

# Robustness of raw waveform speaker embeddings under mismatched conditions

Ge Zhu, Frank Cwitkowitz and Zhiyao Duan

AIRLab, Electrical and Computer Engineering, University of Rochester

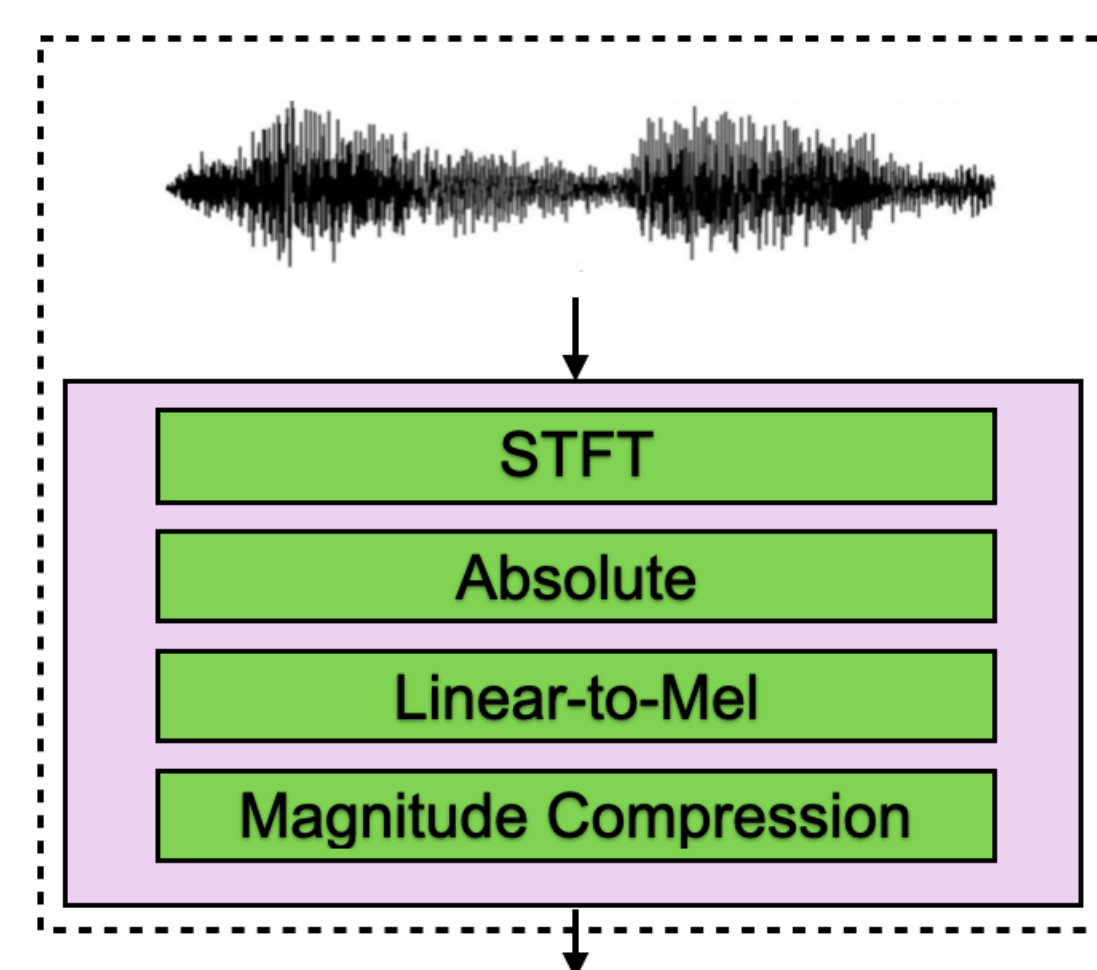
## Abstract

We investigate the cross-dataset speaker verification performance using **raw-waveform based speaker embeddings** and observe a more significant performance degradation compared to spectral based systems. To improve raw-waveform models' cross-dataset performance, we replace the real-valued filters into **analytic filters** to ensure shift invariance; we also apply **variational dropout** to non-parametric filters to prevent them from overfitting irrelevant nuance features. By combining these strategies, we achieve results comparable to spectral based systems on both the VoxCeleb and VOiCEs datasets.

## Time-domain Speaker Embedding

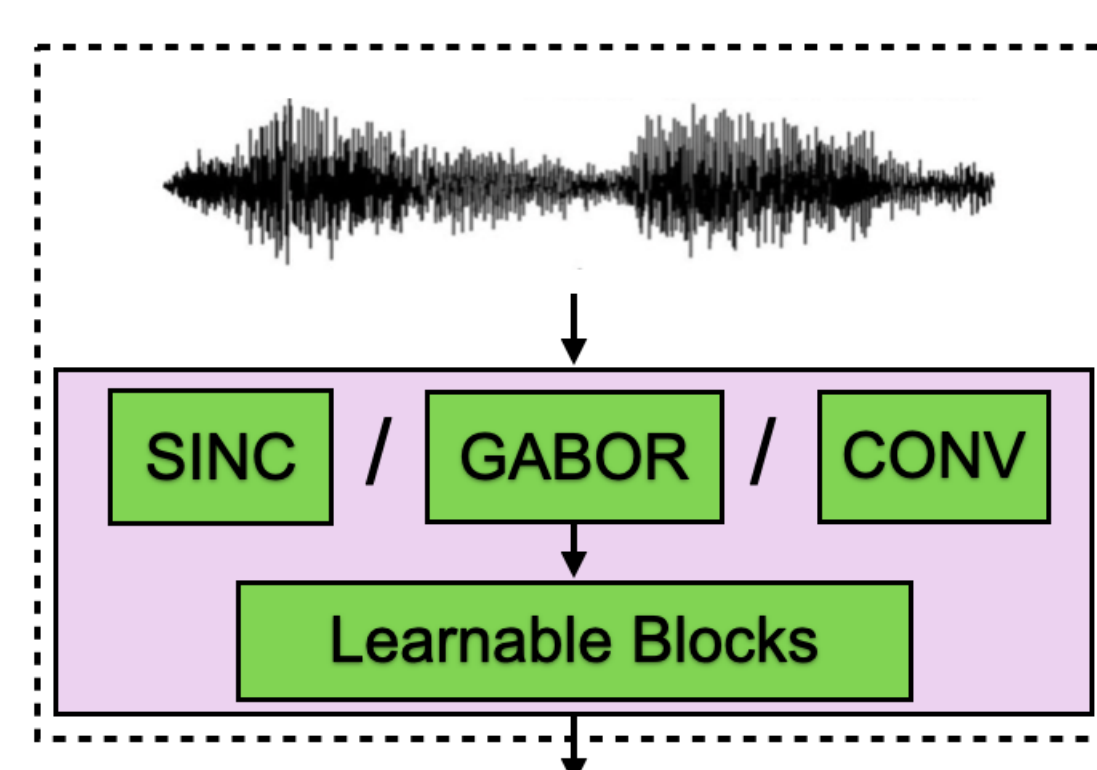
Potential problems for spectral based features:

- Hand-crafted features are not necessarily optimal;
- Mel-spectral transform is lossy.



Two strategies to learn from raw waveforms:

- Non-parametric filterbank with regularization;
- Pre-defined parametric filterbanks.

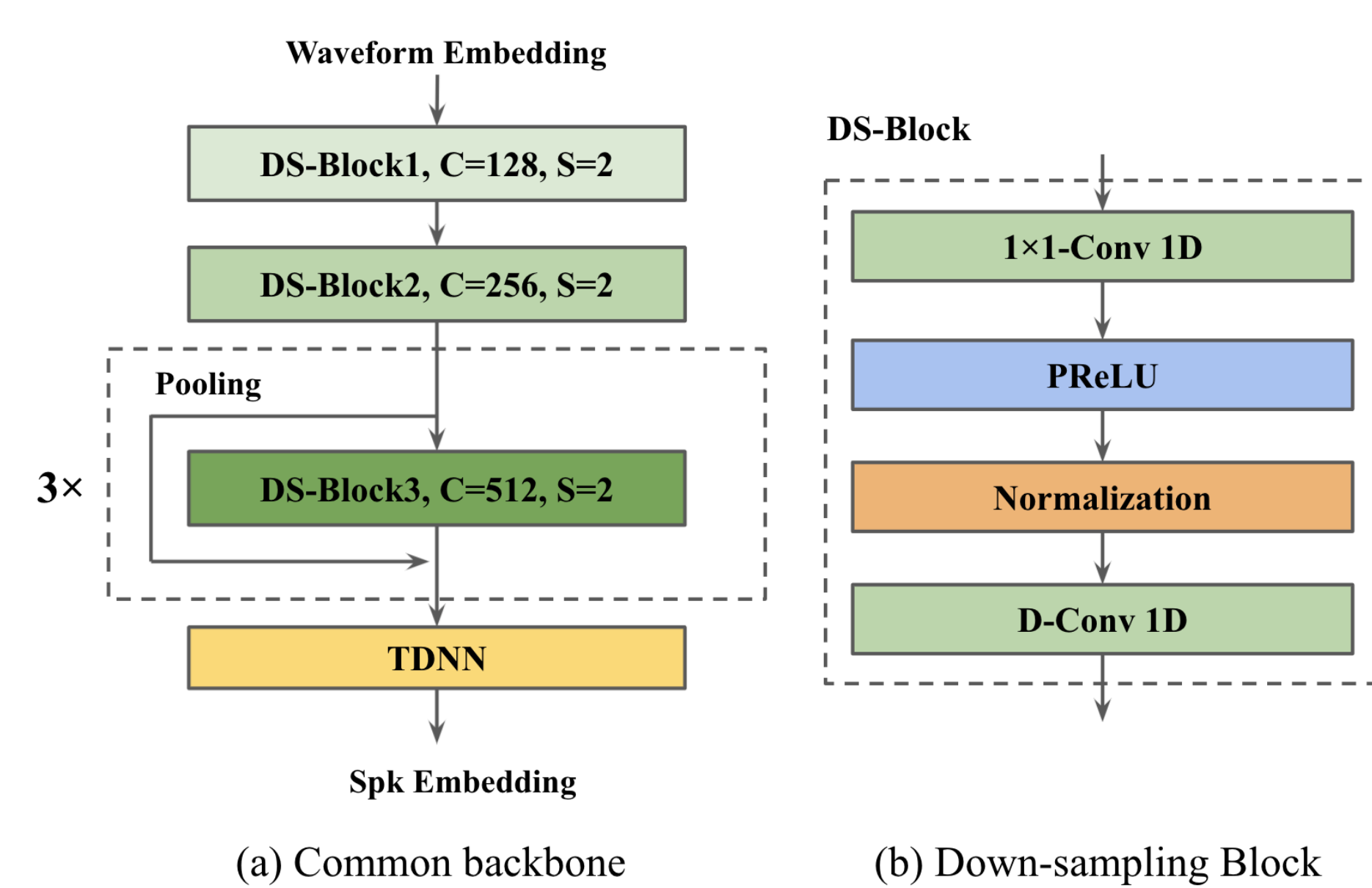


Recent self-supervised speech representation pretraining frameworks, such as wav2vec and WavLM, use waveform as input.

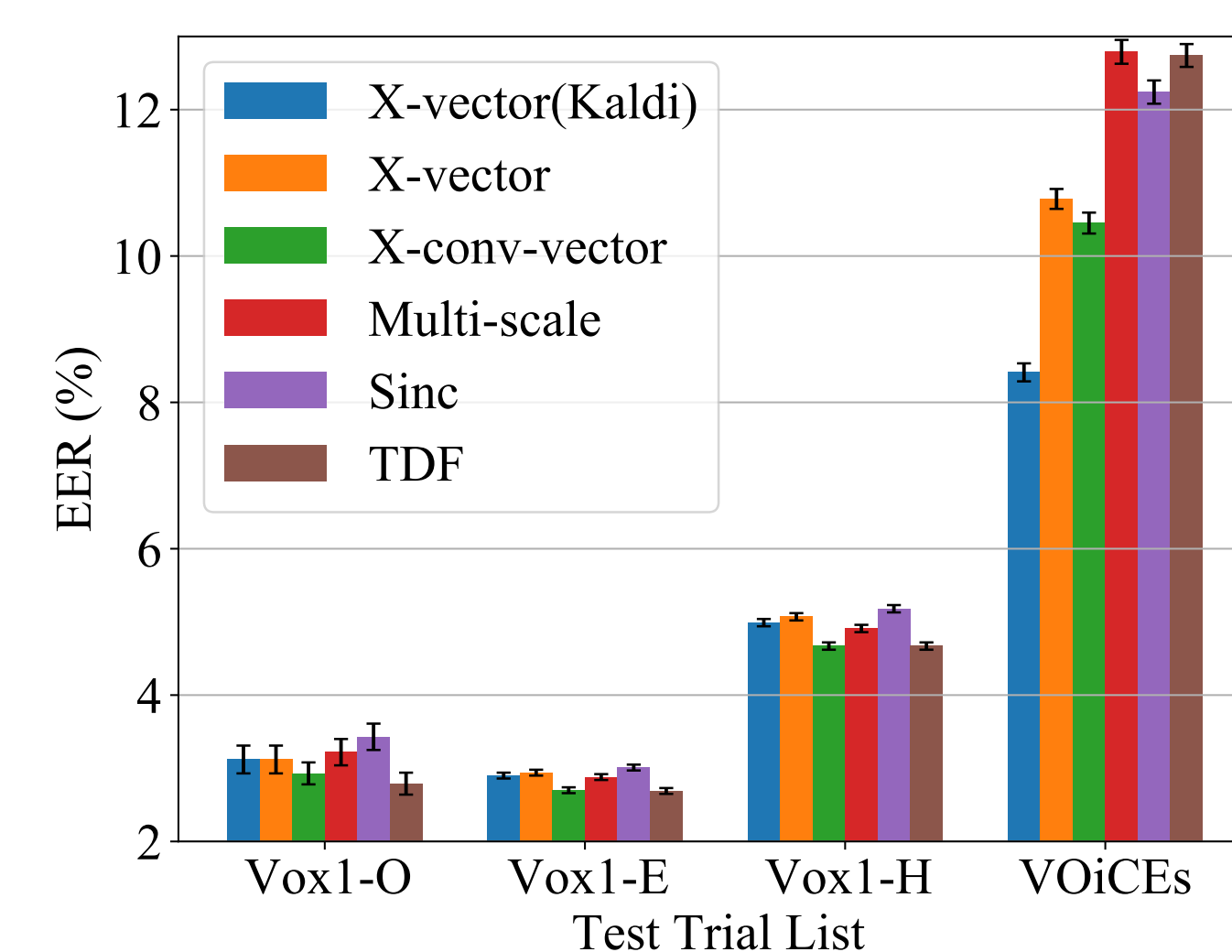
## Channel Mismatch Problem

Comparison of raw-waveform based and mel-spectrum based speaker embeddings under both matched and mismatched conditions.

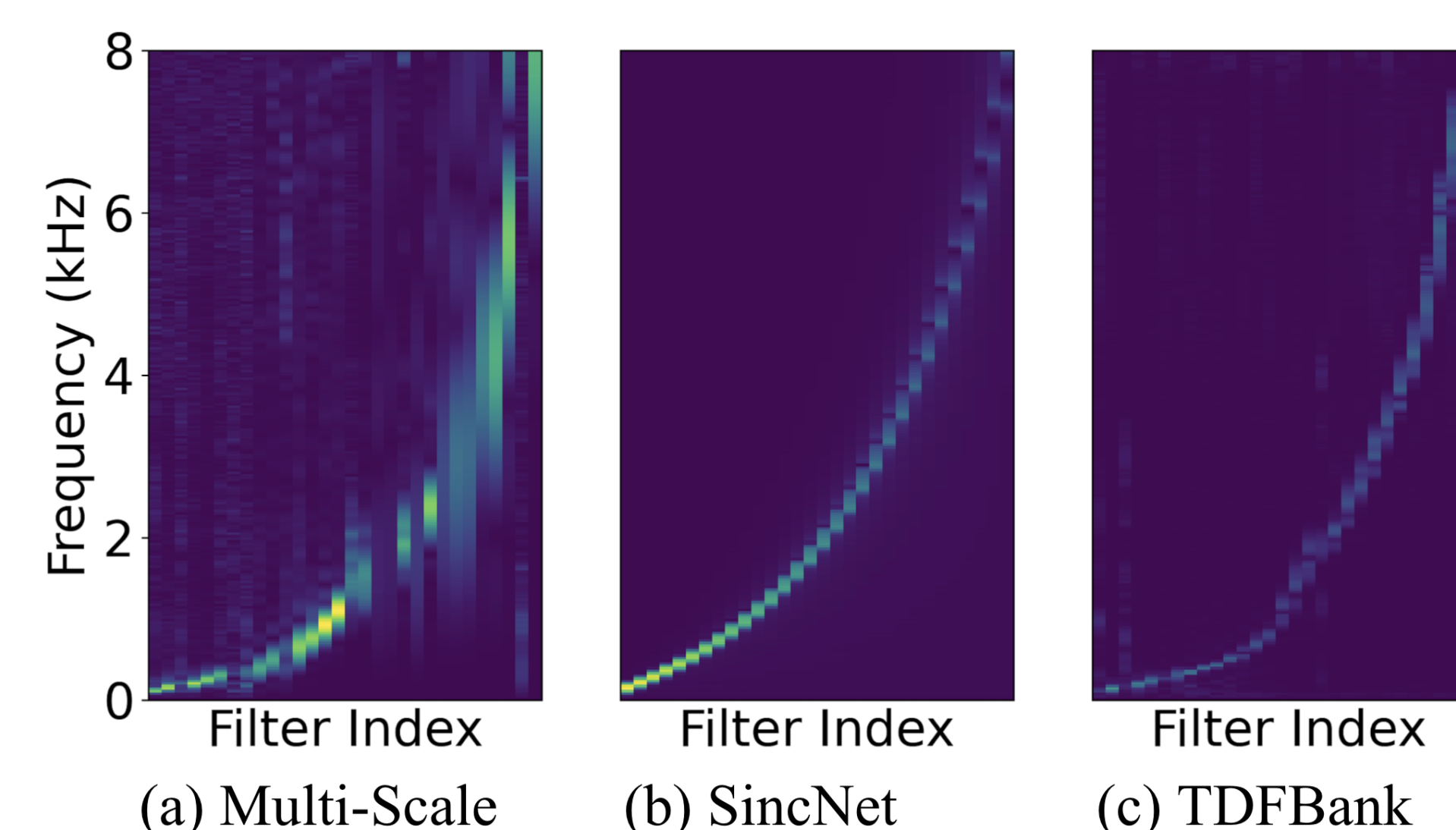
- Dataset: train on noise augmented Vox2, test on in-domain full Vox1 and out-of-domain VOiCEs;
- Audio frontend: Mel-fbank, MFCC, Sinc, TDF;
- Learnable blocks:



- We use equal error rate (EER) to evaluate verification performance scoring with cosine similarity:



- Visualization of learned filters:



## Proposed Strategies

Down-sampled convolutions or pooling layers are not shift-invariant, and they compromise performance on robust classification tasks.

- The modulus of convolution between real-valued input signals  $s(t)$  and analytic filters  $z_a(t)$  are shift-invariant with respect to time:

$$y(t) = |s(t) z_a(t)|$$

- To obtain analytic filter on any given real-valued filter  $s(t)$ , we can apply *Hilbert transform*:

$$z(t) = s(t) + j H\{s(t)\}$$

$$H\{s(t)\} = s(t) \frac{1}{t}$$

Observing learned filter responses trained with noisy datasets, the non-parametric filters tend to overfit the noisy training data, learning task-irrelevant aspects of the recordings.

Variational dropout is a Bayesian regularization technique to help avoid overfitting:

- Dropout can be seen as masking neural network (NN) weights,  $w_{ij} = m_{ij} \cdot w_{ij}$ :

- ① Standard dropout is binary mask  $m_{ij} \sim \text{Bern}(p)$ ;
- ② Gaussian dropout is a ratio mask:

$$m_{ij} \sim N(1, \sigma = p(1-p)).$$

- Equivalently, variational dropout can be seen as applying an independent Gaussian mask parameterized with  $\sigma_{ij}$  to every weight  $w_{ij}$  instead of a fixed parameter  $\sigma$  in Gaussian dropout.

- During training,  $\sigma_{ij}$  is learned through stochastic optimization using an approximated KL-divergence.

- During inference, a threshold is set for  $\sigma_{ij}$ : if it is larger than the threshold, i.e., the corresponding  $w_{ij}$  is stochastic enough,  $w_{ij}$  is then discarded.

The proposed strategies above do not bring extra parameters at inference. In fact, variational dropout can sparsify learned filterbank weights.

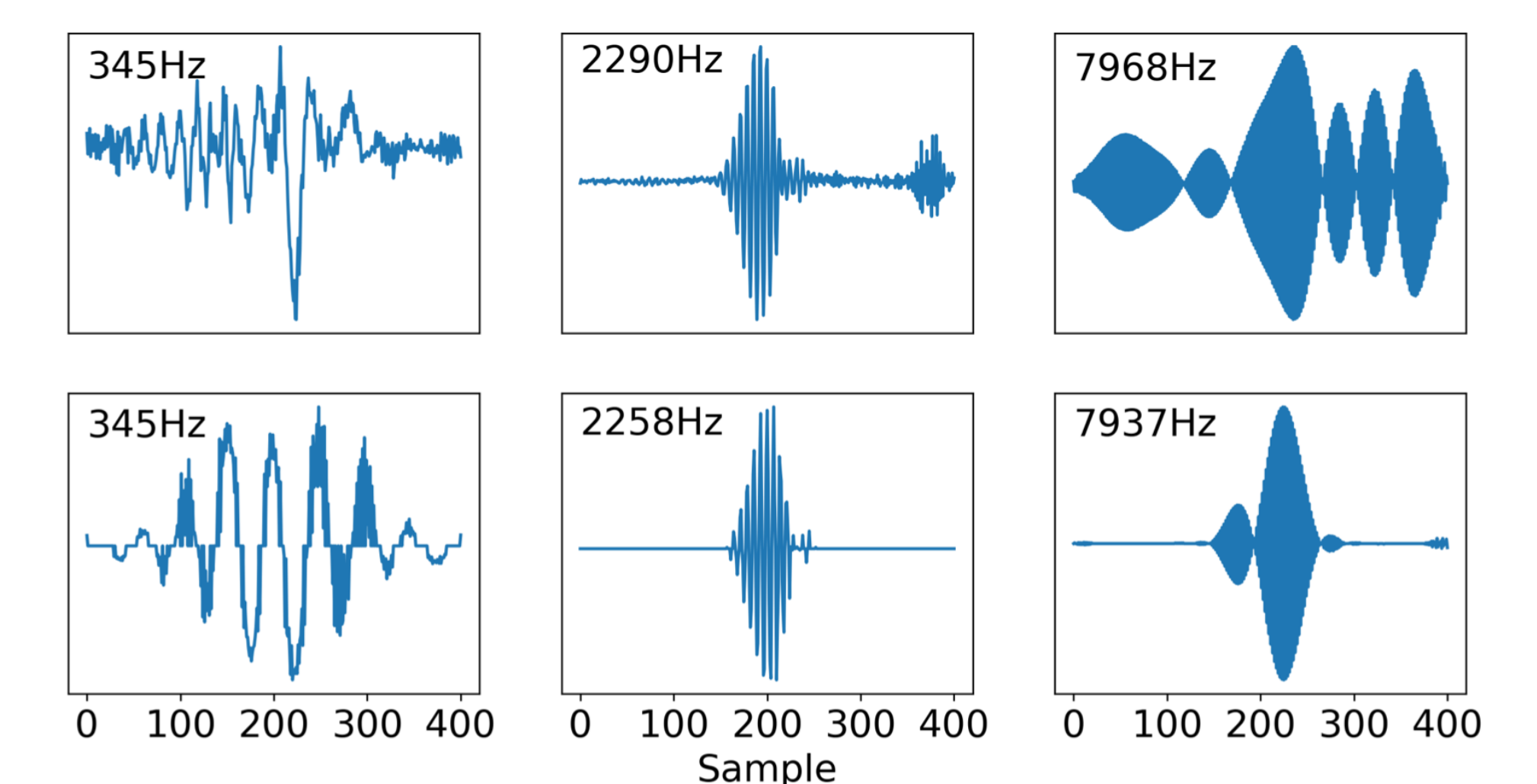
## Results

We repeat the experiments in both matched and mismatched conditions and use PLDA scoring as the backend.

| Frontend     | Vox1-O      | Vox1-E      | Vox1-H      | VOiCEs      |
|--------------|-------------|-------------|-------------|-------------|
| Sinc         | 2.37        | 2.32        | 4.02        | 8.55        |
| Sinc-H       | 2.15        | 2.28        | 3.91        | 8.90        |
| TDF          | <b>1.98</b> | <b>2.19</b> | <b>3.85</b> | <b>8.38</b> |
| TDF-H        | 2.01        | 2.27        | 3.98        | 7.46        |
| TDF-VD       | 1.98        | 2.30        | 4.05        | 7.68        |
| TDF-H-VD     | 1.99        | 2.26        | 3.93        | <b>7.40</b> |
| Mel-Fbank    | <b>2.04</b> | <b>2.17</b> | <b>3.79</b> | 7.10        |
| MFCC (Kaldi) | 2.26        | 2.37        | 4.14        | <b>6.79</b> |

- Analyticity constraint helps non-parametric filters to learn robust representations, but this is not the case for parametric filters.
- Variational dropout improves the performance of non-parametric filterbanks on VOiCEs.

Visualization of learned filters trained on noise augmented VoxCeleb after applying variational dropout: Top row: 'TDF+H' filters. Bottom row: 'TDF+H+VD' filters.



Acknowledgements: This work is funded by a gift from Voice Biometrics Group, an NYS Center of Excellence in Data Science award and National Science Foundation grants 1741472 and DGE-1922591.