 10 Years

- Major advances in speech, NLP, and machine vision
  1) End-to-end neural network models
     - Seq2seq, attention, transformers, conformers; CTC, RNN-T, LAS, …
  2) Learned embedding spaces
     - Enables cross-modal learning (e.g., image captioning)
  3) Self-supervised learning
     - Enables low-resource ASR
- A lot of other supporting things have improved too:
  - Large datasets, challenge tasks, public benchmarks, rapid publishing
  - Compute infrastructure
  - Wide-scale deployment of speech-enabled devices

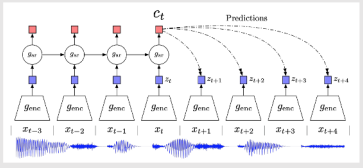
10

Learning from speech and audio

11

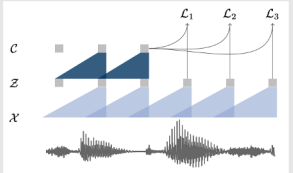
Self-Supervised Learning

Contrastive Learning

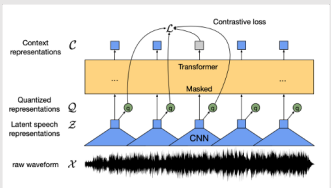


CPC, 2018

wav2vec, 2019

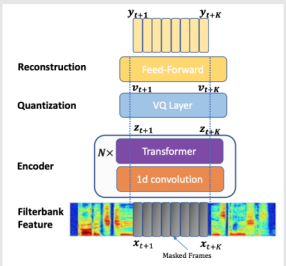


wav2vec 2.0, 2020



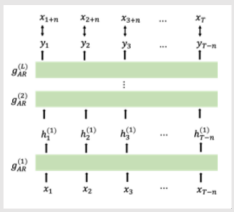
Reconstructive Learning

DeCoAR 2.0, 2020



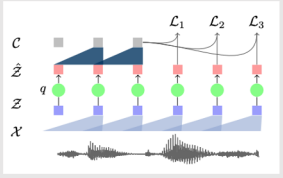
Continuous

APC, 2019

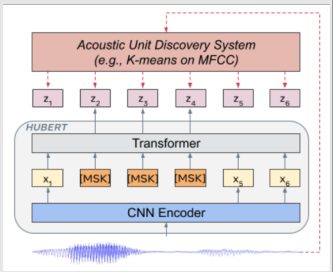


Discrete

vq-wav2vec, 2019



HuBERT, 2021

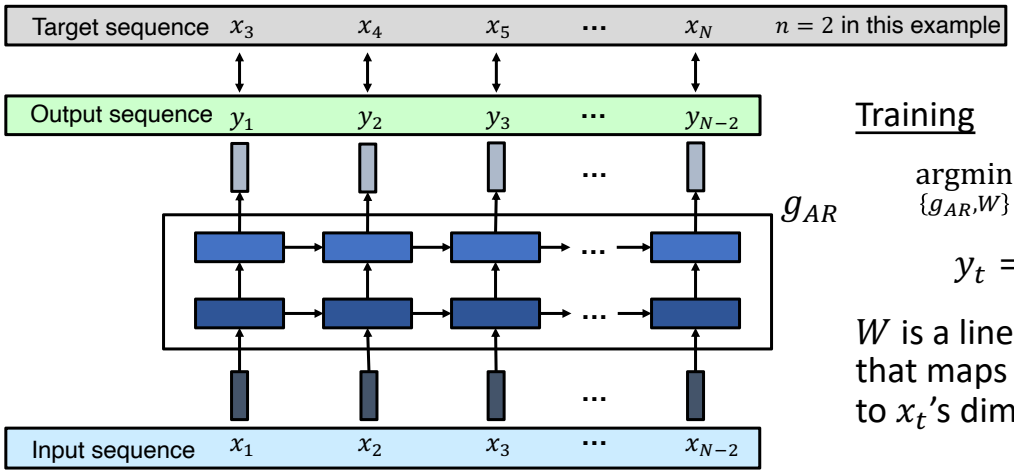


12

Autoregressive Predictive Coding (APC)

Given frames $(\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_t})$, predict frame $\boldsymbol{x_{t+n}}$ that is $\boldsymbol{n}$ steps ahead

- Use autoregressive encoder g_{AR} to summarize history and produce output
- $n \geq 1$ encourages g_{AR} to infer global structure rather than local smoothness



Training

$$\underset{\{g_{AR}, W\}}{\operatorname{argmin}} \sum_{t=1}^{N-n} |x_{t+n} - y_t|,$$

$$y_t = g_{AR}(x_t) \cdot W$$

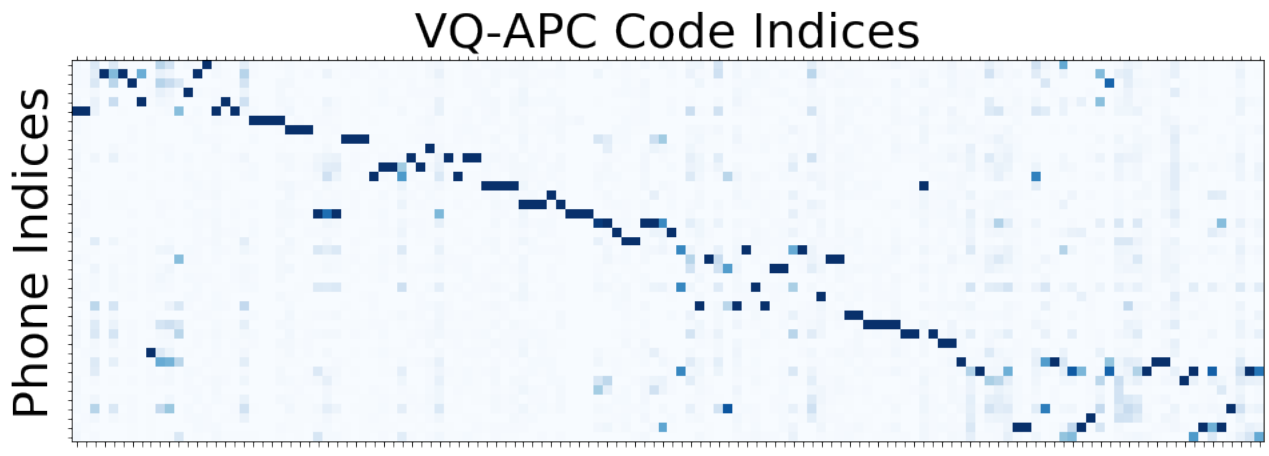
W is a linear transformation that maps g_{AR}'s output back to x_t's dimensionality

Chung & Glass, Generative Pre-Training for Speech with Autoregressive Predictive Coding, ICASSP 2020

13

Visualization of VQ-APC Codebook

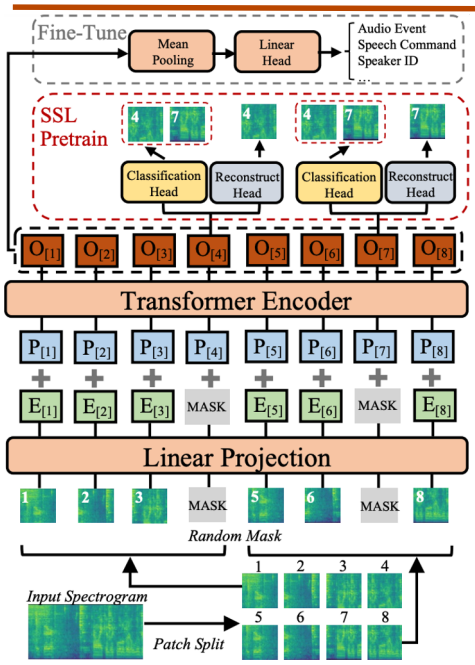
- Conditional probabilities $P(\text{phone}|\text{code})$ estimated from co-occurrence statistics on the training set for codebook size 128, VQ at top layer



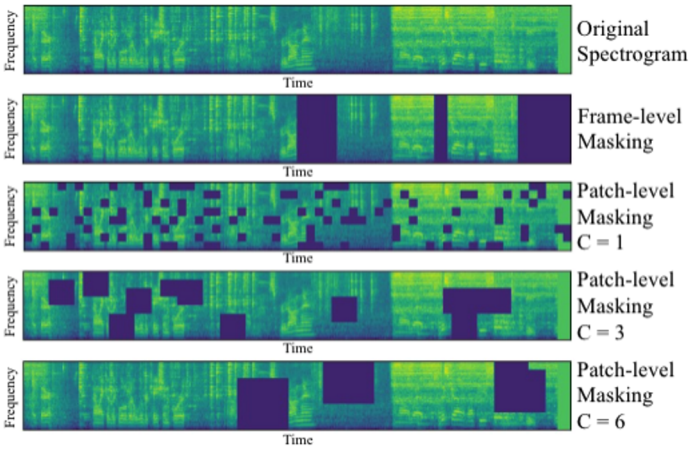
Chung et al., Vector-quantized Autoregressive Predictive Coding, Interspeech, 2020

15

Self-Supervised Audio Representation Learning



- A transformer and patch-based architecture
- Self-supervised learning framework
- Evaluated on several audio & speech tasks



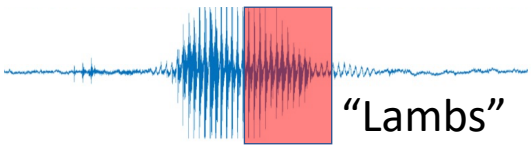
Gong et al., Self-Supervised Audio Spectrogram Transformer, AAAI 2022

16

Learning from speech and vision

18

19



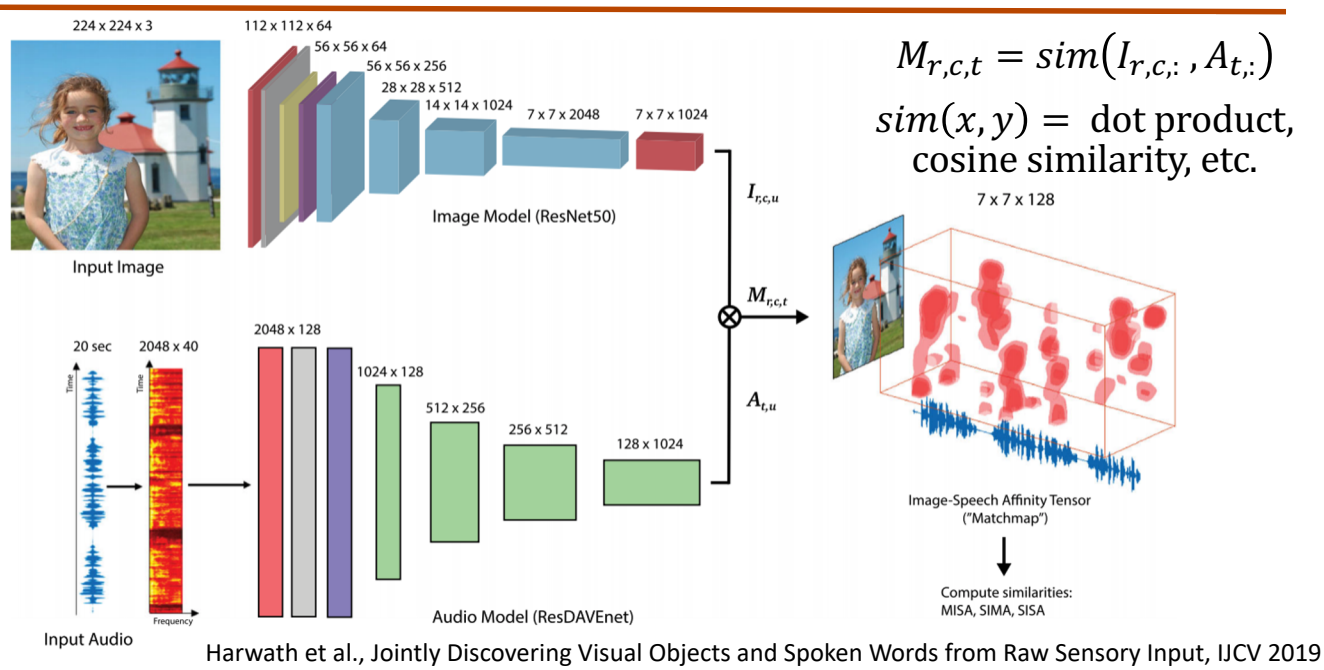
“Lambs”

“Trees”

~~Two white lambs in a meadow.~~ ~~A mountain chain with forests and trees.~~

20

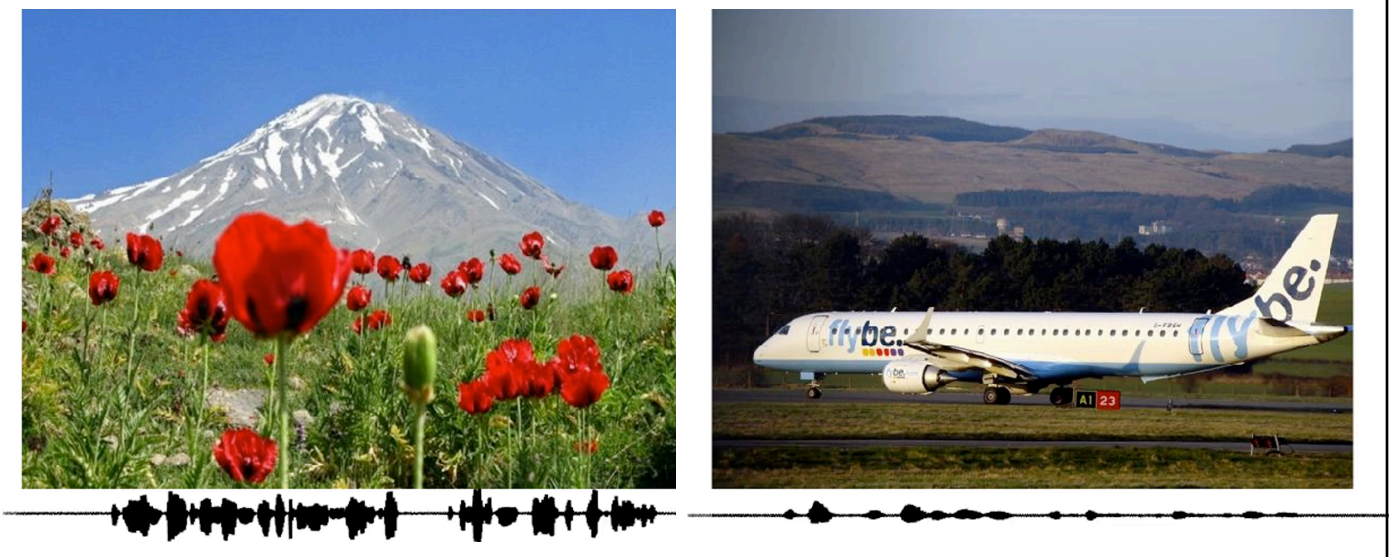
Deep Audio-Visual Embedding Networks



Harwath et al., Jointly Discovering Visual Objects and Spoken Words from Raw Sensory Input, IJCV 2019

21

Matchmap Visualizations

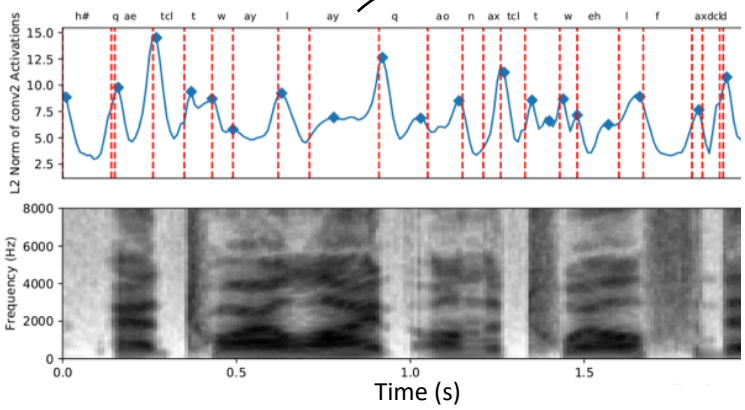


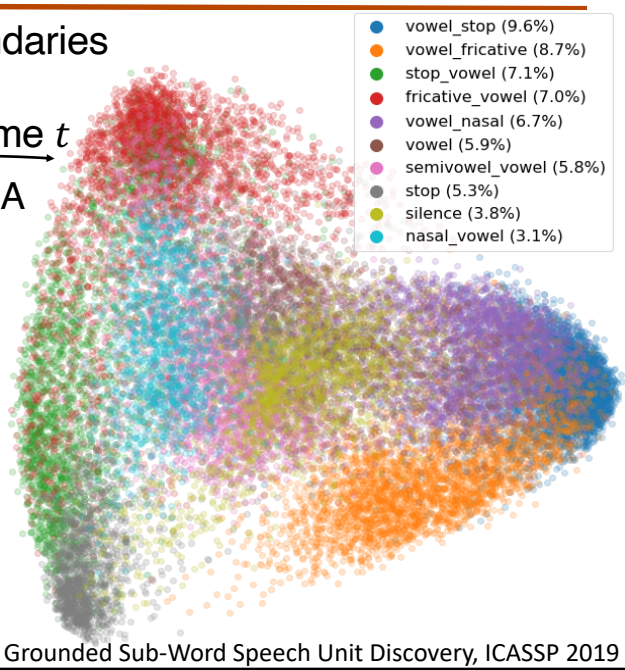
22

Emergence of Diphone-Like Units

Red dashed lines: ground-truth phone boundaries
Blue curve: $f(t) = \|z_t\|_2$
where z_t is vector of all conv2 neurons at time t

Peakpick + PCA

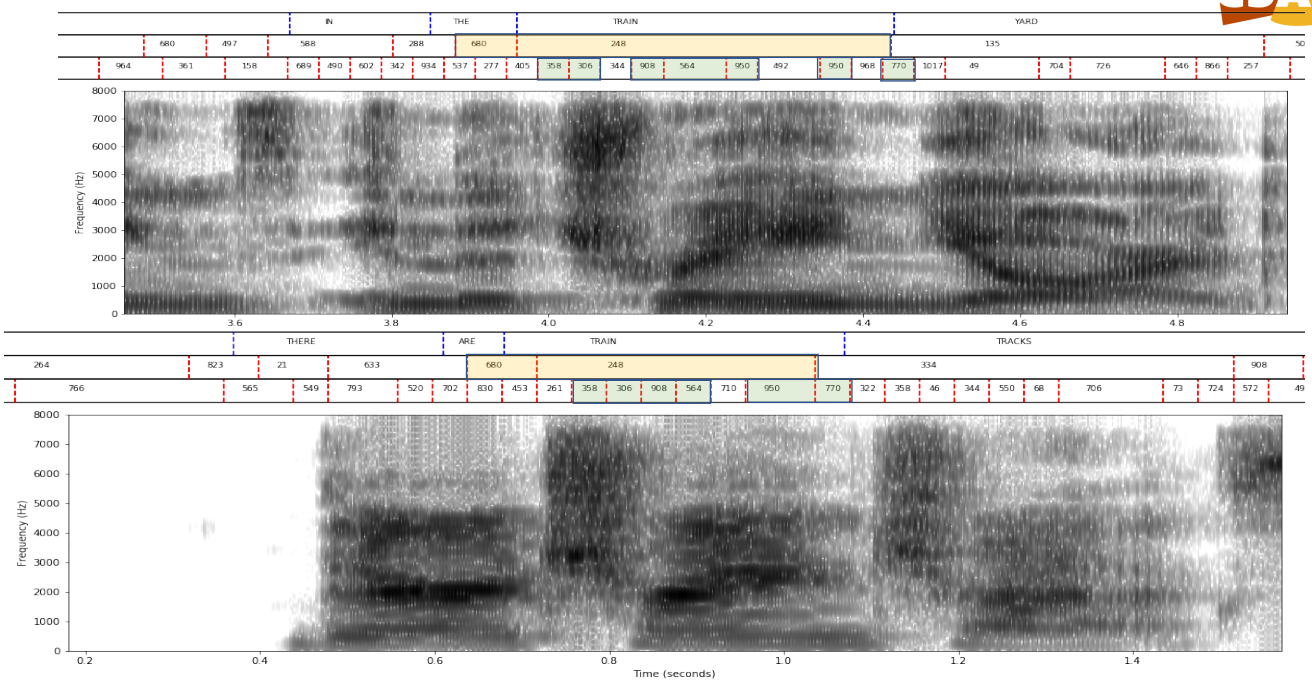




Harwath & Glass, Towards Visually Grounded Sub-Word Speech Unit Discovery, ICASSP 2019

23

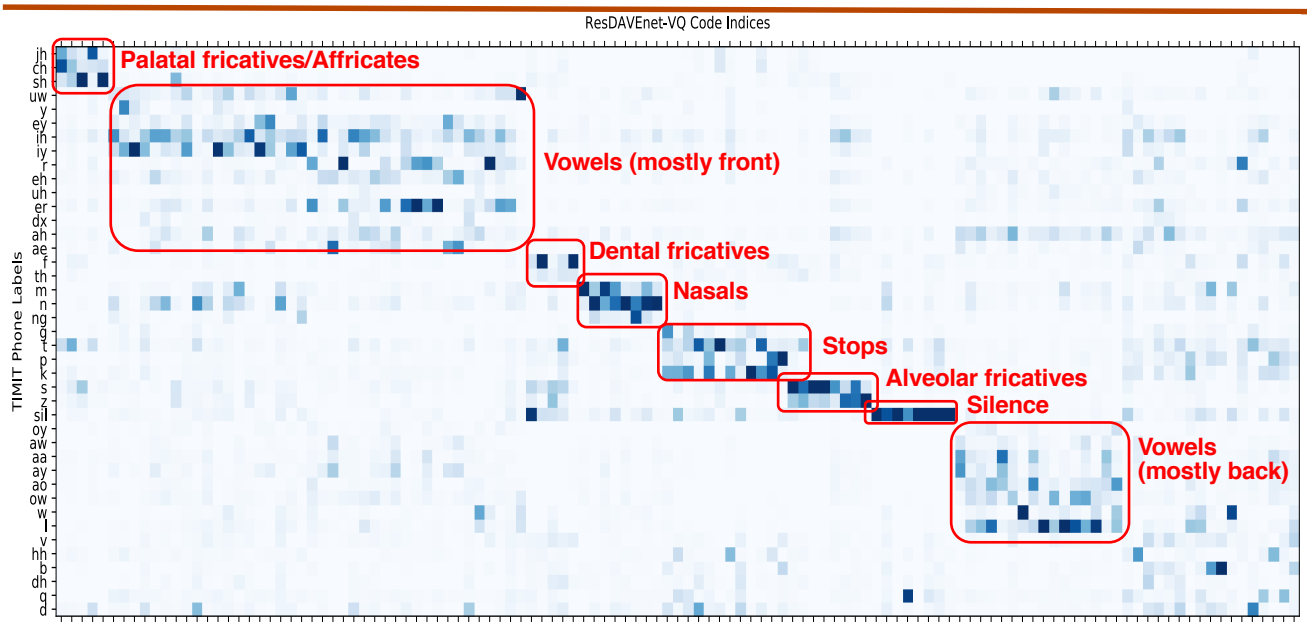
Learning Discrete Speech Representations



Harwath et al., Learning Hierarchical Discrete Linguistic Units from Visually-Grounded Speech," ICLR 2020

24

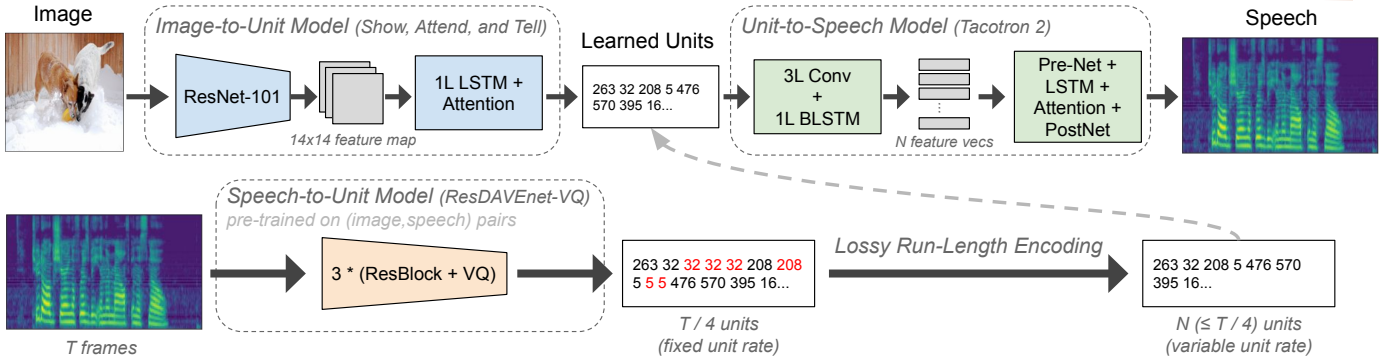
Lower VQ Layers Capture Phones



Frame-level $P(phone|unit)$ computed across TIMIT

25

Learning to Speak from Vision



- Three learning components:
  1. Learn semantic units via audio-visual correspondences
  2. Learn image-to-unit mapping
  3. Learn unit-to-speech mapping

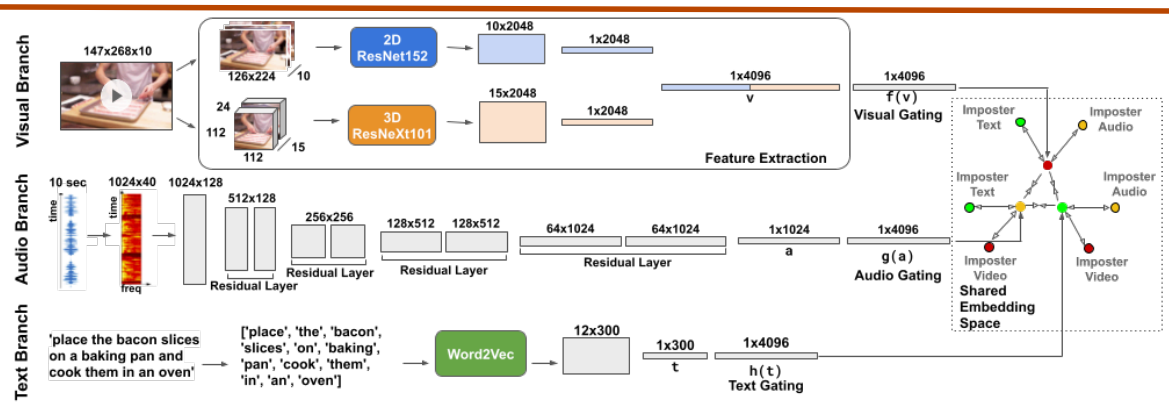
Hsu et al., Text-Free Image-to-Speech Synthesis Using Learned Segmental Units, ACL-IJCNLP 2021

26

Text-Free Image Spoken Caption Examples

27

Learning Representations from Video



- AVLnet: A cross-modal model that incorporates visual, audio, and (optionally) text inputs to learn a joint embedding space
- Learns from unannotated video input (e.g., HowTo100M)
- Evaluated on several image or video and speech retrieval tasks

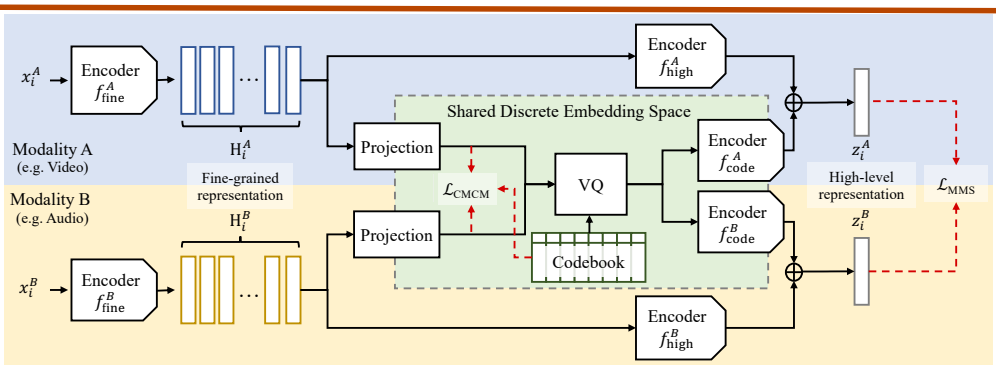
Rouditchenko et al., Learning Audio-Visual Language Representations from Instructional Videos, Interspeech 2021

28

AVLnet Video Retrieval Example

29

Cross-Modal Discrete Representation Learning

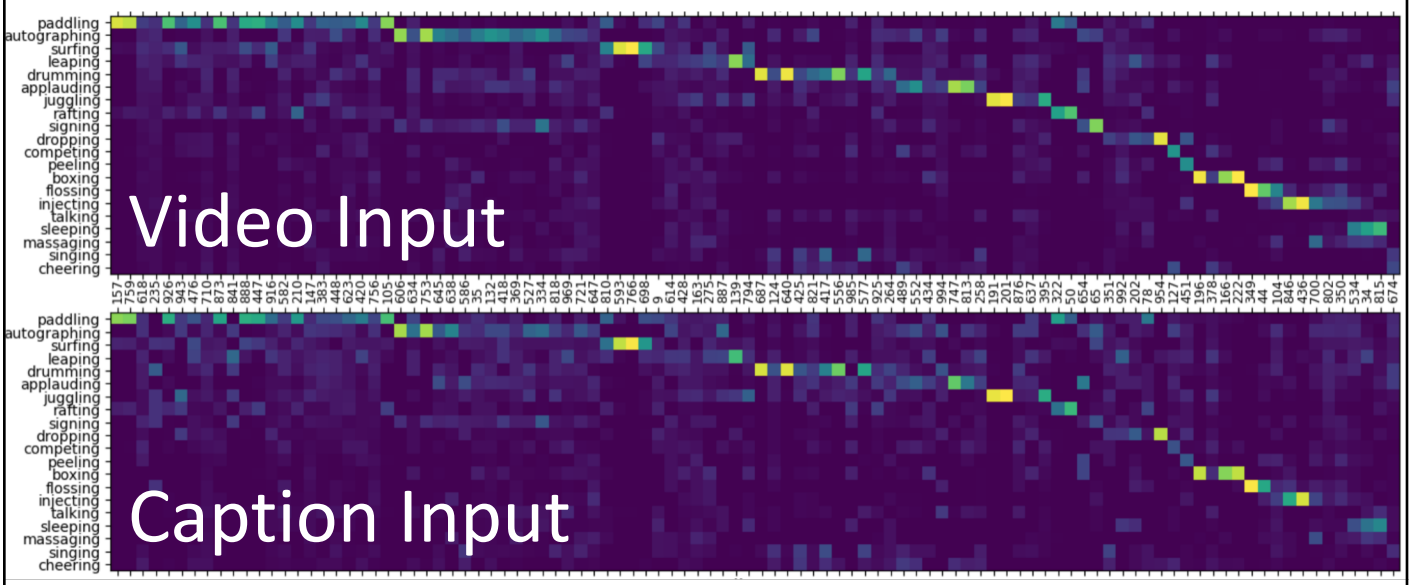


- Self-supervised learning framework to learn a representation that captures fine correspondences of raw audio and visual inputs
- Incorporates a discretized embedding space shared across modalities
  - Uses a cross-modal matching objective of codeword distributions
- Evaluated on several image or video and speech/text tasks

Liu et al., Cross-Modal Discrete Representation Learning, ACL 2022

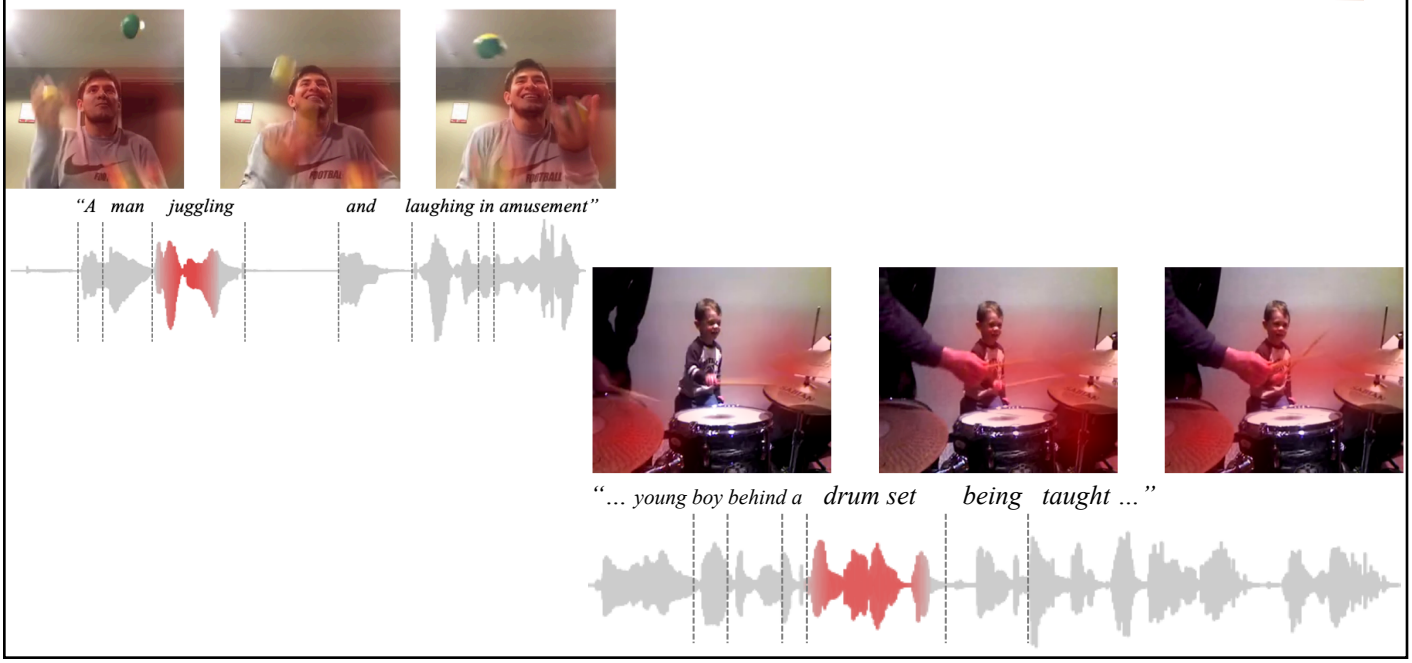
30

Codeword Shared Semantics



31

Audio-Visual Codeword Visualizations



32

Some final thoughts

33

Disentangling Learned Representations

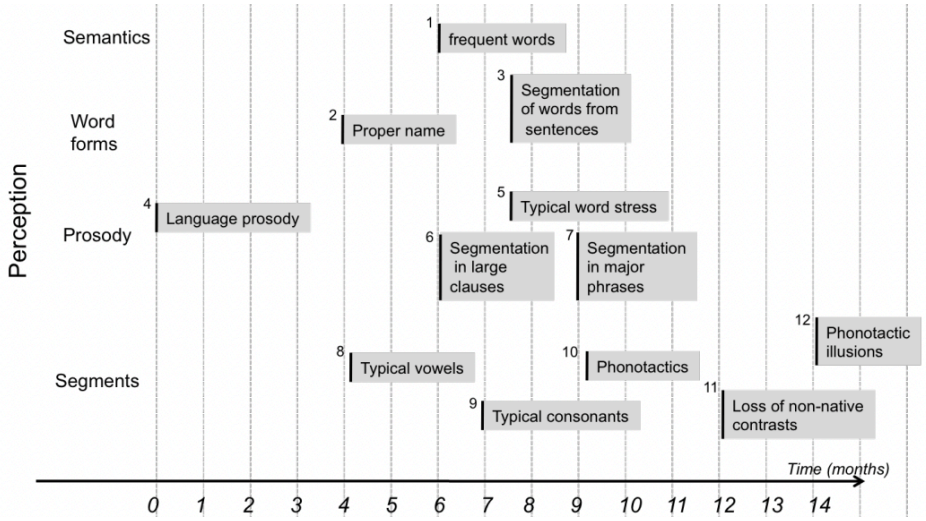
- Learned representations are typically used in a pre-training, fine-tuning learning scenario; can also be used as frozen "feature-vectors"
- Some low-resource problems are not amenable to this paradigm
  - e.g., A small cohort of speakers with a particular health condition
- Self-supervised learning would be more effective if it was able to disentangle the different aspects of the speech signal

  Speech = linguistics + prosody + talker + health + emotion + environment + …
  - Currently, most learned representations mix all of these factors
  - A generalizable talker and environment-embedding would foster generalization and be applicable to all languages

34

Learning More from Less?

- Speech corpora sizes can now be measured in terms of speech "years"
- Humans learn language without supervision with far less data



Dupoux, Cognitive Science in the Era of AI, Cognition, 2018

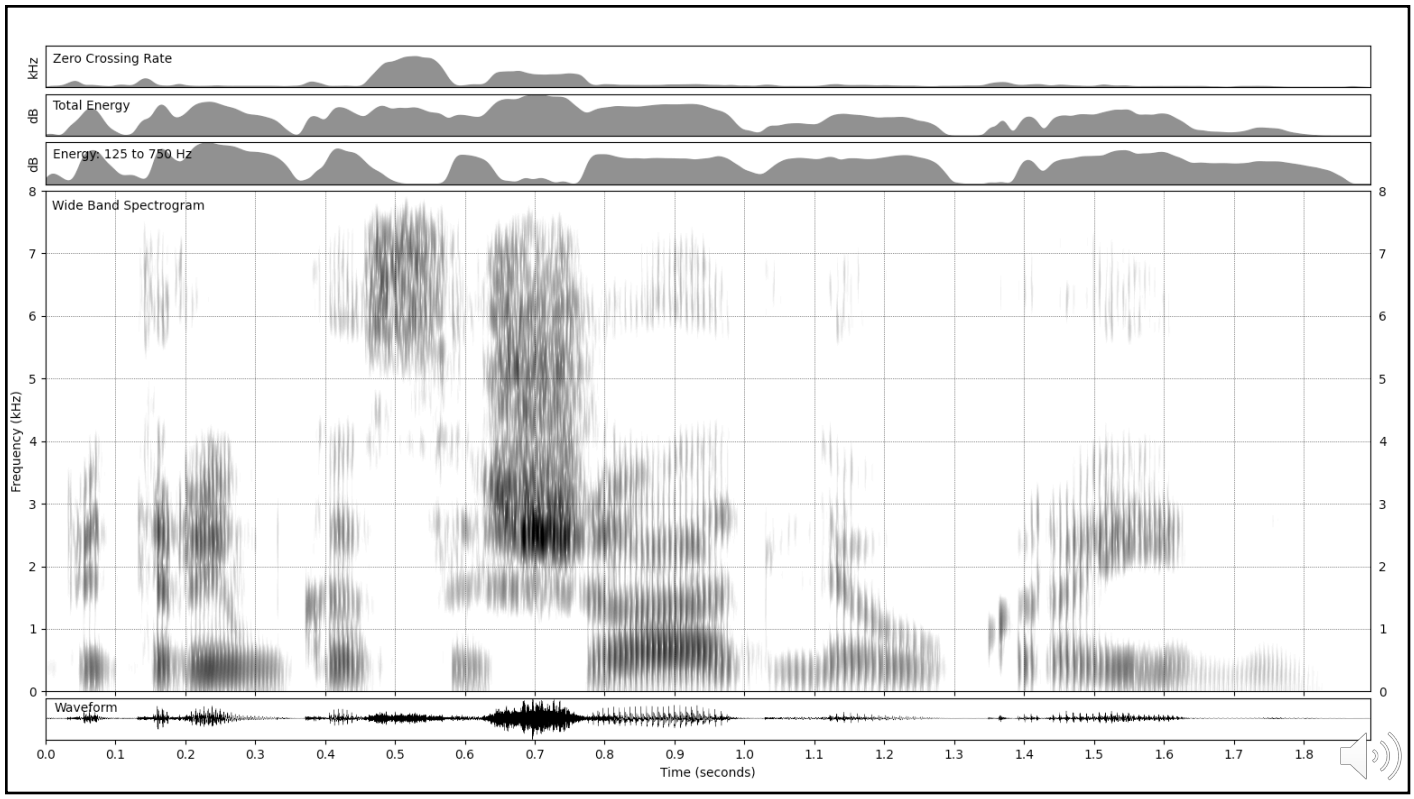
35

Few-shot Learning

- Many languages have very limited or no speech resources
  - Collecting data can be very time consuming
- How much can be learned from a small number of examples?
  - e.g., spoken words and sentence examples?
- In the limit, could we only need to see a single example of a phone to know that it exists in a languages inventory?
- Due to infinite memory and collaborative capabilities, arguably, machines should be able to learn language *faster* than humans
- Disentangled speaker and environmental embeddings would let us better generalize from a small number of examples

https://ojibwe.lib.umn.edu/

36



38

Summary

- The last decade has seen major advances in neural-network architectures, embedding spaces, and self-supervised learning that have spurred major advances in speech, NLP & machine vision
- Current methodology requires significantly more resources than humans which suggests significant improvements are possible
  - Making better use of available data for learning; few-shot learning
- Many exciting opportunities exist for further unsupervised research
  - Disentangled learned representations of speech and audio
  - Spoken language understanding
  - Joint learning with speech synthesis
  - Cross-language learning and speech translation
  - TBD!

39

Acknowledgements


Alex Park

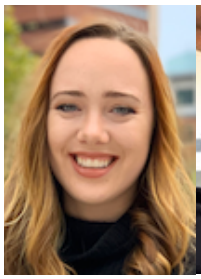

Yaodong Zhang


Chia-Ying Lee


Hao Tang


David Harwath

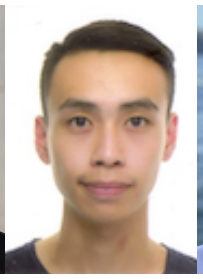

Wei-Ning Hsu


Angie Boggust


Yu-An Chung


Yuan Gong


Andrew Rouditchenko


Alex Liu


Jeff Lai

<u>Co-Authors</u>

SouYoung Jin

Aude Oliva

Adrià Recasens

Antonio Torralba

Kartik Audhkhasi

Brian Chen

Rogerio Feris

Dhiraj Joshi

Brian Kingsbury

Hilde Kuehne

Michael Picheny

Chris Song

Sam Thomas

40