

@Codec-SUPERB, Dec. 3, 2024

Future Directions in Neural Speech Communication Codecs

Minje Kim, Ph.D.

Associate Professor, Siebel School of Computing and Data Science
Visiting Academic at Amazon Lab126

<https://minjekim.com>

minje@illinois.edu



UNIVERSITY OF
ILLINOIS
URBANA - CHAMPAIGN

SIEBEL SCHOOL OF COMPUTING AND DATA SCIENCE
GRAINGER ENGINEERING

Introduction

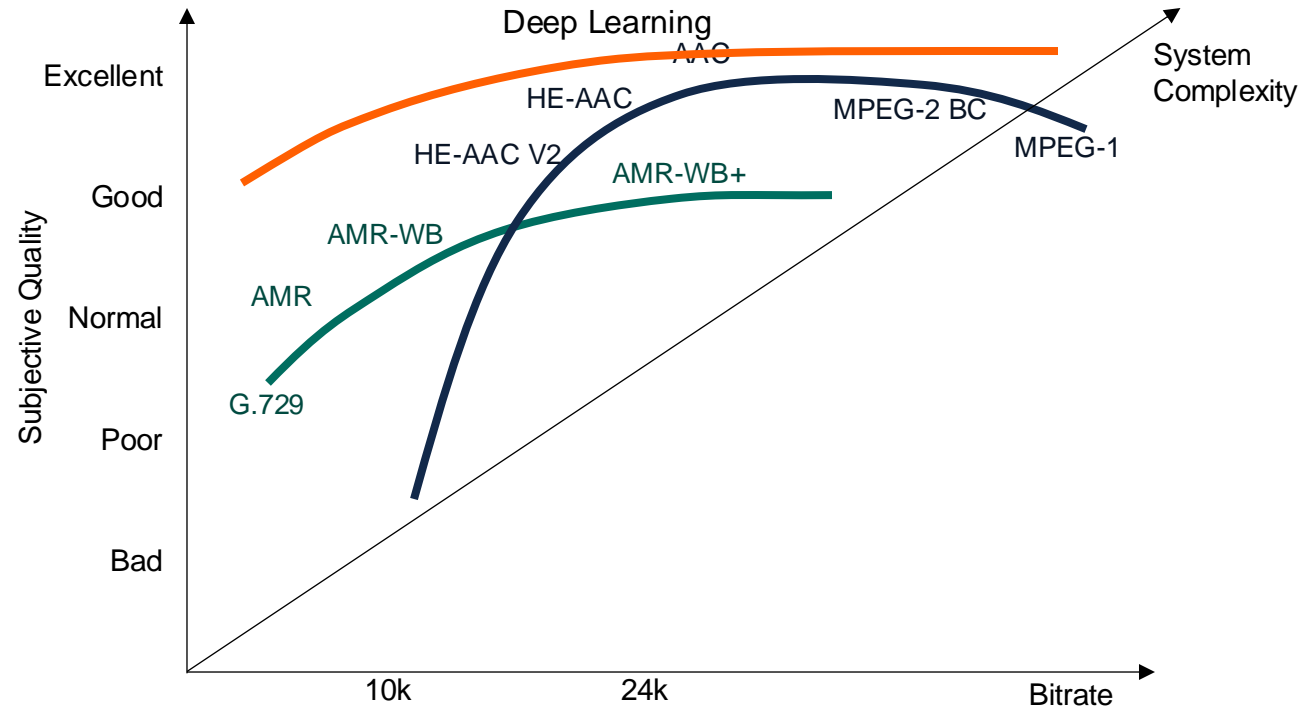
- Deep learning for speech/audio coding

○ Why deep learning-based coding?

- Can expect better coding gain
 - Yes, but not too straightforward
- Can harmonize with other neural nets
 - Speech enhancement, ASR, TTS, etc

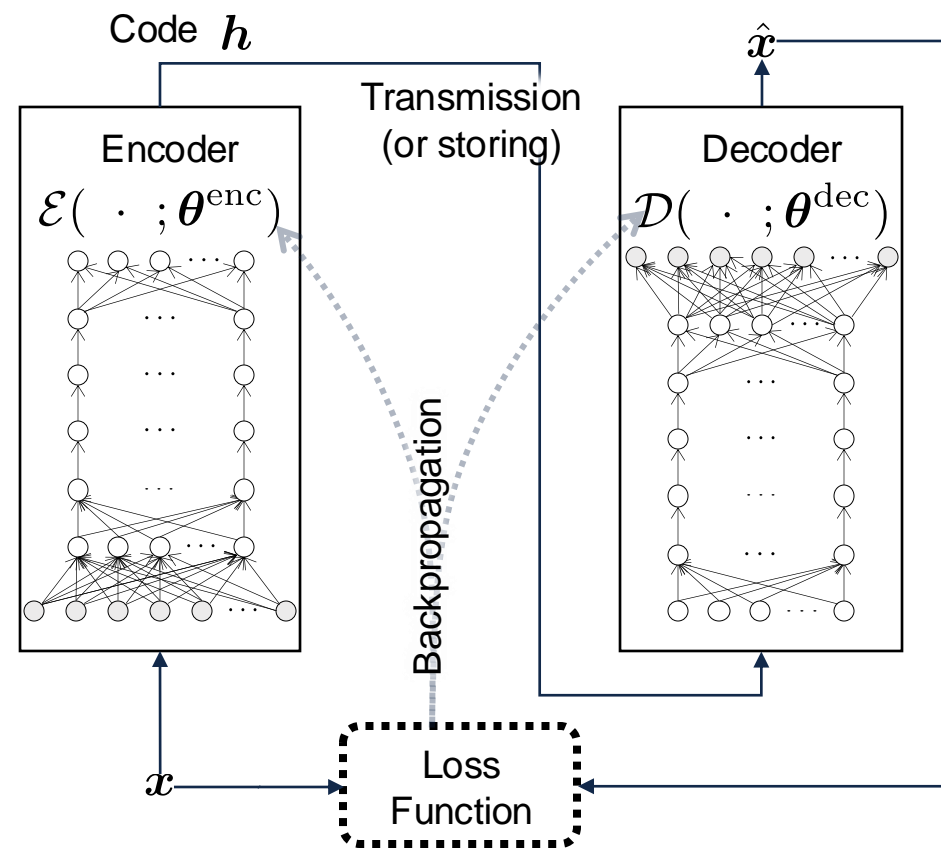
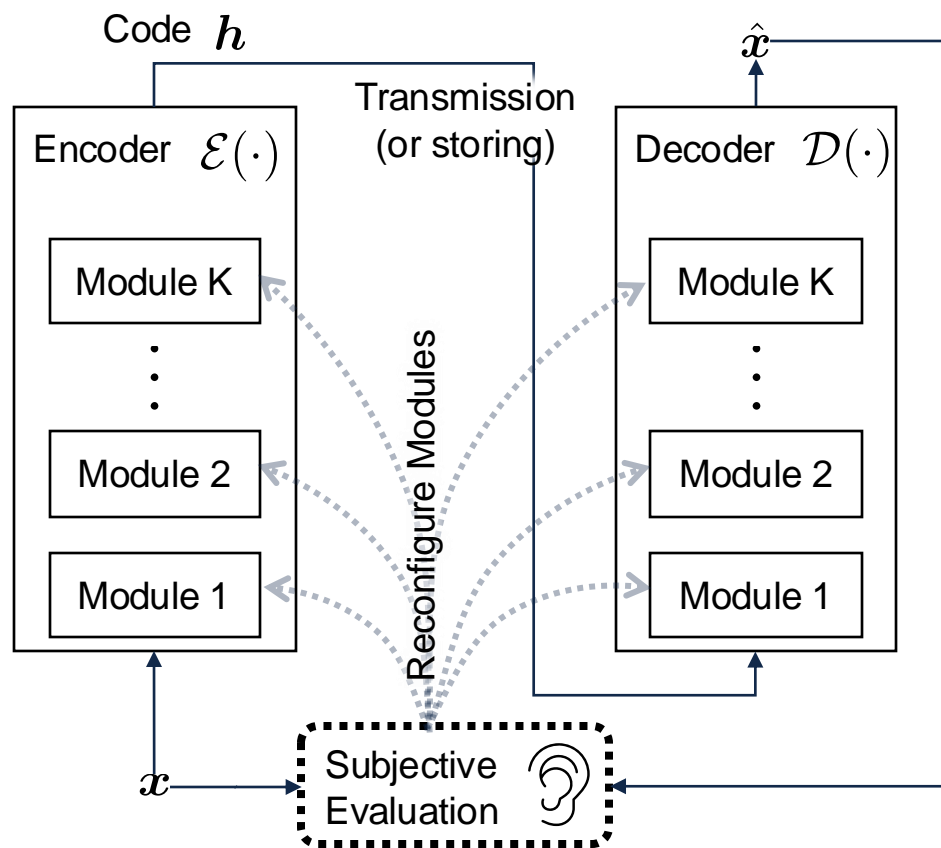
○ Major limitations

- Complexity/delay
- Perceptual loss
- Scalability
- ...



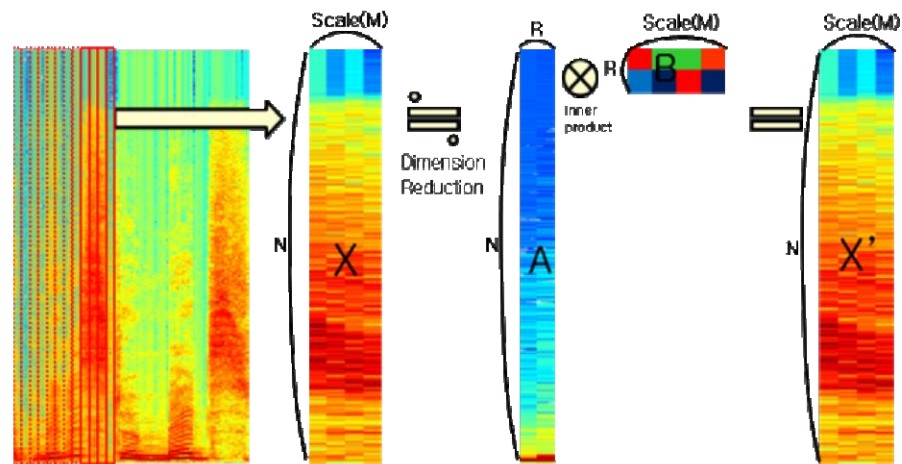
Autoencoders for Audio Coding

- vs. traditional codecs

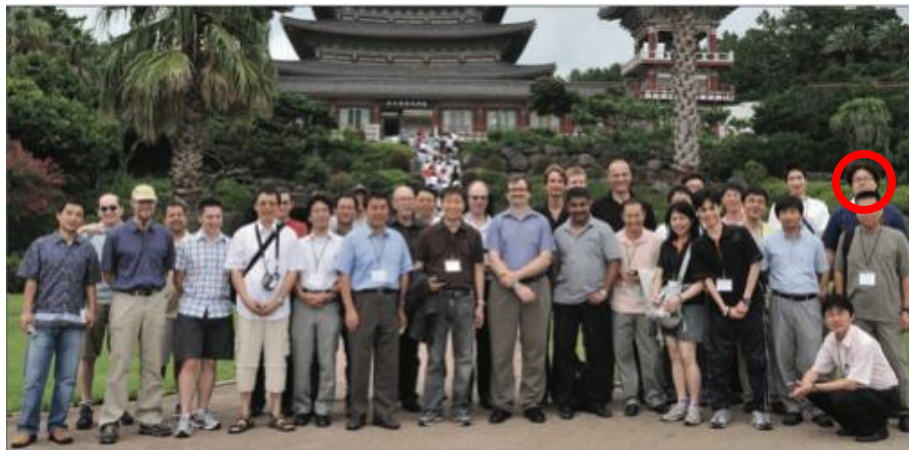


Autoencoders for Audio Coding

- Not a new idea



Non-overlapped sliding window



Audio Engineering Society
AES 34th CONFERENCE, 2008
Jeju Island, Korea

General Information

Program

Registration

Venue and Accommodation

Local Information

Program

HOME > Program > Technical Sessions

CAS2-3

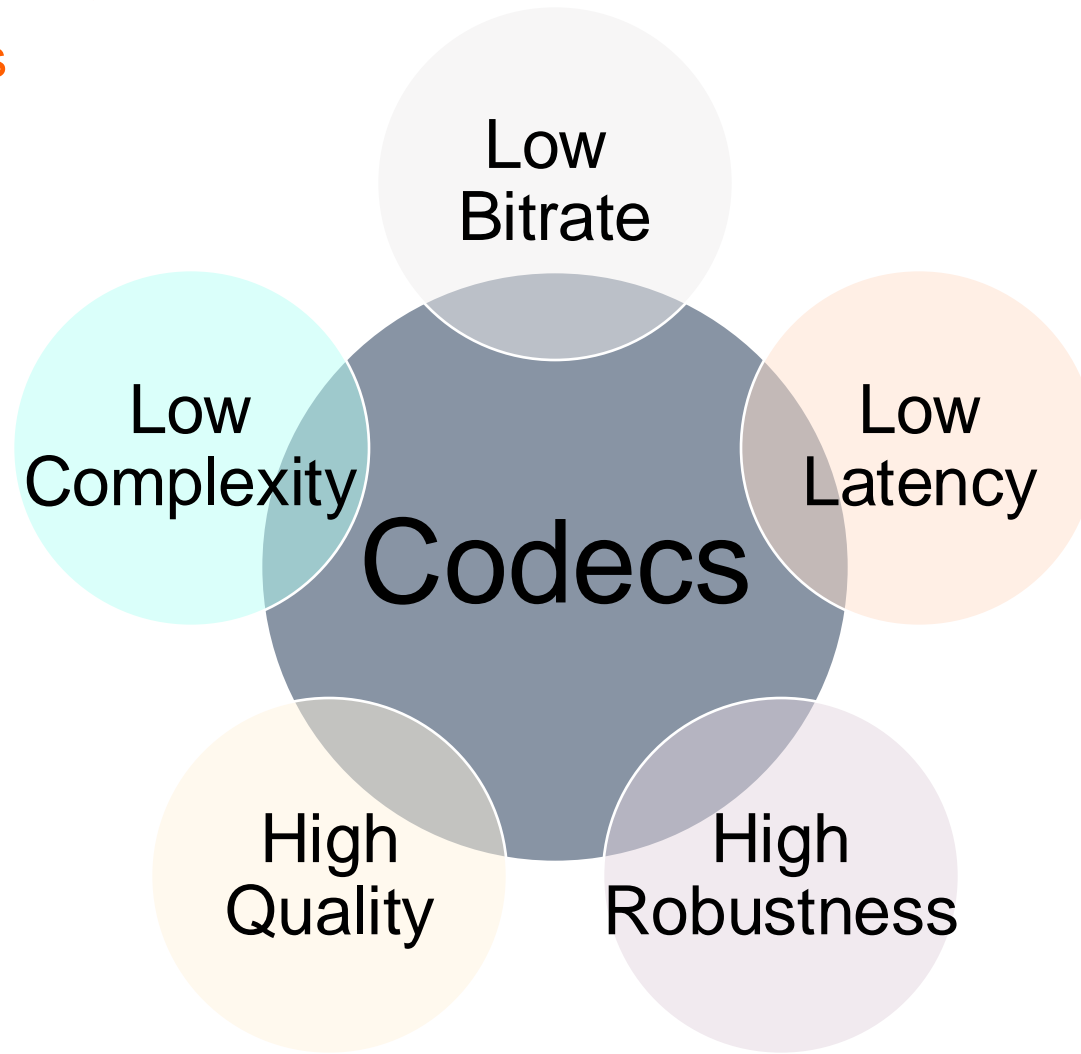
Segmented Dimensionality Reduction Coding on Frequency Domain Signal

Minje Kim, Seungkwon Beack, Taejin Lee, Daeyoung Jang, Kyeongok Kang, Electronics and Telecommunications Research Institute (ETRI), Daejeon, Korea

This paper proposes schemes of compressing frequency domain acoustic signals using dimensionality reduction methods. Dimensionality reduction methods which work on a two-dimensional matrix usually result in high compression ratio since they not only allow us to represent the input matrix with smaller amount of data, but exploit intrinsic information of the original data. Frequency domain signals can be seen as a (number of frequency bands) (number of total frames) input matrix of dimensionality reduction methods. However, in this case, real-time encoding is not possible and encoder-side delay is inevitable which amounts to the length of whole input signal. To minimize the delay this paper proposes a coding scheme which conducts multiple dimensionality reduction on segments of input data frames serially.

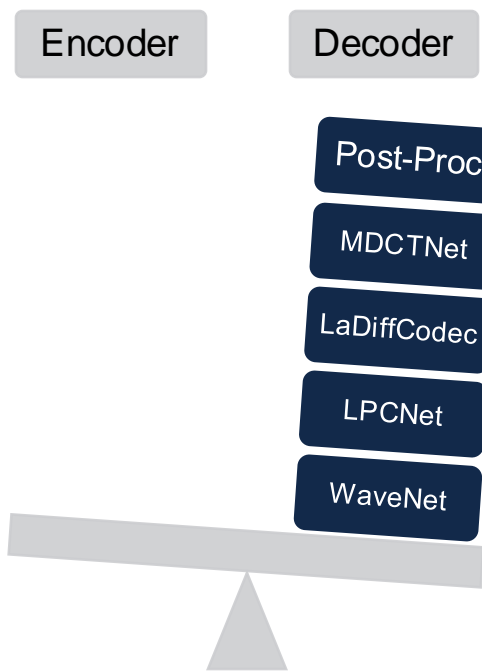
What Is a Good Codec?

- Conflicting attributes

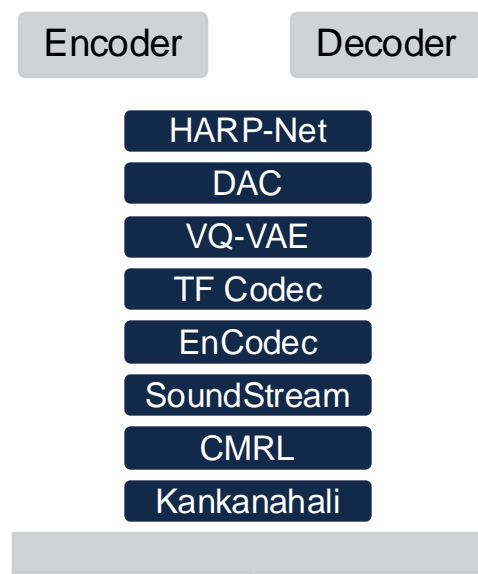


What else?

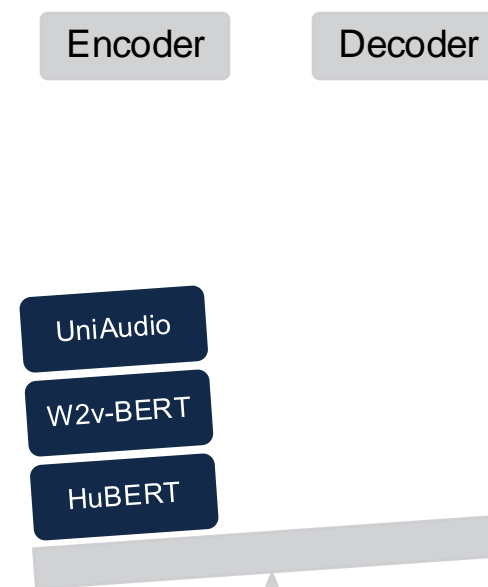
Neural Codecs as Encoder-Decoder Models



- Generative models
- Tradeoff b/w quality, complexity, and bitrate



- Flexible
- Moderate complexity
- Versatile



- Good for other downstream apps

- Psychoacoustics
- Residual Learning
- Predictive Models
- Generative Models
- Source Separation
- Foundational Models
- Personalization
- Language Model

Temporal Prediction of Coded Features

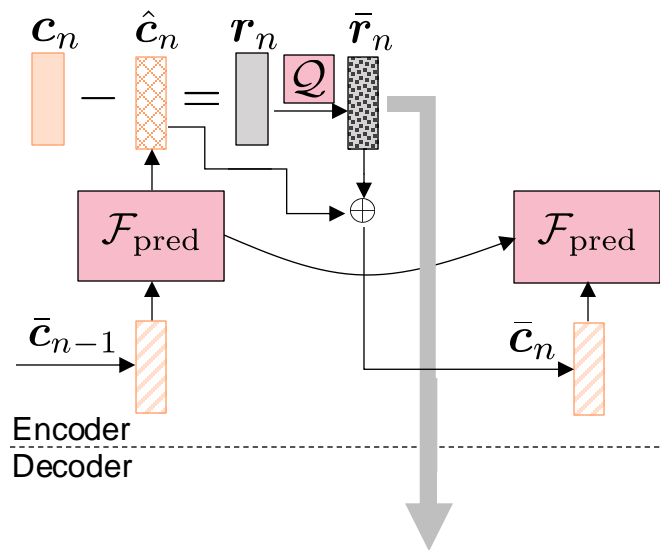
- Feature predictive coding

- Feature prediction function

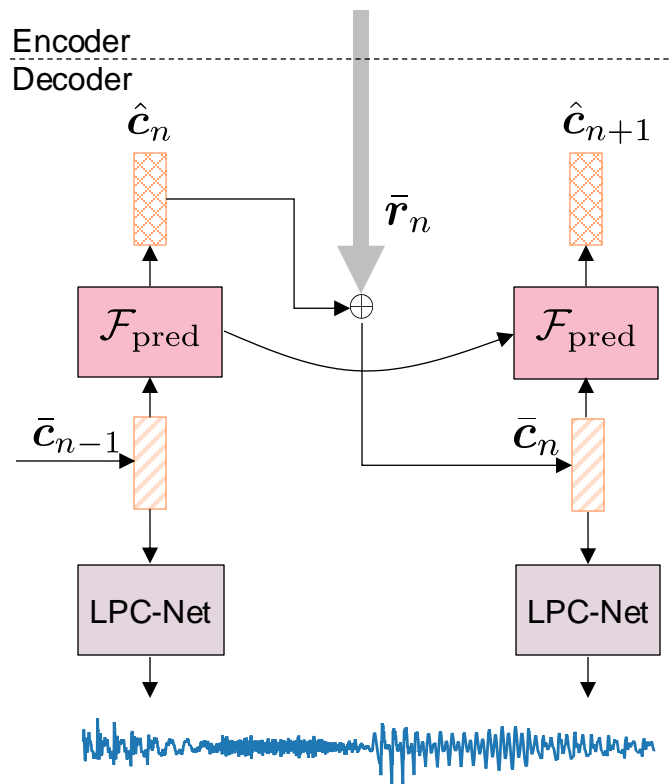
$$\mathcal{F}_{\text{pred}} : \mathbf{C}_{<n} \mapsto \mathbf{C}_n$$

(GRU)

- Encoder



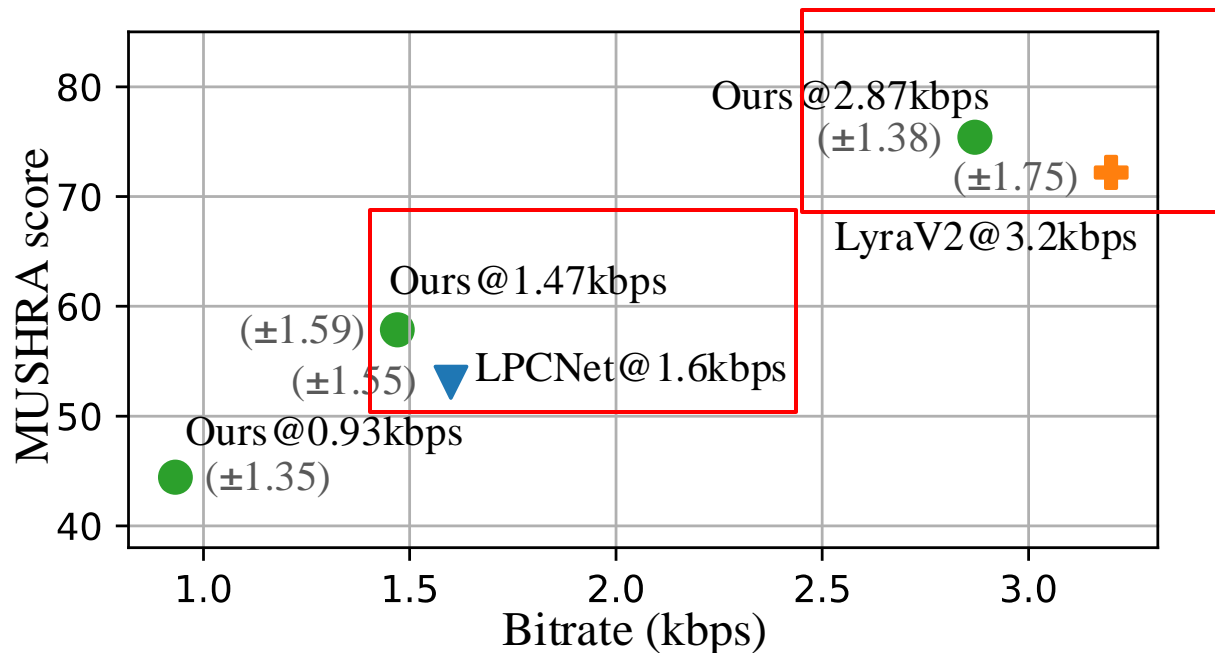
- Decoder



Temporal Prediction of Coded Features

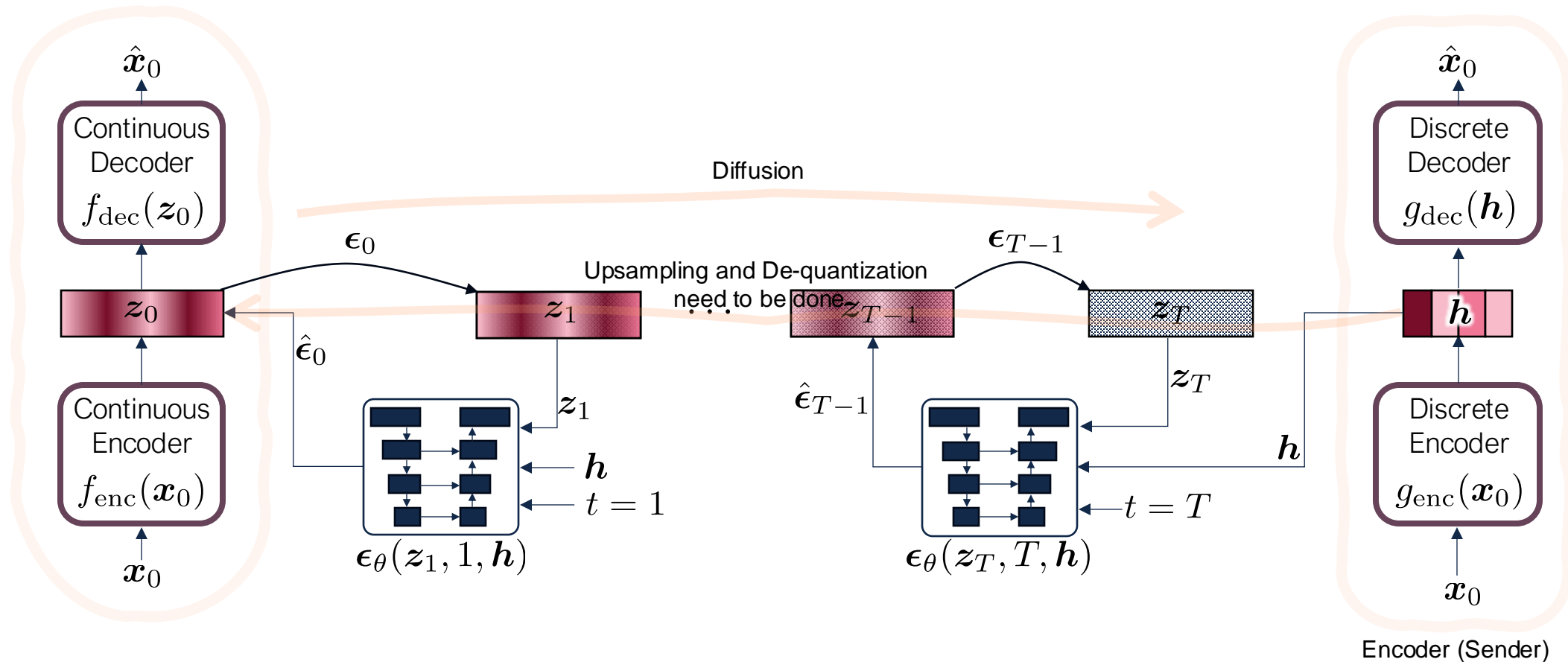
- Subjective Test

- Our model at :
 - 0.93 kbps
 - 1.47 kbps
 - 2.87 kbps
- Baseline models :
 - LPCNet @ 1.6kbps
 - LyraV2 @ 3.2kbps



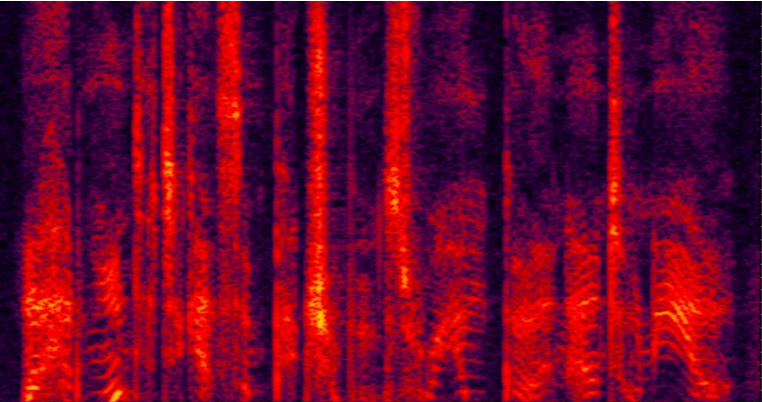
Generative De-Quantization via Latent Diffusion

- The proposed LaDiffCodec architecture

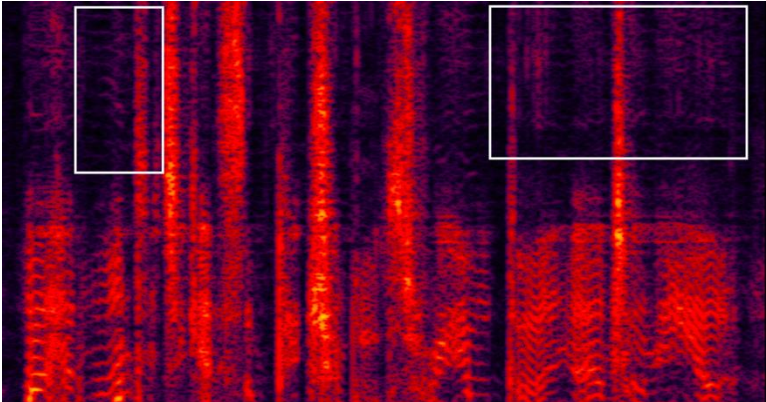


Generative De-Quantization via Latent Diffusion

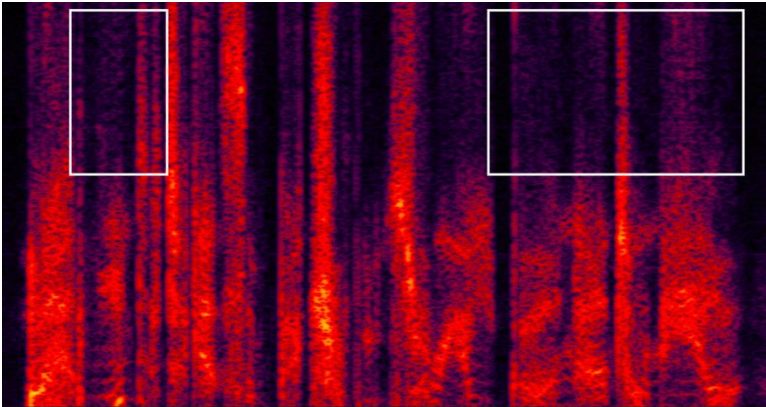
- Results



Reference



EnCodec 3kbps



LaDiffCodec 3kbps



Original



LaDiff Codec



Generative De-Quantization via Latent Diffusion

- Results

- LaDiffCodec prefers large continuous latent dimension

- At no cost of increased BR

Strides	@1kbps	@1.5kbps	@3kbps
[1]	1.18 \pm 0.04	1.20 \pm 0.04	1.77 \pm 0.19
[8]	1.81 \pm 0.15	1.95 \pm 0.15	2.23 \pm 0.17
[4, 8]	1.71 \pm 0.71	2.19 \pm 0.75	2.16 \pm 0.69
[4, 5, 8]	1.66 \pm 0.11	1.71 \pm 0.12	1.84 \pm 0.10
[2, 4, 5, 8]	1.49 \pm 0.09	1.65 \pm 0.13	1.71 \pm 0.12



Reference

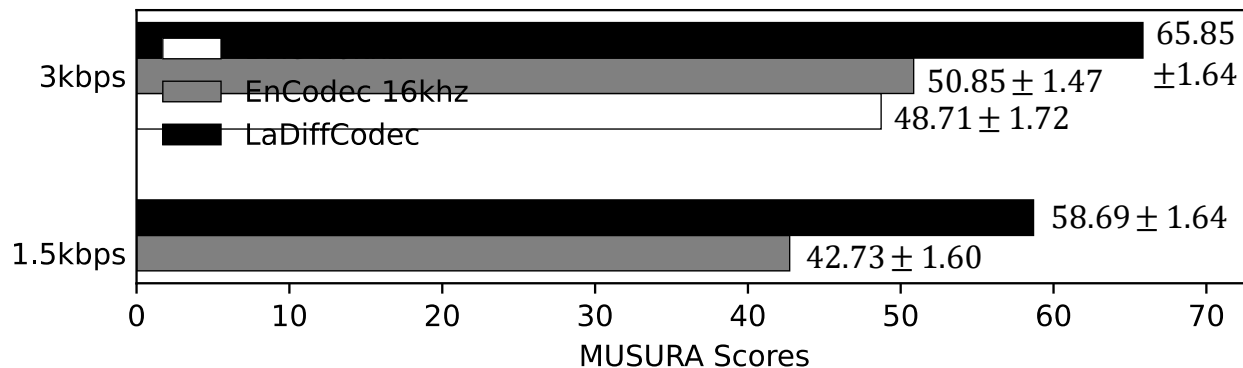


LaDiffCodec 1 kbps
(DDPM 1000 steps)



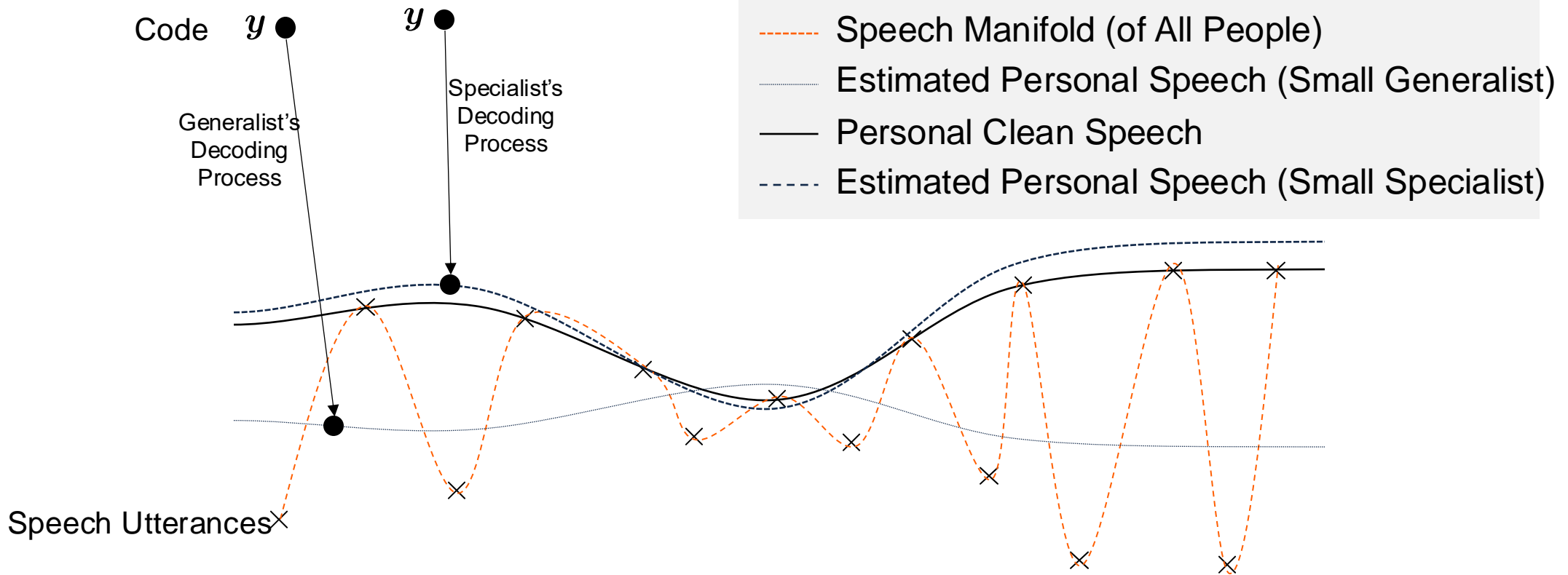
LaDiffCodec 1 kbps
(Midway Infilling 100 steps)

- MUSHRA



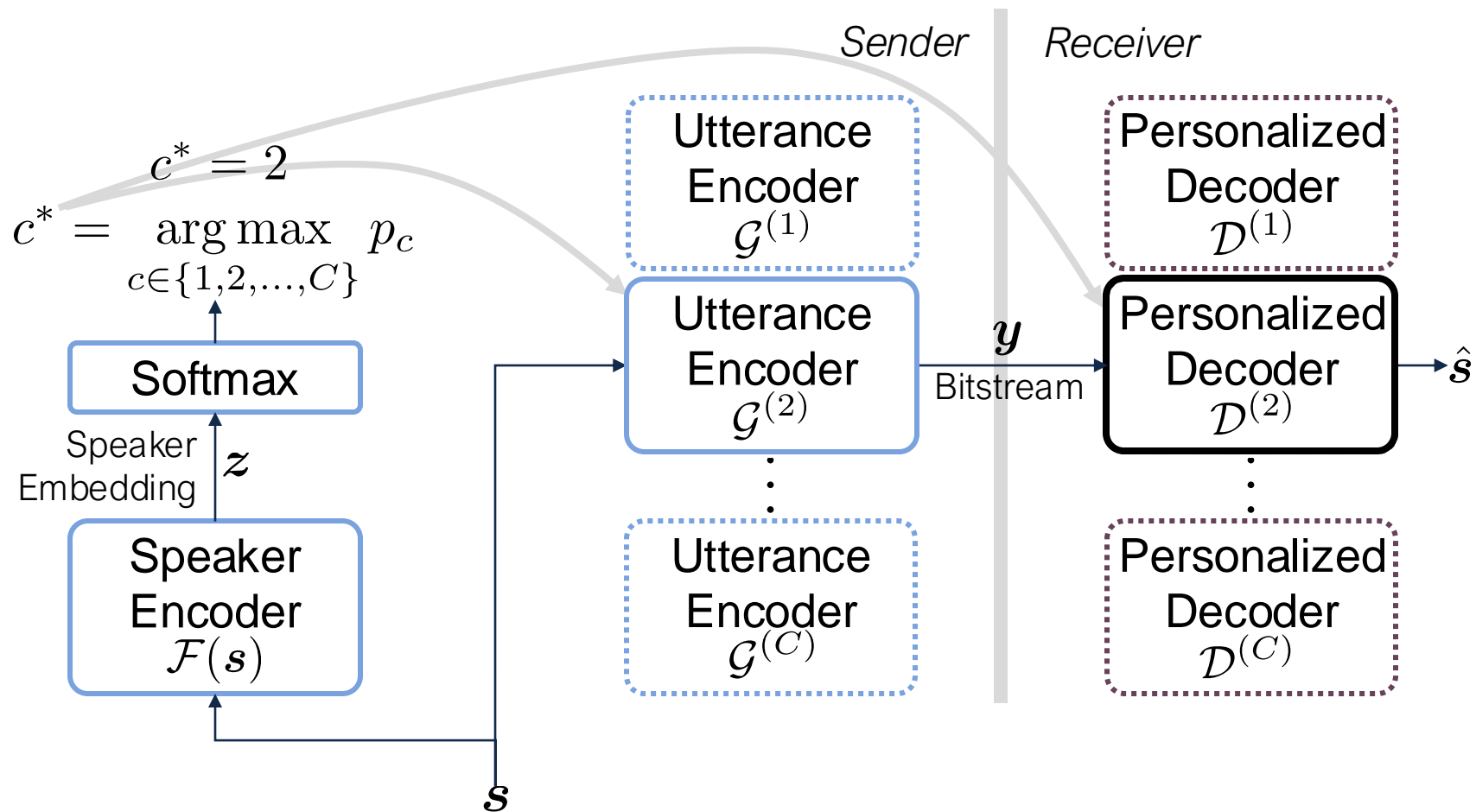
Personalized Neural Speech Codec

- Manifold interpretation



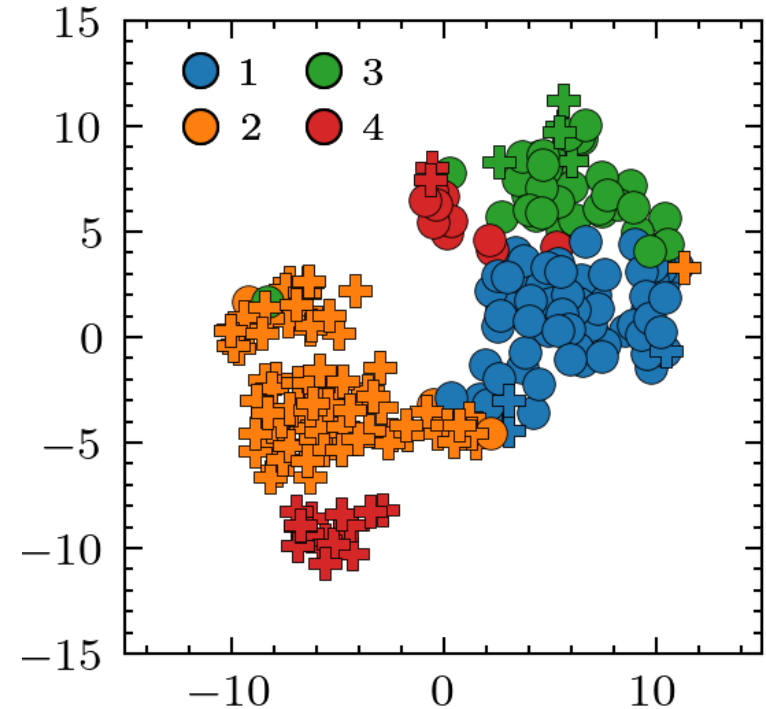
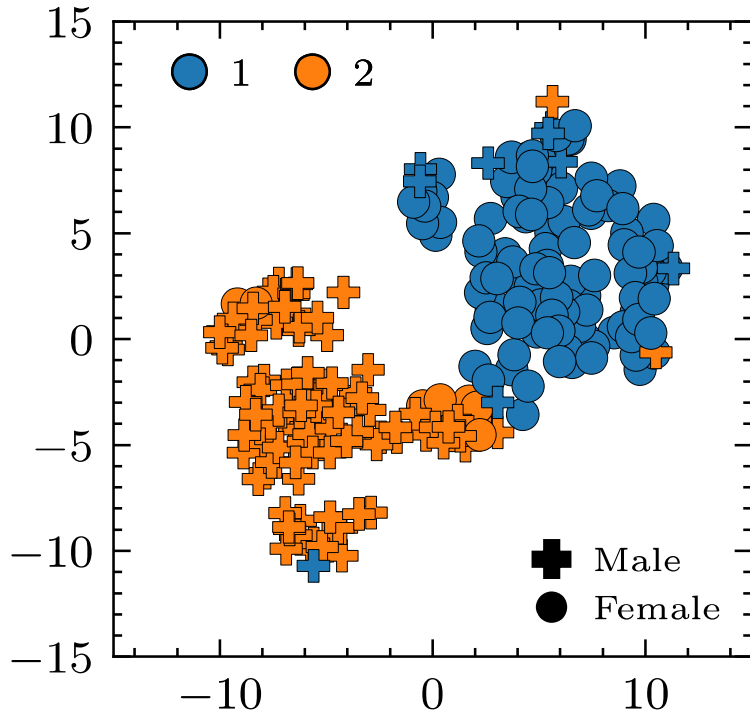
○ Less powerful models still work on the sub-problem!

Personalized Neural Speech Codec



Personalized Neural Speech Codec

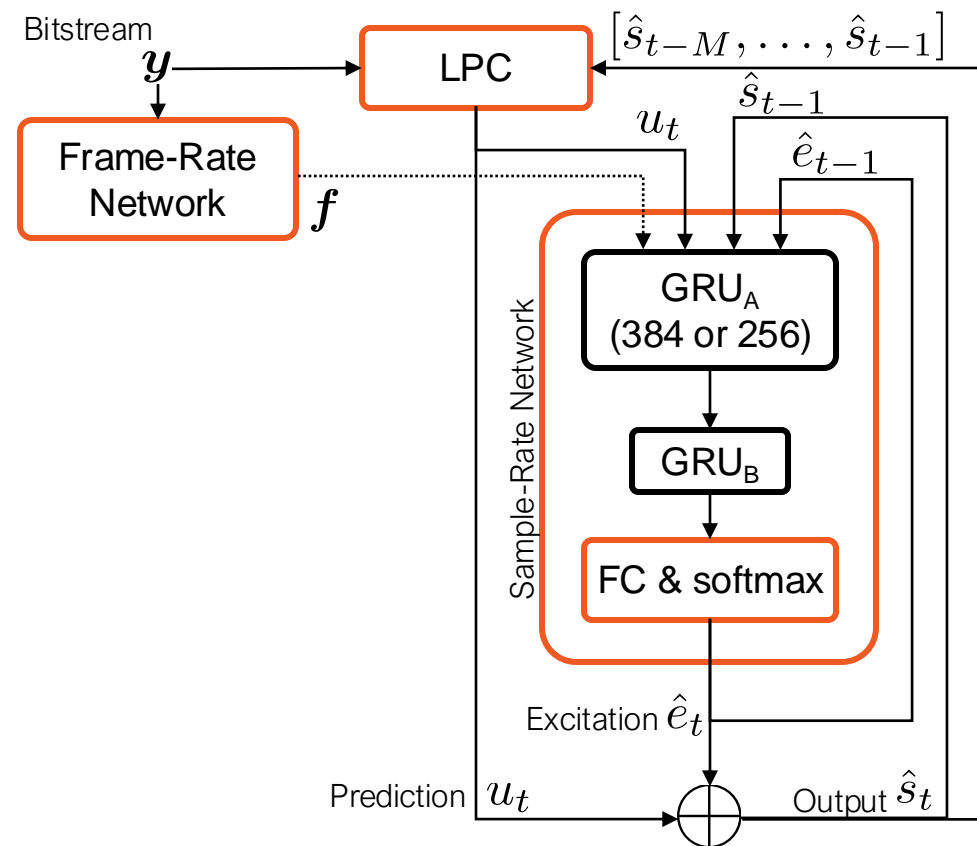
- Speaker groups



Personalized Neural Speech Codec

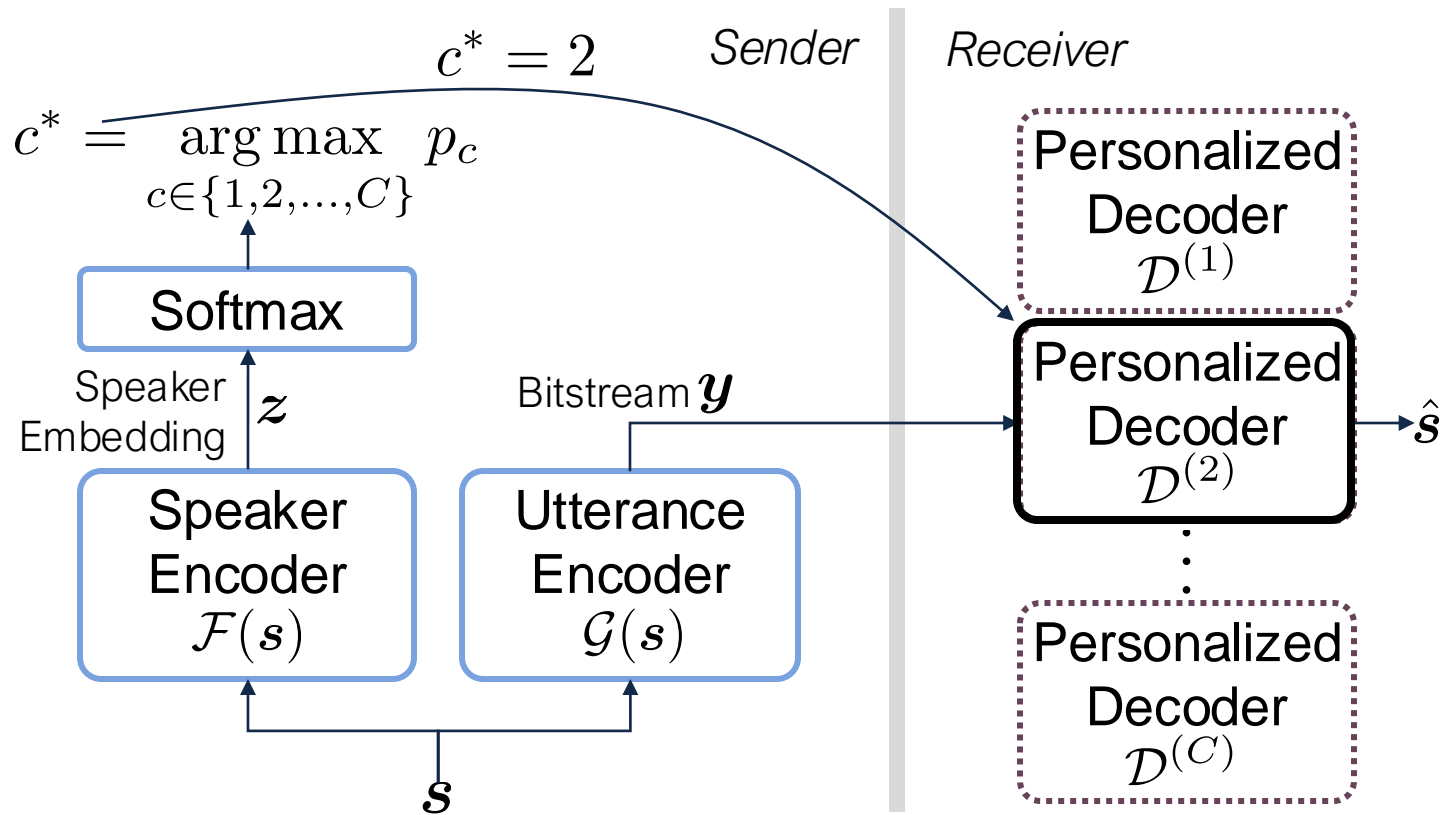
- LPCNet

- Low bitrate speech vocoder (1.6 kbps)
- Speech Features
 - 18 Bark-scale cepstral coefficients
 - 2 pitch parameters (period, correlation)
- A decoder-heavy architecture
 - Frame-rate network
 - Sample rate network
 - WaveRNN
 - Conditioned on processed frame features
 - Autoregressively sample synthesis



Personalized Neural Speech Codec

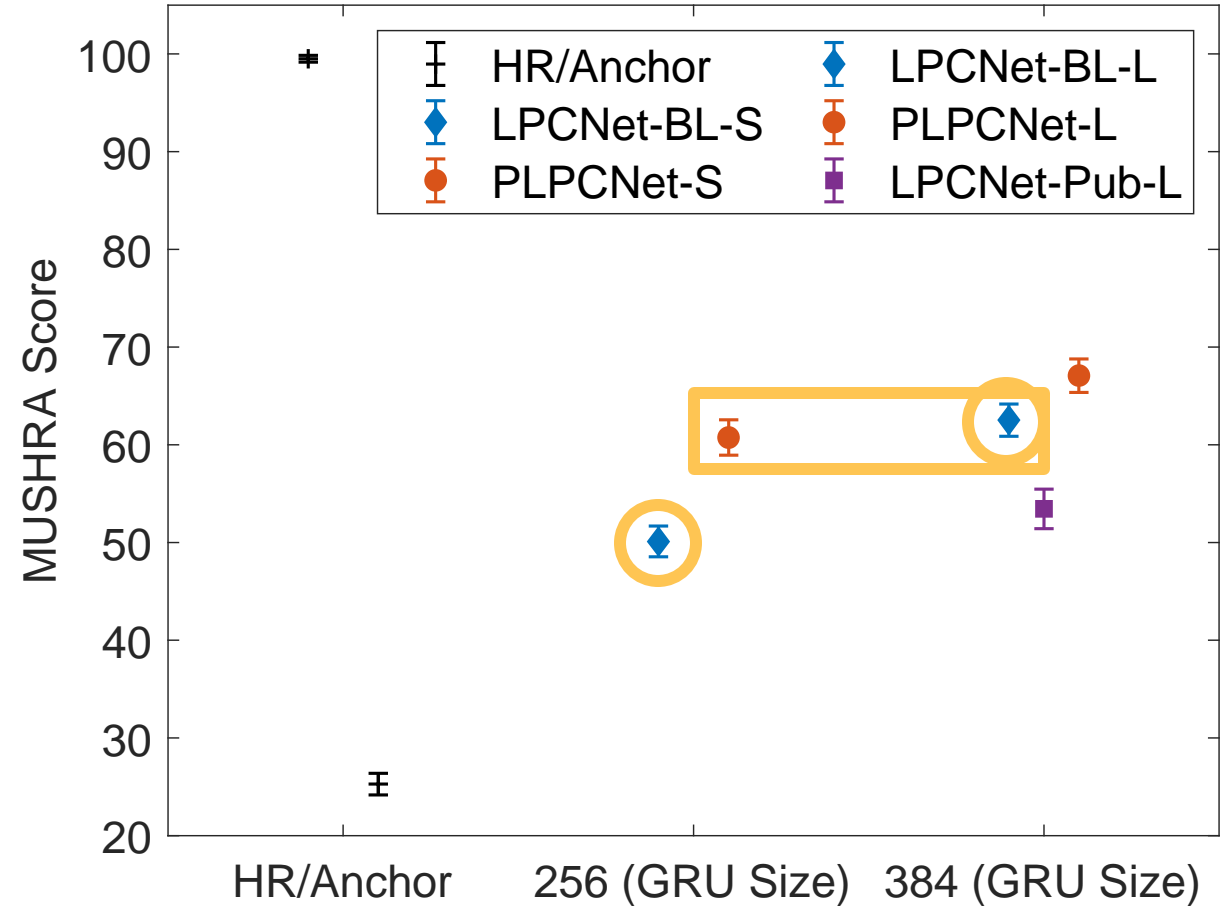
- Personalized LPCNet (1.6 kbps)



Personalized Neural Speech Codec

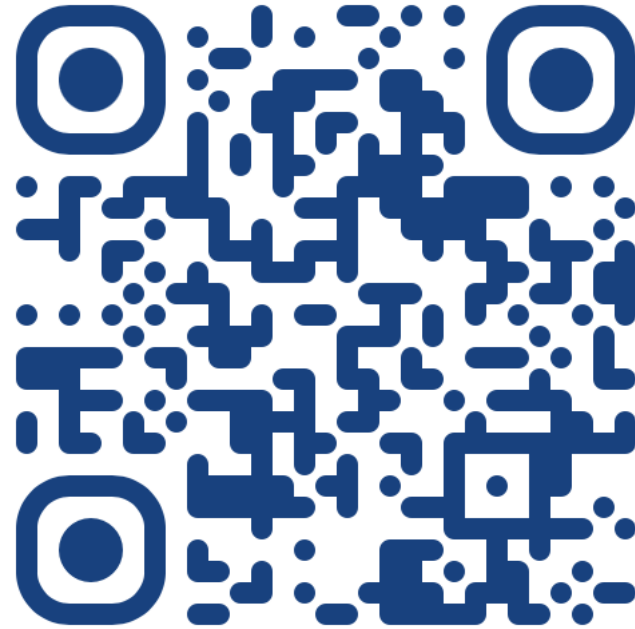
- MUSHRA Test

- Personalized LPCNet outperforms
 - The baseline at the same bitrate
- The small personalized model is on par with the large LPCNet



Recap

- Neural speech codecs are very successful in terms of coding gain
 - Neural audio codecs are getting there
- Deep learning is computationally heavy
 - Speech and audio coding are typically used for real-time communication or streaming applications
 - Foundational models are particularly heavy
- Predictive models, residual learning, and generative models help
- Personalization is a new concept that traditional codecs couldn't handle
- NSAC is versatile
 - Can be combined with other applications
 - From the feature learning perspective
 - Discretized features for the downstream tasks



Minje Kim and Jan Skoglund,
“Neural Speech and Audio Coding,”
IEEE Signal Processing Magazine (to appear)



UNIVERSITY OF
ILLINOIS
URBANA-CHAMPAIGN

Thank You!

(Q&A)

Minje Kim, Ph.D.
<https://minjekim.com>
minje@illinois.edu



Haici Yang



Darius Petermann



Anastasia Kuznetsova