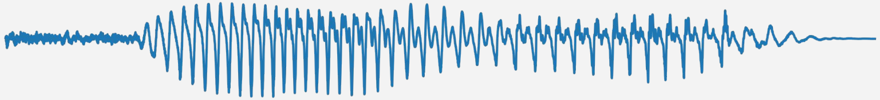
Vector Quantized Neural Networks for Acoustic Unit Discovery



Benjamin van Niekerk, Leanne Nortje, Herman Kamper

The Generative Factors of Speech

HH / Y / UW / M / ER

HUMOUR

Content:

- Discrete phonetic units.
- ≅44 phonemes in English.

Prosody:

- Rhythm
- Intonation
- Stresses

Timbre:

- Quality of a particular voice.
- Characterized by frequency spectrum.

The Generative Factors of Speech

HH / Y / UW / M / ER

HUMOUR

Content:

- Discrete phonetic units.
- ≅44 phonemes in English.

Prosody:

- Rhythm
- Intonation
- Stresses

Timbre:

- Quality of a particular voice.
- Characterized by frequency spectrum.

The Generative Factors of Speech

HH / Y / UW / M / ER

HUMOUR

Content:

- Discrete phonetic units.
- ≅44 phonemes in English.

Prosody:

- Rhythm
- Intonation
- Stresses

Timbre:

- Quality of a particular voice.
- Characterized by frequency spectrum.

The Generative Factors of Speech

HH / Y / UW / M / ER

HUMOUR

Content:

- Discrete phonetic units.
- ≅44 phonemes in English.

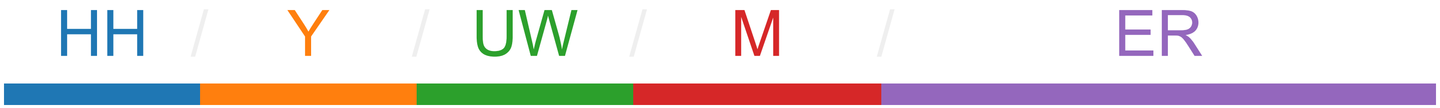
Prosody:

- Rhythm
- Intonation
- Stresses

Timbre:

- Quality of a particular voice.
- Characterized by frequency spectrum.

The Generative Factors of Speech



HH / Y / UW / M / ER

Content:

- Discrete phonetic units.
- ≅44 phonemes in English.

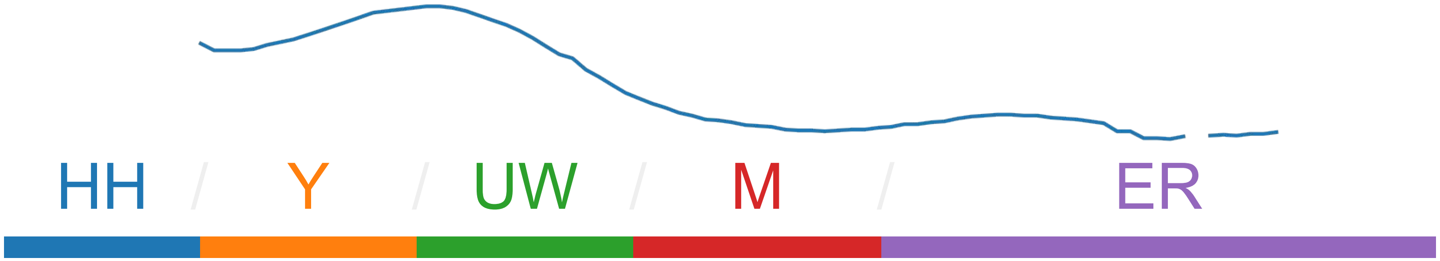
Prosody:

- Rhythm
- Intonation
- Stresses

Timbre:

- Quality of a particular voice.
- Characterized by frequency spectrum.

The Generative Factors of Speech



Content:

- Discrete phonetic units.
- ≅44 phonemes in English.

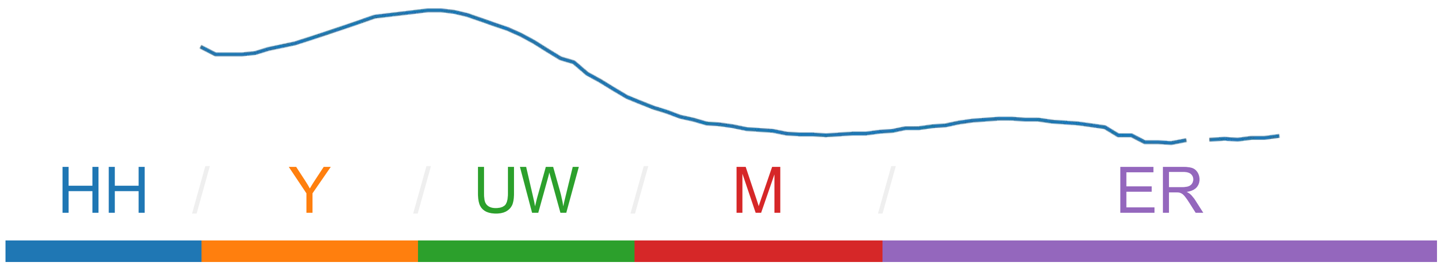
Prosody:

- Rhythm
- Intonation
- Stresses

Timbre:

- Quality of a particular voice.
- Characterized by frequency spectrum.

The Generative Factors of Speech



Content:

- Discrete phonetic units.
- ≅44 phonemes in English.

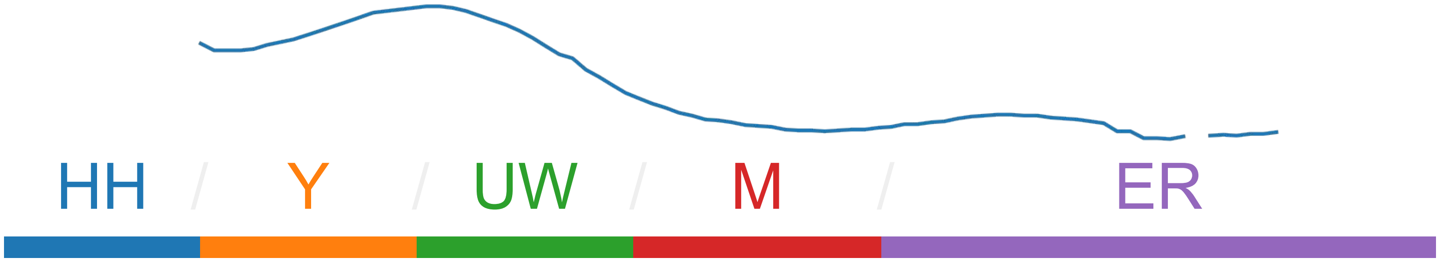
Prosody:

- Rhythm
- Intonation
- Stresses

Timbre:

- Quality of a particular voice.
- Characterized by frequency spectrum.

The Generative Factors of Speech



Content:

- Discrete phonetic units.
- ≅44 phonemes in English.

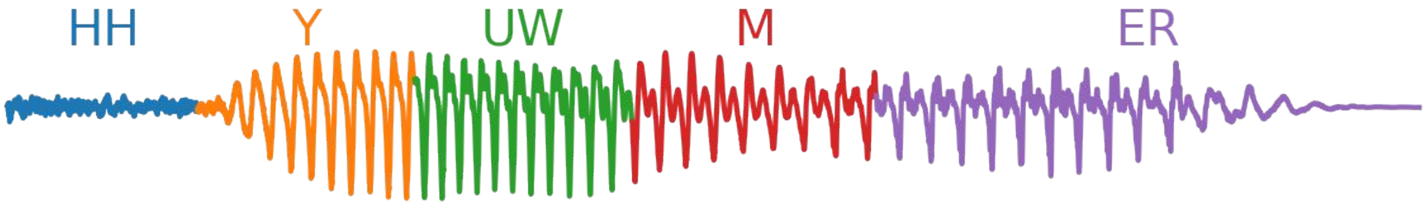
Prosody:

- Rhythm
- Intonation
- Stresses

Timbre:

- Quality of a particular voice.
- Characterized by frequency spectrum.

The Generative Factors of Speech



Content:

- Discrete phonetic units.
- ≅44 phonemes in English.

Prosody:

- Rhythm
- Intonation
- Stresses

Timbre:

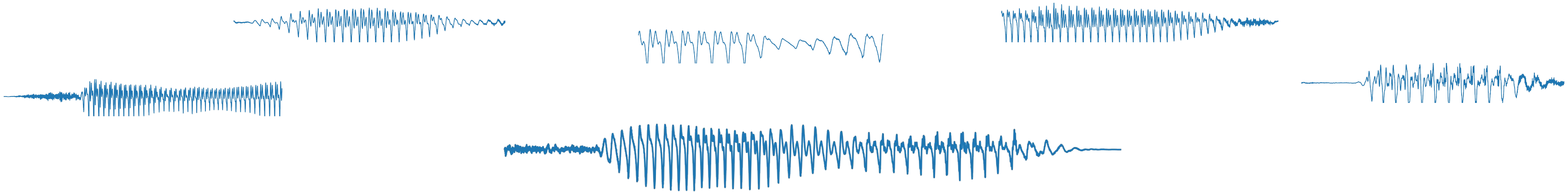
- Quality of a particular voice.
- Characterized by frequency spectrum.

What is Acoustic Unit Discovery?

The goal is to learn **discrete** representations of speech that separate phonetic content from the other factors.
…all without any labels or annotations!

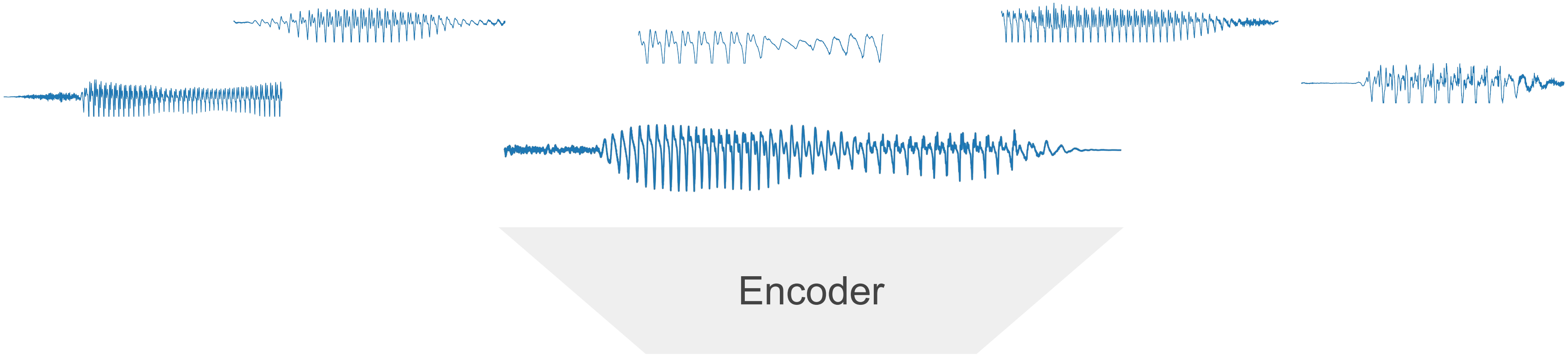
What is Acoustic Unit Discovery?

The goal is to learn **discrete** representations of speech that separate phonetic content from the other factors.
…all without any labels or annotations!



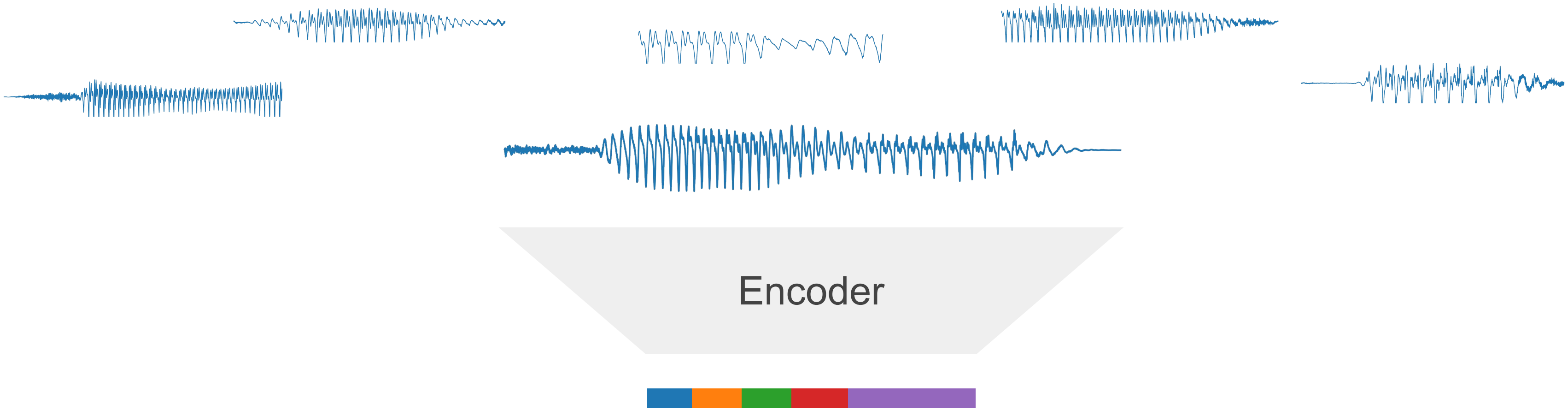
What is Acoustic Unit Discovery?

The goal is to learn **discrete** representations of speech that separate phonetic content from the other factors.
…all without any labels or annotations!



What is Acoustic Unit Discovery?

The goal is to learn **discrete** representations of speech that separate phonetic content from the other factors.
…all without any labels or annotations!



Applications

Bootstrap training of **low-resource** speech systems:

- Automatic speech recognition
- Text-to-speech
- Non-parallel voice conversion

Applications

Bootstrap training of **low-resource** speech systems:

- Automatic speech recognition
- Text-to-speech
- Non-parallel voice conversion

Applications

Bootstrap training of **low-resource** speech systems:

- Automatic speech recognition
- Text-to-speech
- Non-parallel voice conversion

Applications

Bootstrap training of **low-resource** speech systems:

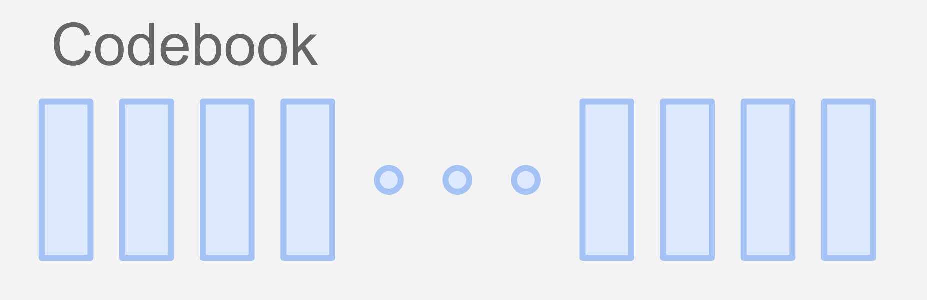
- Automatic speech recognition
- Text-to-speech
- Non-parallel voice conversion

But, how do we learn **discrete** representations using neural networks?

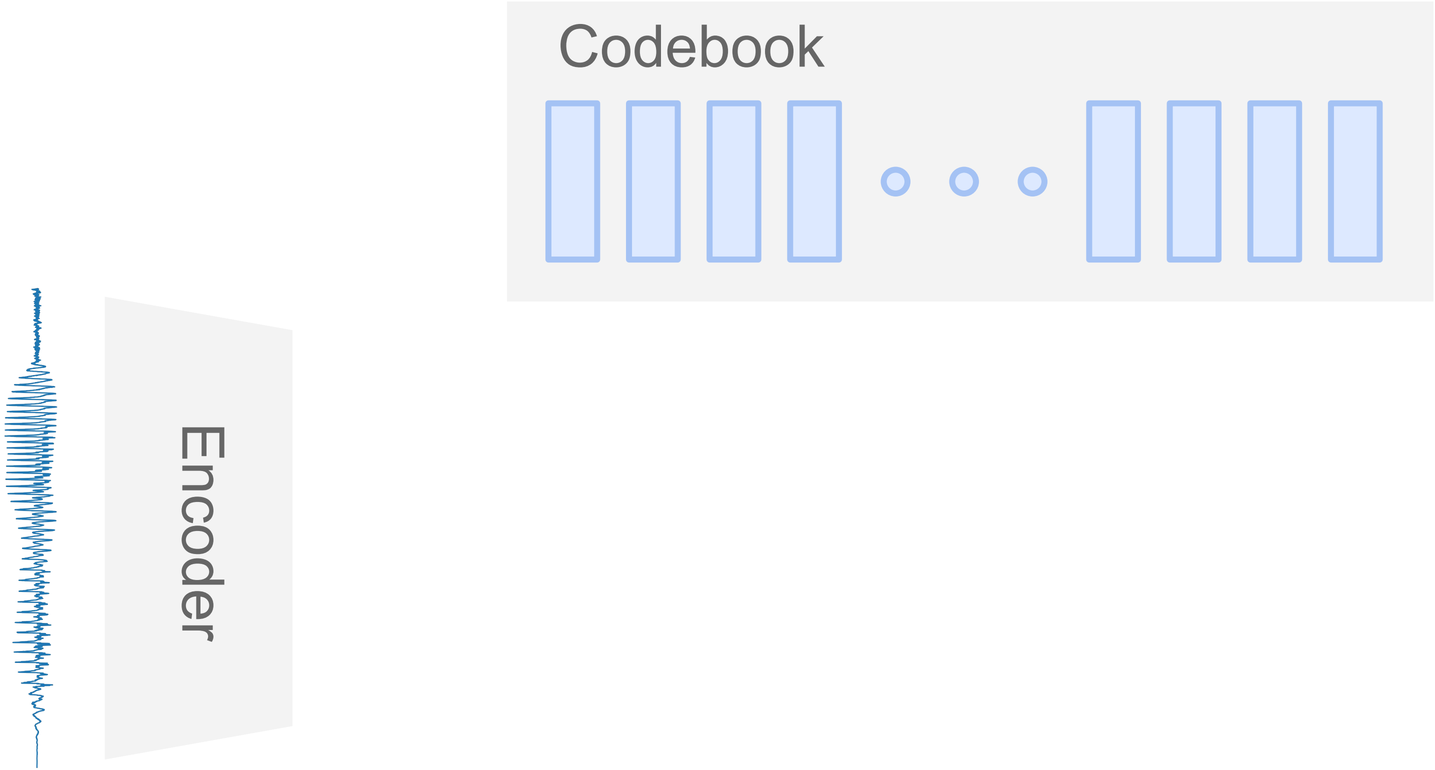
But, how do we learn **discrete** representations using neural networks?

A. van den Oord, O. Vinyals, and K. Kavukcuoglu. “Neural discrete representation learning.” *Advances in Neural Information Processing Systems*. 2017.

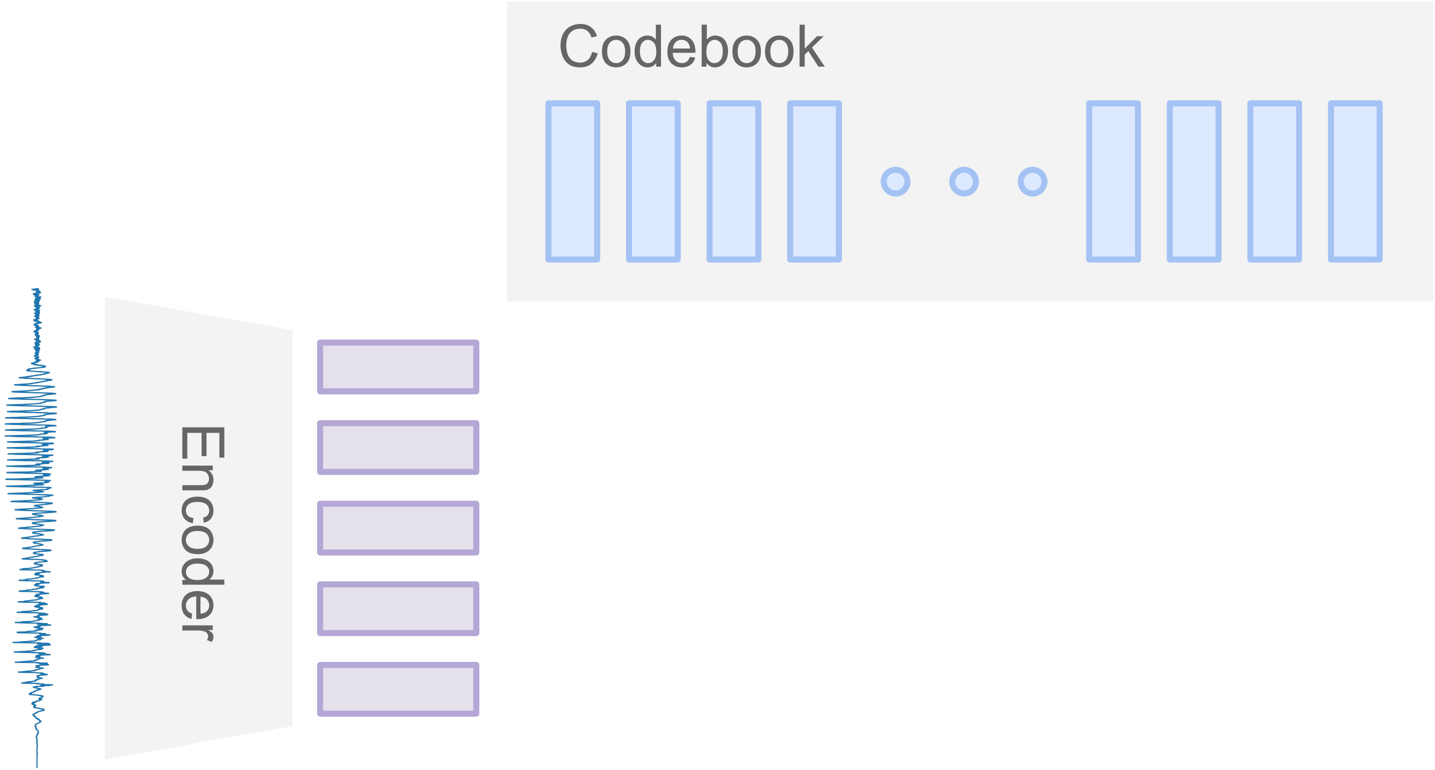
Vector Quantization Layer



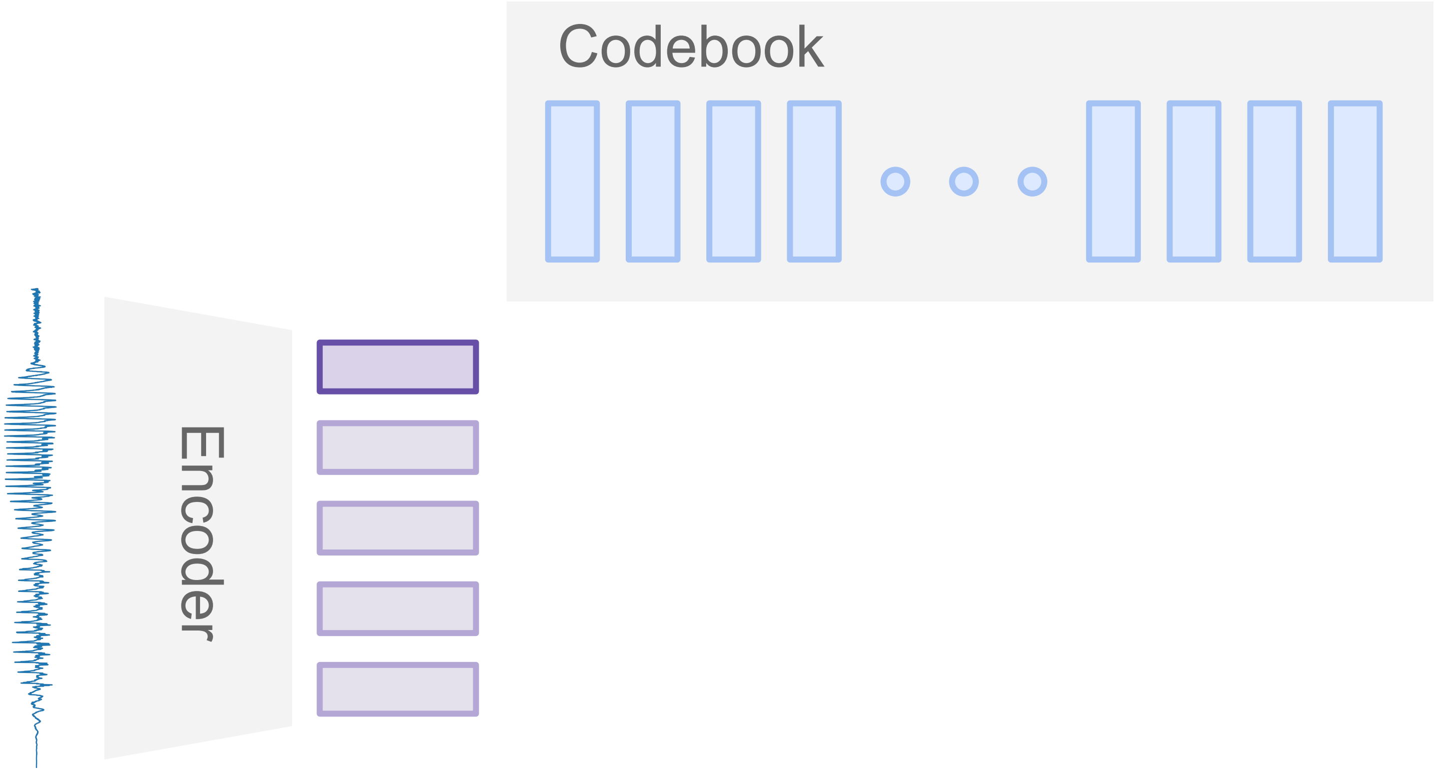
Vector Quantization Layer



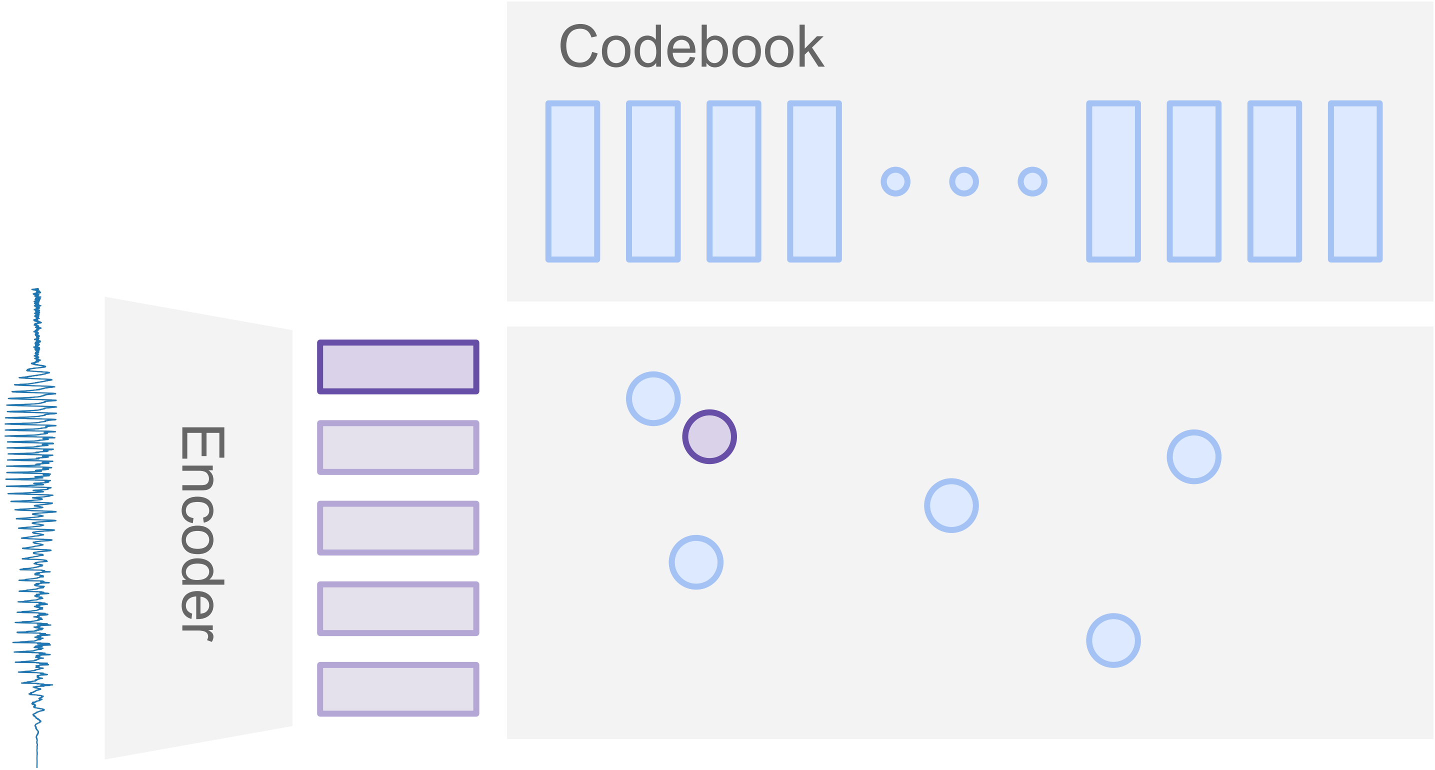
Vector Quantization Layer



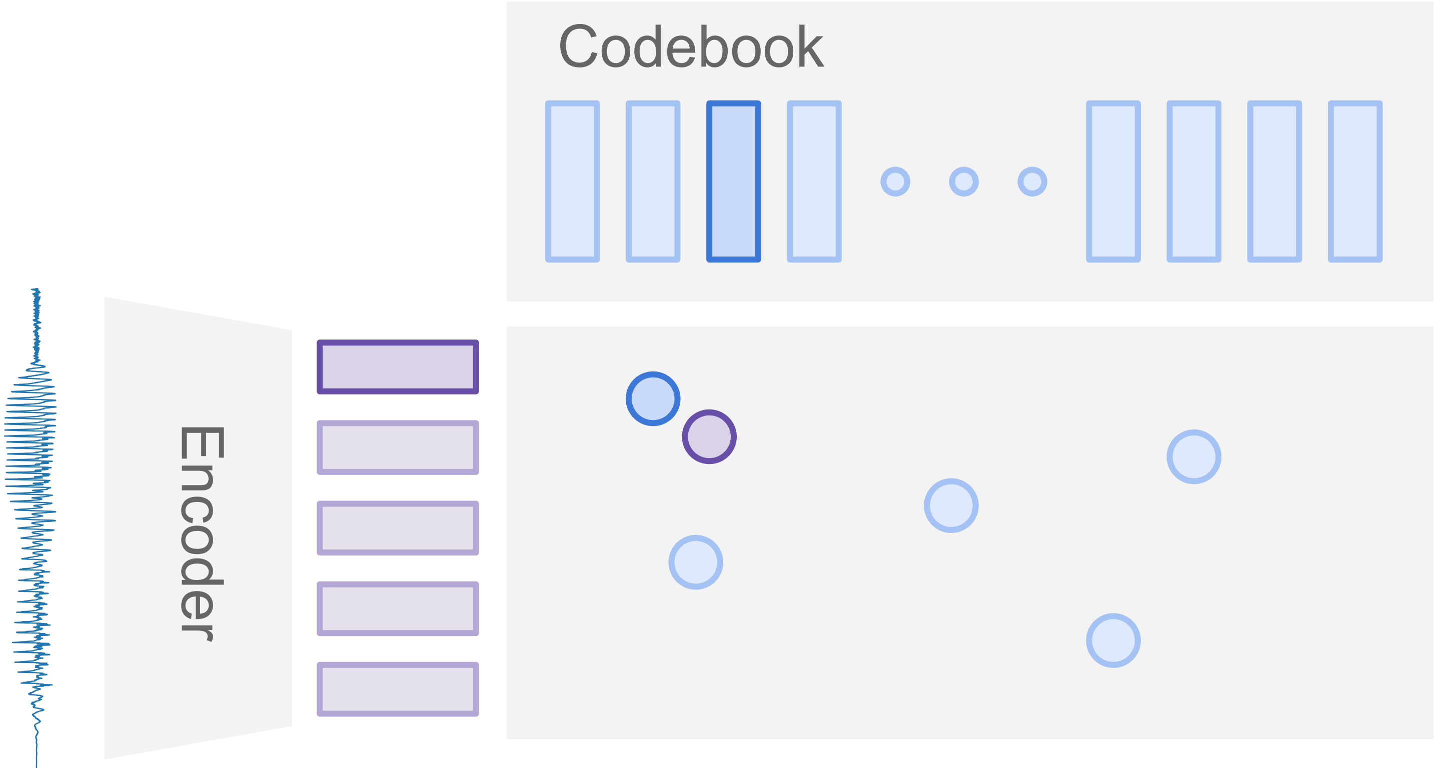
Vector Quantization Layer



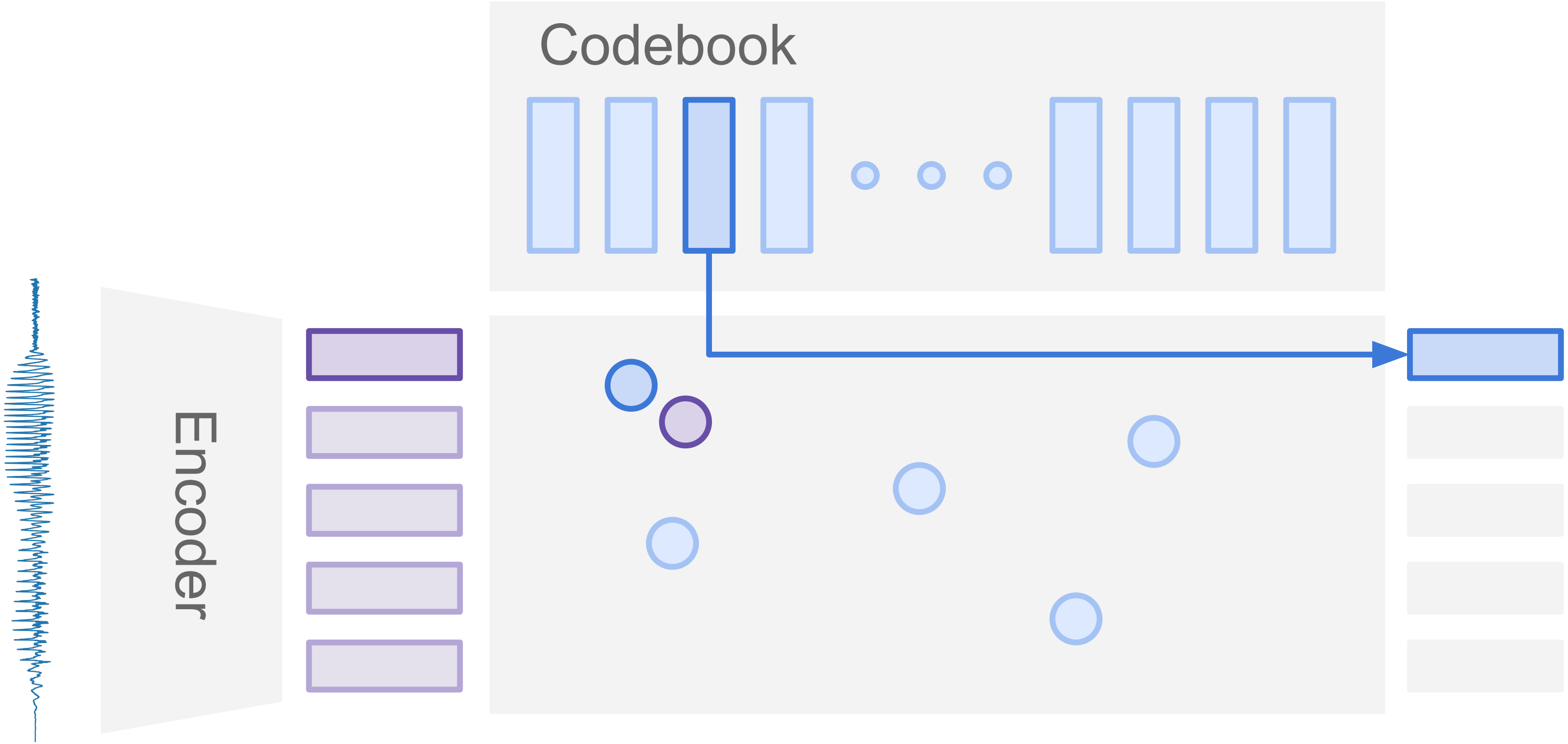
Vector Quantization Layer



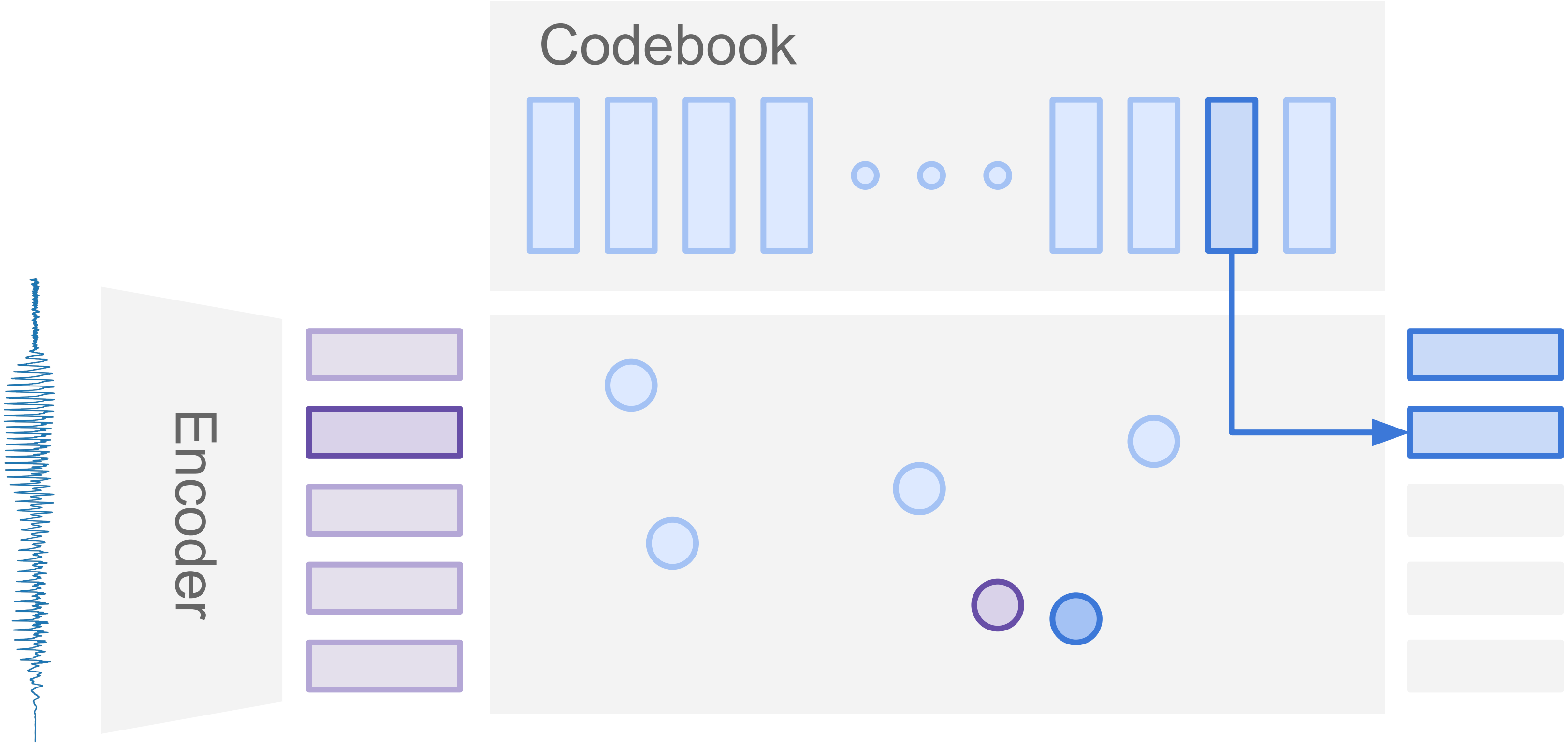
Vector Quantization Layer



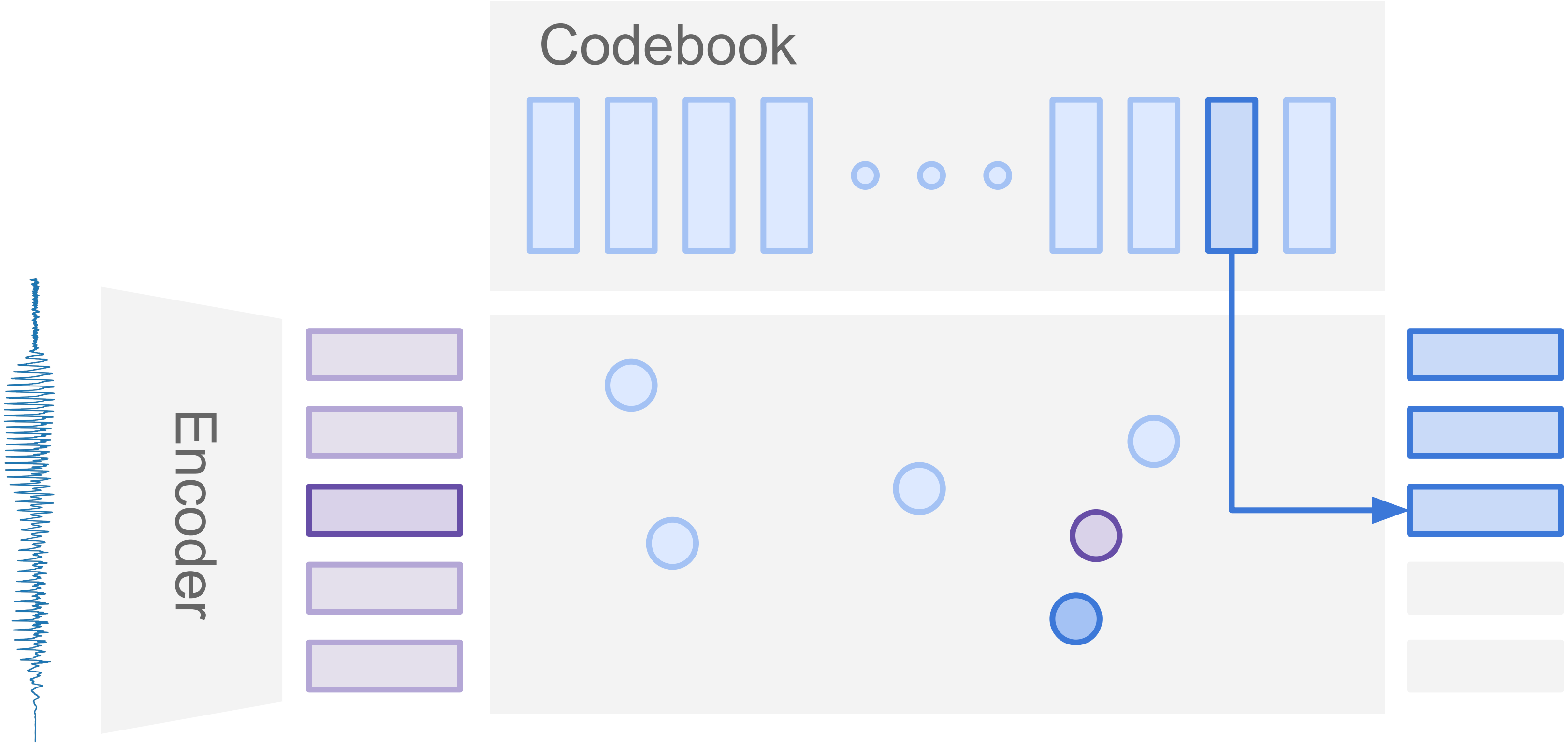
Vector Quantization Layer



Vector Quantization Layer

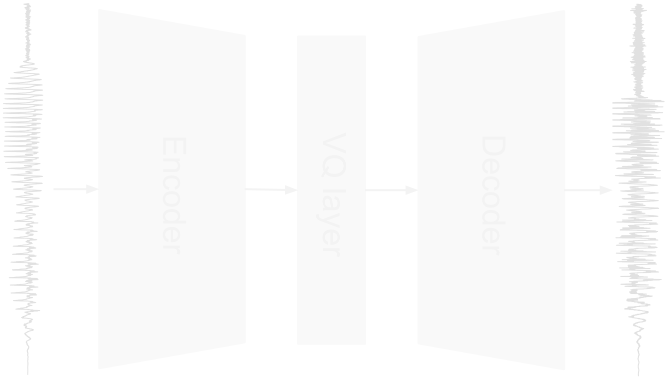


Vector Quantization Layer



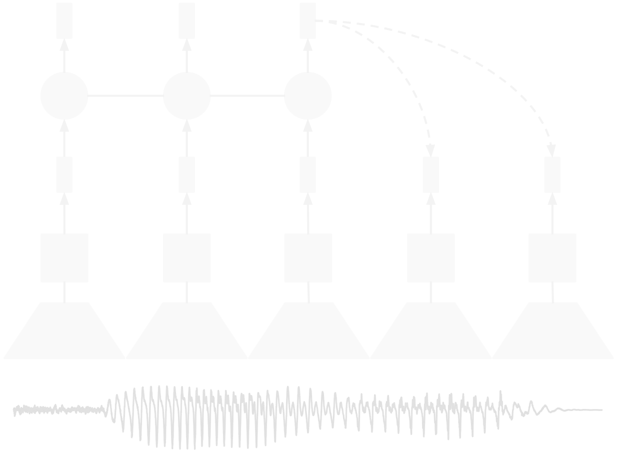
Our contribution: we propose and compare two models for acoustic unit discovery in the *ZeroSpeech 2020 Challenge*.

1. A Vector-Quantized Variational Autoencoder (VQ-VAE)



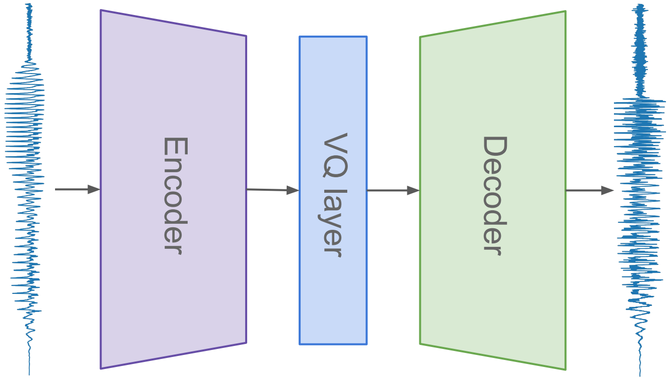
Inspired by: J. Chorowski, et al. "Unsupervised speech representation learning using wavenet autoencoders." IEEE/ACM transactions on audio, speech, and language processing. 2019.

2. A combination of Vector-Quantization and Contrastive Predictive Coding (VQ-CPC)



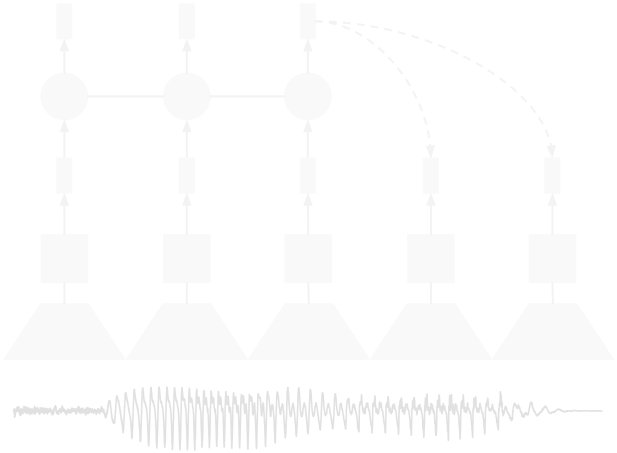
Our contribution: we propose and compare two models for acoustic unit discovery in the *ZeroSpeech 2020 Challenge*.

1. A Vector-Quantized Variational Autoencoder (VQ-VAE)



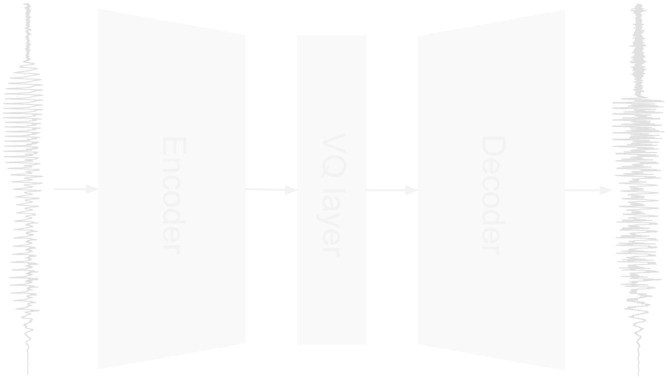
Inspired by: J. Chorowski, et al. “Unsupervised speech representation learning using wavenet autoencoders.” IEEE/ACM transactions on audio, speech, and language processing. 2019.

2. A combination of Vector-Quantization and Contrastive Predictive Coding (VQ-CPC)



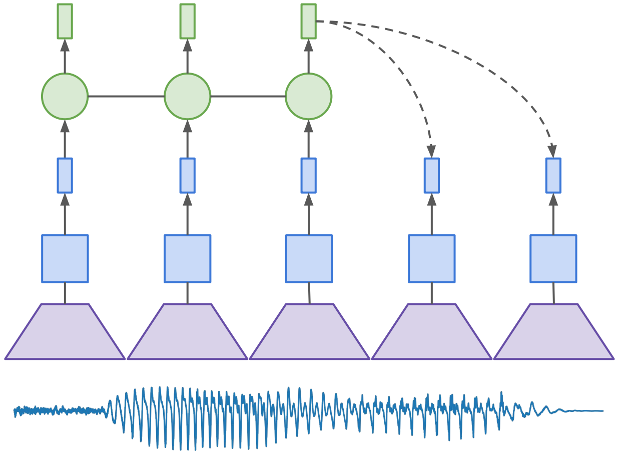
Our contribution: we propose and compare two models for acoustic unit discovery in the *ZeroSpeech 2020 Challenge*.

1. A Vector-Quantized Variational Autoencoder (VQ-VAE)



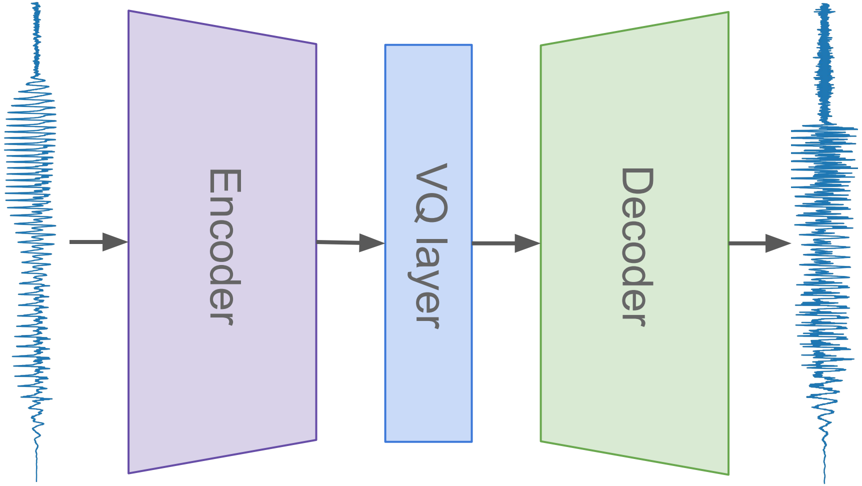
Inspired by: J. Chorowski, et al. “Unsupervised speech representation learning using wavenet autoencoders.” IEEE/ACM transactions on audio, speech, and language processing. 2019.

2. A combination of Vector-Quantization and Contrastive Predictive Coding (VQ-CPC)

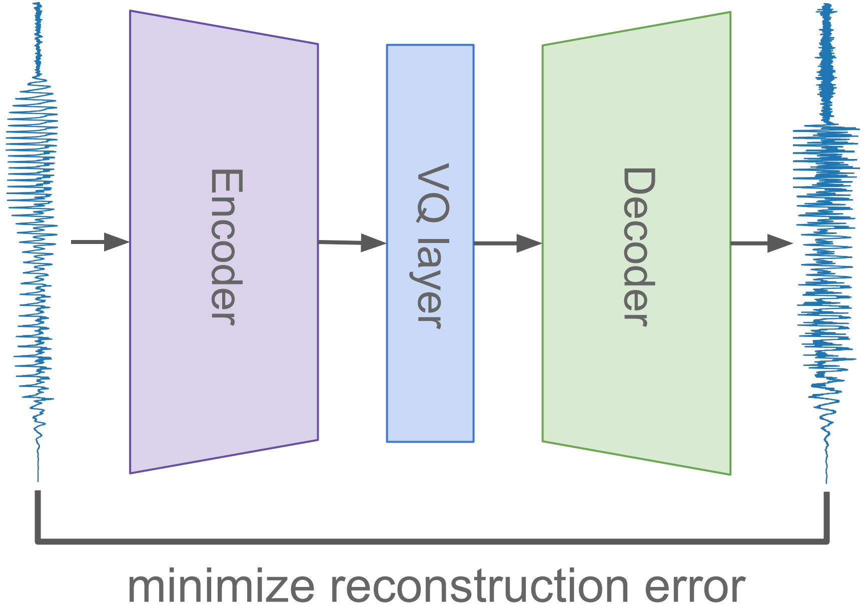


Inspired by: A. van den Oord, et al. “Representation Learning with Contrastive Predictive Coding.” 2018.

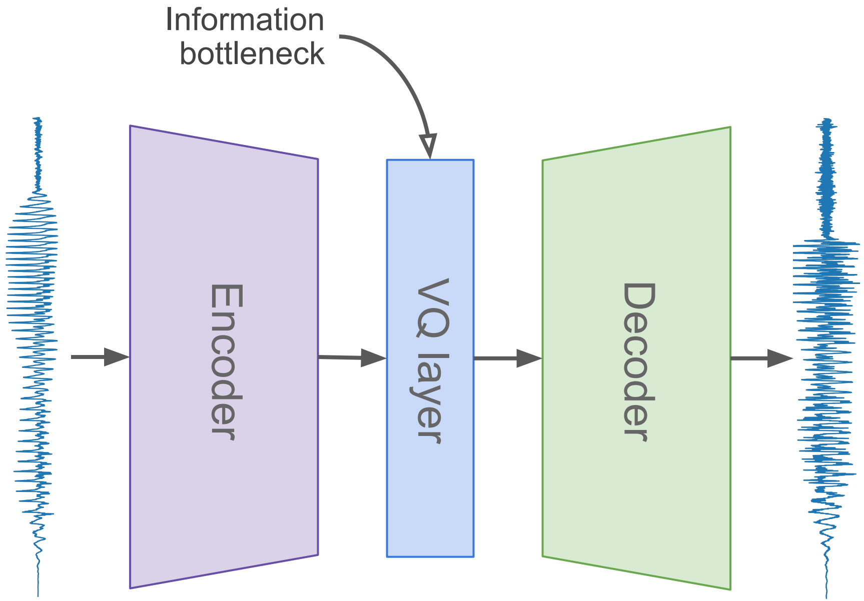
Vector-Quantized Variational Autoencoder



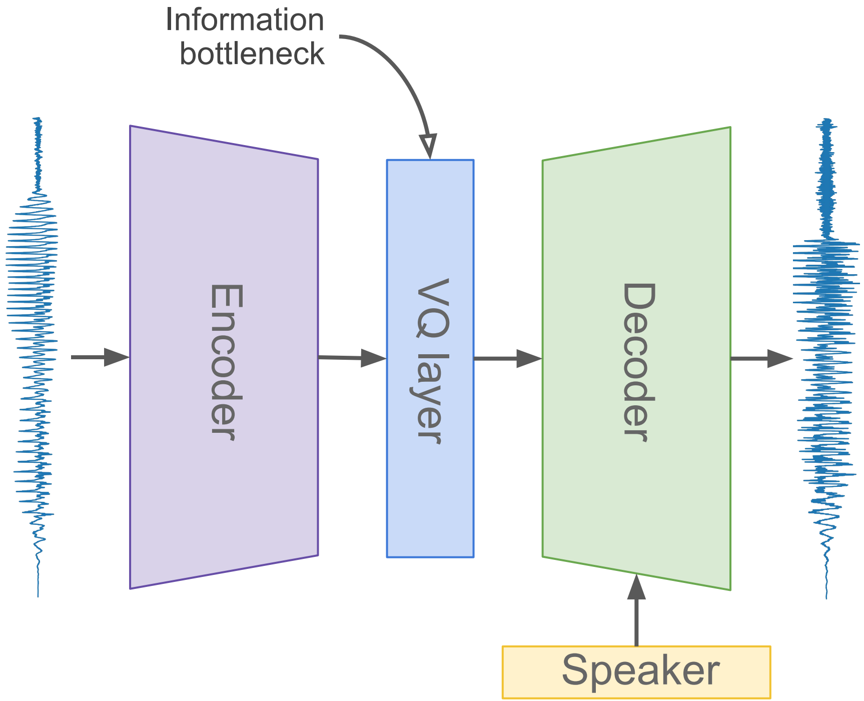
Vector-Quantized Variational Autoencoder



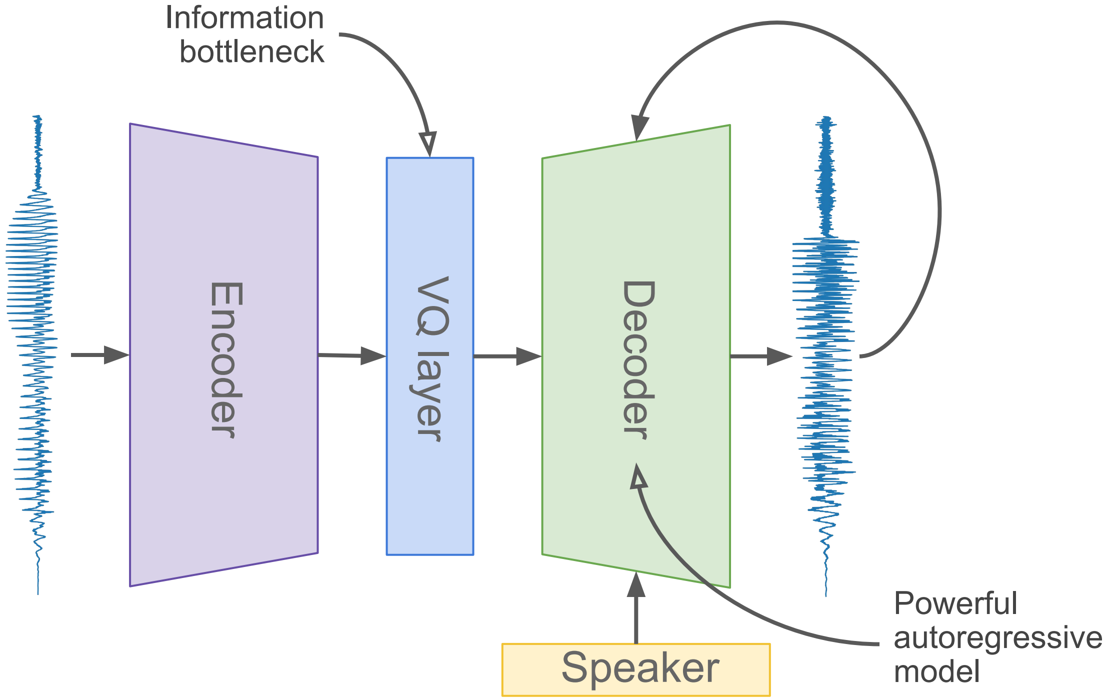
Vector-Quantized Variational Autoencoder


Information bottleneck

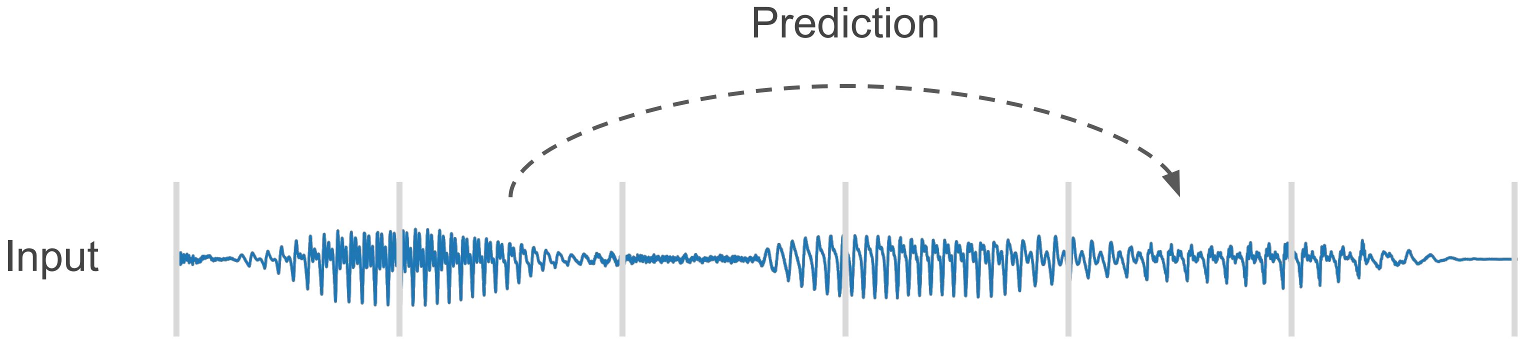
Vector-Quantized Variational Autoencoder



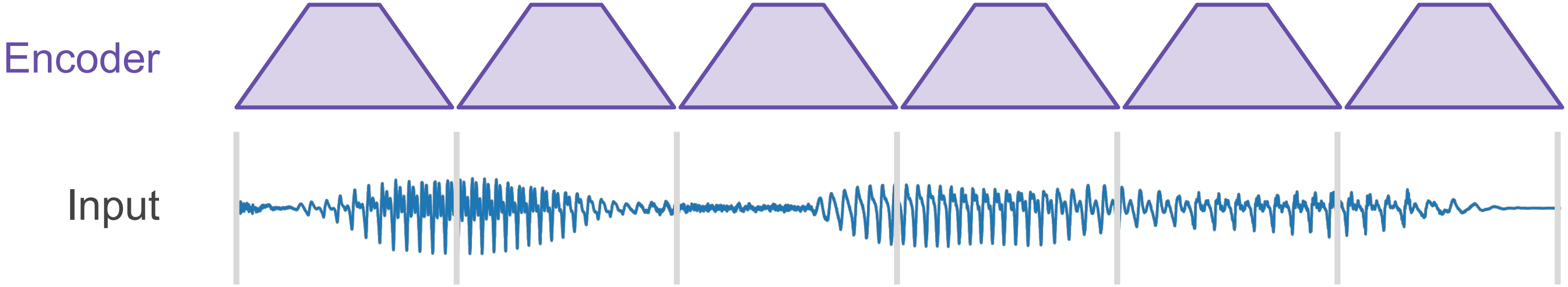
Vector-Quantized Variational Autoencoder



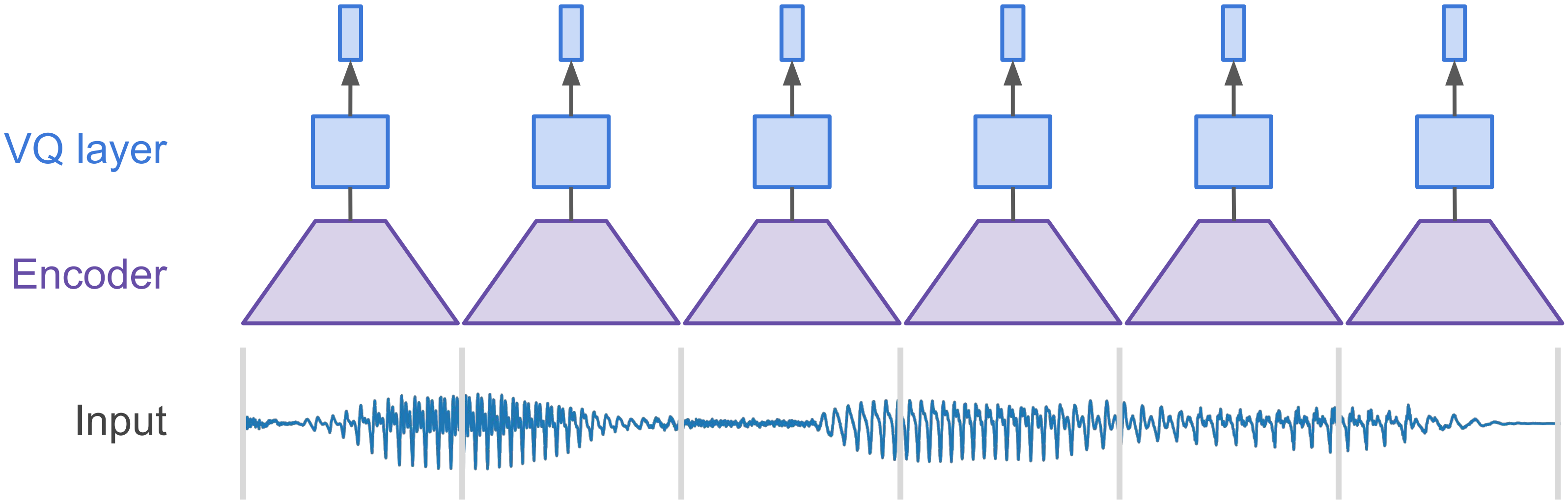
Vector-Quantized Contrastive Predictive Coding



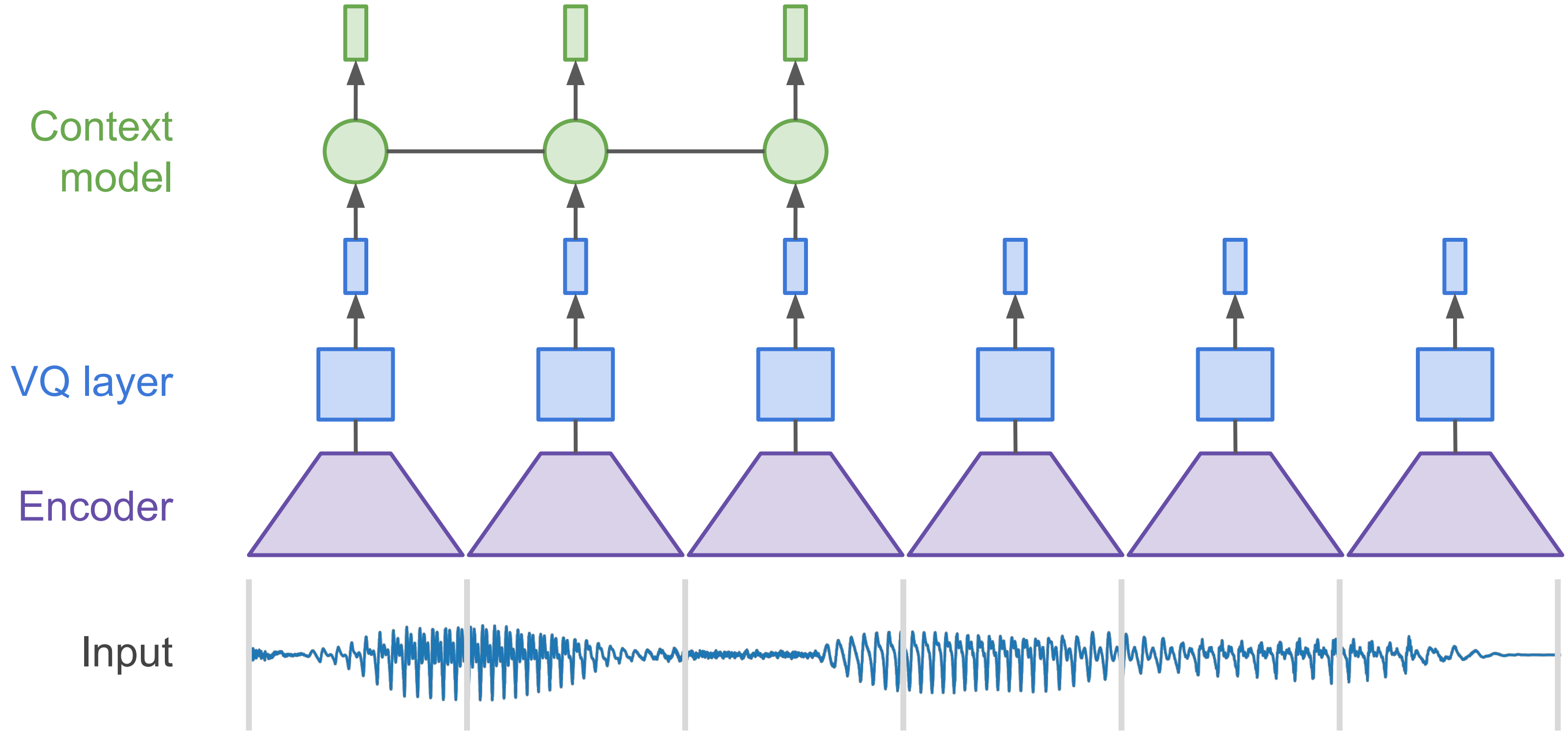
Vector-Quantized Contrastive Predictive Coding



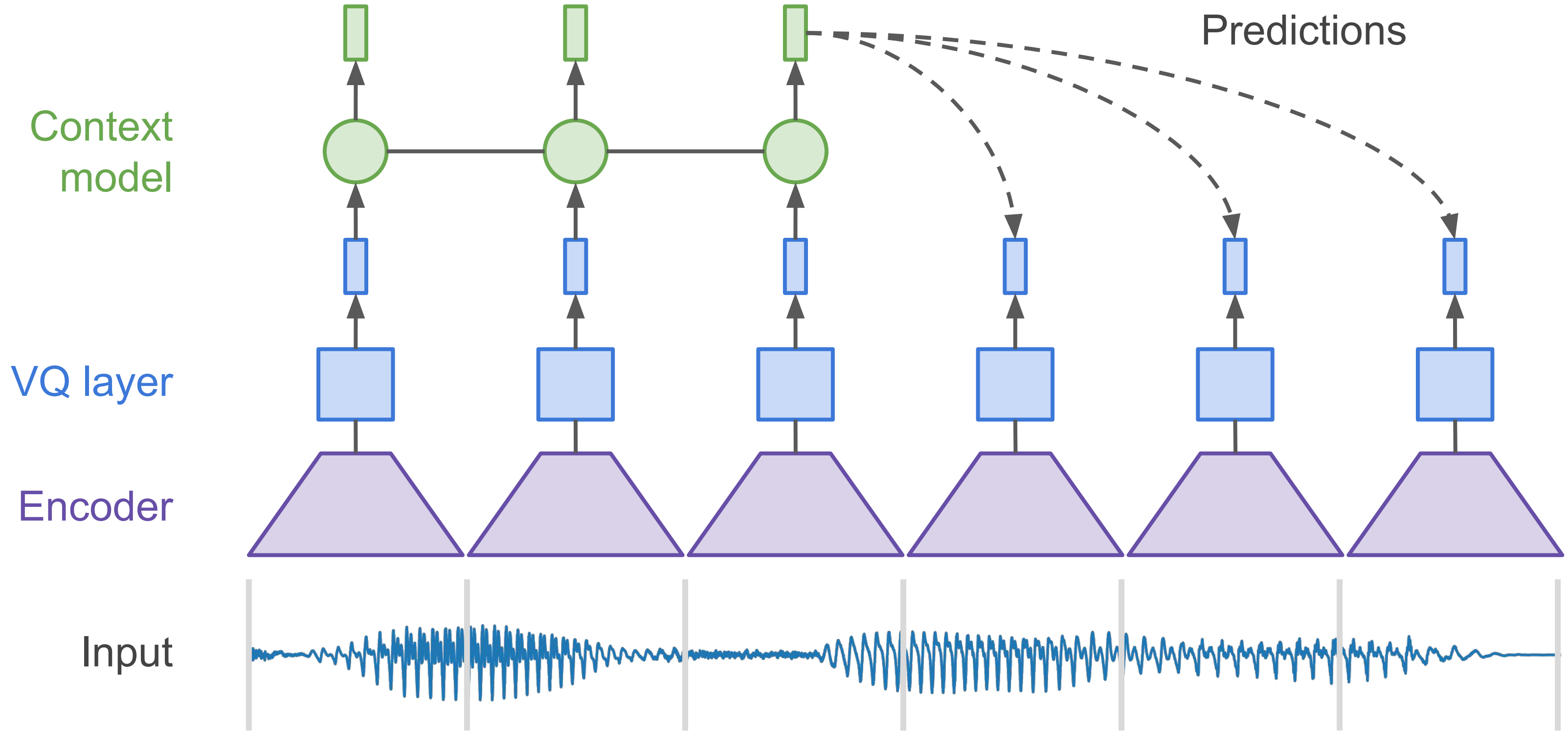
Vector-Quantized Contrastive Predictive Coding



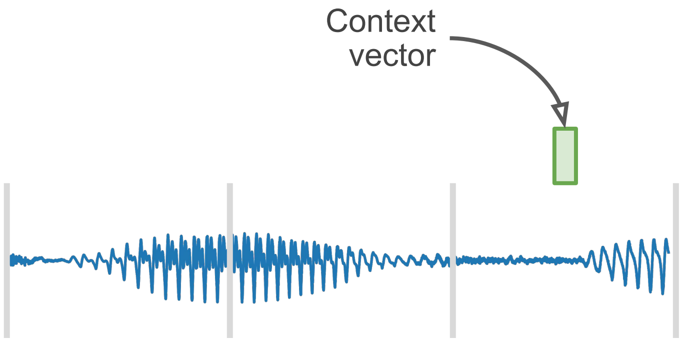
Vector-Quantized Contrastive Predictive Coding



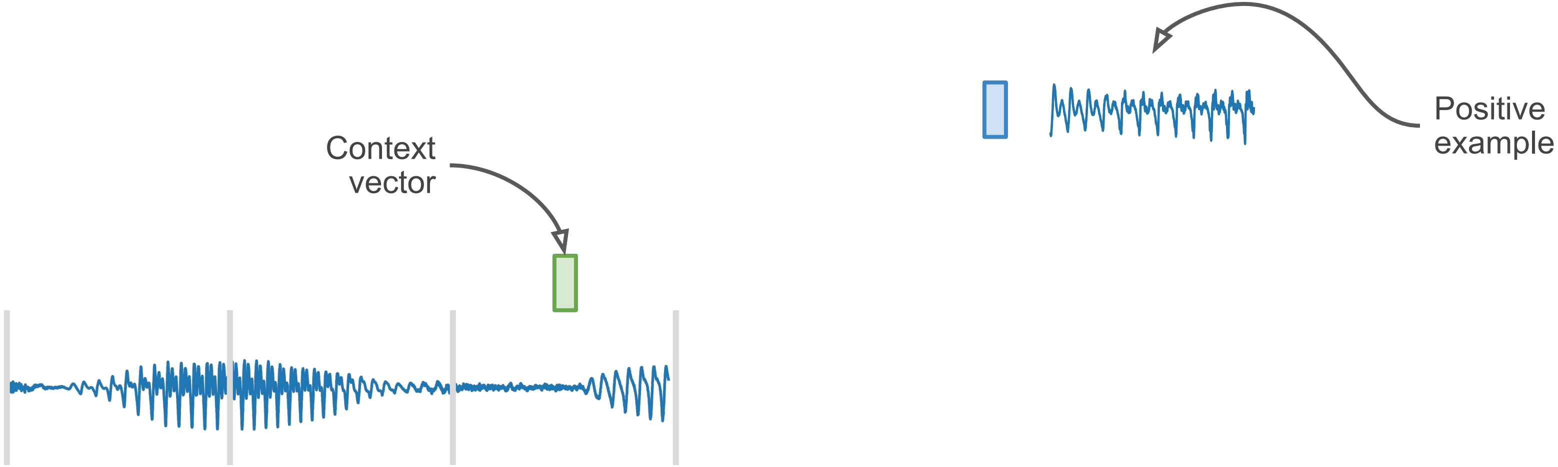
Vector-Quantized Contrastive Predictive Coding



Vector-Quantized Contrastive Predictive Coding



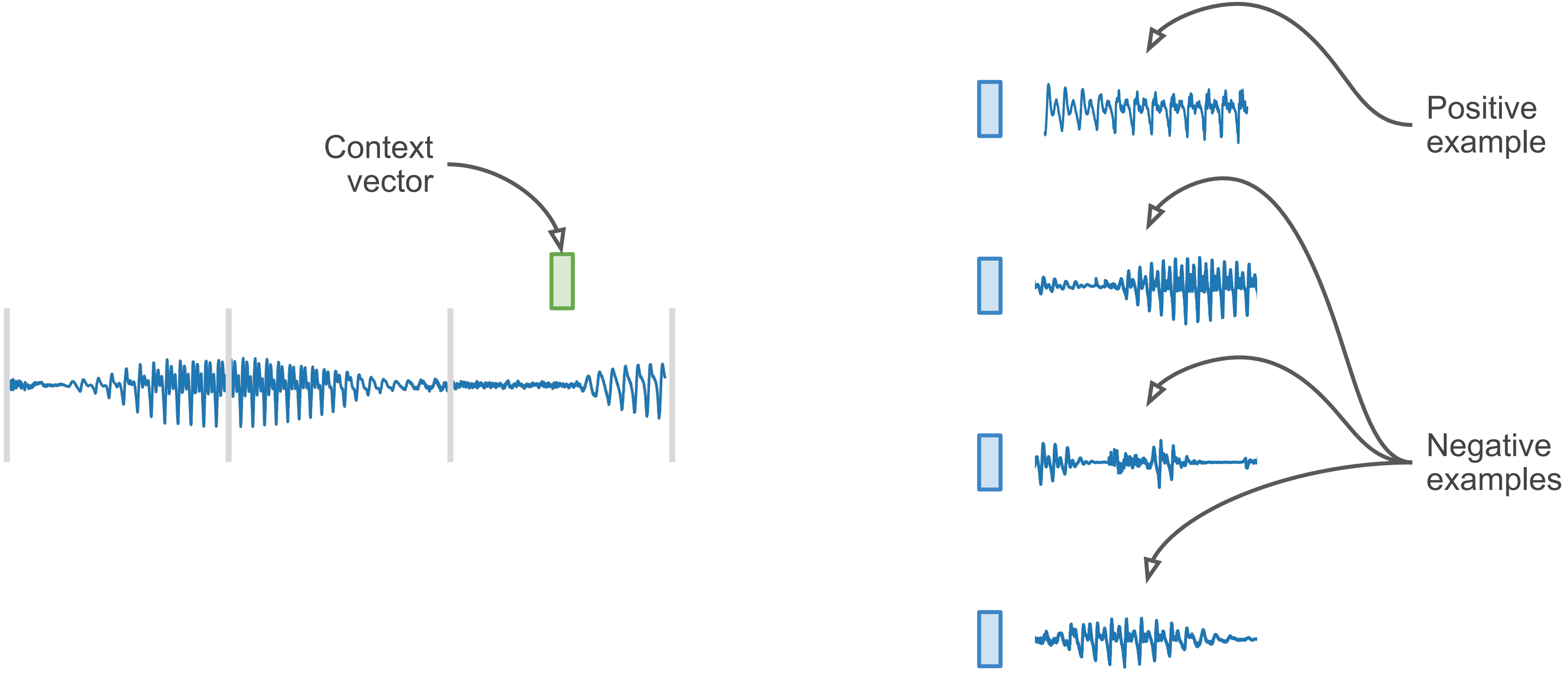
Vector-Quantized Contrastive Predictive Coding



Context vector

Positive example

Vector-Quantized Contrastive Predictive Coding

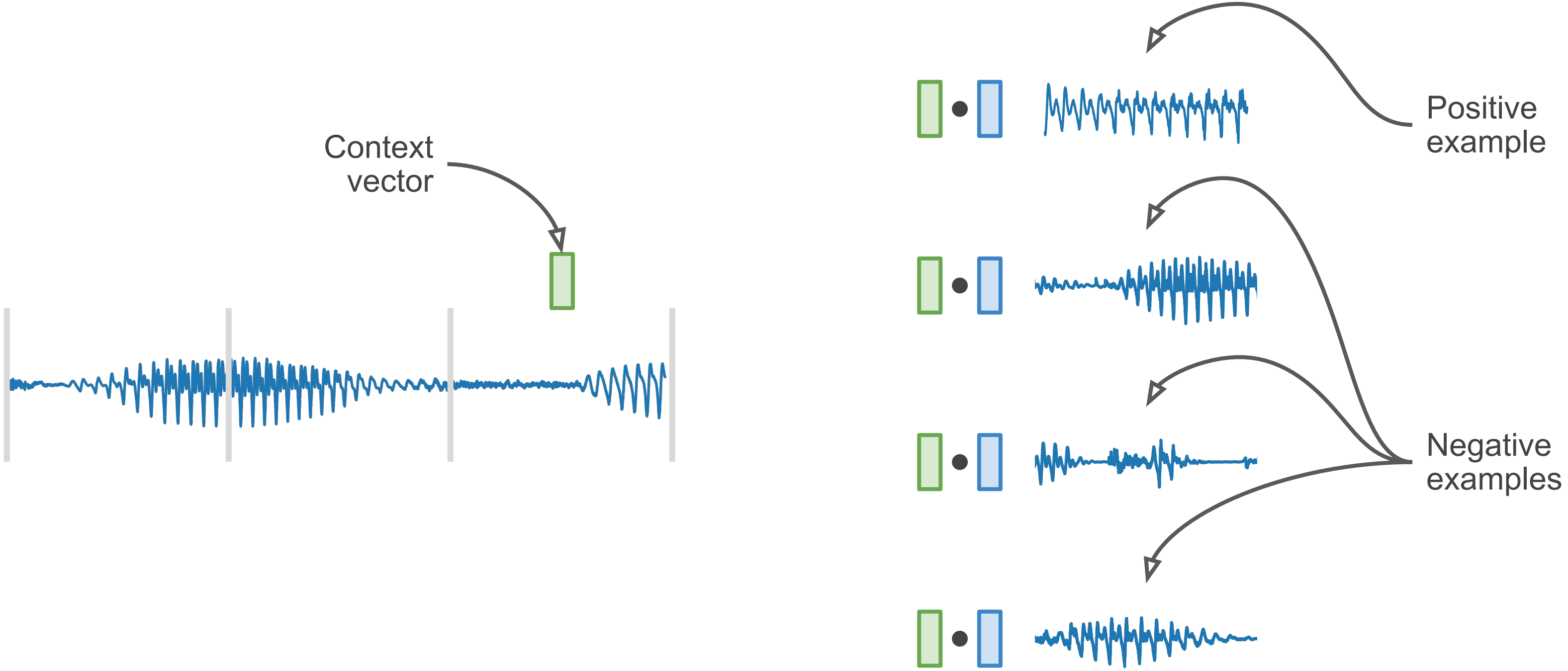


Context vector

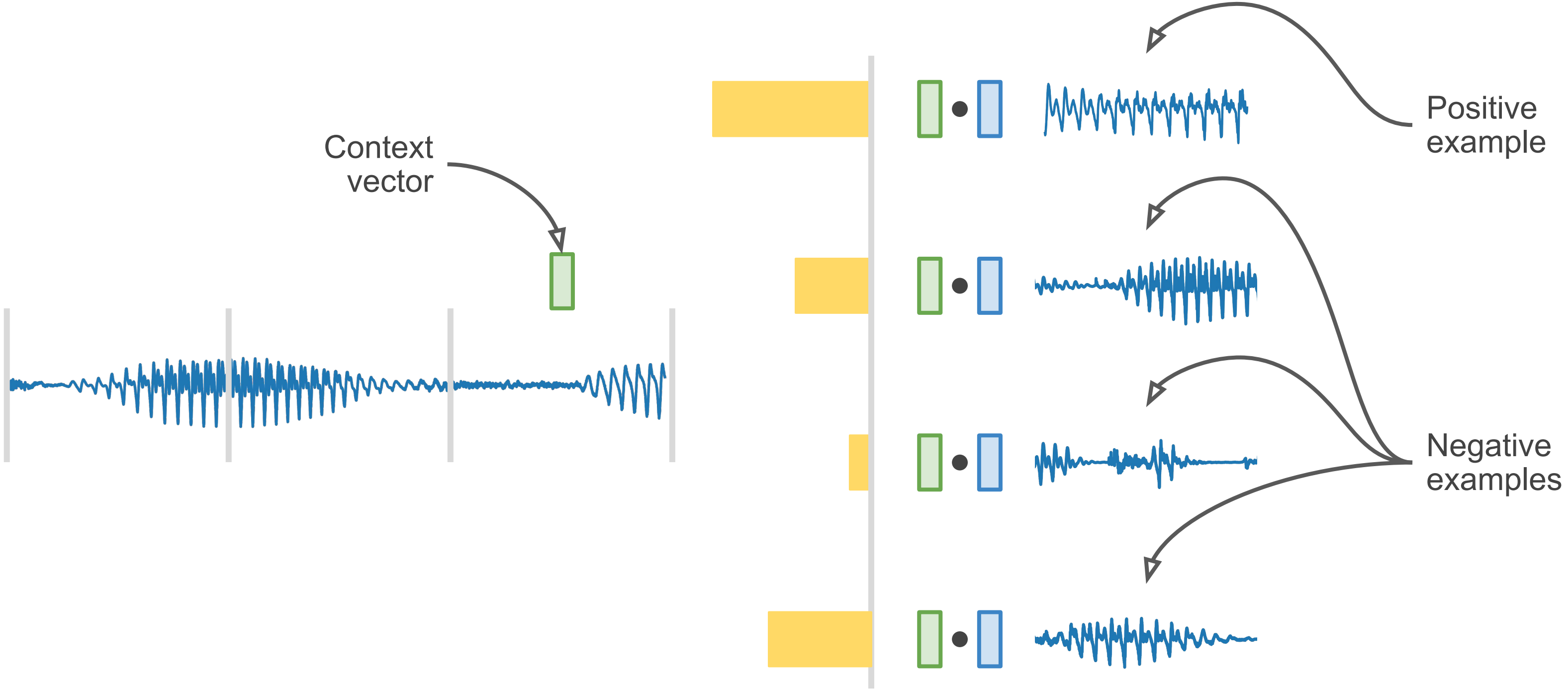
Positive example

Negative examples

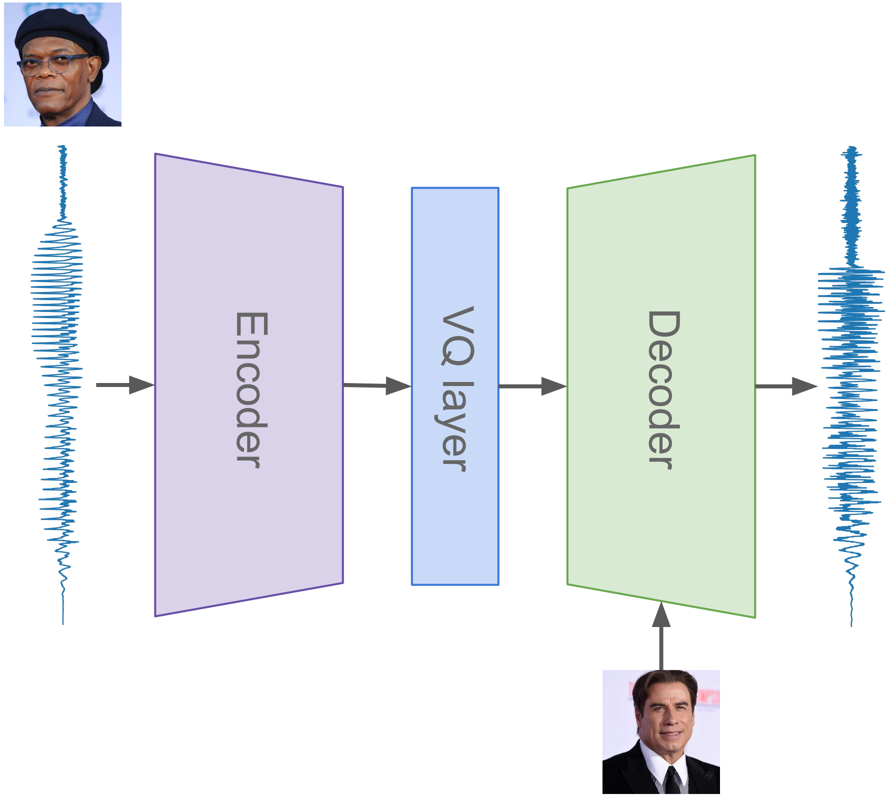
Vector-Quantized Contrastive Predictive Coding



Vector-Quantized Contrastive Predictive Coding



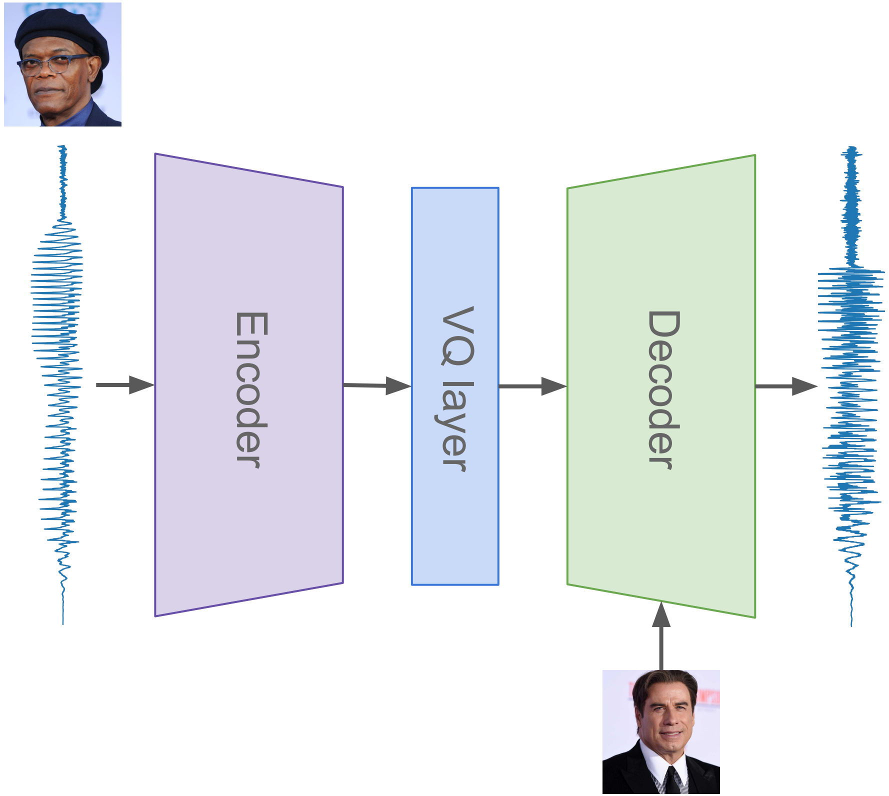
Evaluation - Voice Conversion



Evaluation Metrics:

- Speaker similarity (1-5 scale).
- Intelligibility (character error rate).
- Mean opinion score (1-5 scale).

Evaluation - Voice Conversion



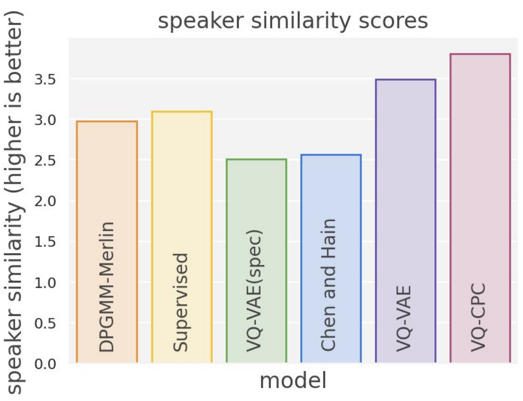
Evaluation Metrics:

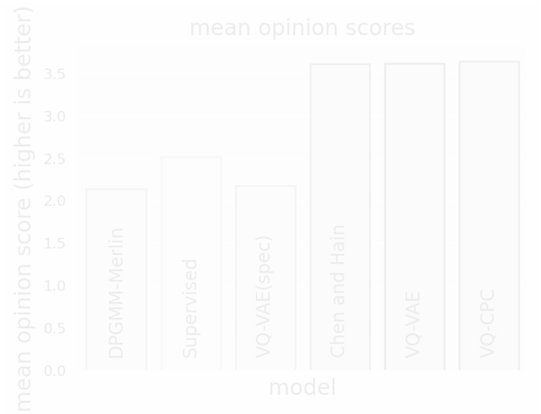
- Speaker similarity (1-5 scale).
- Intelligibility (character error rate).
- Mean opinion score (1-5 scale).

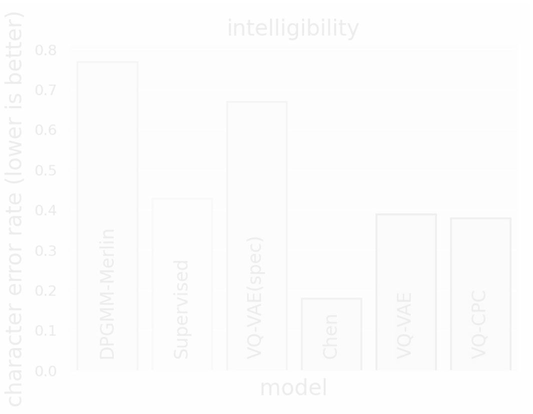
Evaluation - Voice Conversion

Source	Converted	Target	Other Conversion
🔊	🔊	🔊	🔊
🔊	🔊	🔊	🔊

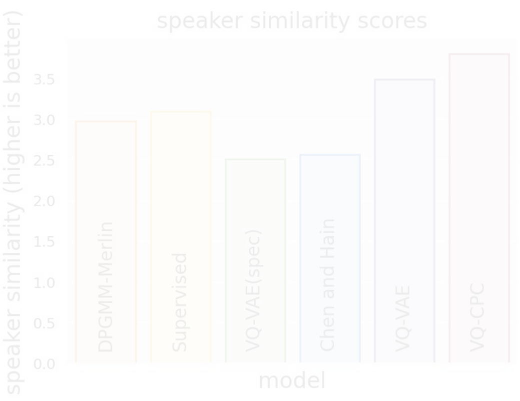
Evaluation - Voice Conversion

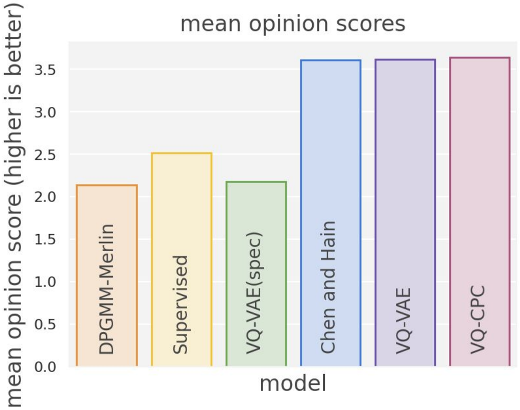


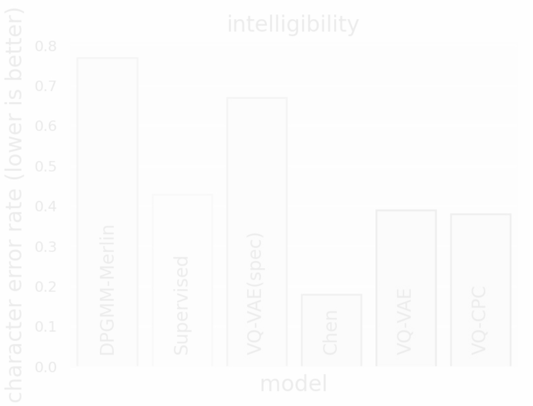




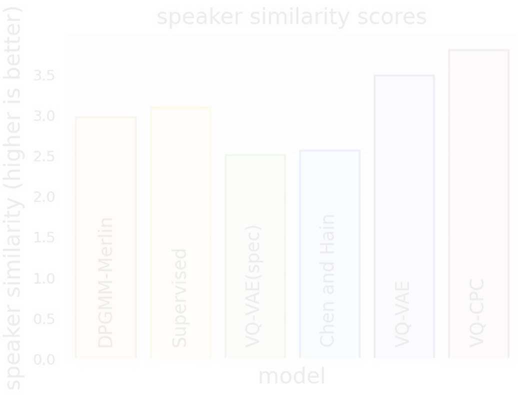
Evaluation - Voice Conversion

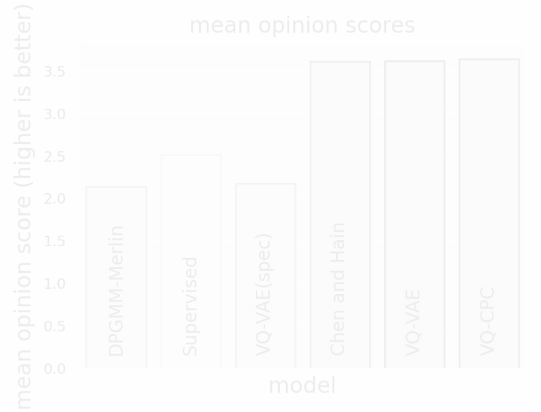


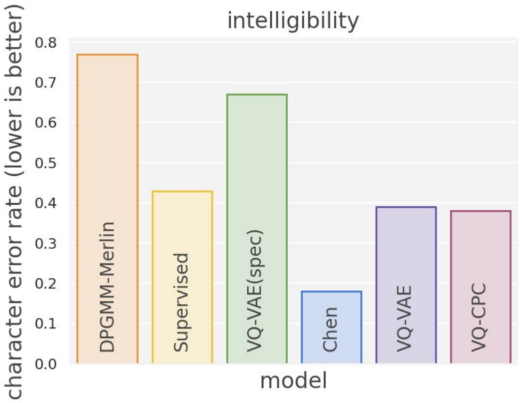




Evaluation - Voice Conversion

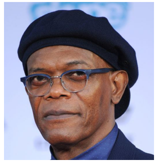




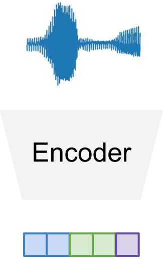


Evaluation - ABX Score

Triphone A:

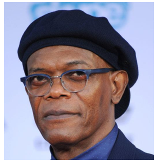


beg

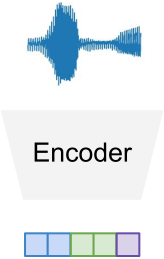


Evaluation - ABX Score

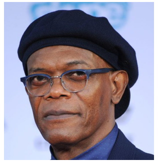
Triphone A:



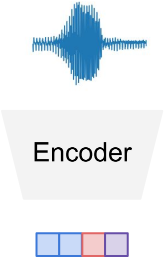
beg



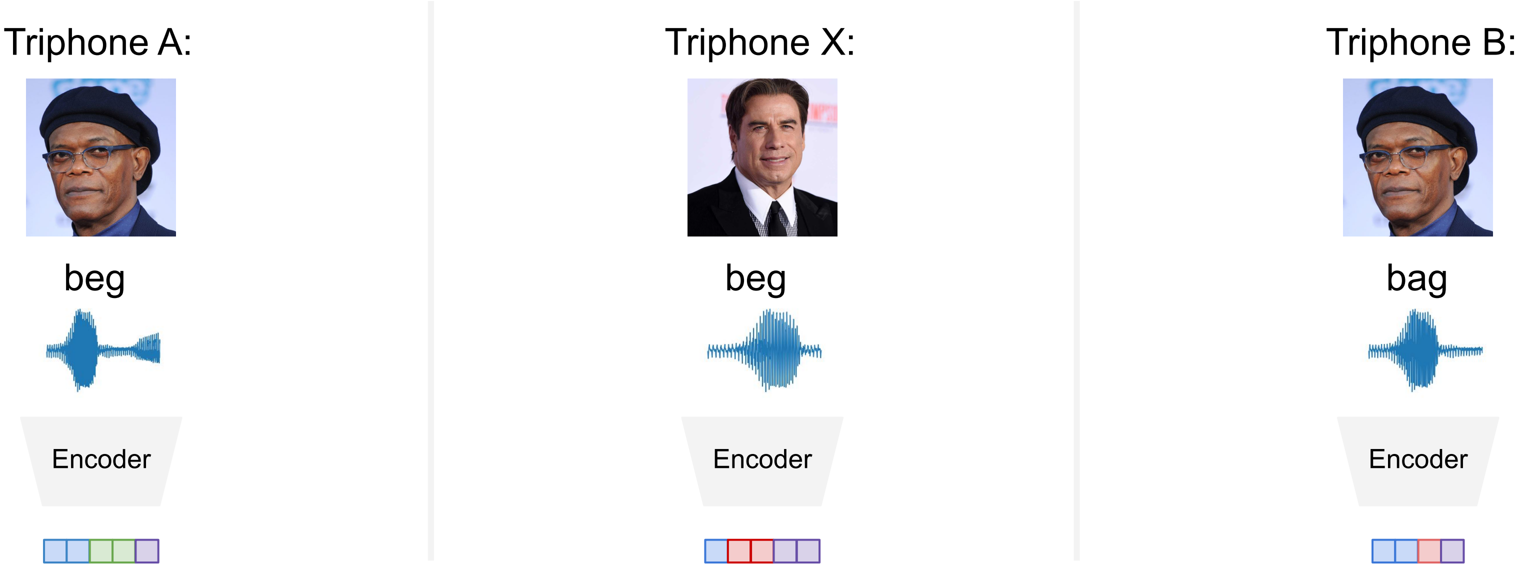
Triphone B:



bag



Evaluation - ABX Score



Triphone A:

beg

Encoder

Triphone X:

beg

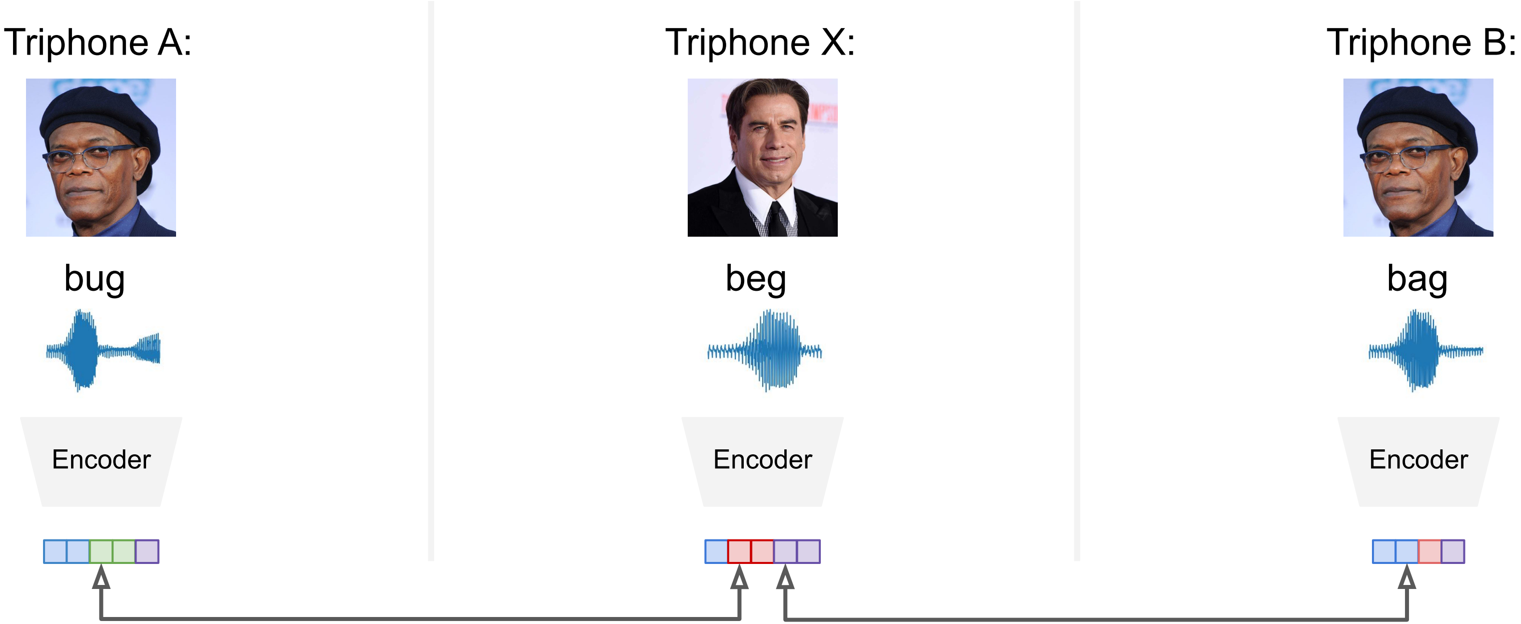
Encoder

Triphone B:

bag

Encoder

Evaluation - ABX Score



Triphone A:

bug

Encoder

Triphone X:

beg

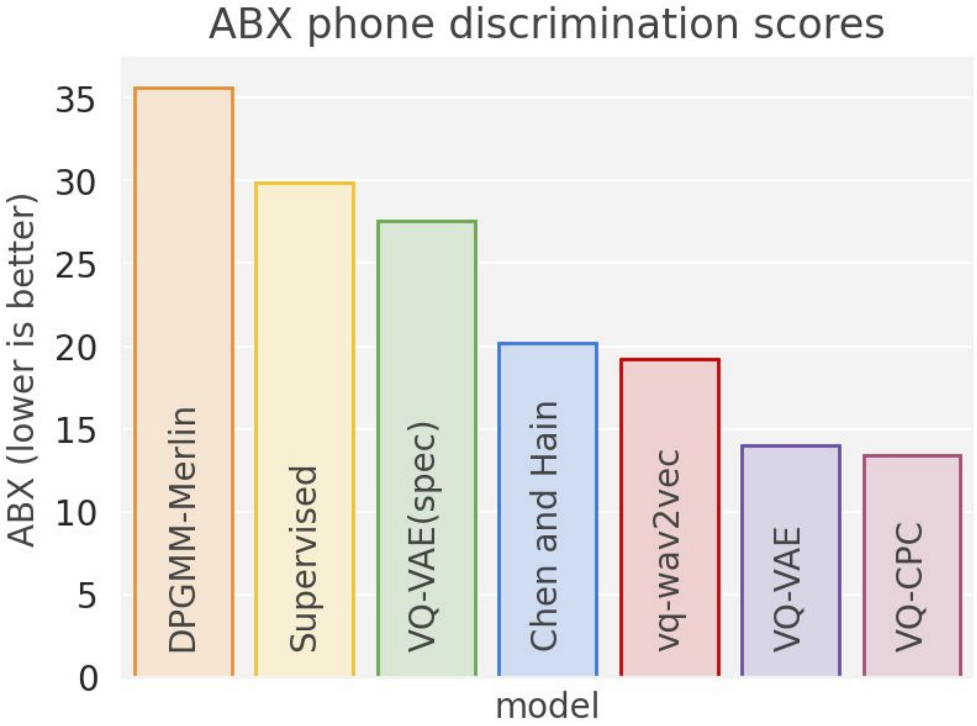
Encoder

Triphone B:

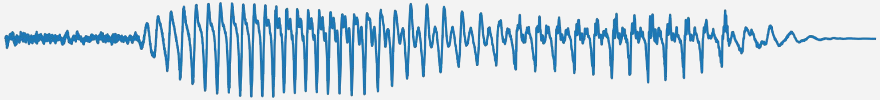
bag

Encoder

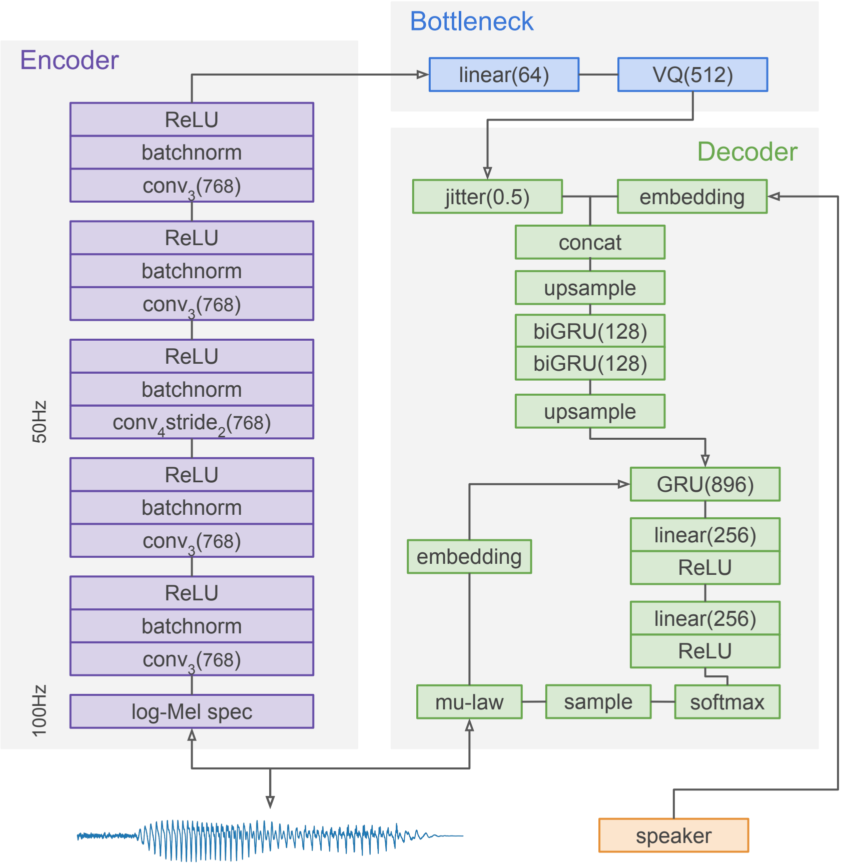
Evaluation - ABX Score



Questions?



Vector Quantized Variational Autoencoder


Encoder
ReLU
batchnorm
$conv_3(768)$
ReLU
batchnorm
$conv_3(768)$
ReLU
batchnorm
$conv_4stride_2(768)$
ReLU
batchnorm
$conv_3(768)$
ReLU
batchnorm
$conv_3(768)$
log-Mel spec
50Hz
100Hz
Bottleneck
linear(64)
VQ(512)
Decoder
jitter(0.5)
embedding
concat
upsample
biGRU(128)
biGRU(128)
upsample
GRU(896)
embedding
linear(256)
ReLU
linear(256)
ReLU
mu-law
sample
softmax
speaker