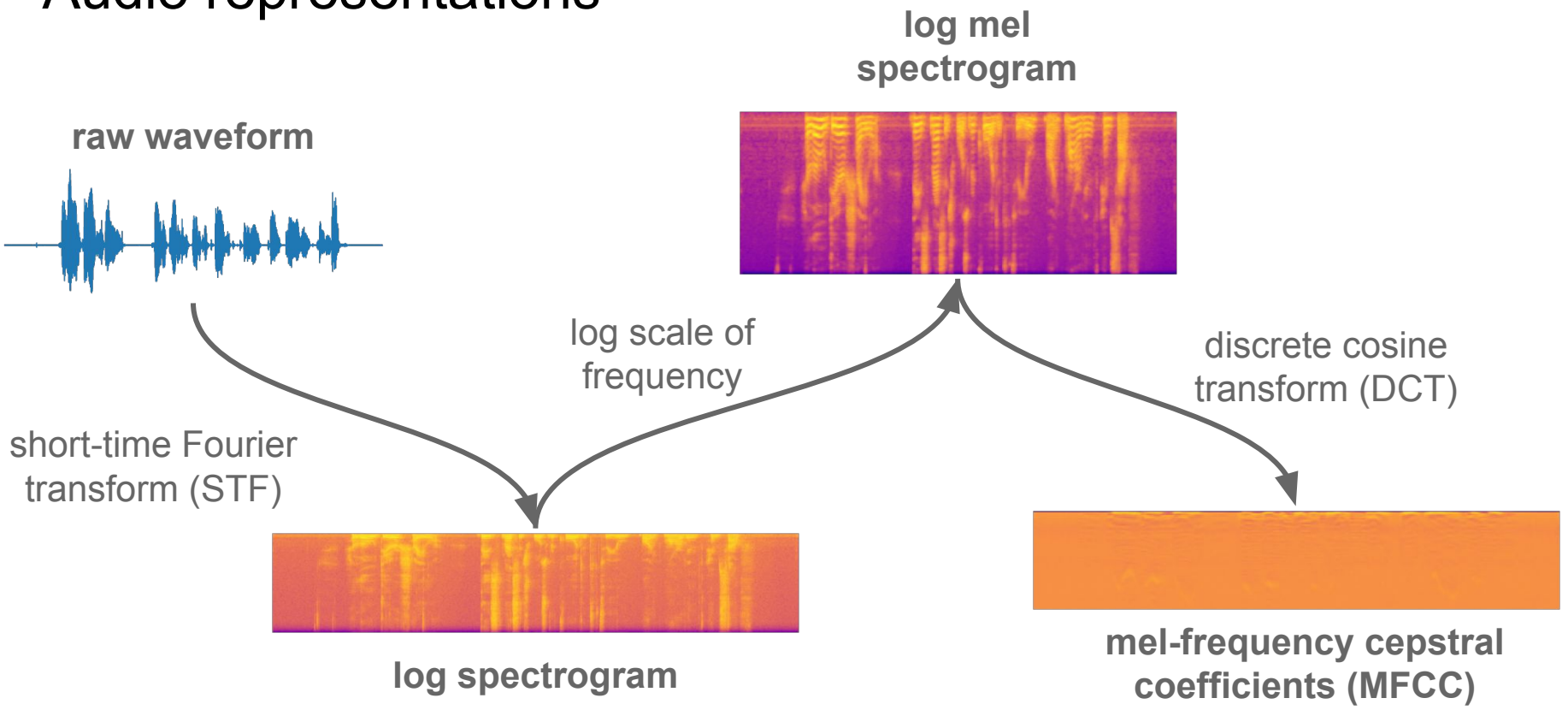
# Self-supervised learning for audio

Ildus Sadrtdinov, 07.03.23

# What makes audio different?

- Audio signal is continuous (same as images)
- Audio signal is sequential (same as texts)
- Audio signal has high-frequency (even more than images)
- There are many informative feature representations available for audio

# Audio representations



# Approaches to self-supervised learning

- Autoregressive (GPT-family)
- Contrastive (SimCLR, BYOL, CLIP)
- Masked modelling (*BERT, MAE)

# SSL methods for audio

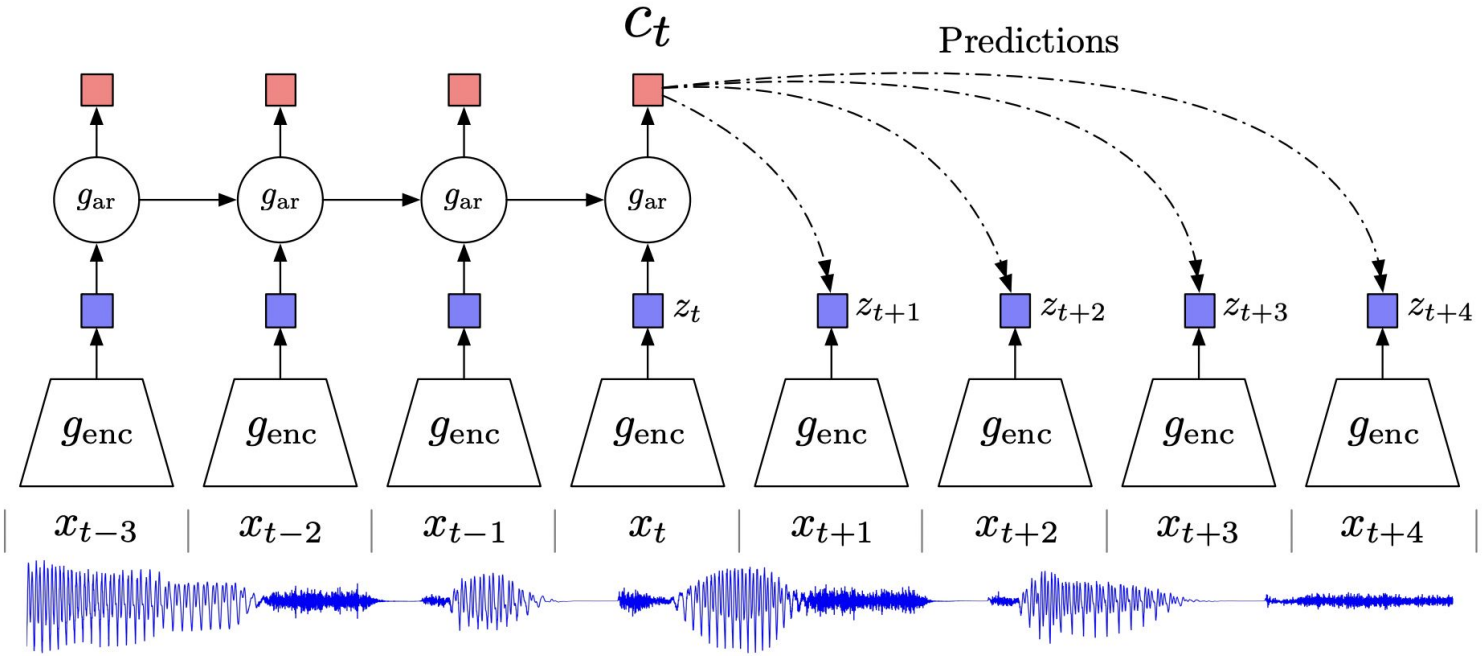
| Model | Speech | Input format | Framework | Encoder | Loss | Inspired by |
|---|---|---|---|---|---|---|
| **LIM [36]** | ✓ | raw waveform | (d) | SincNet | BCE, MINE or NCE loss | SimCLR |
| **COLA [36]** | ✗ | log mel-filterbanks | (d) | EfficientNet | InfoNCE loss | SimCLR |
| **CLAR [33]** (semi) | ✗ | raw waveform<br>log mel-spectrogram | (d) | 1D ResNet-18<br>ResNet-18 | NT-Xent<br>+ cross-entropy | SimCLR |
| **Fonseca et al. [36]** | ✗ | log mel-spectrogram | (d) | ResNet, VGG, CRNN | NT-Xent loss | SimCLR |
| **Wang et al. [88]** | ✗ | raw waveform<br>+ log mel-filterbanks | (d) | CNN<br>ResNet | NT-Xent loss<br>+ cross-entropy | SimCLR |
| **BYOL-A [89]** | ✗ | log mel-filterbanks | (b) | CNN | MSE loss | BYOL |
| **Speech2Vec [48]** | ✓ | mel-spectrogram | (a) | RNN | MSE loss | Word2Vec |
| **Audio2Vec [91]** | ✓✗ | MFCCs | (a) | CNN | MSE loss | Word2Vec |
| **Carr [67]** | ✓ | MFCCs | (a) | Context-free network | Fenchel-Young loss | - |
| **Ryan [68]** | ✗ | constant-Q transform<br>spectrogram | (a) | AlexNet | Triplet loss | -<br>- |
| **Mockingjay [92]** | ✓ | mel-spectrogram | (a) | Transformer | L1 loss | BERT |
| **TERA [93]** | ✓ | log mel-spectrogram | (a) | Transformer | L1 loss | BERT |
| **Audio ALBERT [94]** | ✓ | log mel-spectrogram | (a) | Transformer | L1 loss | BERT |
| **DAPC [95]** | ✓ | spectrogram | (a) | Transformer | Modified MSE loss<br>+ orthogonality penalty | BERT |
| **PASE [96]** | ✓ | raw waveform | (a) | SincNet + CNN | L1, BCE loss | BERT |
| **PASE+ [97]** | ✓ | raw waveform | (a) | SincNet + CNN + QRNN | MSE, BCE loss | BERT |
| **CPC [40]** | ✓ | raw waveform | (a) | ResNet + GRU | InfoNCE loss | - |
| **CPC v2 [59]** | ✓ | raw waveform | (a) | ResNet + Masked CNN | InfoNCE loss | - |
| **CPC2 [98]** | ✓ | raw waveform | (a) | ResNet + LSTM | InfoNCE loss | - |
| **Wav2Vec [84]** | ✓ | raw waveform | (a) | 1D CNN | Contrastive loss | - |
| **VQ-Wav2Vec [85]** | ✓ | raw waveform | (a) | 1D CNN + BERT | Contrastive loss | BERT |
| **Wav2Vec 2.0 [81]** | ✓ | raw waveform | (a) | 1D CNN + Transformer | Contrastive loss | BERT |
| **HuBERT [99]** | ✓ | raw waveform | (c) | 1D CNN + Transformer | Contrastive loss | BERT |

# Plan

- Contrastive Predictive Coding (CPC)
- Wav2Vec 2.0
- HUBERT
- Multi-format contrastive learning
- BYOL-A (BYOL for audio)

van der Oord et al., 2018

# Contrastive Predictive Coding (CPC)




$$f_k(x_{t+k}, c_t) \propto \frac{p(x_{t+k}|c_t)}{p(x_{t+k})} \qquad f_k(x_{t+k}, c_t) = \exp\left(z_{t+k}^T W_k c_t\right) \qquad \mathcal{L}_N = -\mathop{\mathbb{E}}_X \left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$

van der Oord et al., 2018

# Contrastive Predictive Coding (CPC)

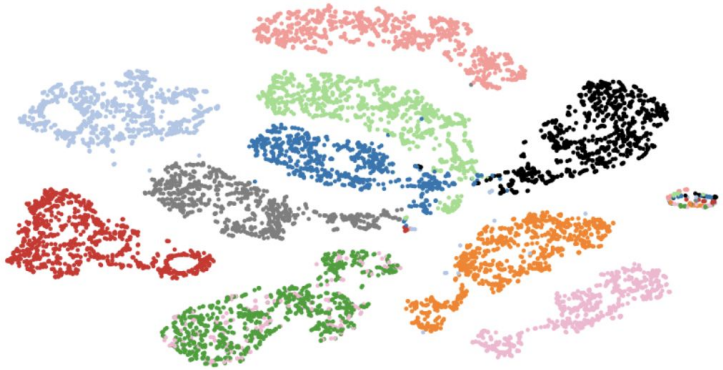


Figure 2: t-SNE visualization of audio (speech) representations for a subset of 10 speakers (out of 251). Every color represents a different speaker.

| Method | ACC |
|---|---|
| **Phone classification** | |
| Random initialization | 27.6 |
| MFCC features | 39.7 |
| CPC | 64.6 |
| Supervised | 74.6 |
| **Speaker classification** | |
| Random initialization | 1.87 |
| MFCC features | 17.6 |
| CPC | 97.4 |
| Supervised | 98.5 |

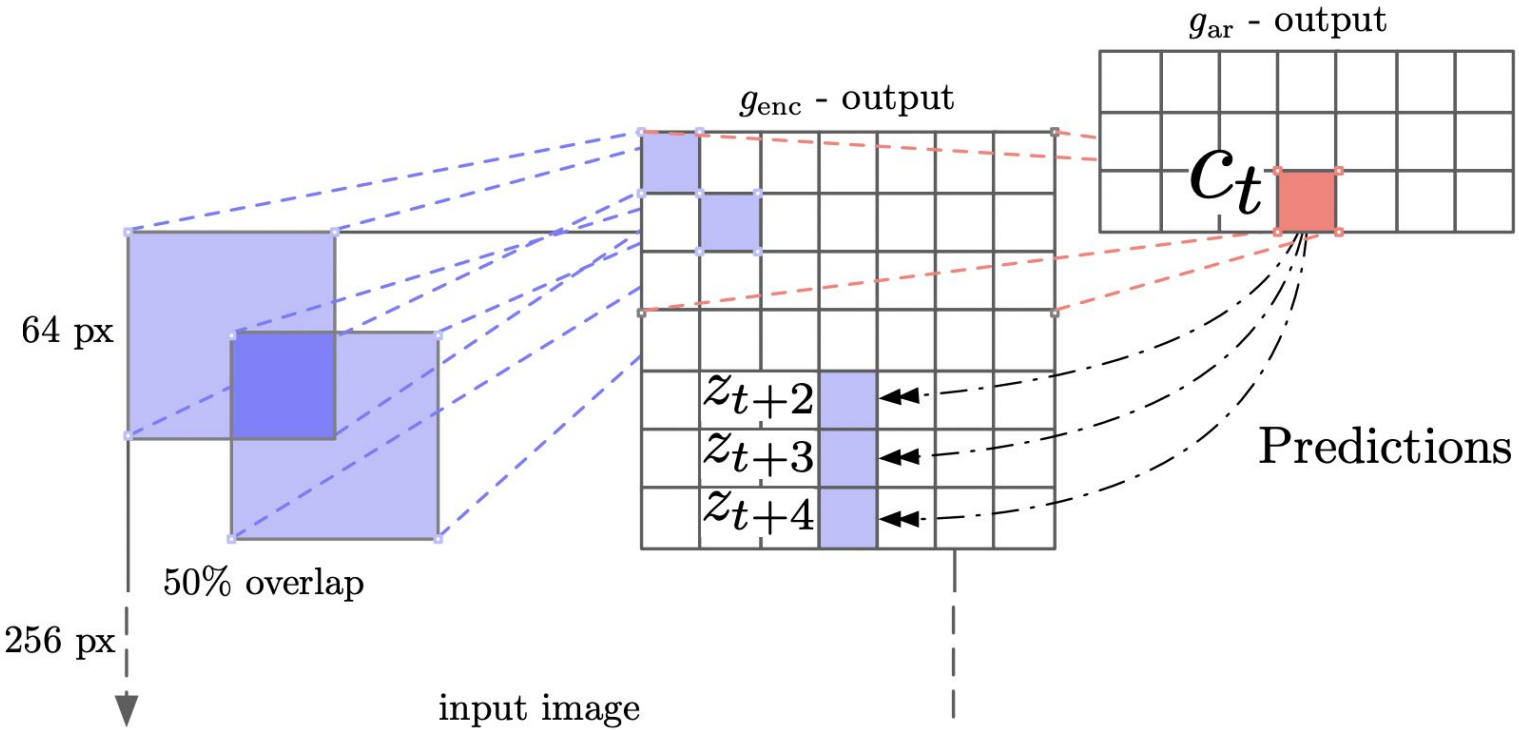
Table 1: LibriSpeech phone and speaker classification results. For phone classification there are 41 possible classes and for speaker classification 251. All models used the same architecture and the same audio input sizes.

| Method | ACC |
|---|---|
| **#steps predicted** | |
| 2 steps | 28.5 |
| 4 steps | 57.6 |
| 8 steps | 63.6 |
| 12 steps | 64.6 |
| 16 steps | 63.8 |
| **Negative samples from** | |
| Mixed speaker | 64.6 |
| Same speaker | 65.5 |
| Mixed speaker (excl.) | 57.3 |
| Same speaker (excl.) | 64.6 |
| Current sequence only | 65.2 |

Table 2: LibriSpeech phone classification ablation experiments. More details can be found in Section 3.1.

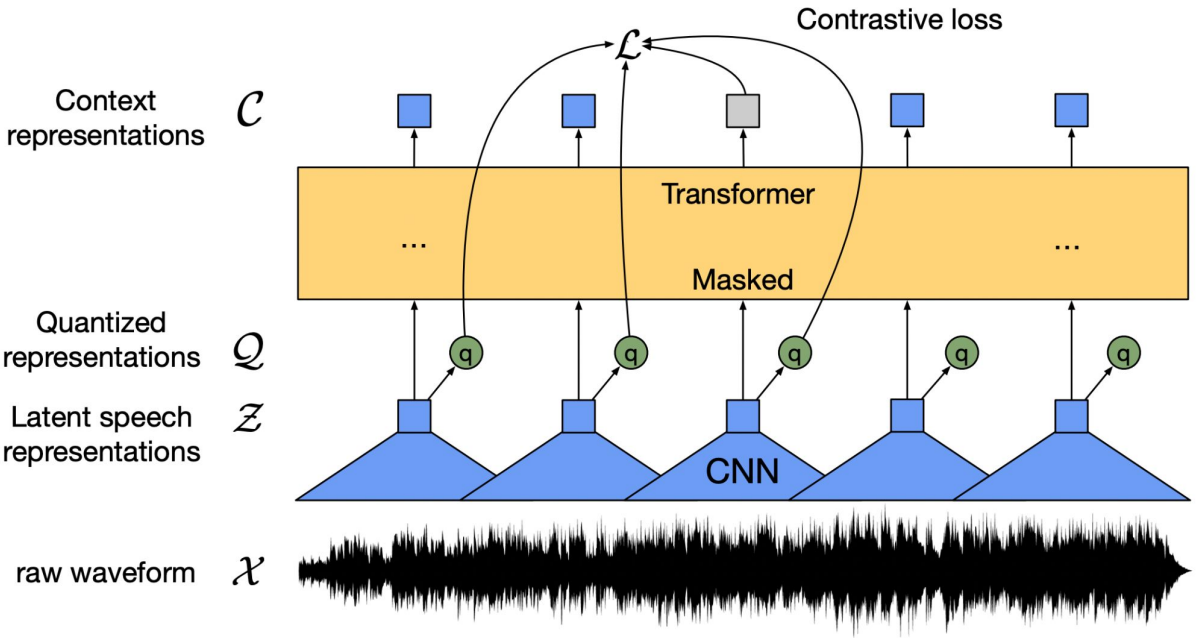
van der Oord et al., 2018

# Contrastive Predictive Coding (CPC)



Baevski et al., 2020

# Wav2Vec 2.0

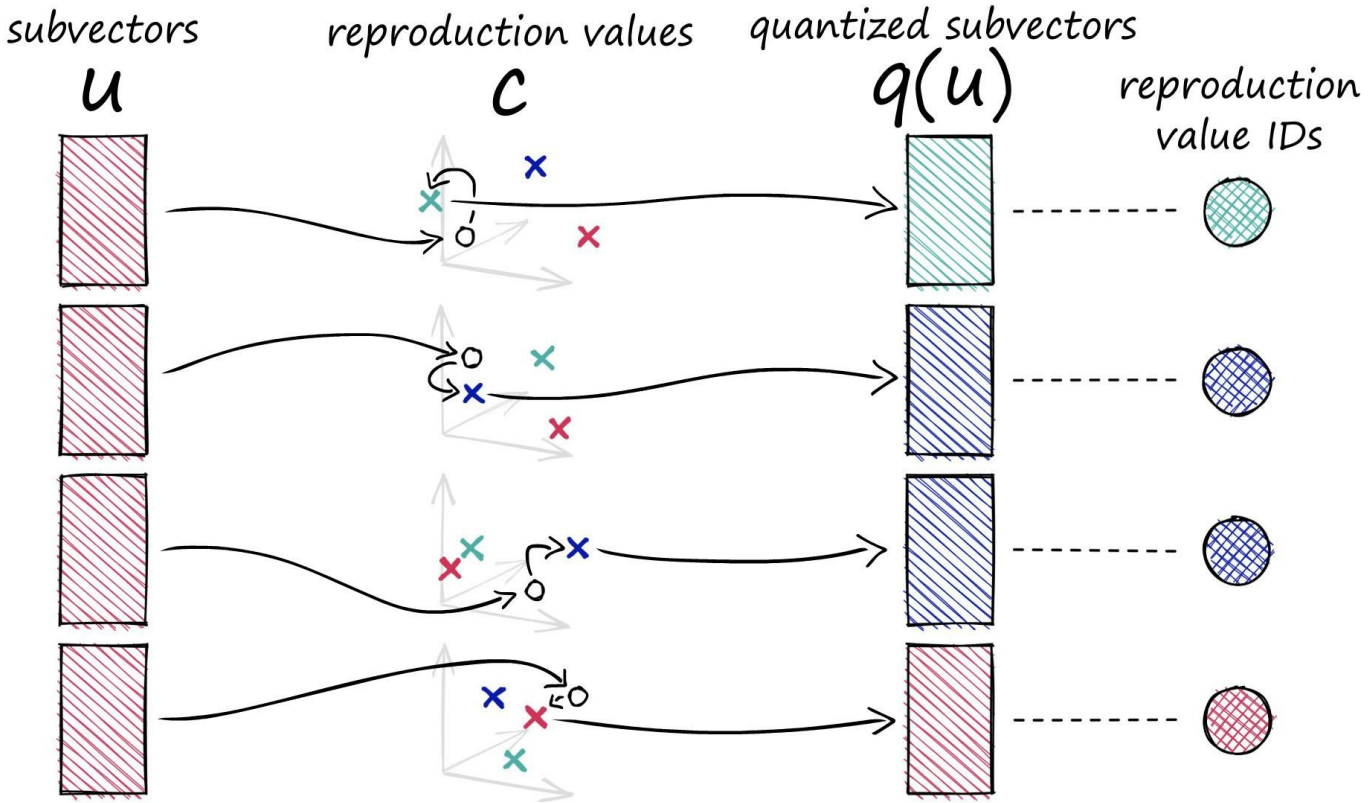



$$\mathcal{L} = \mathcal{L}_m + \alpha \mathcal{L}_d$$

$$\mathcal{L}_m = -\log \frac{\exp(sim(\mathbf{c}_t, \mathbf{q}_t)/\kappa)}{\sum_{\tilde{\mathbf{q}} \sim \mathbf{Q}_t} \exp(sim(\mathbf{c}_t, \tilde{\mathbf{q}})/\kappa)}$$

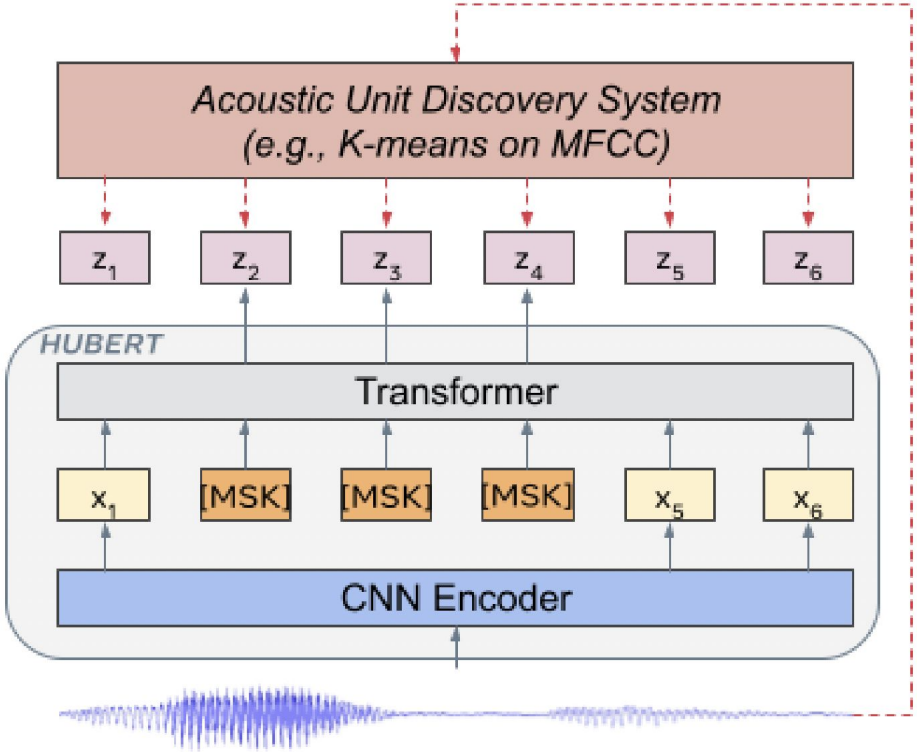
https://www.pinecone.io/

# Product Quantization



Hsu et al., 2021

# HUBERT

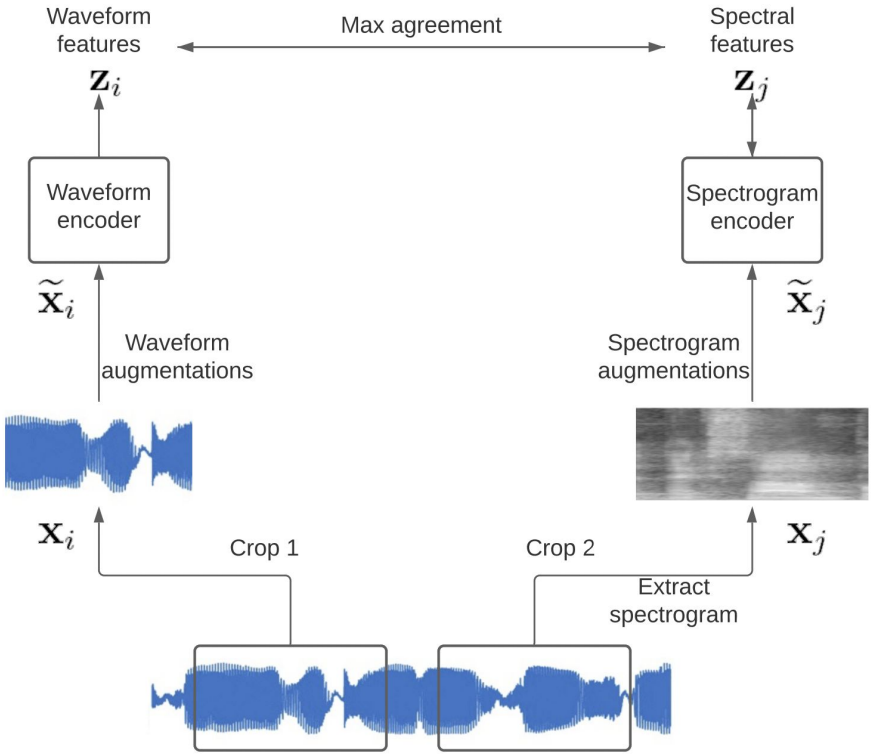



$$L = \alpha L_m + (1-\alpha) L_u$$

$$p_f^{(k)}(c \mid \tilde{X}, t) = \frac{\exp(\mathrm{sim}(A^{(k)} o_t, e_c)/\tau)}{\sum_{c'=1}^{C} \exp(\mathrm{sim}(A^{(k)} o_t, e_{c'})/\tau)}$$

Wang and van der Oord., 2021

# Multi-format contrastive learning




$$L_{i,j} = -\log \frac{\exp\left(\text{sim}\left(\mathbf{z}_i, \mathbf{z}_j\right)/\tau\right)}{\sum_{k \neq i} \exp\left(\text{sim}\left(\mathbf{z}_i, \mathbf{z}_k\right)/\tau\right)}$$

Wang and van der Oord., 2021

# Multi-format contrastive learning

**Audio mixing** Small additive noise of any sort will not alter the original categories of the audio. Given two audio clips $\mathbf{x_1}$ and $\mathbf{x_2}$, the mixed-up version is

$$\hat{\mathbf{x}}_\mathbf{1} = \alpha \mathbf{x}_1 + (1 - \alpha)\mathbf{x}_2 \tag{2}$$

where $\hat{\mathbf{x}}_\mathbf{1}$ inheritances labels from $\mathbf{x_1}$. In this work, $\alpha$ is samples from $\beta(5, 2)$ distribution. This simulates various realistic noise conditions.

**Time masking** $t$ consecutive time steps $[t_0, t_0 + t)$ of the audio can be dropped out and it should not change the event classes, where $t_0$ is randomly sampled. This can be applied both to raw audio and spectrograms.

**Frequency masking** A small amount of $f$ frequency components $[f_0, f_0 + f)$ on the spectrogram can be masked out without losing semantic information.

**Frequency shift** One can apply the truncated shift in frequency to the spectrograms by an integer number sampled from $[-F, F]$, where $F$ is the maximum shift size. Missing values after the shift are set to zero energy. Intuitively, this is a less expensive alternative of changing the pitch of the audio.

Wang and van der Oord., 2021

# Multi-format contrastive learning

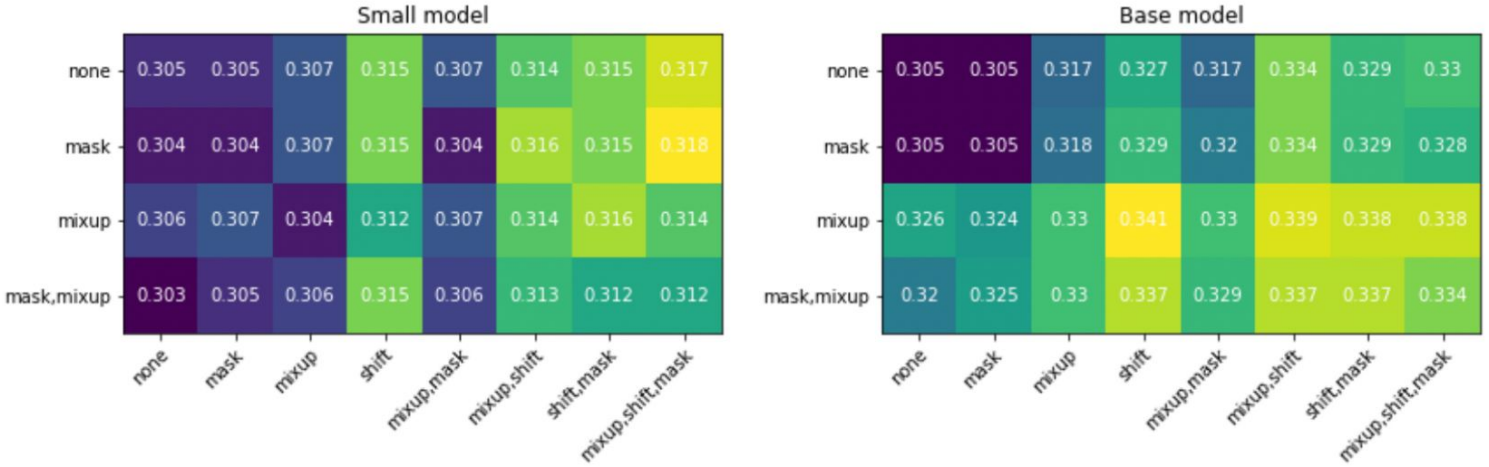



Figure 3: Validation mAP of the raw-audio-vs-log-mel models with different combinations of raw audio (along rows) and spectrogram (along columns) augmentations.

Wang and van der Oord., 2021

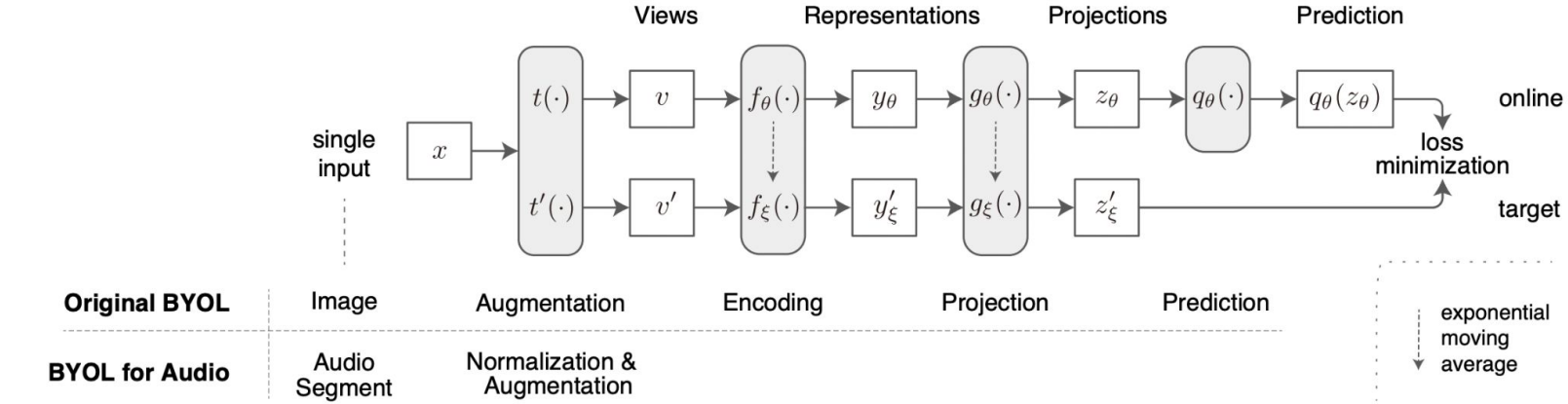
# Multi-format contrastive learning

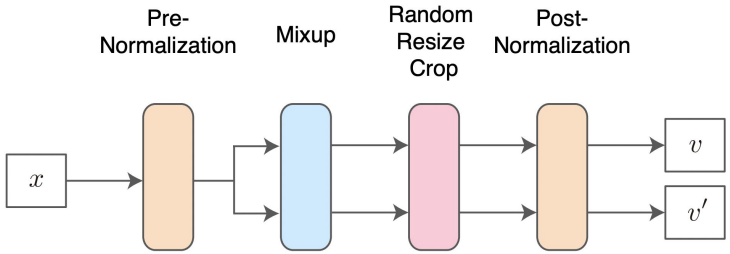
Table 3: Test performance of shallow model classification on AudioSet with fixed representations.

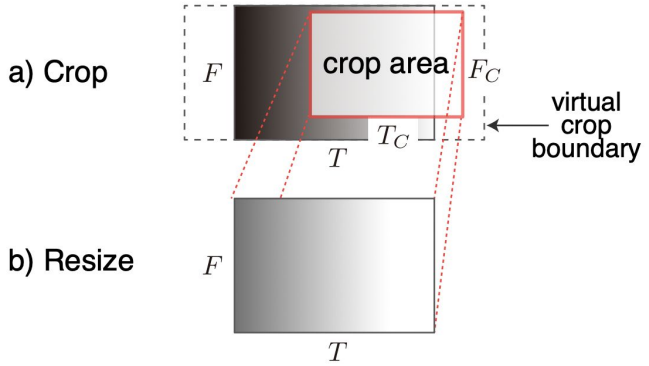
| Model | Train inputs | Eval inputs | Test mAP |
|---|---|---|---|
| Triplet [20] | log-mel | log-mel | 0.244 |
| $L^3$ [22] | log-mel + video | log-mel | 0.249 |
| CPC [21] | waveform | waveform | 0.277 |
| $C^3$ [26] | log-mel + video | log-mel | 0.285 |
| MMV [28] | log-mel + video + text | log-mel | 0.309 |
| Ours | log-mel | log-mel | 0.329 |
| Ours | waveform | waveform | 0.336 |
| Ours | waveform + log-mel | log-mel | 0.368 |
| Ours | waveform + log-mel | waveform | 0.355 |
| Ours | waveform + log-mel | waveform + log-mel | **0.376** |
| Supervised [19] | waveform + log-mel | waveform + log-mel | 0.439 |

Niizumi et al., 2021

# BYOL-A



Niizumi et al., 2021

# BYOL-A

TABLE II
ABLATIONS OF BYOL-A AUGMENTATION MODULE WITH ACCURACY RESULTS, PRETRAINED WITH 1/10 AUDIOSET

| Augmentation blocks used | NS | US8K | VC1 | VF | SPCV2/12 | SPCV2 | Average | Degradation |
|---|---|---|---|---|---|---|---|---|
| Mixup+RRC (BYOL-A) | **71.2%** | 77.0% | 31.0% | 83.1% | **84.5%** | 87.2% | **72.3%** | |
| Mixup+Gaussian+RRC | 69.5% | 74.3% | 25.2% | **84.0%** | 82.8% | **87.4%** | 70.5% | BYOL-A -1.8 |
| Gaussian+RRC | 69.7% | 73.1% | 29.2% | 83.1% | 78.0% | 83.1% | 69.3% | BYOL-A -3.0 |
| RRC | 69.4% | **77.1%** | **34.5%** | 80.3% | 71.4% | 77.4% | 68.4% | BYOL-A -3.9 |
| Mixup | 55.6% | 69.4% | 22.3% | 78.3% | 75.8% | 82.0% | 63.9% | BYOL-A -8.4 |
| Gaussian | 29.5% | 31.2% | 0.9% | 57.9% | 9.4% | 10.3% | 23.2% | BYOL-A -49.1 |

TABLE III
ABLATIONS OF NORMALIZATION BLOCKS WITH AVERAGE ACCURACY RESULTS, PRETRAINED ON 1/10 AUDIOSET

| Method | Average | Degradation |
|---|---|---|
| BYOL-A | **72.3%** | |
| w/o Post-Norm | 72.1% | BYOL-A -0.2 |
| w/o Pre-Norm (mixup $\alpha = 0.05$) | 70.5% | BYOL-A -1.8 |
| w/o Pre-Norm (mixup $\alpha = 0.1$) | 70.3% | BYOL-A -2.0 |
| w/o Pre-Norm (mixup $\alpha = 0.4$) | 68.9% | BYOL-A -3.4 |