Audio Codec Enhancement with Generative Adversarial Networks

ARIJIT BISWAS

DAI JIA

Presented at ICASSP 2020

Motivation – I

- Low-bitrate audio coding introduces unavoidable coding artifacts, with significant impact on the quality of:
  - ➢ Speech
  - ➢ Dense transient events, e.g. applause

- Traditional signal processing-based coded audio restoration tools do exist:
  - Targeted at specific artifacts
  - Require specialized knowledge about codec and its settings
  - Do not provide significant audio quality improvement

© 2020 DOLBY LABORATORIES, INC.

Motivation – II

- Deep (conditioned) generative models have opened up exciting opportunities
  - Novel samples created by generative models are suited to restore lost information (e.g. by intelligent gap filling) due to quantization and coding

- Currently generative models for coded audio restoration are based on auto-regressive models (e.g. WaveNet, RNN-based LPCNet)
  - Decoded parameters for conditioning → not an end-to-end system
  - \+ Significant quality boost: demonstrated for speech (coded with speech codec)
  - – Complex: due to autoregressive nature of the model

© 2020 DOLBY LABORATORIES, INC.

Goals

- Backwards compatible improvement of low-bit rate audio codec
- Improve the quality of coded speech and applause signals
- Employ deep generative model
  - End-to-end system: operating directly on decoded audio samples
  - One-shot enhancement

© 2020 DOLBY LABORATORIES, INC.

Proposal

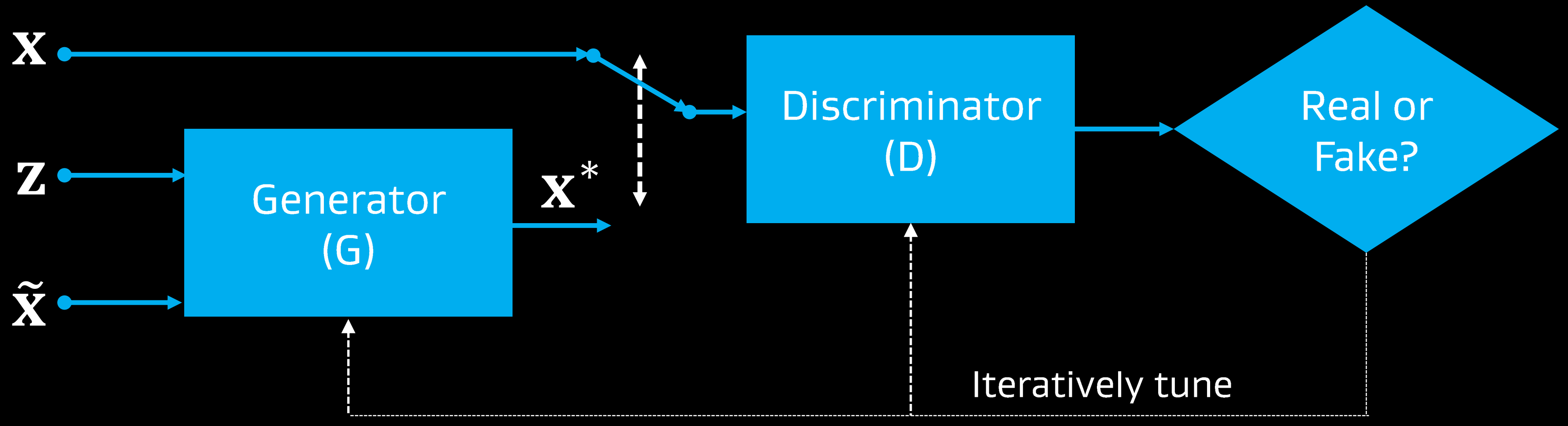
Audio Codec Enhancement with Generative Adversarial Networks (GAN)

- First contribution demonstrating:
  - Adversarial framework to enhance coded audio
  - Deep learning based coded applause restoration

© 2020 DOLBY LABORATORIES, INC.

GAN Training Setup

- Let $\mathbf{x}$ be unencoded audio and $\tilde{\mathbf{x}}$ be the decoded audio. Goal: $\mathbf{x}^*$ is enhanced audio
- For our problem, estimation of $\mathbf{x}^*$ is dependent on $\tilde{\mathbf{x}}$



© 2020 DOLBY LABORATORIES, INC.

Conditional GAN* Training Setup

- Let $\mathbf{x}$ be unencoded audio and $\tilde{\mathbf{x}}$ be the decoded audio. Goal: $\mathbf{x}^*$ is enhanced audio
- For our problem, estimation of $\mathbf{x}^*$ is dependent on $\tilde{\mathbf{x}}$

- Train D with both signals as input: enables D to learn conditional classification task.
- Same principle was also employed in SEGAN** on which our contribution is based on.

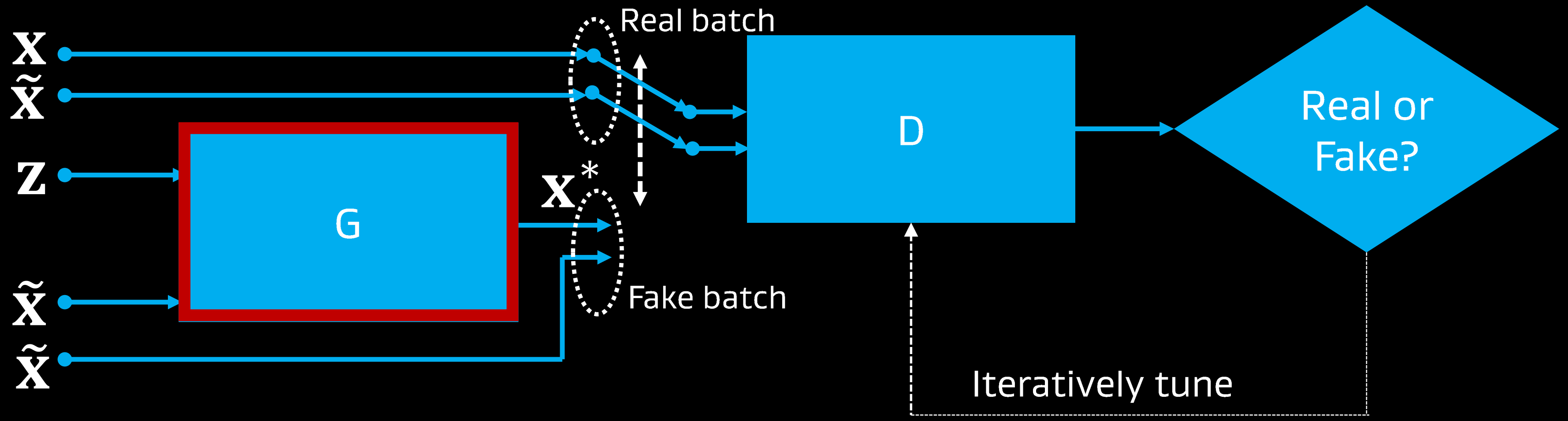
*P. Isola, et al., "Image-to-Image Translation with Conditional Adversarial Networks," *CVPR 2017.*

**S. Pascual, et al., "SEGAN: Speech Enhancement Generative Adversarial Network," *Interspeech 2017.*

© 2020 DOLBY LABORATORIES, INC.

Deep "Coded Audio Enhancer" (DCAE) Training – Step I (a)

- Let $\mathbf{x}$ be unencoded audio and $\tilde{\mathbf{x}}$ be the decoded audio. Goal: $\mathbf{x}^*$ is enhanced audio
- G is fixed then train D to recognize unencoded audio as real

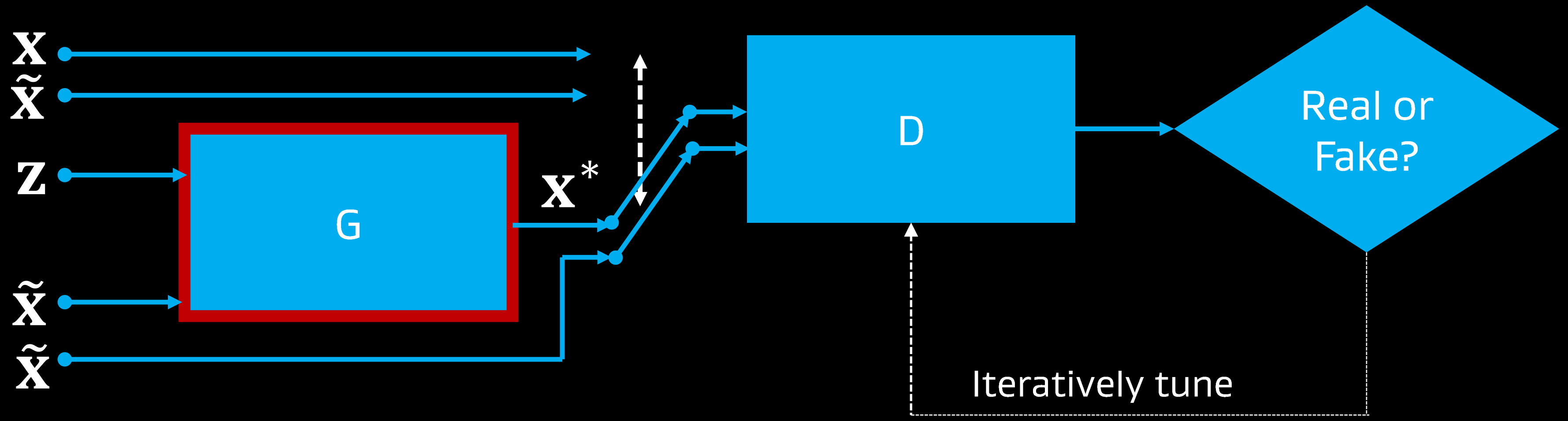


$$\mathcal{L}_D = \frac{1}{2}\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}\sim p_{data}(\mathbf{x},\tilde{\mathbf{x}})}[(D(\mathbf{x},\tilde{\mathbf{x}})-1)^2]$$

© 2020 DOLBY LABORATORIES, INC.

Deep “Coded Audio Enhancer” (DCAE) Training – Step I (b)

- Let $\mathbf{x}$ be unencoded audio and $\tilde{\mathbf{x}}$ be the decoded audio. Goal: $\mathbf{x}^*$ is enhanced audio
- Keep G fixed: train D to recognize generated audio $\mathbf{x}^*$ as fake

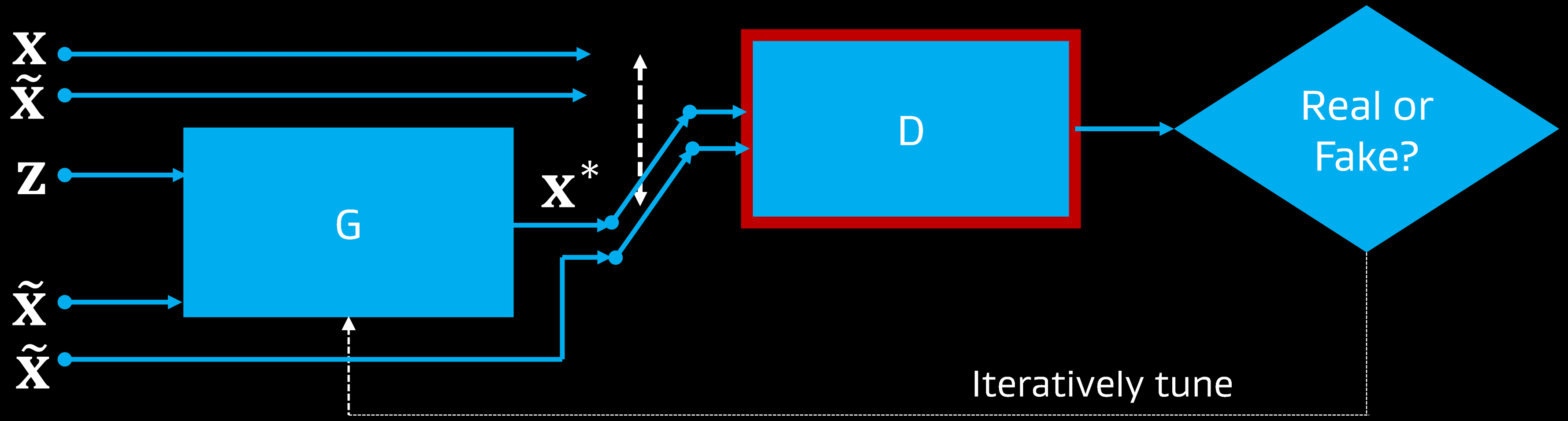


$$\mathcal{L}_D = \frac{1}{2}\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}\sim p_{data}(\mathbf{x},\tilde{\mathbf{x}})}[(D(\mathbf{x},\tilde{\mathbf{x}})-1)^2] + \frac{1}{2}\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z}),\tilde{\mathbf{x}}\sim p_{data}(\tilde{\mathbf{x}})}[D(\mathbf{x}^*,\tilde{\mathbf{x}})^2]$$

© 2020 DOLBY LABORATORIES, INC.

Deep “Coded Audio Enhancer” (DCAE) Training – Step II

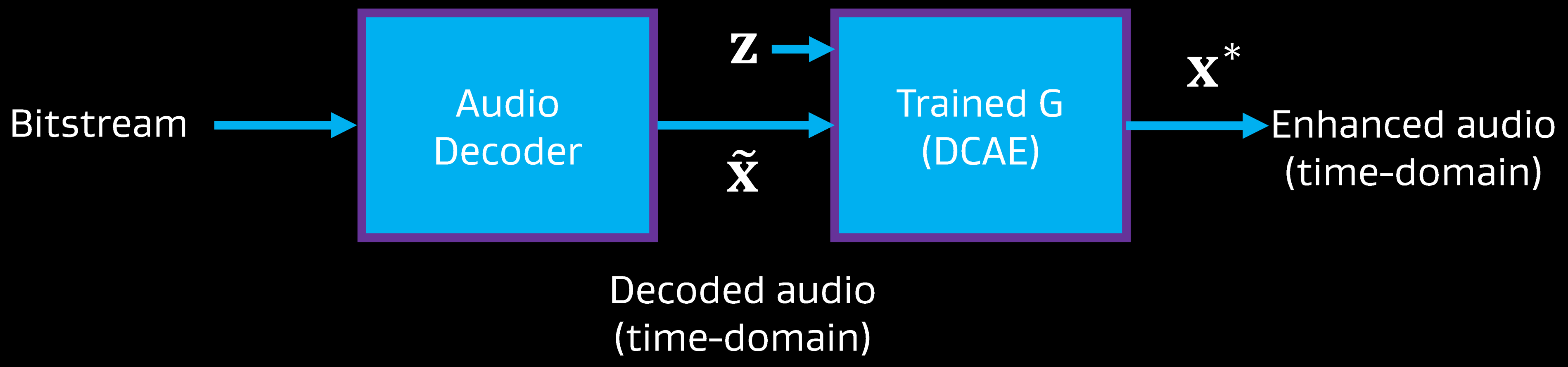
- Let $\mathbf{x}$ be unencoded audio and $\tilde{\mathbf{x}}$ be the decoded audio. Goal: $\mathbf{x}^*$ is enhanced audio
- D is fixed then train G so that D recognizes $\mathbf{x}^*$ as real



$$\mathcal{L}_G = \overbrace{\frac{1}{2}\,\mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z}),\tilde{\mathbf{x}}\sim p_{data}(\tilde{\mathbf{x}})}\left[\left(D(\mathbf{x}^*,\tilde{\mathbf{x}})-1\right)^2\right]}^{\text{Adversarial Loss}} + \lambda \overbrace{\|\mathbf{x}^* - \mathbf{x}\|_1}^{L1\text{-norm}}$$

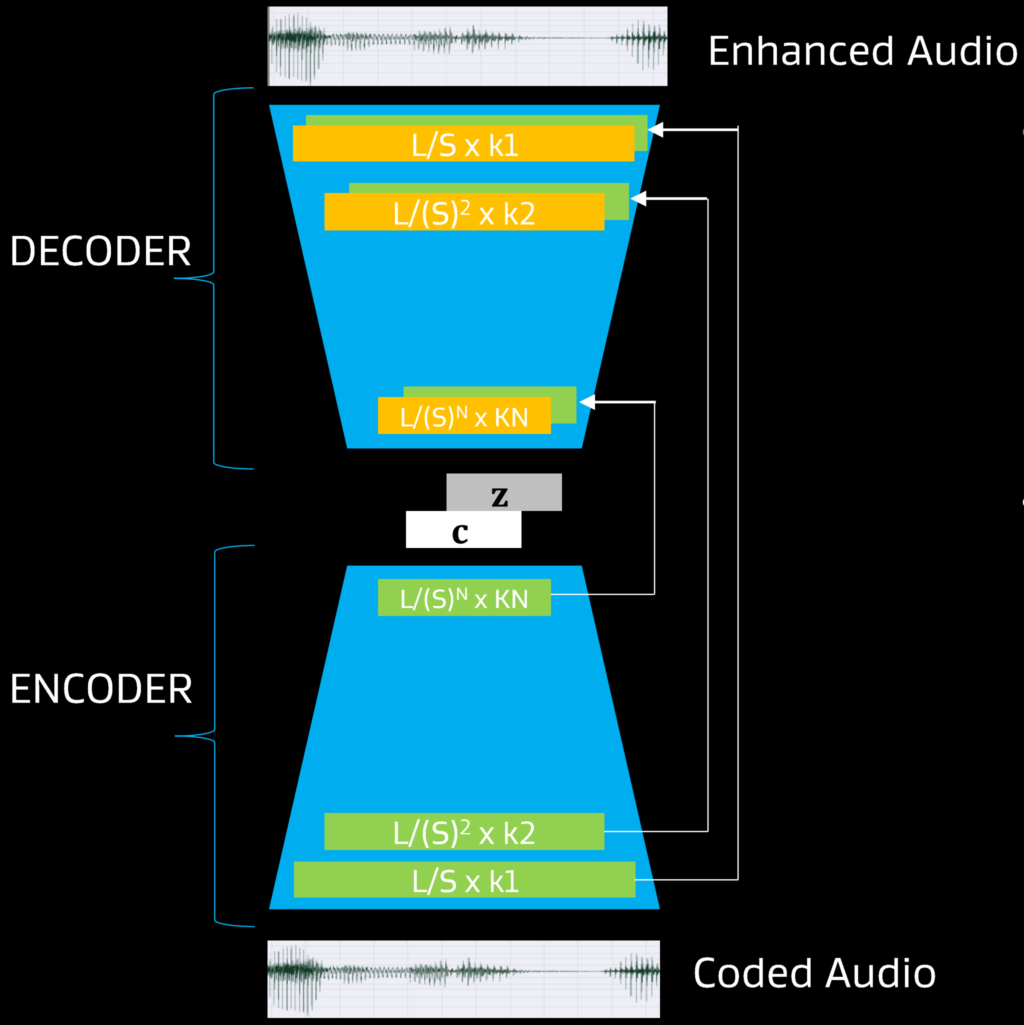
© 2020 DOLBY LABORATORIES, INC.

Audio Codec Enhancement



© 2020 DOLBY LABORATORIES, INC.

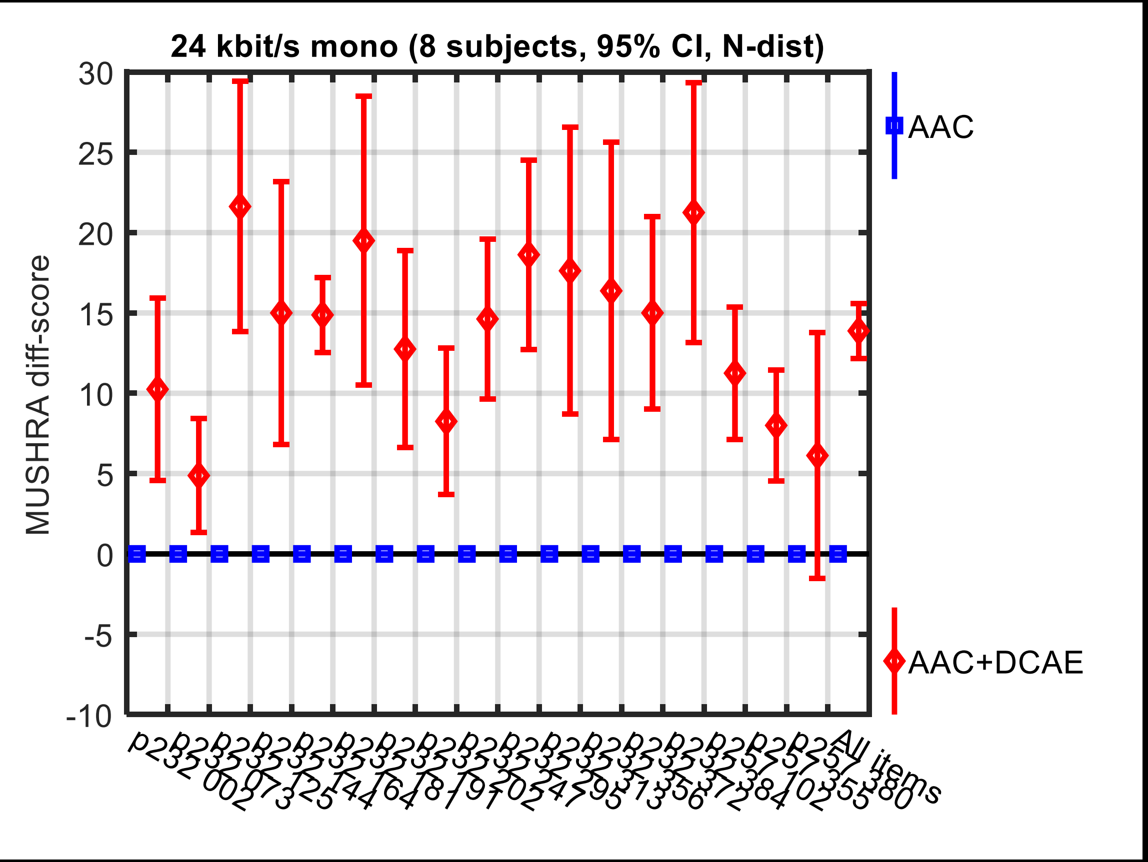
Generator



- 1D fully convolutional auto-encoder with non-linear activations
  - Bottleneck: $\mathbf{c}$
  - $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ concatenated at bottleneck: adds stochastic behavior to generator predictions
- Skip connections
  - Generated audio maintains fine structure of the coded audio

© 2020 DOLBY LABORATORIES, INC.

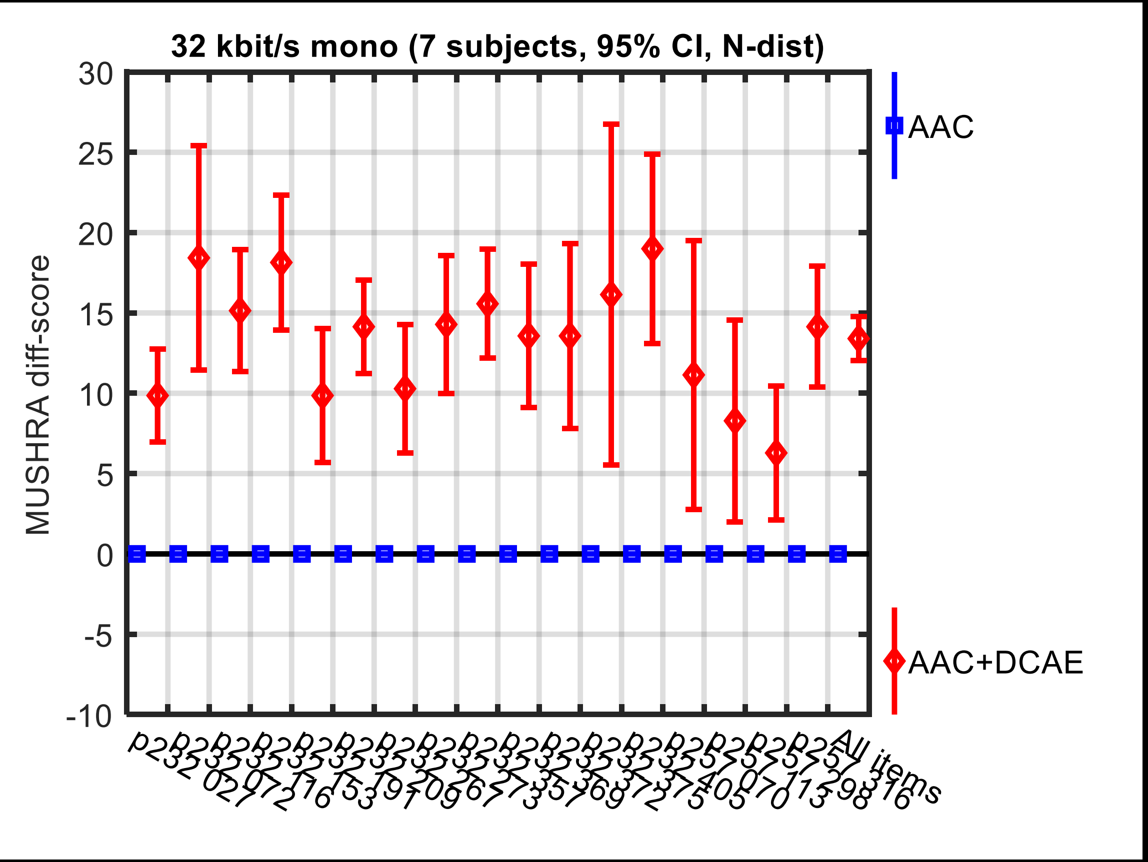
Listening Test – AAC @ 24 kbit/s Mono (Speech – VCTK Test Set)



Trained on VCTK training set: 28 speakers (14 male, 14 female) with mix of regional English accents

© 2020 DOLBY LABORATORIES, INC.

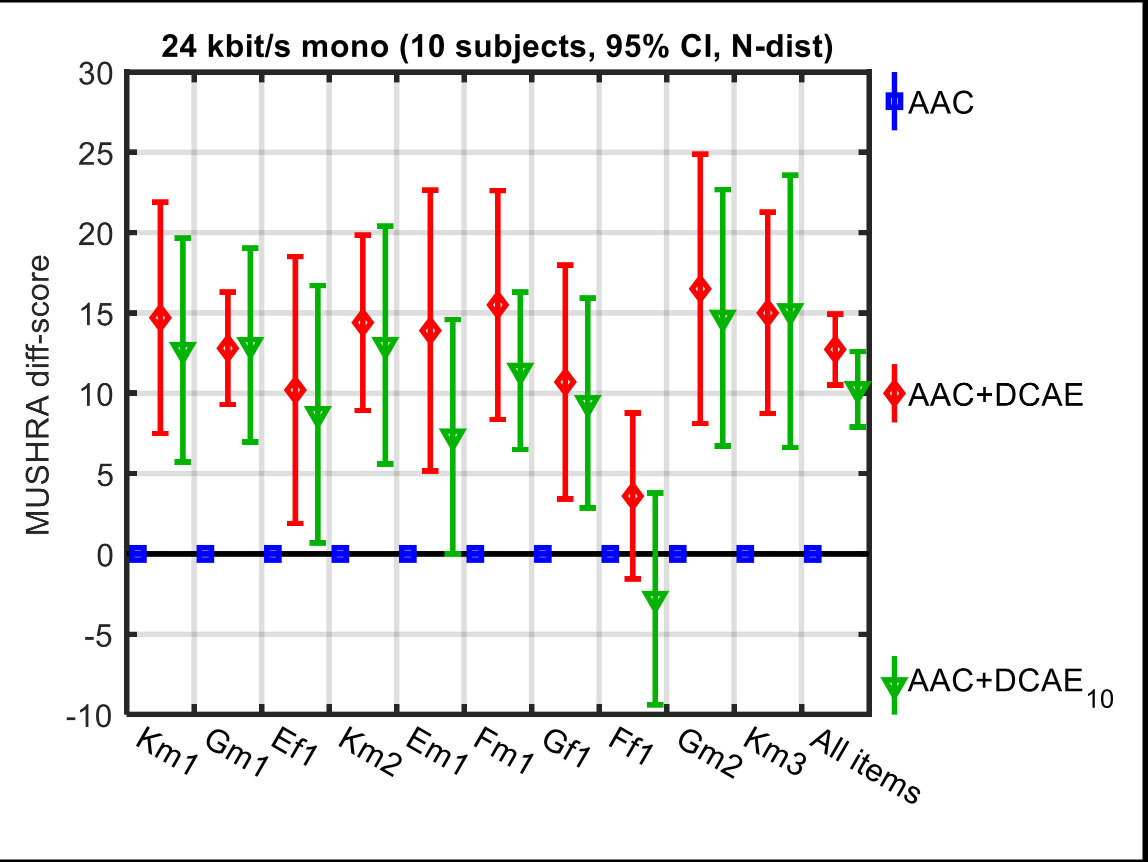
Listening Test – AAC @ 32 kbit/s Mono (Speech – VCTK Test Set)



Trained on VCTK training set: 28 speakers (14 male, 14 female) with mix of regional English accents

© 2020 DOLBY LABORATORIES, INC.

Listening Test – AAC @ 24 kbit/s Mono (Speech – Out-of-Domain Test Set)

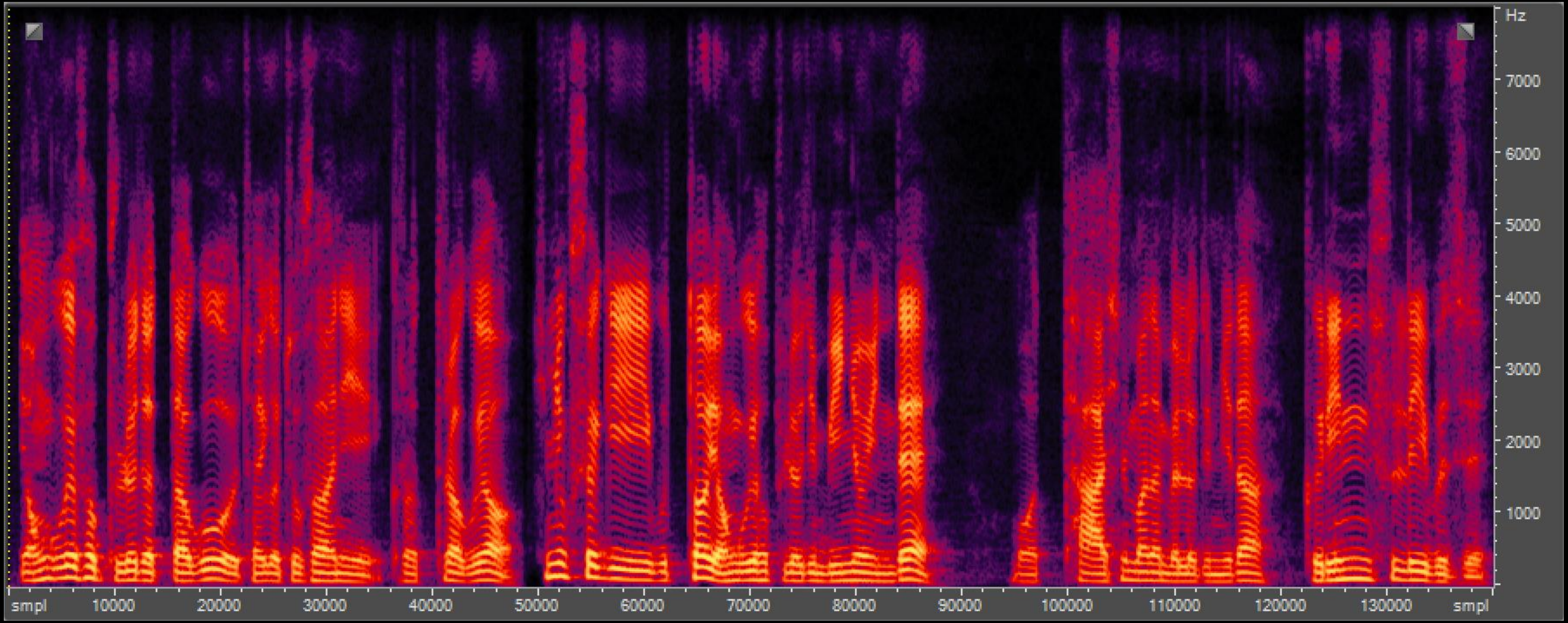


$DCAE_{10}$ is a smaller model

Trained on VCTK training set: 28 speakers (14 male, 14 female) with mix of regional English accents

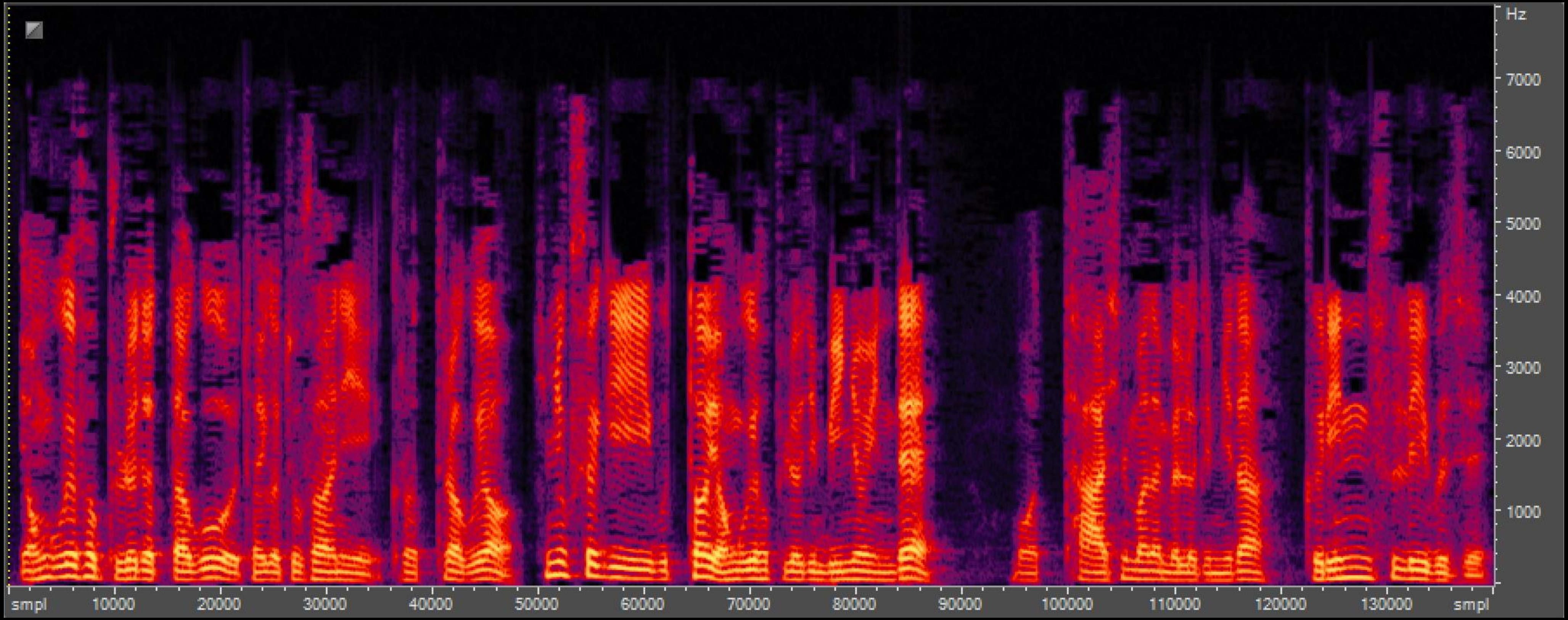
© 2020 DOLBY LABORATORIES, INC.

Unencoded Speech



© 2020 DOLBY LABORATORIES, INC.

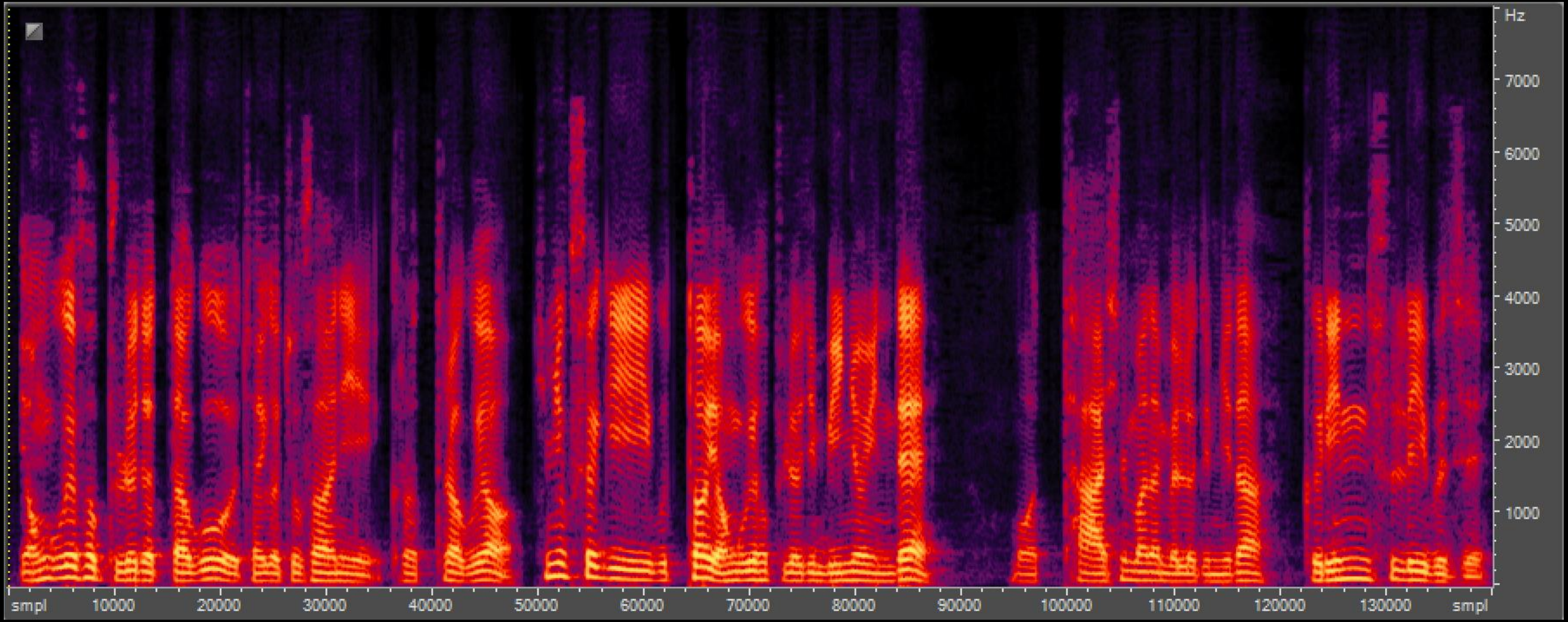
AAC @ 24 kbit/s



Note the spectral gaps

© 2020 DOLBY LABORATORIES, INC.

AAC @ 24 kbit/s + DCAE



Note spectro-temporal noise shaping + spectral gap filling

© 2020 DOLBY LABORATORIES, INC.

Coded Applause Enhancement

- Prepared an in-house applause dataset: includes 4 hours of applause snippets with high perceptual entropy (→ high coding difficulty)

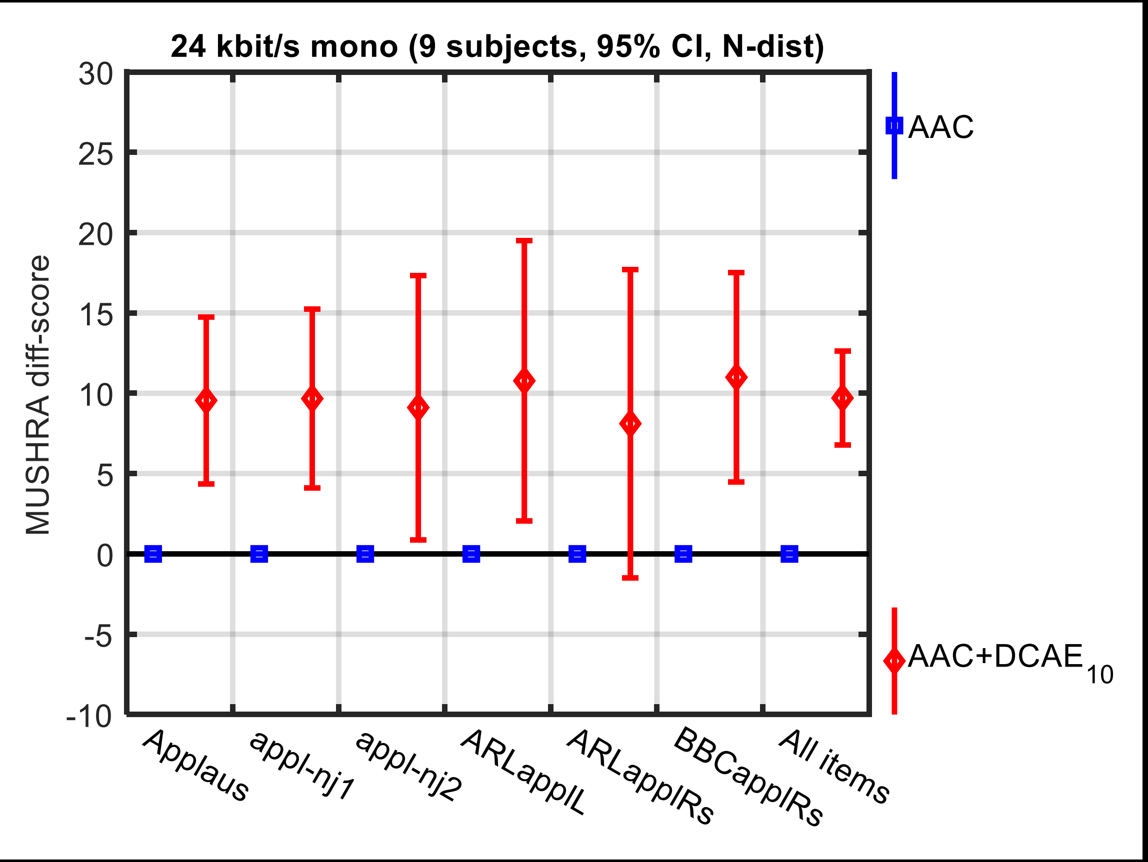
- Tricky to balance transient prominence without making applause signals sound “dry/artificial”
- Solution:
  - Decay λ from early epochs (as soon as GAN training has stabilized)

$$\mathcal{L}_G = \frac{1}{2}\mathbb{E}_{\mathbf{z}\sim p_z(\mathbf{z}),\tilde{\mathbf{x}}\sim p_{data}(\tilde{\mathbf{x}})}[(D(\mathbf{x}^*,\tilde{\mathbf{x}})-1)^2] + \lambda\|\mathbf{x}^* - \mathbf{x}\|_1$$

  - Intuition: incorporate a little bit more stochastic behavior in the generator output to make use of the noise latent z

© 2020 DOLBY LABORATORIES, INC.

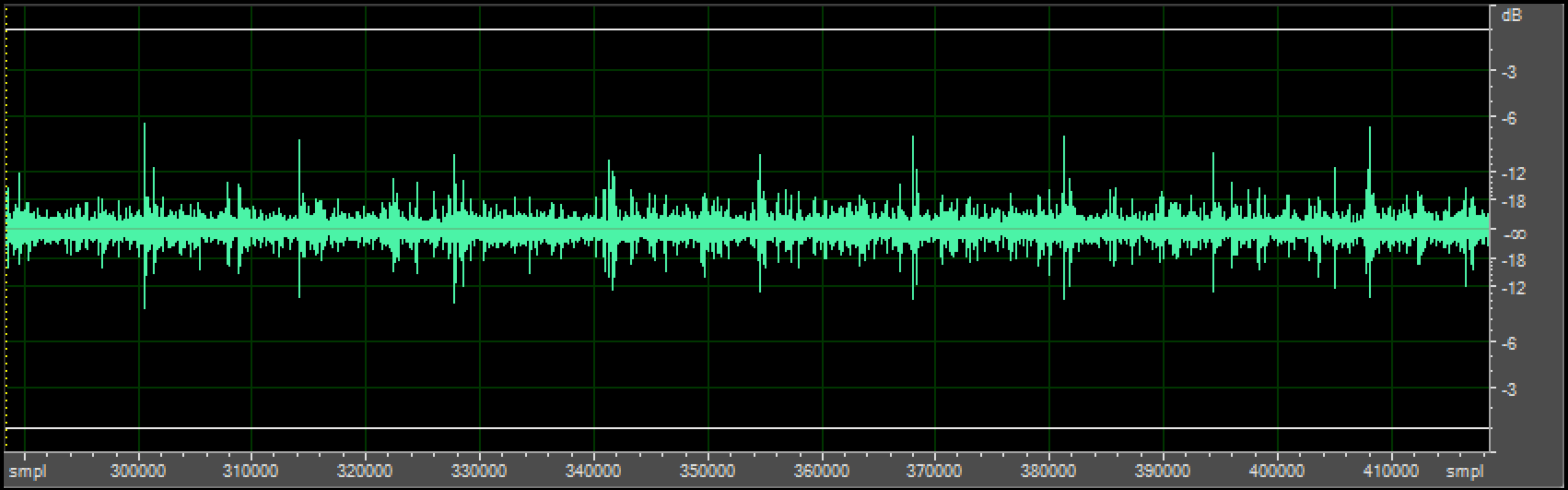
Listening Test – AAC @ 24 kbit/s Mono (Applause – In-house Test Set)



Trained on in-house applause dataset: includes 4 hours of applause snippets with high perceptual entropy

© 2020 DOLBY LABORATORIES, INC.

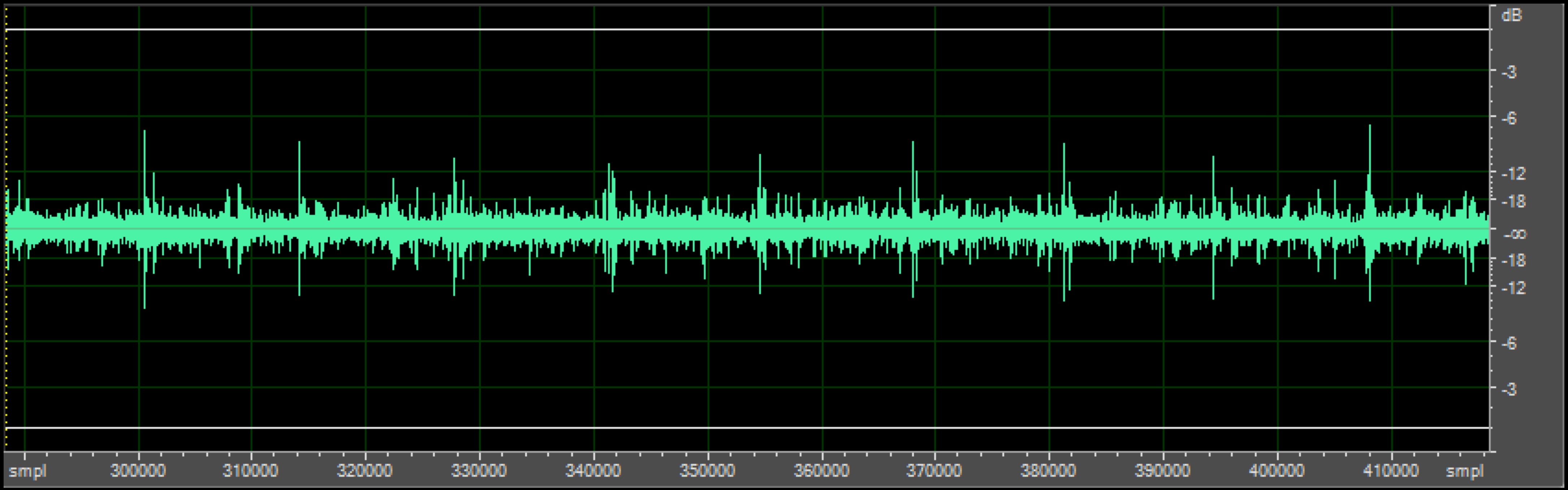
Unencoded Applause



Section from middle of the “Applaus” excerpt

© 2020 DOLBY LABORATORIES, INC.

AAC @ 24 kbit/s

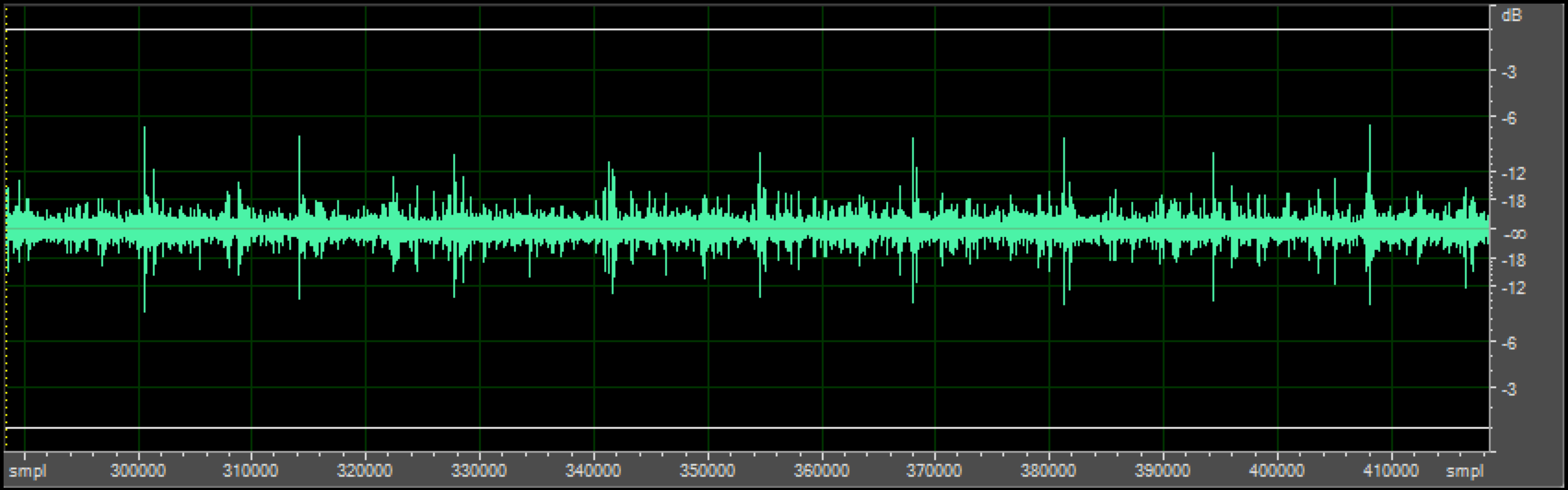


Note noise in between transients are slightly amplified and transients are slightly attenuated

© 2020 DOLBY LABORATORIES, INC.

AAC @ 24 kbit/s + $DCAE_{10}$

Model simply performed a transient-to-noise ratio restoration



Transients and noise are very slightly (between 0 and 1 dB) amplified and attenuated, respectively

© 2020 DOLBY LABORATORIES, INC.

Conclusions

- Proposed GAN-based coded audio enhancer
- Demonstrated significant quality improvement for coded speech and applause signals
- Provides one-shot enhancement
    - Un-optimized PyTorch implementation of our best performing model for speech and applause runs at 5x and 7x real-time, respectively, on a CPU.

© 2020 DOLBY LABORATORIES, INC.