Wav2vec: Learning the structure of speech from audio

Learning the structure of speech from raw audio

Ritwika Mukherjee, PhD

01 **Audio data**
The curse of high-dimensionality

02 **Self-supervised learning**
All you need is a good representation

03 **Wav2vec**
What does the model do?

04 **Wav2vec 2.0**
The evolution of Wav2vec

Visualizing audio

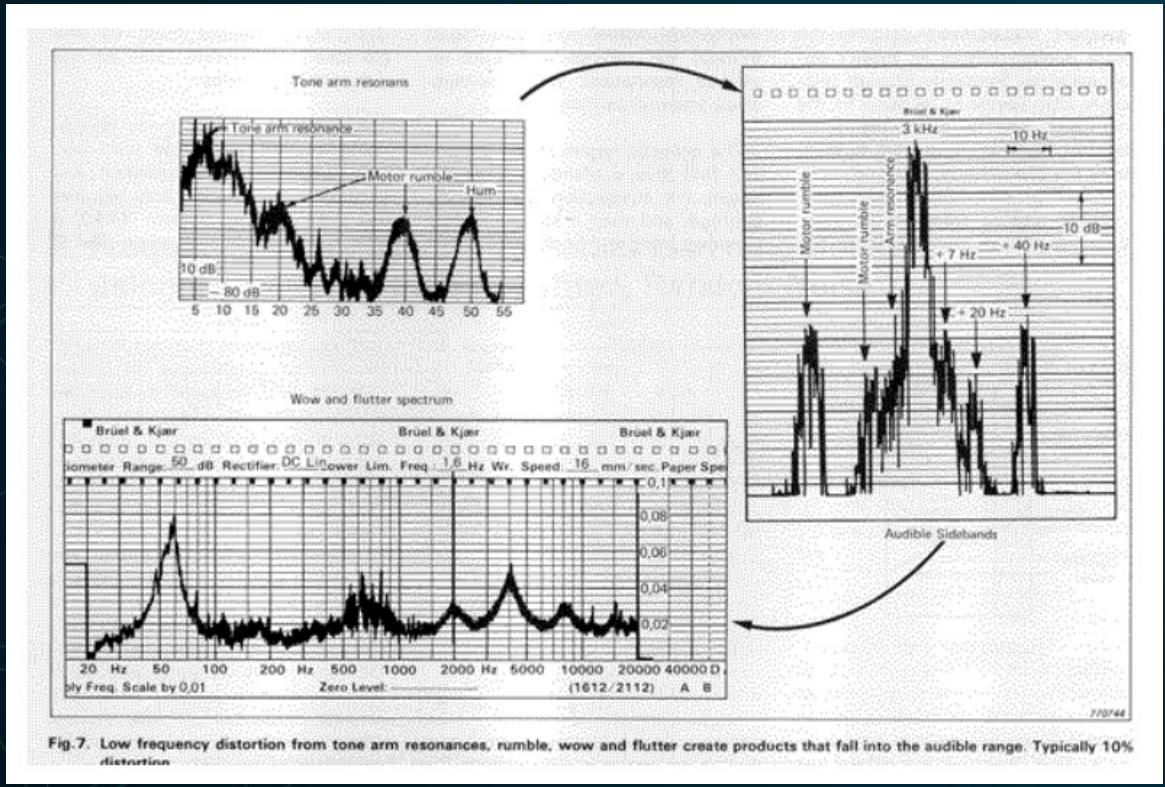


Fig.7. Low frequency distortion from tone arm resonances, rumble, wow and flutter create products that fall into the audible range. Typically 10% distortion

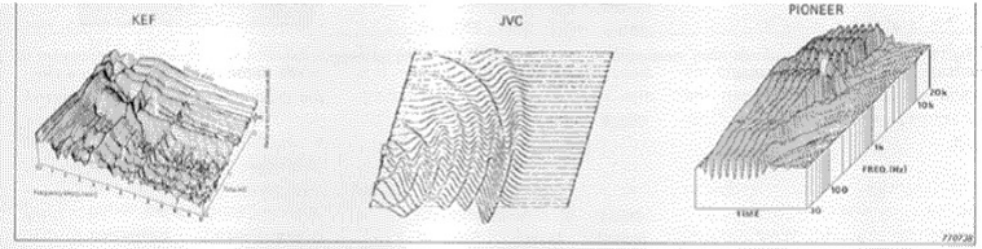


Fig.17. Examples of how the latest years' 3D plots have increased the subjective visibility

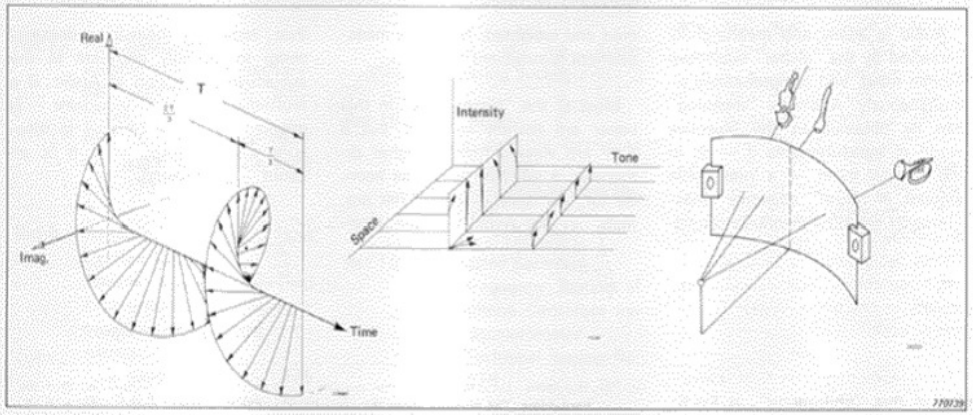


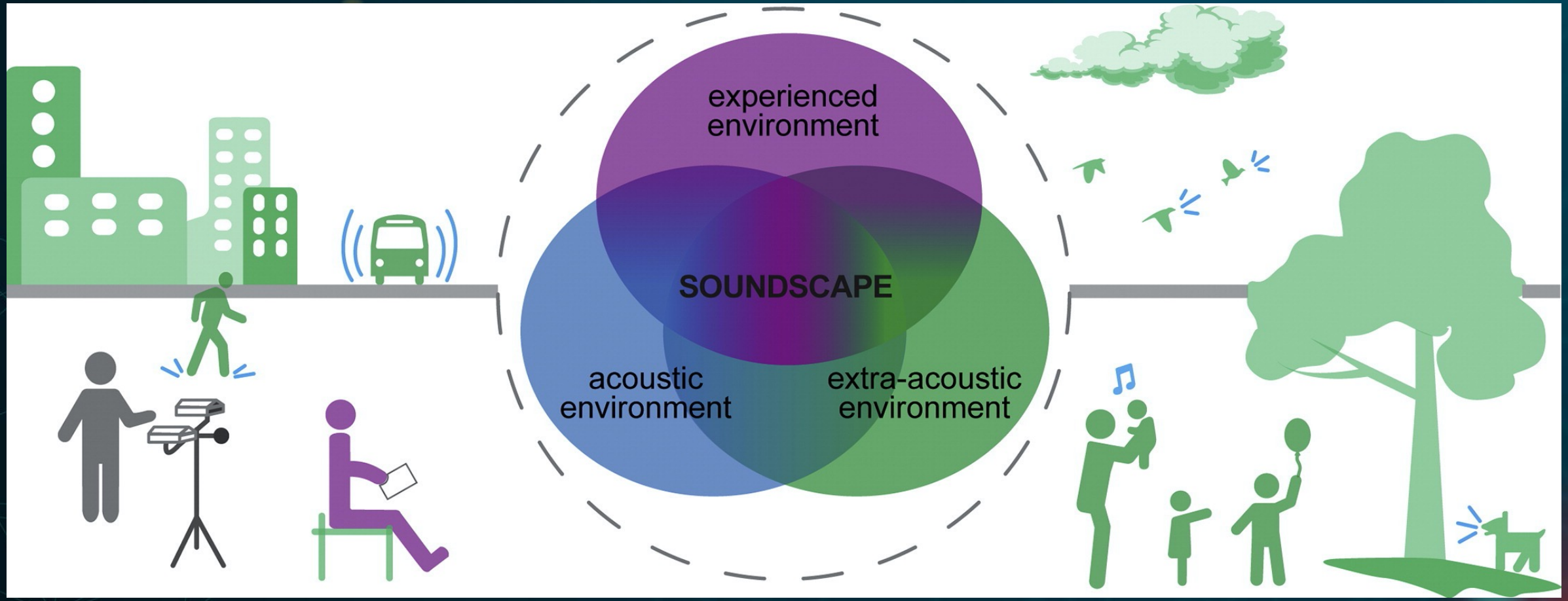
Fig.18. Examples of possible future 3D-plots

https://www.linkwitzlab.com/multidimensionalaudio.pdf

Audio is easily and meaningfully perceived by the "global" subjective human mind and comprehended simultaneously. A similar "meaning" can be obtained in the objective world of measurements if - as in the human mind - a reasonable amount of "local" objective measurements are simultaneously considered and weighted. No single measurement is sufficient. Most of the "dimensions" in this paper are literally very old, but if they are viewed from a higher dimension a meaning might be seen. So far, we have all been "flat animals" in the Audio domain. If a multidimensional viewpoint is adopted, we might be able to measure and interpret what it is all about - good sound.

From “Multidimensional Audio” by Henning Moller, Bruel, and Kjaer

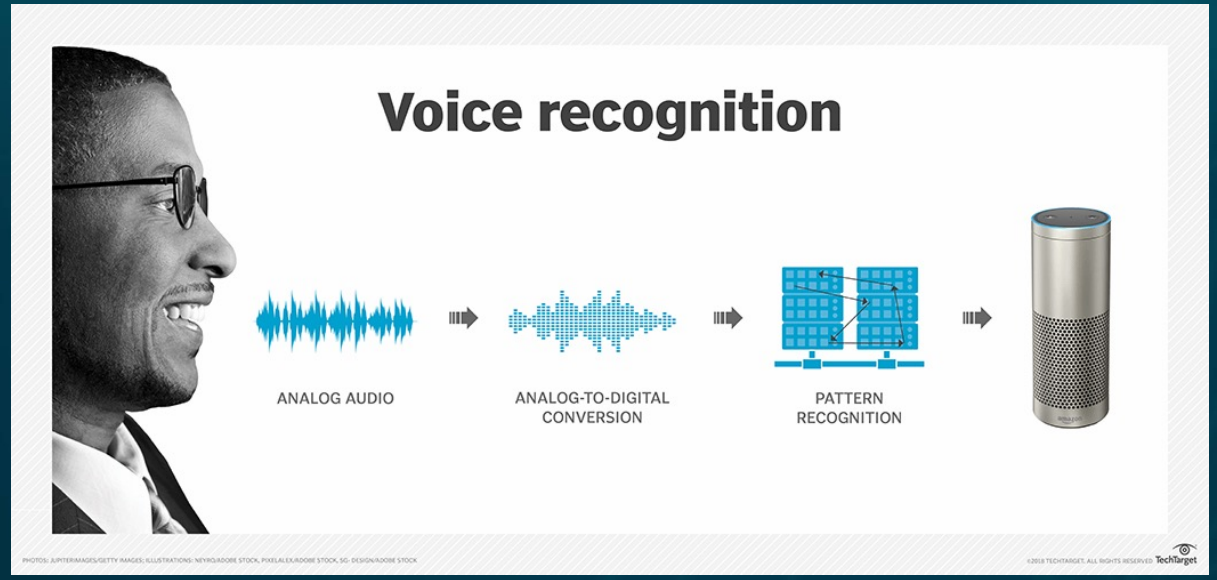
The curse of multi-dimensionality in audio data

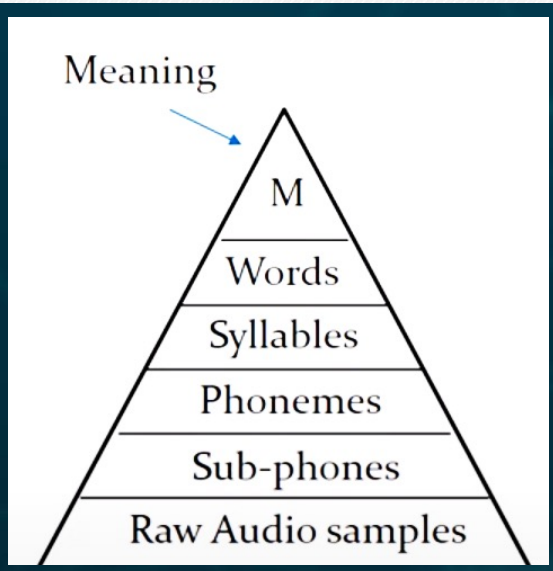


https://www.sciencedirect.com/science/article/abs/pii/S0048969716327486

Audio data, and in particular Speech

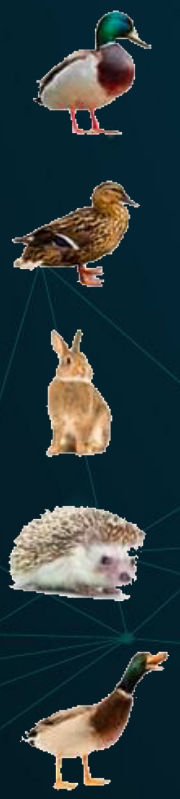
- Multidimensional time-series composition
- A continuous signal that captures many aspects of voice without clear segmentation of words or units
- It has highly variable lengths and a hierarchical structure

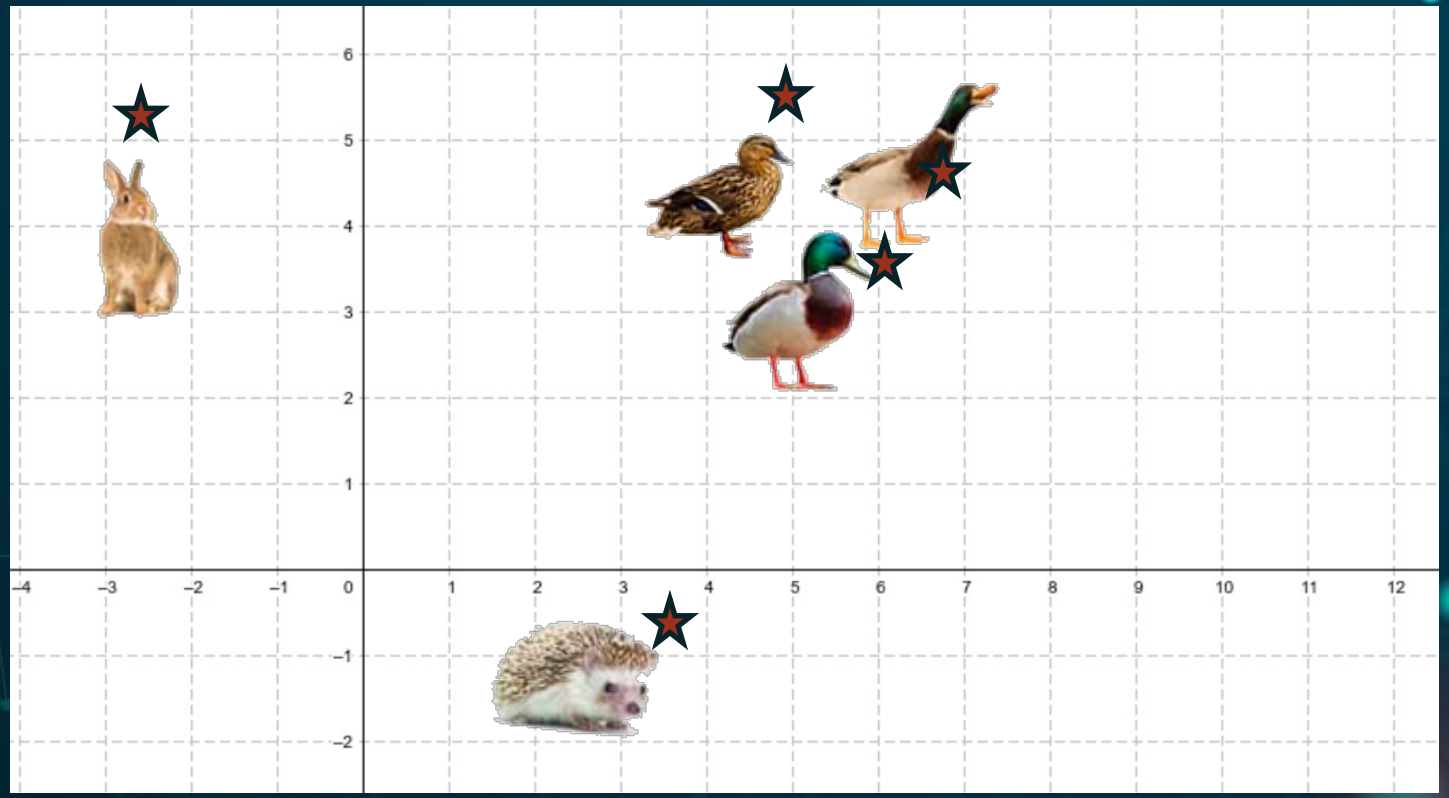




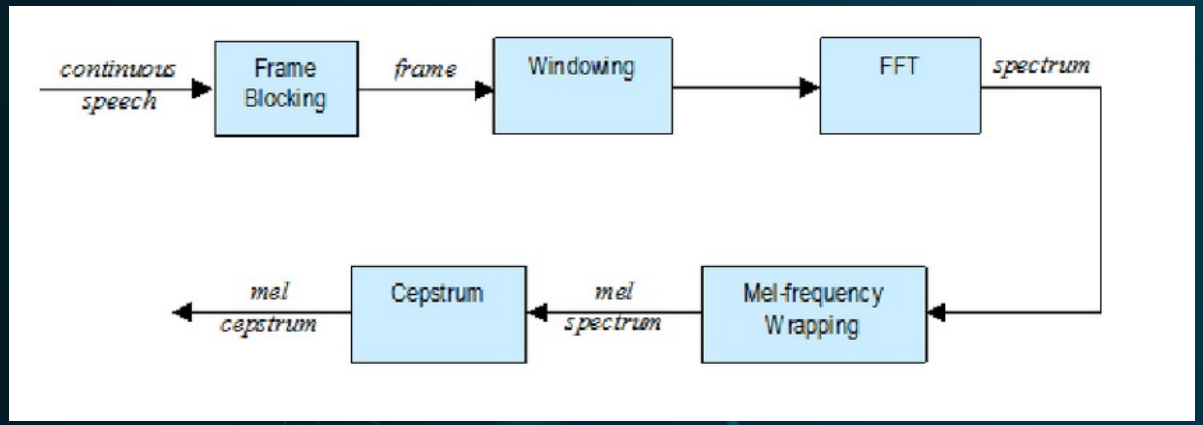
All you need is a good representation

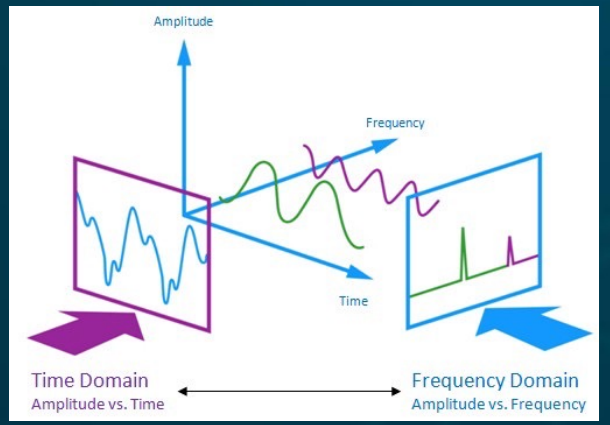
- Transform or extract machine-readable features from input data



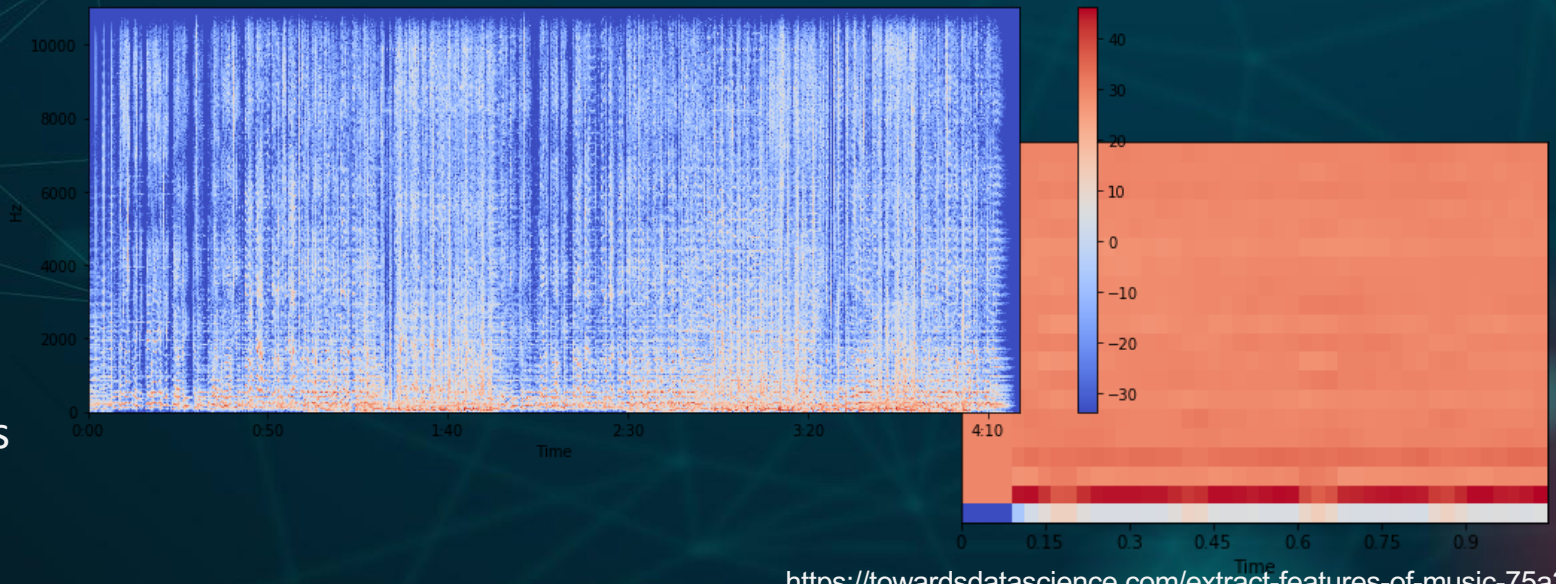


Features of audio data



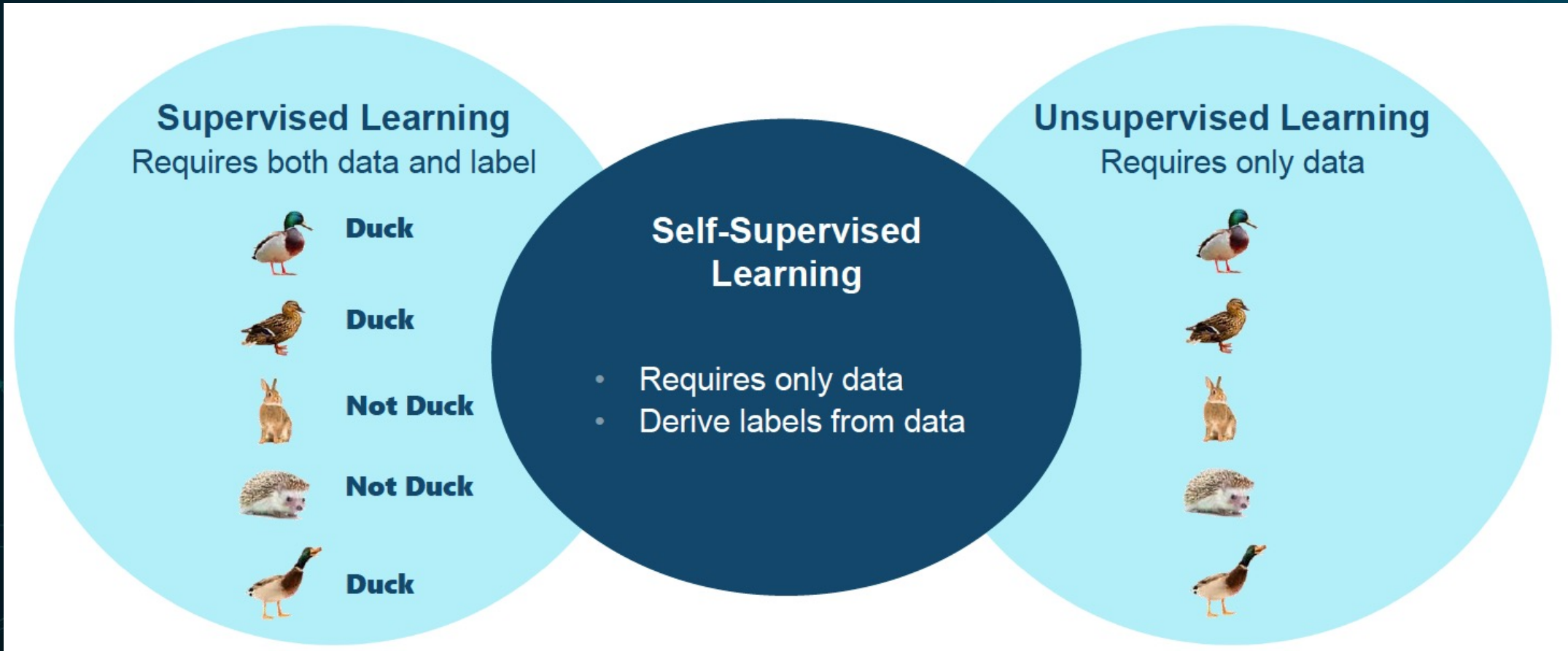


- **Time-series features**: zero crossing rate, energy, spectral centroid, spectral rolloff
- **Frequency features**: short-term fourier transform, mel-frequency cepstral coefficients



https://towardsdatascience.com/extract-features-of-music-75a3f9bc265d

Self-supervised learning



Self-supervised learning in speech recognition

- Traditional speech recognition models were primarily trained on annotated speech audio with transcriptions; large amounts of it are rare
- Self-supervision provides a way to leverage unannotated data to build better systems
- Good representation learning requires features that are less specialized towards solving a single supervised task

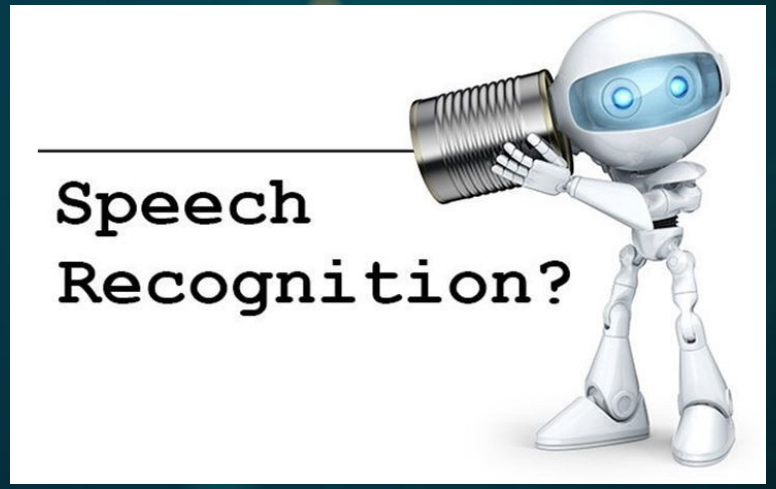
Facebook AI's Wav2vec

WAV2VEC: UNSUPERVISED PRE-TRAINING FOR SPEECH RECOGNITION

Steffen Schneider, Alexei Baevski, Ronan Collobert, Michael Auli
Facebook AI Research

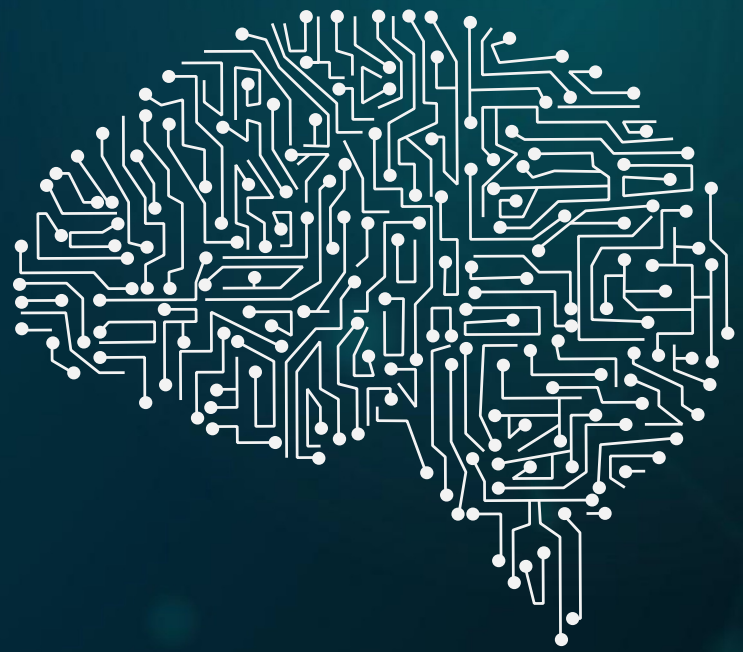
ABSTRACT

We explore unsupervised pre-training for speech recognition by learning representations of raw audio. wav2vec is trained on large amounts of unlabeled audio data and the resulting representations are then used to improve acoustic model training. We pre-train a simple multi-layer convolutional neural network optimized via a noise contrastive binary classification task. Our experiments on WSJ reduce WER of a strong character-based log-mel filterbank baseline by up to 36 % when only a few hours of transcribed data is available. Our approach achieves 2.43 % WER on the nov92 test set. This outperforms Deep Speech 2, the best reported character-based system in the literature while using two orders of magnitude less labeled training data.[1]



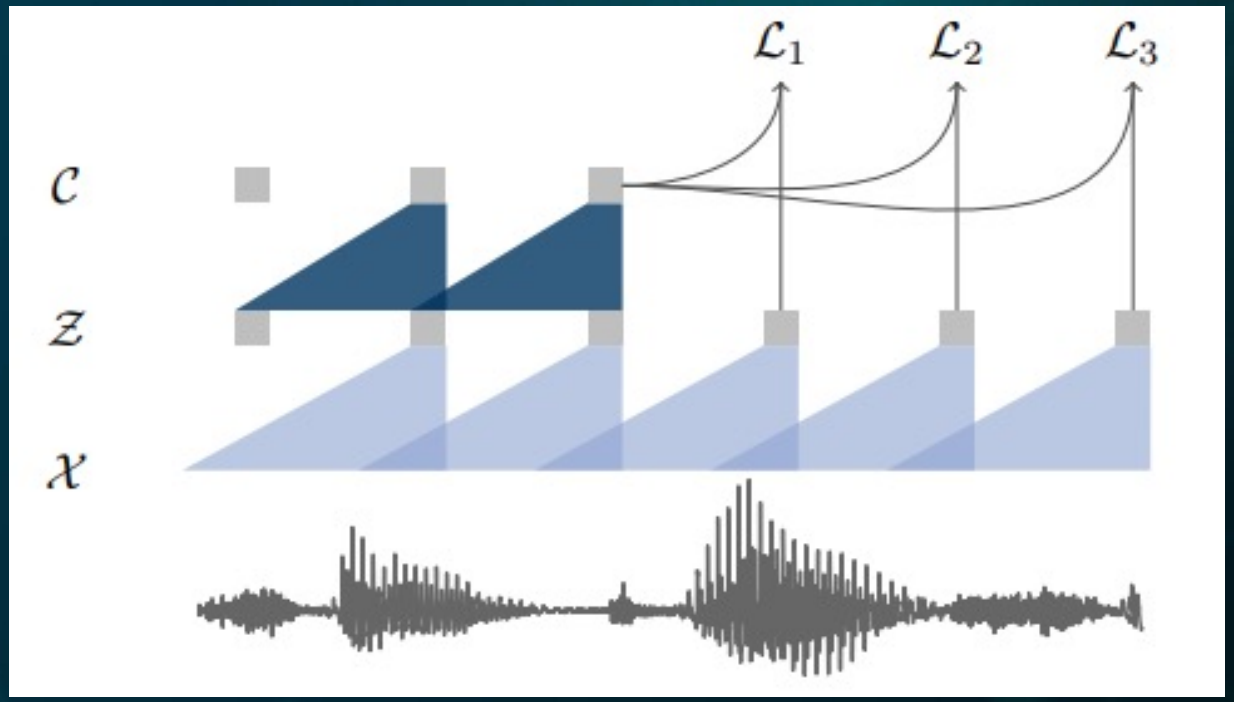
Wav2vec: Speech recognition through self-supervision

- A new self-supervised approach that beats traditional ASR systems that rely solely on transcribed audio
- A more accurate and versatile approach for transcribing proper names and other words that are outside of ASR systems' lexicons
- Superior to past models (Deep Speech 2) because they learn meaningful high-level representations through contrastive predictive coding



Wav2vec architecture

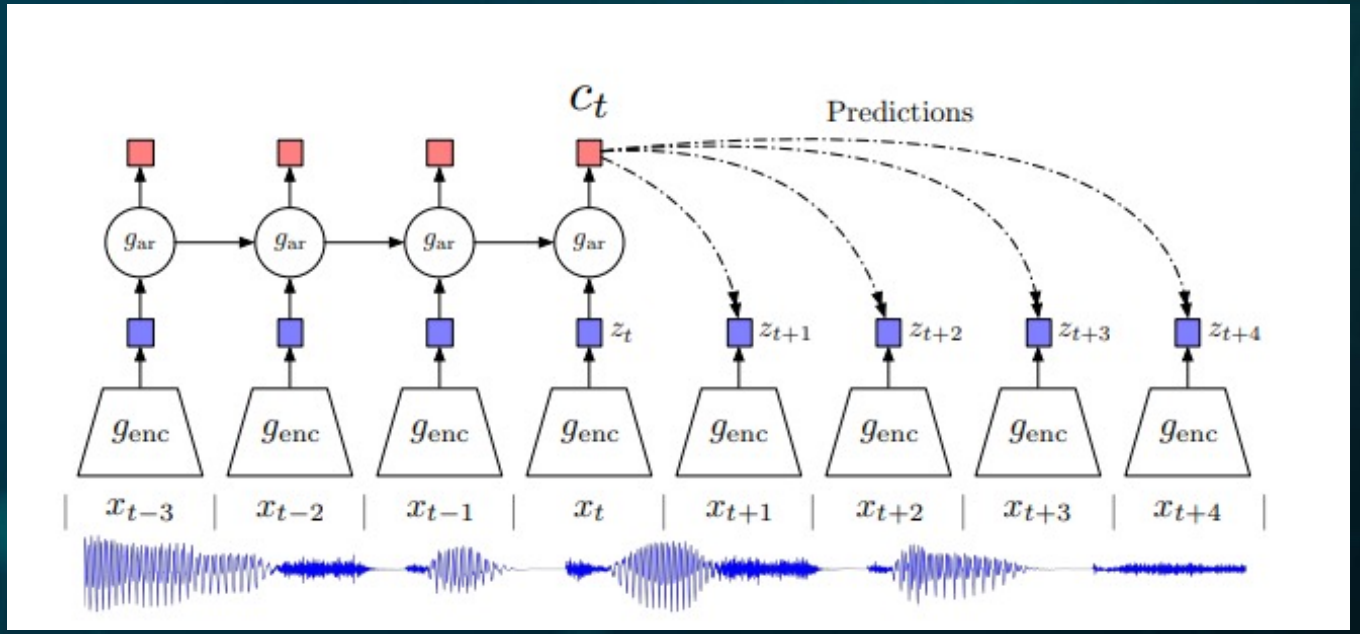
- Two multilayer convolutional neural networks stacked on top of each other
- The encoder network maps raw audio input to a representation, where each vector covers about 30 ms of speech
- The context network uses those vectors to generate its own representations covering a larger span
- The model then uses these representations to solve a self-supervised prediction task



Contrastive predictive coding

A universal unsupervised learning approach to extract useful representations from high-dimensional data

- Learn representations by predicting the future in latent space by using powerful autoregressive models
- Employ probabilistic contrastive loss which induces the latent space to capture information that is maximally useful to predict future samples
- Made tractable by using negative sampling

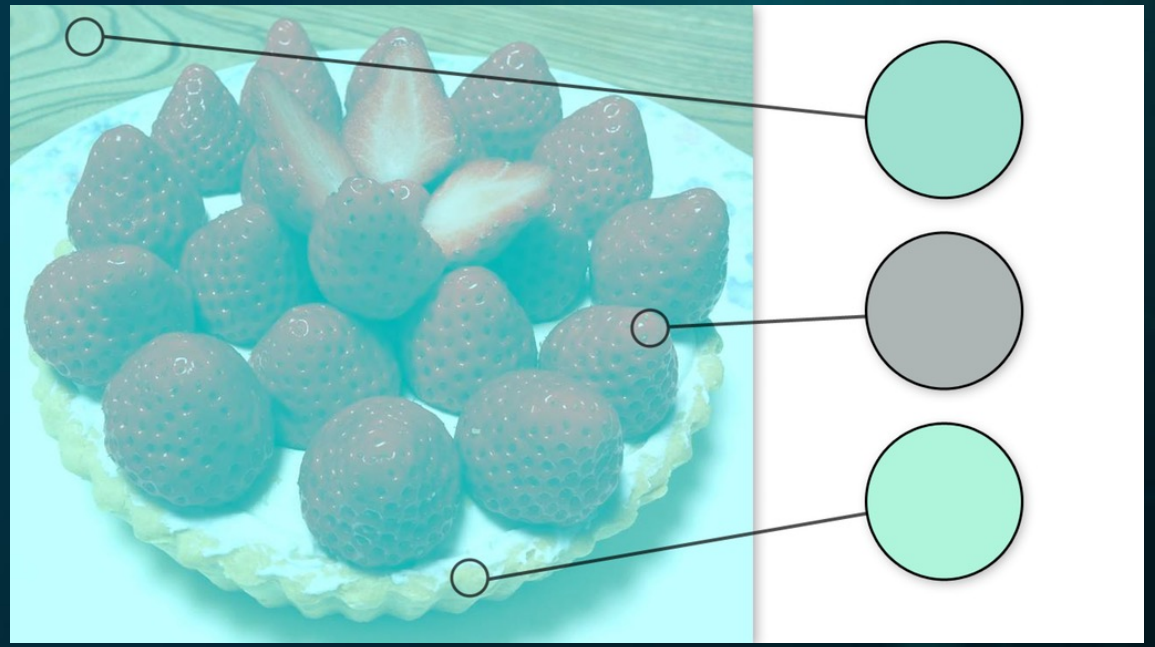


Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." arXiv:1807.03748 (2018)

Self-supervised learning – Hierarchical predictive coding

Theorized in neuroscience for the brain to constantly update mental models of the environment – update and revise based on predictions being made about the environment

- Self-supervision occurs by extracting meaning from the signal itself -- apply known transforms to the input data and use resulting outcomes as targets
- Predictive coding is an old technique used in signal processing for data compression
- E.g., learning word representations to predict neighboring words, or for images, predict color from gray-scale

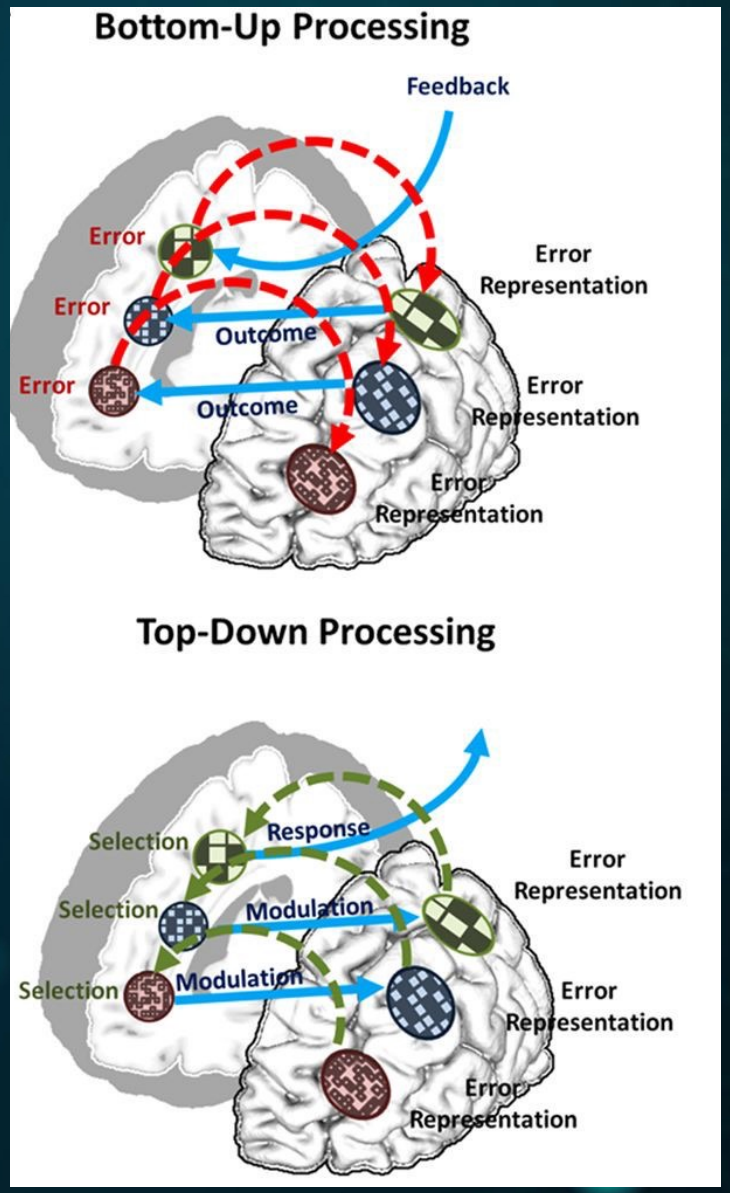


Hierarchical Predictive coding

Update and revise based on predictions being made about the environment

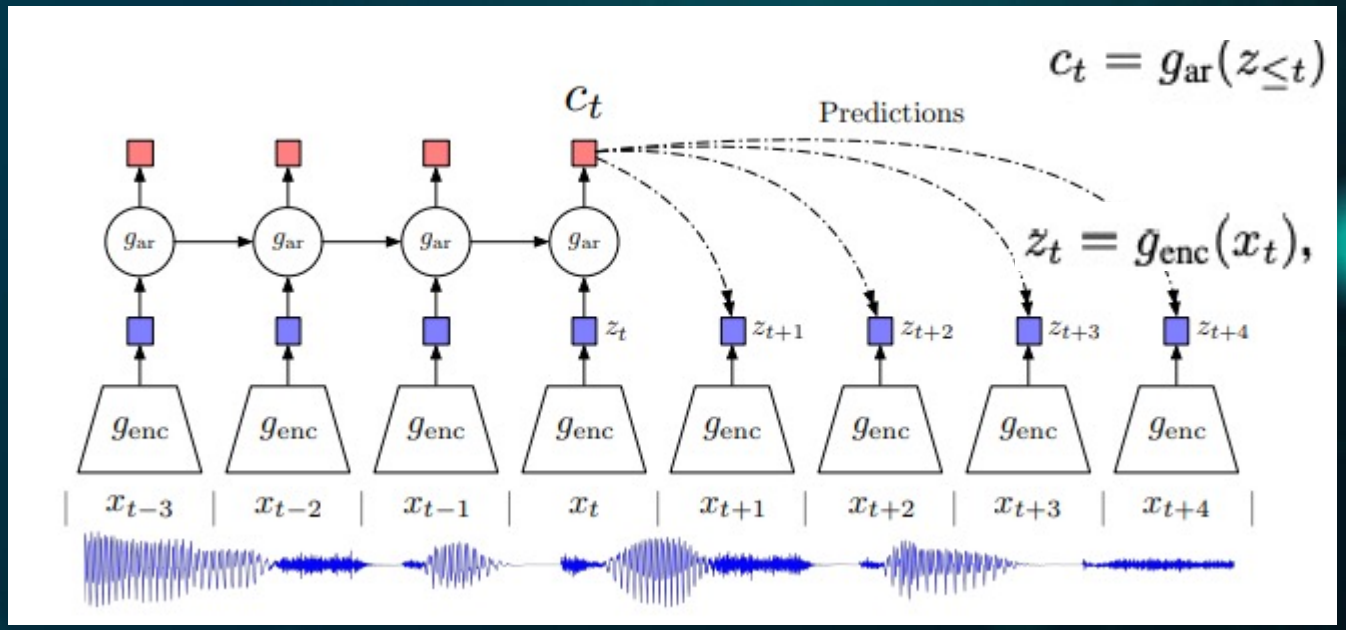
- Uses multiple layers of predictive assumptive models
- Constructs plausible versions of sensory input data for itself using what was previously learned about the probabilistic outcome patterns

The mind works more like a computer graphic program rather than a standard pattern recognition model or classic classifier



Contrastive predictive coding architecture

- Predict many steps into the future using autoregressive models
- Use noise-contrastive estimation to minimize loss such that the predicted samples are well separated from “negative” or “distractor” samples and are closest to the true samples
- Applications in speech, images, natural language and reinforcement learning



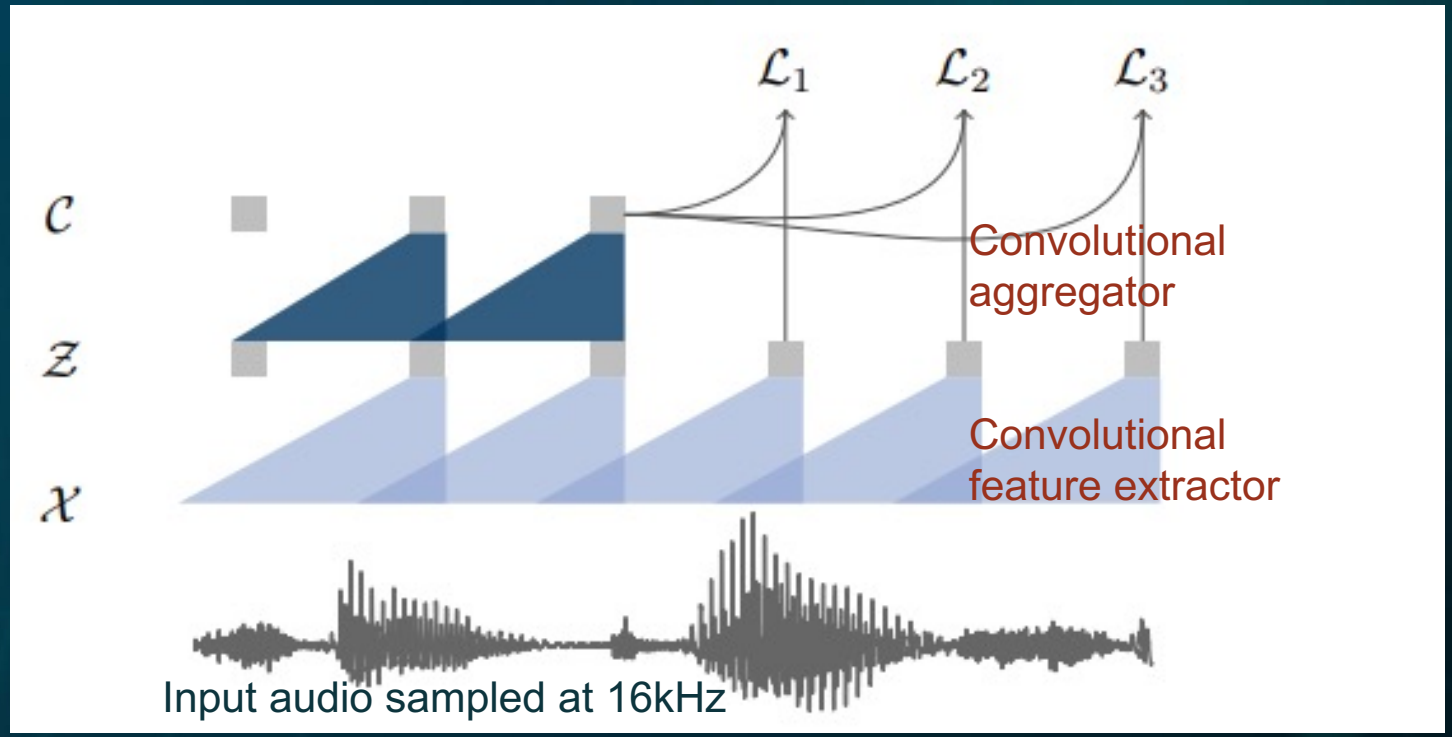
Paper: https://arxiv.org/pdf/1807.03748.pdf

Wav2vec architecture

Optimize loss to obtain context representations of audio

Produces a context tensor after convolutions of embeddings

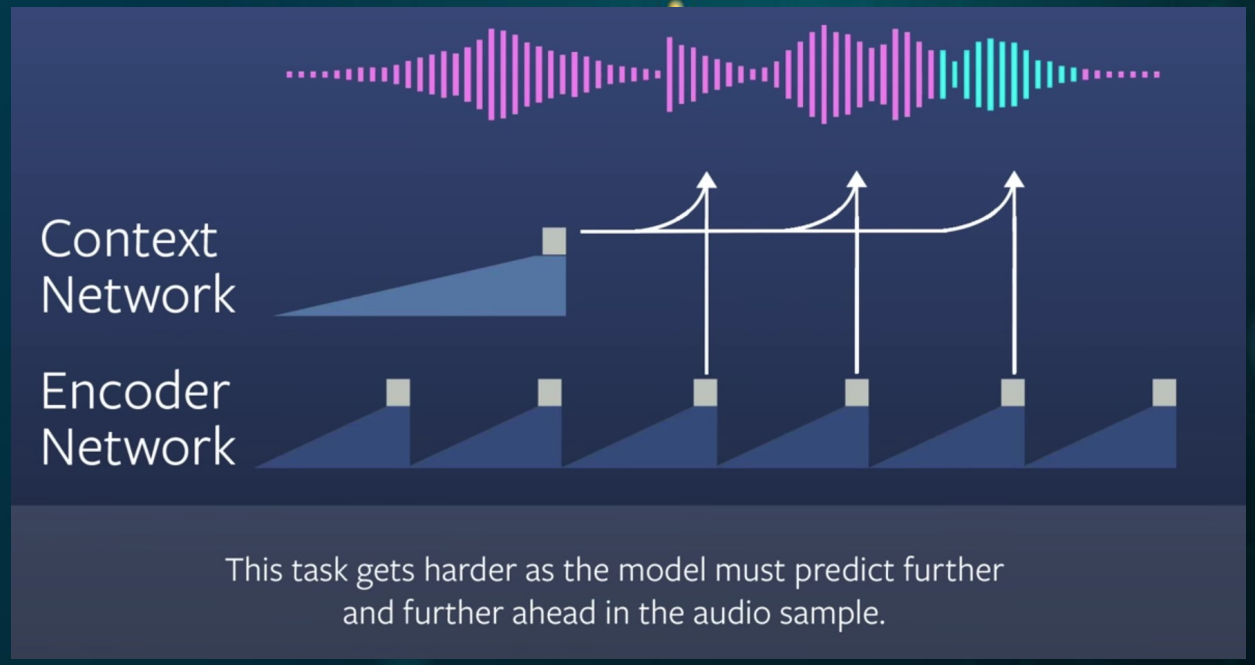
Low-freq feature embedding in latent space
(encodes every 32 ms with window stride of 10 ms)



Schneider, Steffen, et al. "wav2vec: Unsupervised pre-training for speech recognition." arXiv:1904.05862 (2019)

Wav2vec for character-level predictions

Wav2vec for character-level predictions



Wav2vec: Two multilayer convolutional neural networks stacked on top of each other

$\mathcal{Z}$ Space (Encoder)

covers 30ms Segments (16Khz Audio)

5 Layer CNN
Kernel sizes (10,8,4,4,4)
Stride (5,4,2,2,2)
Channels (512)
Group Norm + ReLU nonlinearity

C Space (Context)

210ms receptive field

9 Layer CNN
Kernel size (3 for all)
Stride (1)
Channels (512)
Group Norm + ReLU nonlinearity

Paper: https://arxiv.org/pdf/1904.05862.pdf?
Model: https://github.com/pytorch/fairseq/tree/master/examples/wav2vec

Contrastive predictive coding enhances mutual information of original signal with context embeddings

- Extract the underlying latent variables the inputs have in common and slow features, which maximize the mutual information of observations over long time horizons
- Model a density ratio which preserves the mutual information between x_{t+k} and c_t

Density ration

$$f_k(x_{t+k}, c_t) \propto \frac{p(x_{t+k}|c_t)}{p(x_{t+k})}$$

context latent representation

Applying a log-bilinear model where a linear transformation is used for a prediction with different Wk,

$$f_k(x_{t+k}, c_t) = \exp\left(z_{t+k}^T W_k c_t\right)$$

Paper: https://arxiv.org/pdf/1807.03748.pdf

Optimizing loss

Maximize separation of the predicted embedding (closest to the true value) from random samples

InfoNCE Loss

$$\mathcal{L}_{\mathrm{N}} = -\underset{X}{\mathbb{E}}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$

HIGH

LOW

where,

$$f_k(x_{t+k}, c_t) = \exp\left(z_{t+k}^T W_k c_t\right),$$

Interpretation: The network learns a representation which improves the probability of predicting the ‘true’ segment from the ‘distractor’ segments.

Optimizing loss

Contrastive loss takes the output of the network for a positive example and calculates its distance to an example of the same class and contrasts that with the distance to negative examples

Context latent representation(weighted sum of the previously aggregated features)

Predicted embedding k steps ahead

Distractor/random embeddings drawn randomly

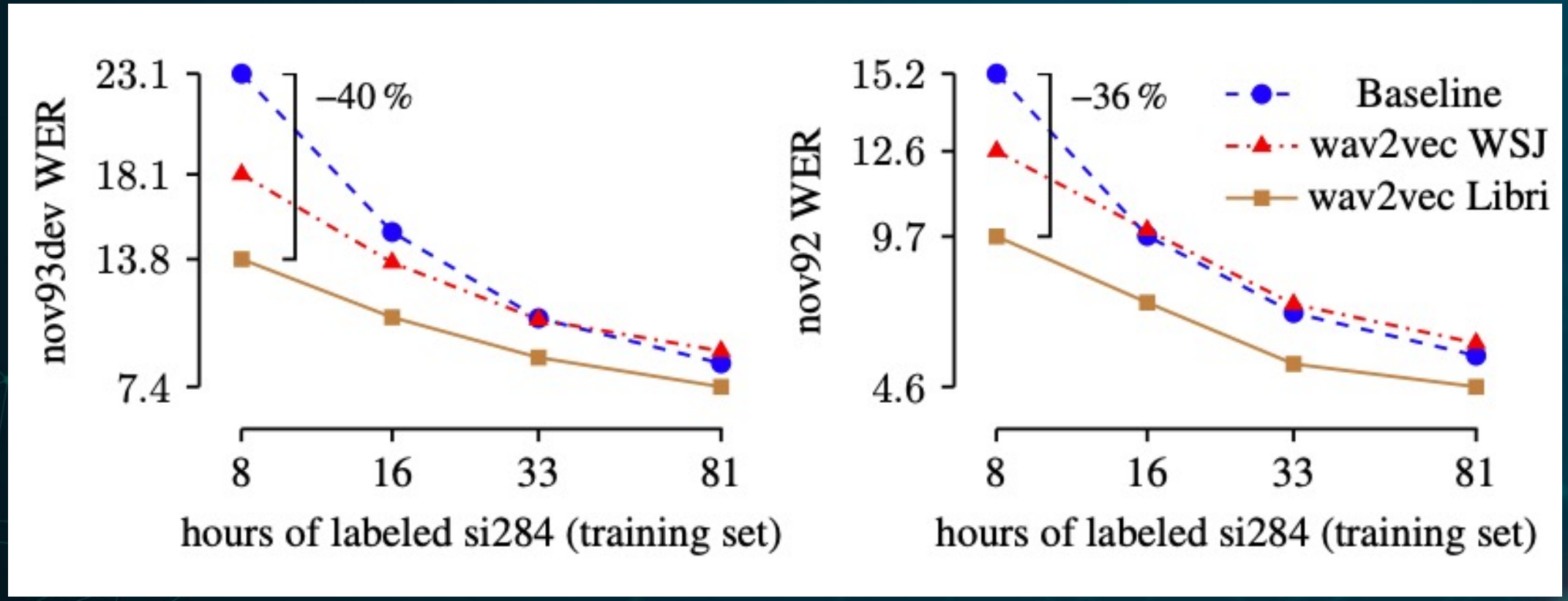
$$\mathcal{L}_k = -\sum_{i=1}^{T-k}\left(\log\sigma(\mathbf{z}_{i+k}^{\top}h_k(\mathbf{c}_i)) + \lambda\,\mathbb{E}_{\tilde{\mathbf{z}}\sim p_n}\left[\log\sigma(-\tilde{\mathbf{z}}^{\top}h_k(\mathbf{c}_i))\right]\right)$$

Density ration which preserves mutual information between true future samples and context

—

Density ration which preserves mutual information between “distractor” samples and context

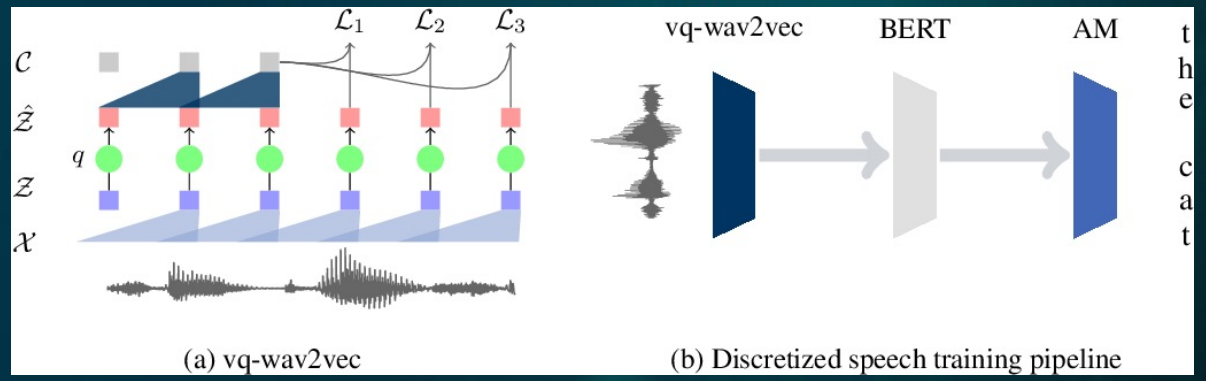
Wav2vec performance



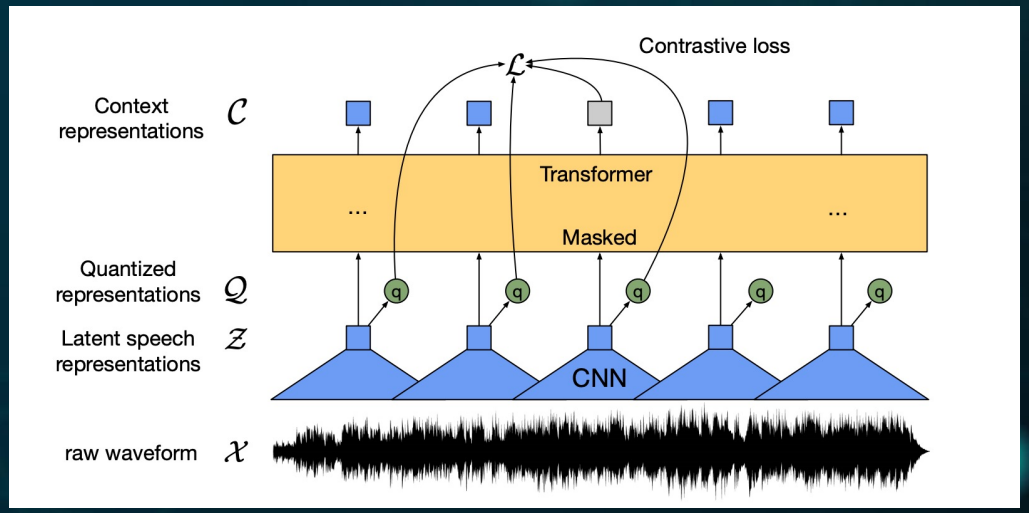
Evolution of Wav2vec

Update and revise based on predictions being made about the environment

- Vector Quantized Wav2vec (vq-wav2vec): Learning “discrete” speech representations (allows transfer learning from NLP models)



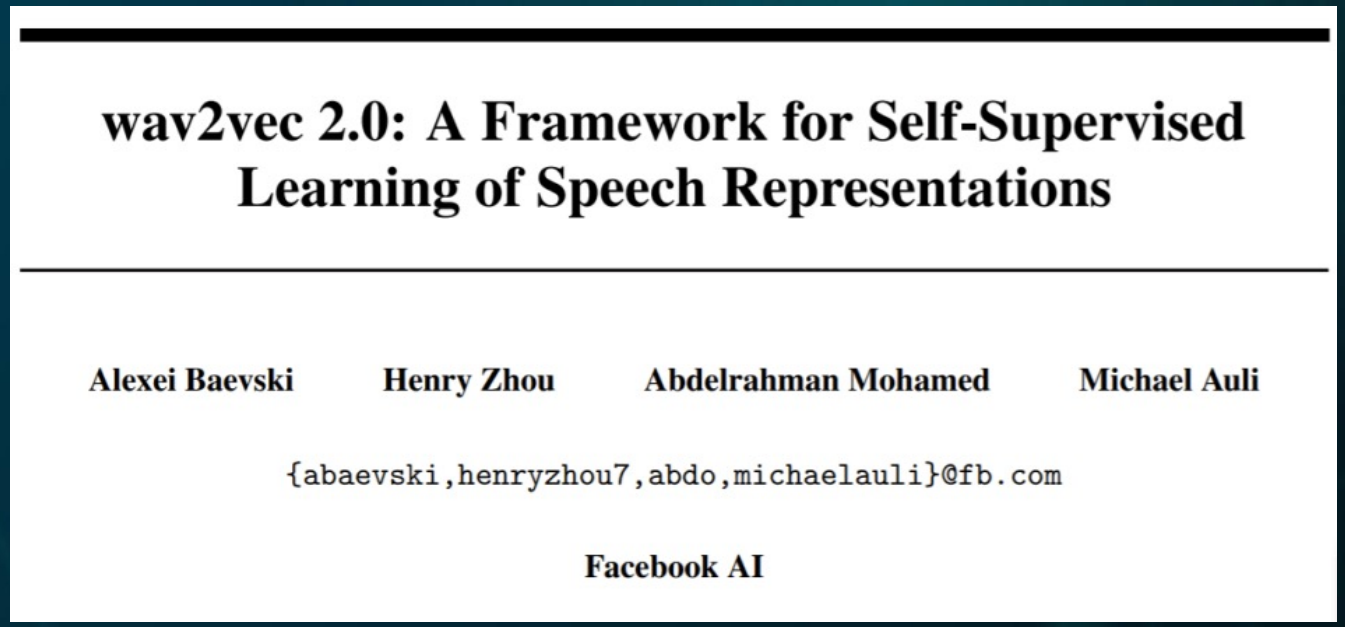
- Wav2vec 2.0: which is essentially Wav2vec with transformers



Papers: https://arxiv.org/abs/2006.11477
https://arxiv.org/abs/1910.05453

Wav2vec 2.0: An end-to-end version of vq-wav2vec

- Trained to predict correct speech units for masked parts of the audio while learning future speech units
- The model pre-trained on LibriVox (LV-60k) and fine-tuned on only 10 minutes of labeled data achieves a word error rate of 5.2/8.6 on the Librispeech clean/other test sets
- Implications in speech recognition models in many languages, dialects, and domains that previously required much more transcribed audio data



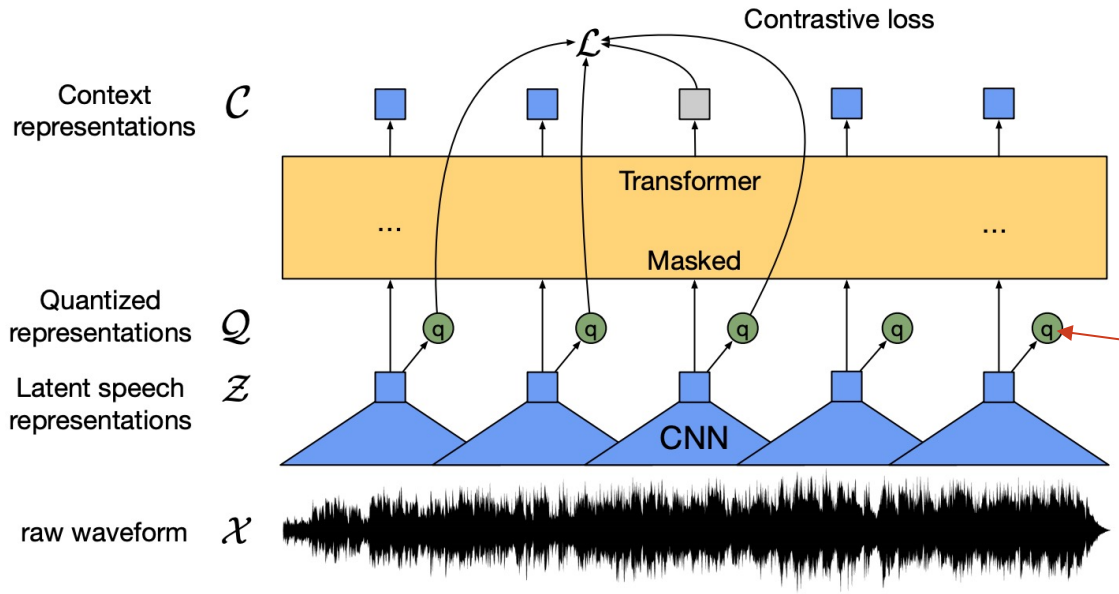
wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations

Alexei Baevski **Henry Zhou** **Abdelrahman Mohamed** **Michael Auli**

{abaevski,henryzhou7,abdo,michaelauli}@fb.com

Facebook AI

Wav2vec 2.0 architecture



The quantizer chooses a speech unit for the latent audio representation from an inventory of learned units

Figure 1: Illustration of our framework which jointly learns contextualized speech representations and an inventory of discretized speech units.

Current model performances

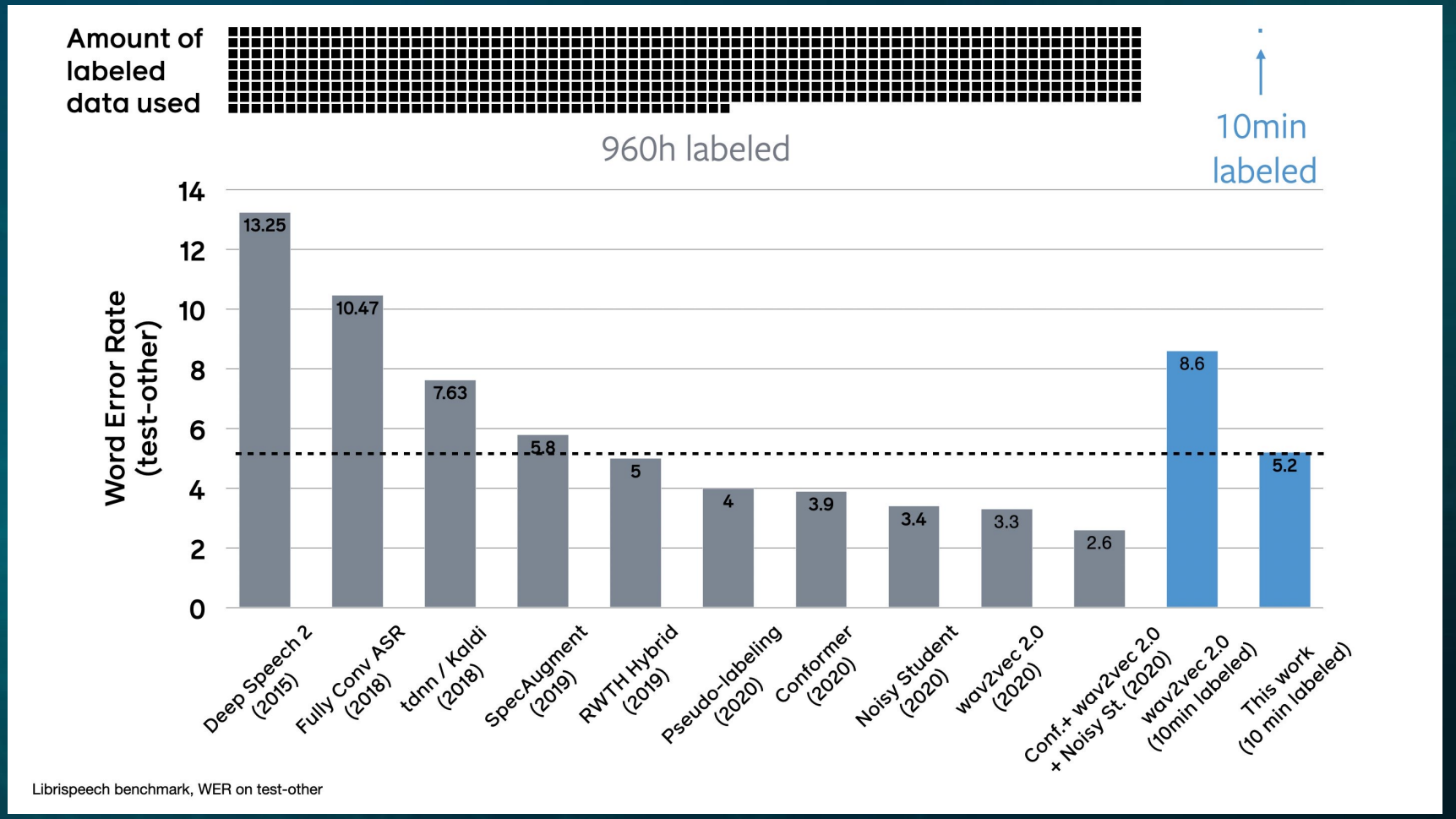
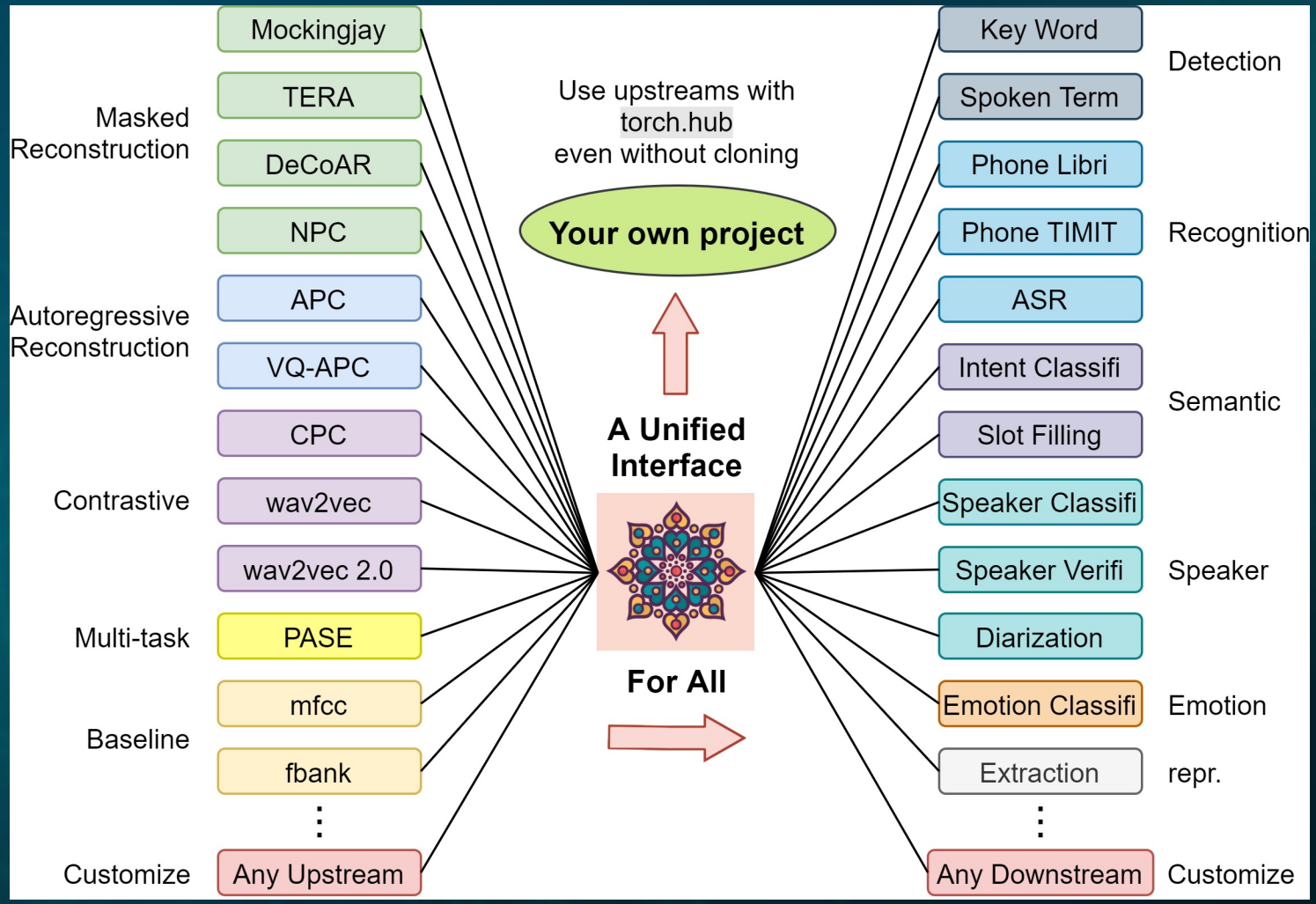


Image from Michael Auli's Tweet
Paper: https://arxiv.org/abs/2010.11430
Model: https://github.com/pytorch/fairseq/tree/master/examples/wav2vec

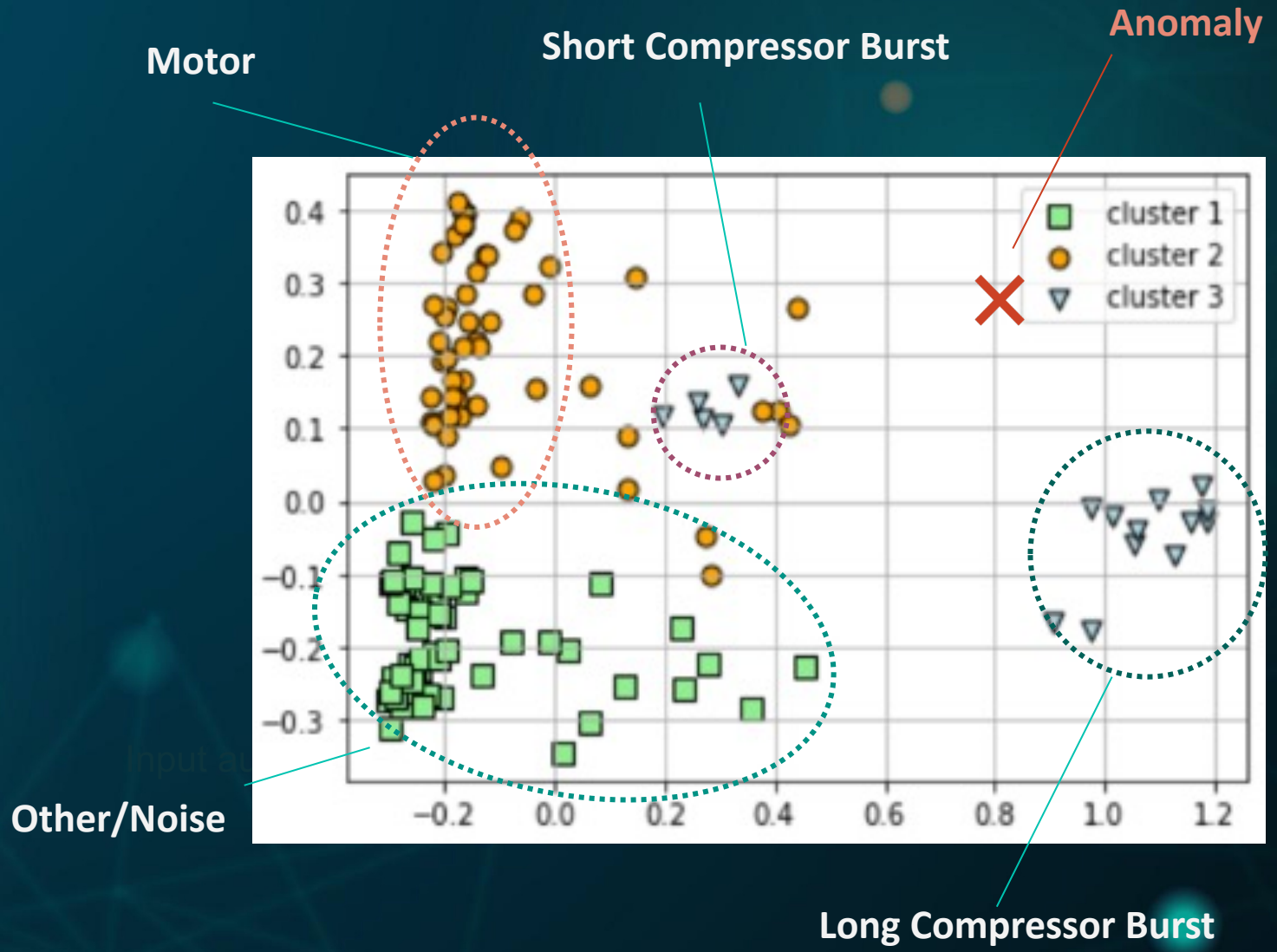
Other model architectures and self-supervised speech toolkit



https://github.com/s3prl/s3prl

Wav2vec features of machine acoustics: Latent space visualization

- The latent space represents some unknown encoded features
- Visualize using a PCA to further reduce the data to two dimensions
- Even 2 dimensions provides interesting separation
- Points further away could be anomalies



Example usage

Example usage:

```
import torch
import fairseq

cp_path = '/path/to/wav2vec.pt'
model, cfg, task = fairseq.checkpoint_utils.load_model_ensemble_and_task([cp_path])
model = model[0]
model.eval()

wav_input_16khz = torch.randn(1,10000)
z = model.feature_extractor(wav_input_16khz)
c = model.feature_aggregator(z)
```

https://github.com/pytorch/fairseq/tree/master/examples/wav2vec#pre-trained-models-1

Wav2vec: Learning the structure of speech from audio

Learning the structure of speech from raw audio

Ritwika Mukherjee, PhD