InSE-NET: A Perceptually Coded Audio Quality Model based on CNN

—

GUANXIN JIANG, ARIJIT BISWAS, CHRISTIAN BERGLER, ANDREAS MAIER

151ST AES CONVENTION, OCTOBER 20-23, 2021

Deep Learning-based Speech/Audio Quality Predictors

- Mainly deals with:
  1. Non-intrusive quality measurements
  2. Speech at lower (e.g., 16-kHz) sampling rate
  3. Models are fed with either time-domain signals or spectral domain signals (e.g., spectrograms and Mel-scale spectrograms).
- For a comprehensive list, see the references listed in [1].

- None of the work deals with predicting the quality of coded audio.

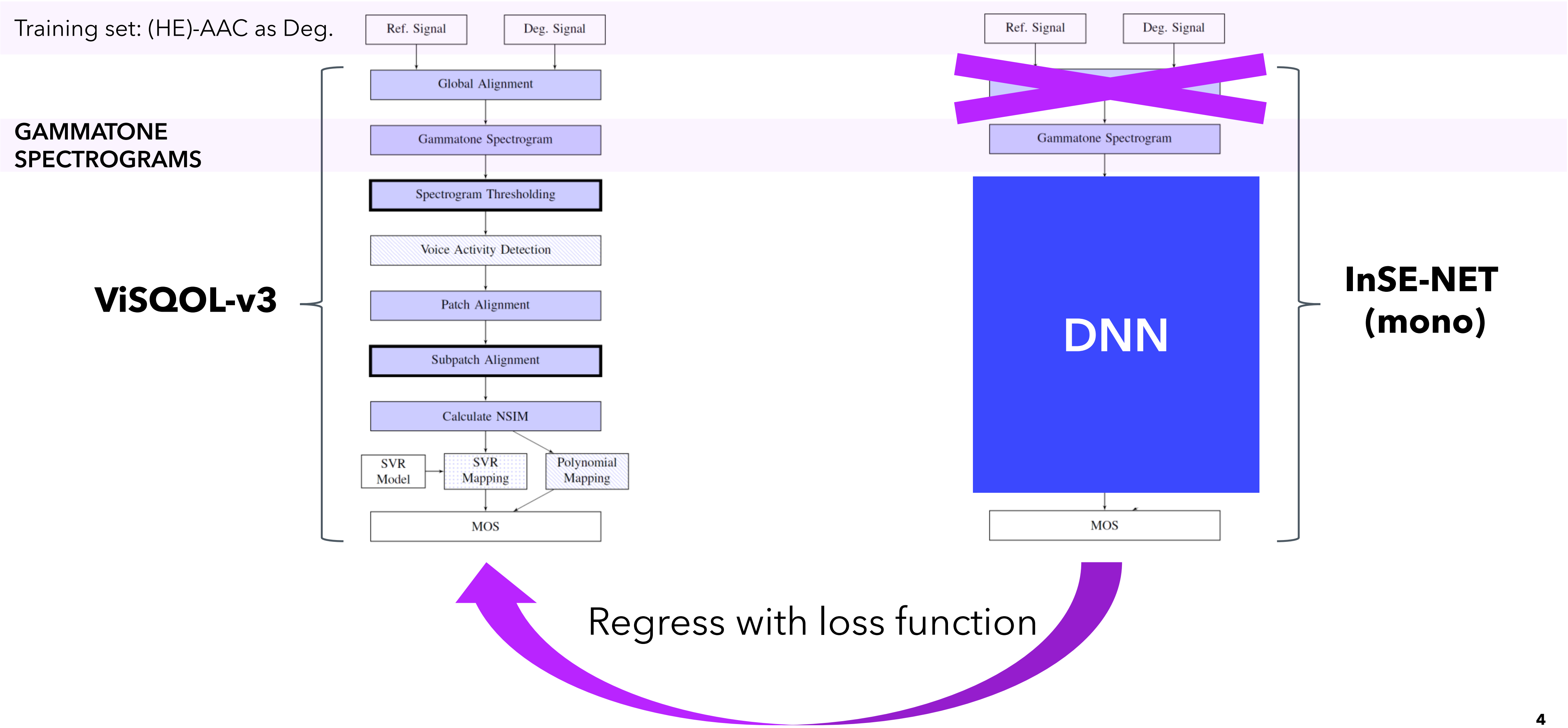
[1] J. Serrà, et al., "SESQA: semi-supervised learning for speech quality assessment," *ICASSP 2021*.

Our contributions

- Intrusive (or full-reference) coded audio quality predictor, designed to operate on:
  1. General audio signal at 48-kHz sampling rate
  2. Gammatone spectrograms (a perceptually-motivated spectrogram representation)
  3. Completely utilize programmatically generated data.

- Mimicking the quality score predicted by a state-of-the-art objective quality metric (ViSQOL-v3) with a deep neural network (DNN), followed by improving over it.

ViSQOL-v3 to InSE-NET

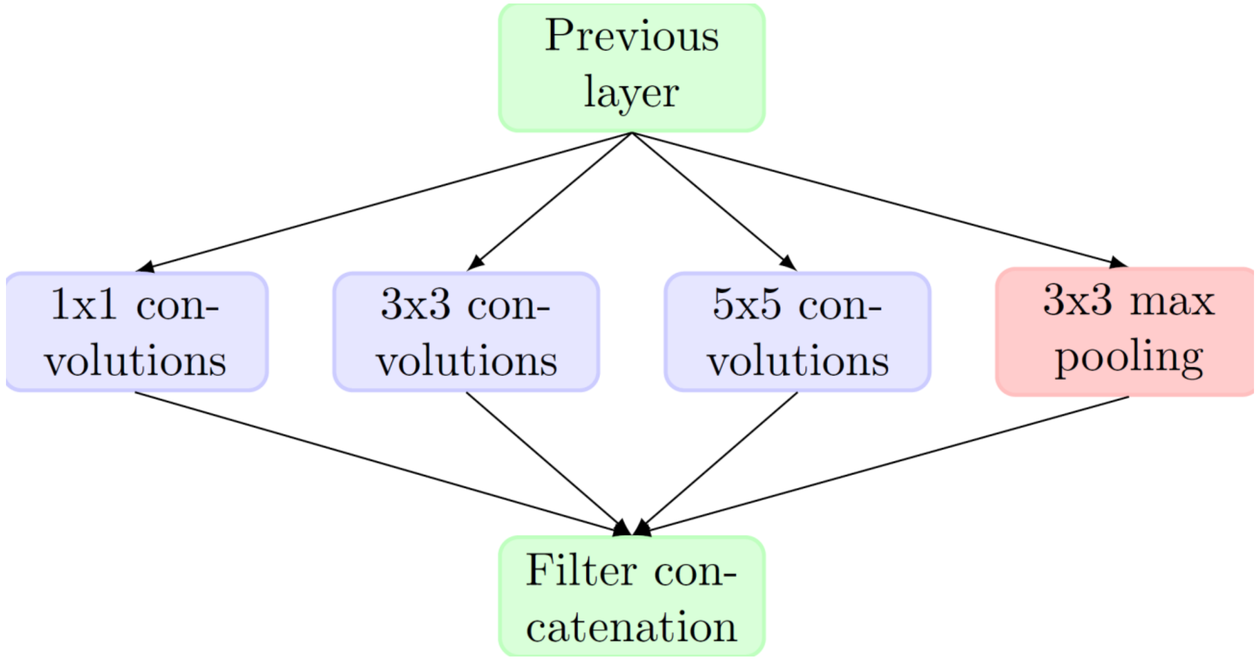


Training Data

- Clean (i.e., reference/un-encoded) data (12h)
  - 4500 music excerpts (10h) from 10 different genres
  - 900 speech excerpts (2h)
- Degraded data
  - 16, 20, 24, 32, 40, 48 kbps (coded, i.e. encoded-decoded with HE-AAC)
  - 80, 96, 128 kbps (coded with AAC)
  - 3.5 and 7.0-kHz low-pass filtered versions of clean
- Label: ViSQOL -v3 MOS as ground truth

Inception Block*

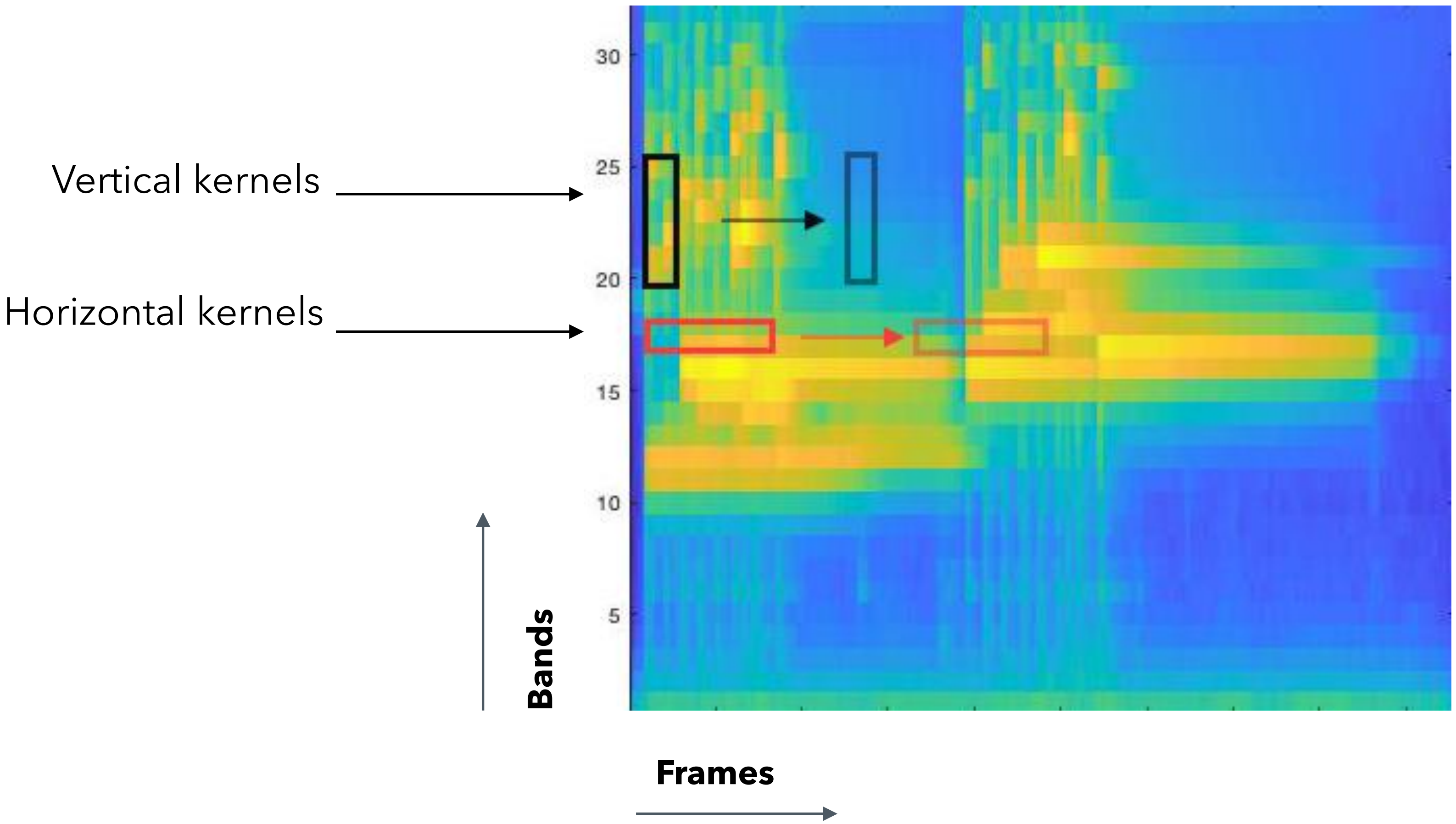
- Adapts to different receptive field size
- Structure with four parallel branches:
    - 1 x 1 conv
    - n x n kernel
    - m x m kernel
    - Max pooling
- Concatenate the outputs of each kernel along the channel axis



*Szegedy, Christian, et al. "Rethinking the inception architecture for computer vision." *CVPR*. 2016.

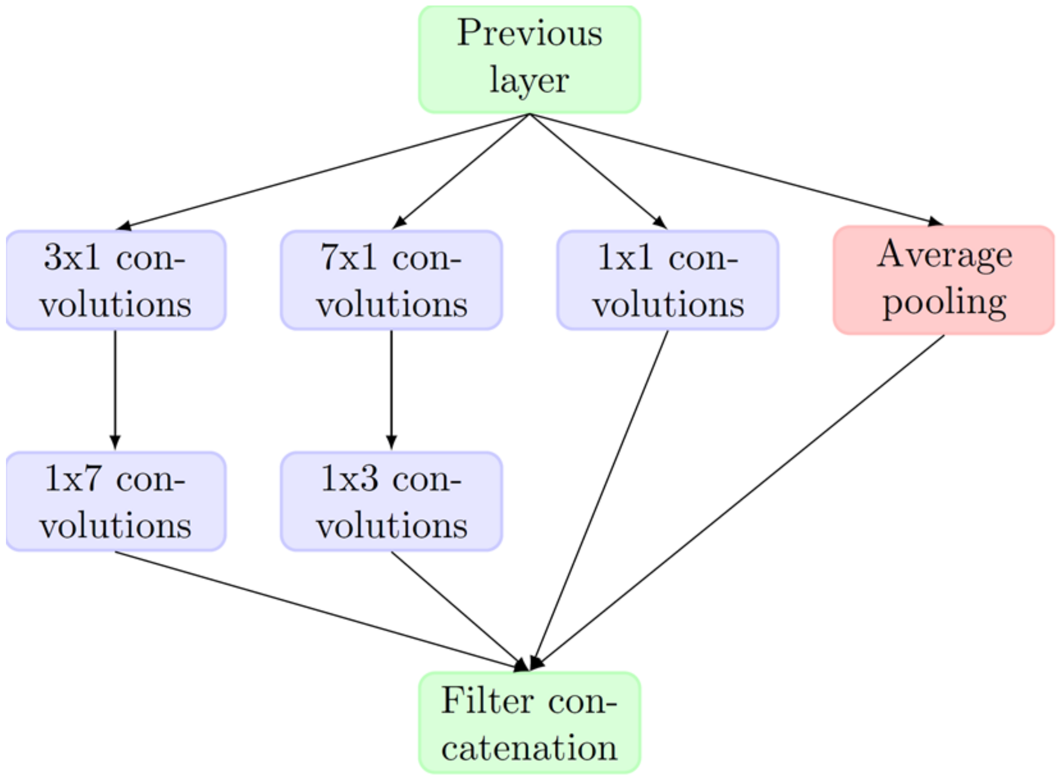
Horizontal and Vertical Kernels

Gammatone Spectrogram



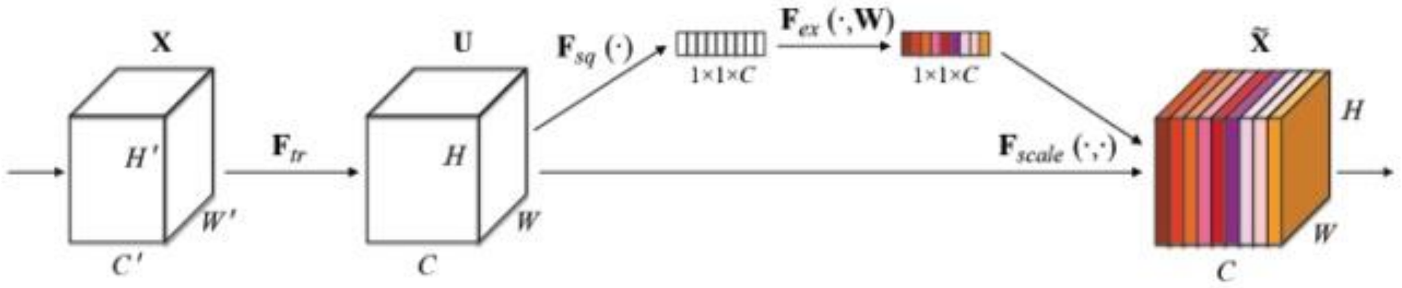
Modified Inception Block for Audio

- Replace the square-shaped kernel with vertical & horizontal rectangular-shaped kernels (3x7, 7x3, 3x5, 5x3)
- Split the kernel into smaller ones to reduce the number of parameters
  - 3 x 7 kernel (21 param) into 3 x 1 and 1 x 7 (10 param)
- Replace max pooling by average pooling



Squeeze & Excitation (SE) Layer*

- A special attention mechanism along channel axis
  - Squeeze: use 1 x 1 conv to squeeze information along time and frequency
  - Excitation: use 2 following fully connected layers and a sigmoid to boost those channels of more importance

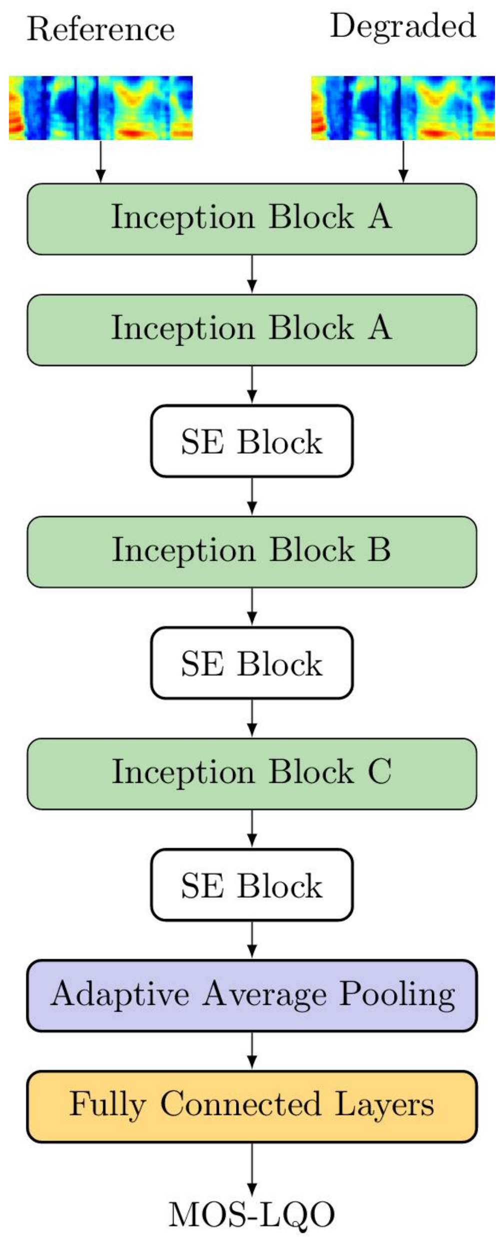


*Hu, Jie, Li Shen, and Gang Sun. "Squeeze-and-excitation networks." *CVPR*. 2018.

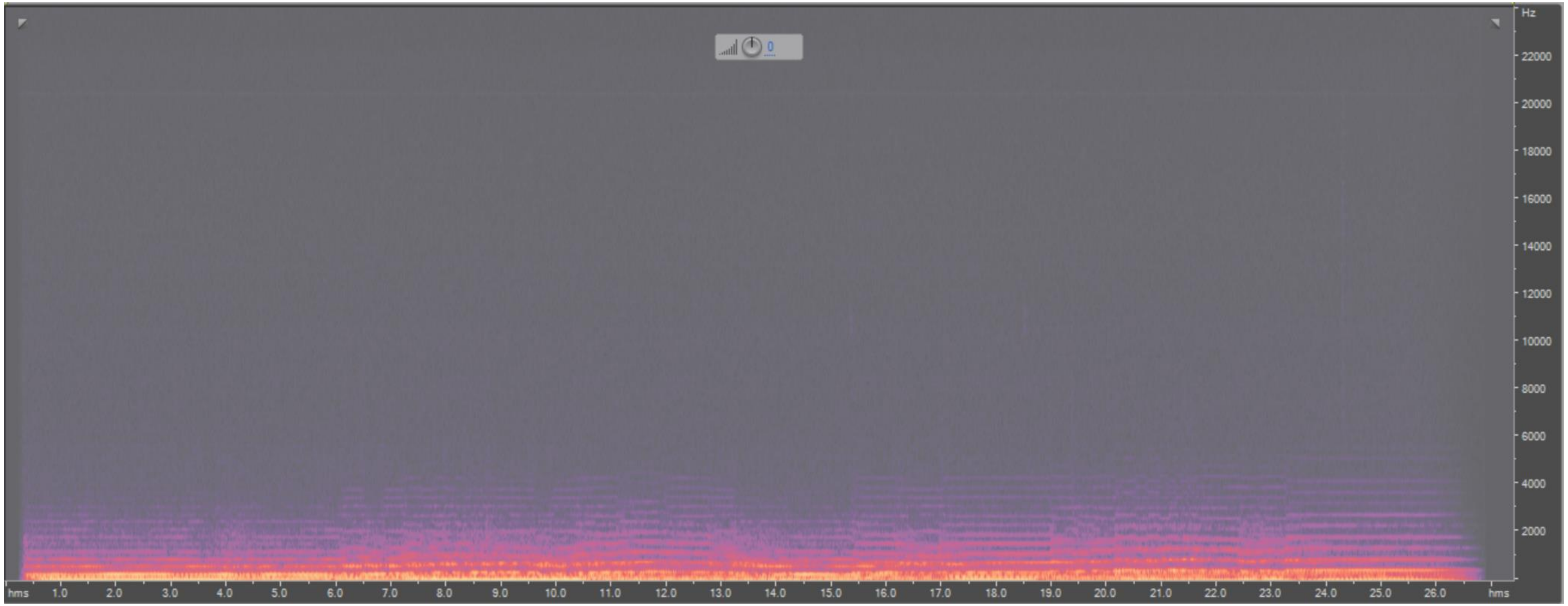
InSE-NET Architecture

Two major changes:

- Removed the head layer
- Replace max pooling with average pooling
- Reason:
    - Head layer in original Inception was designed to extract features from images
    - In our case, Gammatone spectrogram can be already viewed as a feature representation for audio



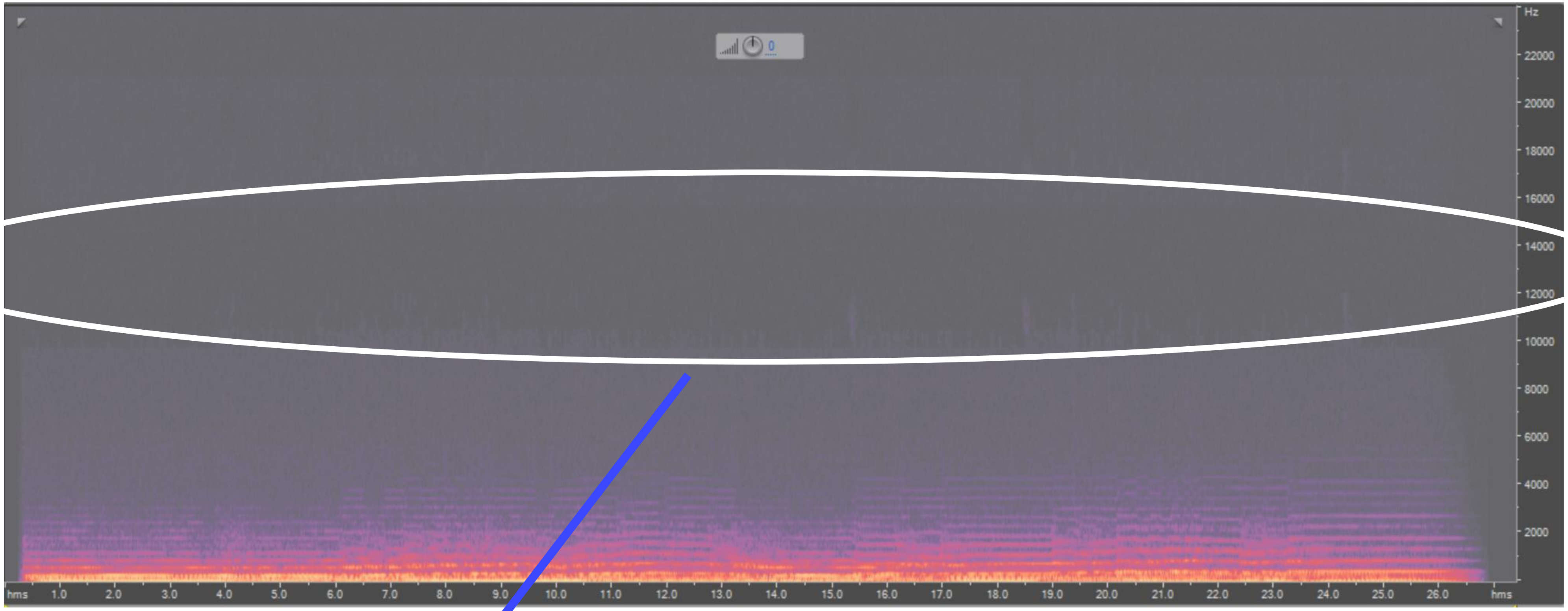
Unencoded audio



Excerpt with very low energy content in high-frequency bands

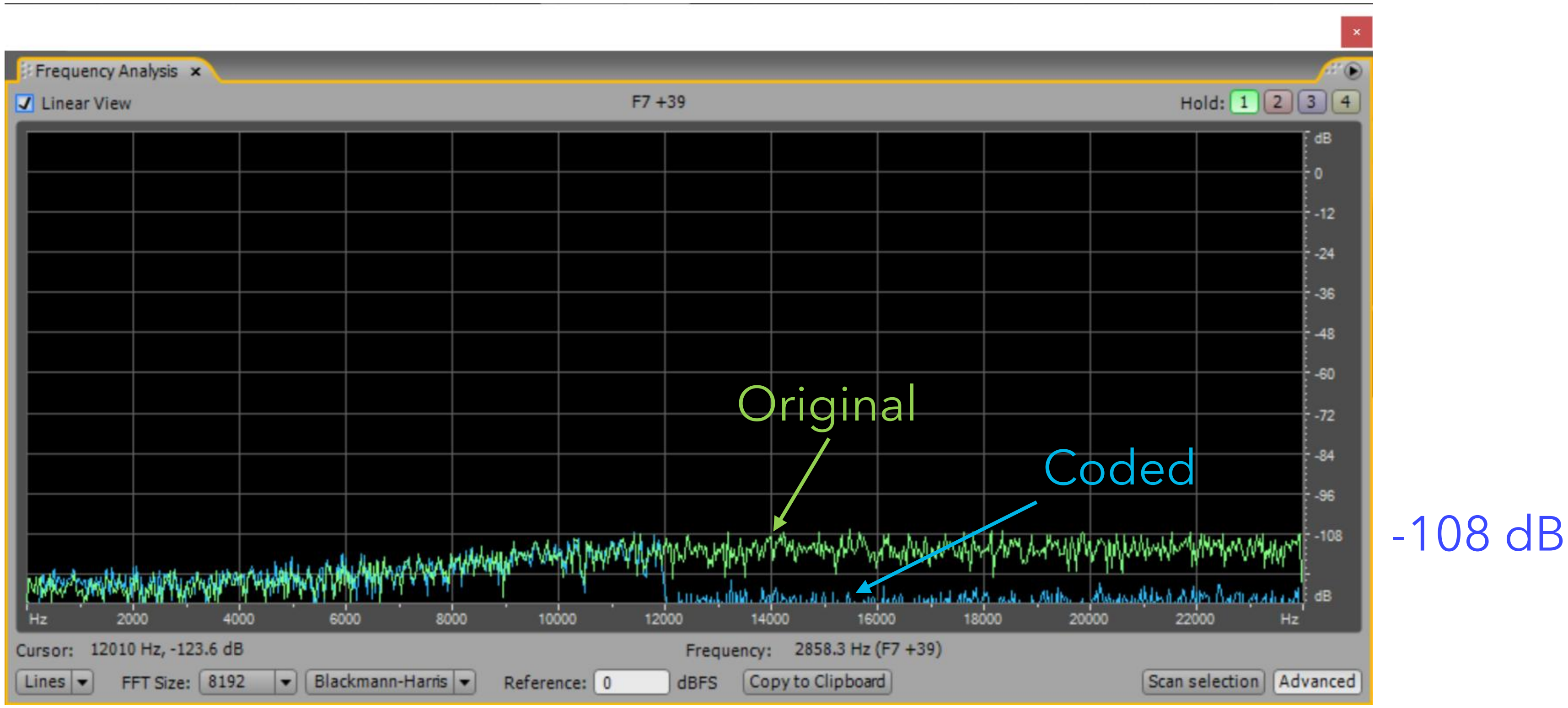
Coded at a very high bitrate

Predicted MOS is low even though there is no audible difference



Reason: visible but inaudible spectral hole in high-frequency region

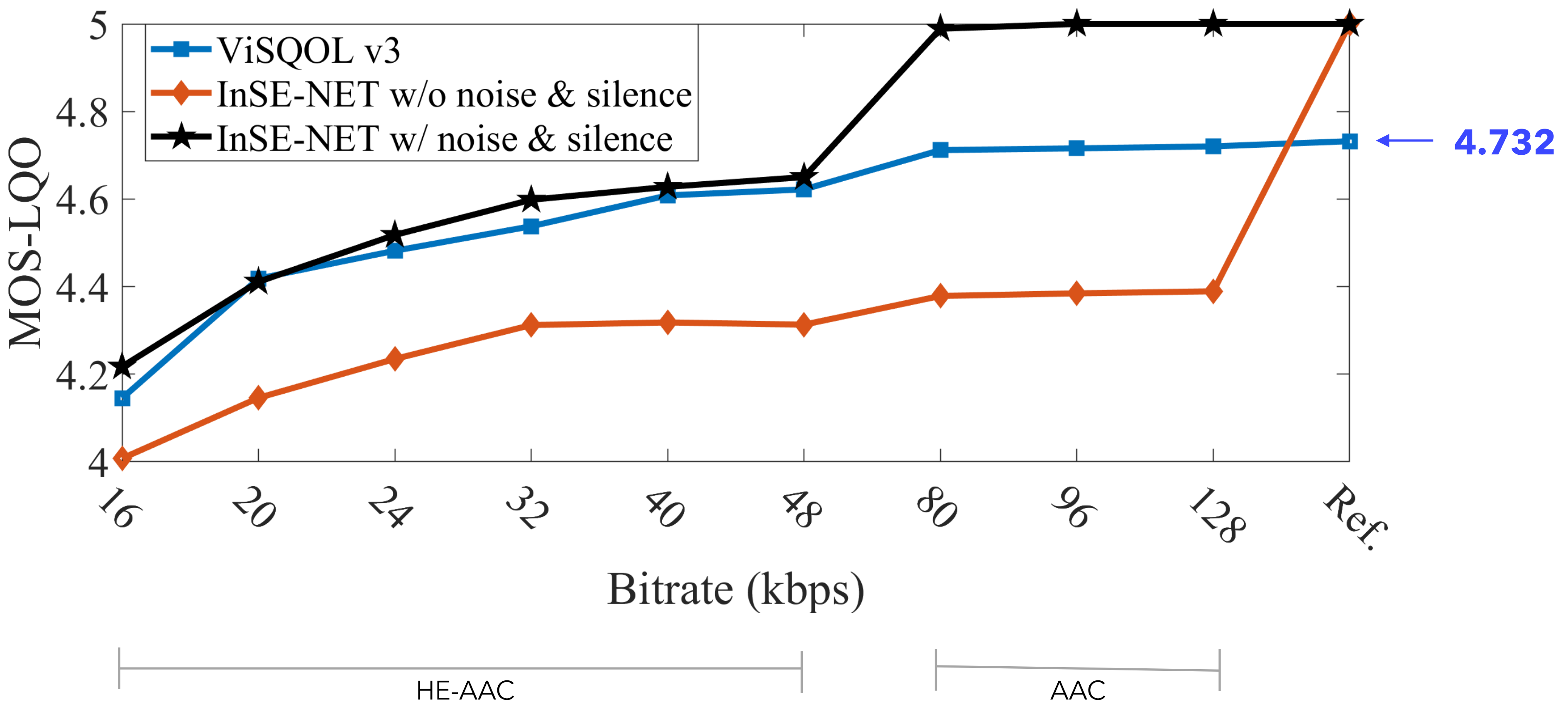
Train with visibly different but perceptually equivalent pairs (code at high bitrates and label MOS as 5)



An example of a visibly different but perceptually equivalent pair

Training with additional synthetic data

An example of (an easy to code) Korean speech excerpt.



Mono MPEG USAC Verification Listening Tests

Mono low-rates

Pearson's correlation coefficient

Spearman's Rank correlation coefficient

	R_p	R_s
PEAQ Advanced	0.650	0.700
ViSQOL-v3	0.810	0.840
InSE-NET (mono)	0.830	0.835

For the Siefried02 excerpt:

48.5% improvement in correlation coefficients

Codecs included in the MUSHRA tests were: AMR-WB+, HE-AAC-v1, and USAC.

Mono MPEG USAC Verification Listening Tests

Mono low-rates

	ViSQOL-v3		InSE-NET	
Codecs	R_p	R_s	R_p	R_s
AMR-WB+	0.877	0.862	0.889	0.856
HE-AAC	0.836	0.792	0.853	0.791
USAC	0.853	0.881	0.873	0.881

Codecs included in the MUSHRA tests were AMR-WB+, HE-AAC-v1, and USAC.

Stereo MPEG USAC Verification Listening Tests

	Low Bitrates		*High Bitrates*	
	$\boldsymbol{R_p}$	$\boldsymbol{R_s}$	$\boldsymbol{R_p}$	$\boldsymbol{R_s}$
ViSQOL v3	0.777	0.782	0.825	0.906
InSE-NET	0.806	0.788	0.847	0.895

Codecs included in the MUSHRA tests were: AMR-WB+, HE-AAC-v1, and USAC.
*ViSQOL-v3 compares the mid-signal: $M = \frac{1}{2}(L + R)$
**Signals fed to the model for comparison are the mid-signal.
.

Conclusions

- We demonstrate mimicking a state-of-the-art coded audio quality metric with a deep neural network called InSE-NET followed by improving over it.
- Synthetic data augmentation can steer the model to predict accurately.
- Listening tests should further improve the accuracy of the prediction.

—

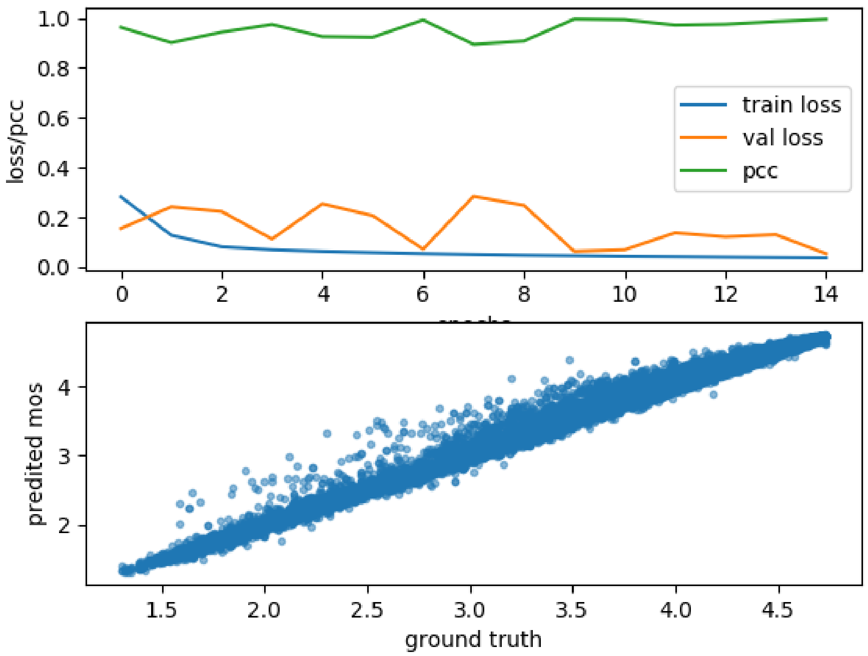
THANK YOU

APPENDIX

Training Dynamics

IN + SE (w/o Head) with L1-loss

$$L_1 = \frac{1}{N}\sum_{i}^{N} \mid M - \hat{M}_i \mid$$



IN + SE (w/o Head) with Smooth L1-loss

$$L_1, smooth = \frac{1}{N}\sum_{i} z_i$$

$$where \quad z_i = \begin{cases} \frac{1}{2}(M - \hat{M}_i)^2 & \text{if} \mid M - \hat{M}_i \mid < 1 \\ \mid M - \hat{M}_i \mid - \frac{1}{2} & \text{otherwise} \end{cases}$$