Incorporating Symbolic Sequential Modeling for Speech Enhancement

Chien-Feng Liao, Yu Tsao,
Xugang Lu, Hisashi Kawai

Outline

- Introduction
- Methodology
  - U-Net
  - Symbolic Encoder
  - Multi Head Attention
- Architecture
- Experiments
- Conclusion

Introduction

Even in a noisy environment, as long as the listener can understand what the speaker is saying, the defective voice signal can be restored.

In other words, with the aid of the language model, the damage caused by interference noise can be effectively suppressed.

Therefore, this paper attempts to use VQ-VAE's Symbolic Book to construct acoustic units, and then Transformer's Multi Head Attention (MHA) uses acoustic features to extract speech content to help improve the effect of Speech Enhancement.

Methodology

U-Net

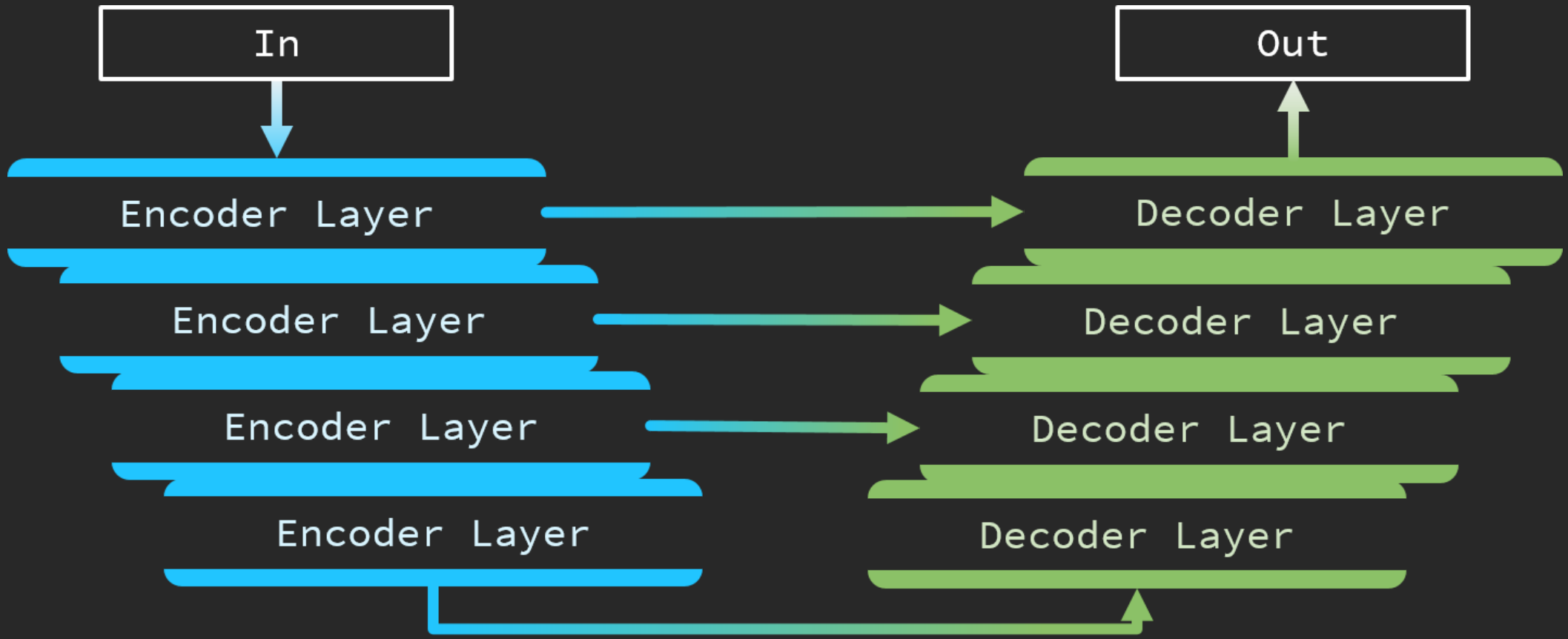
+

VQ-VAE

+

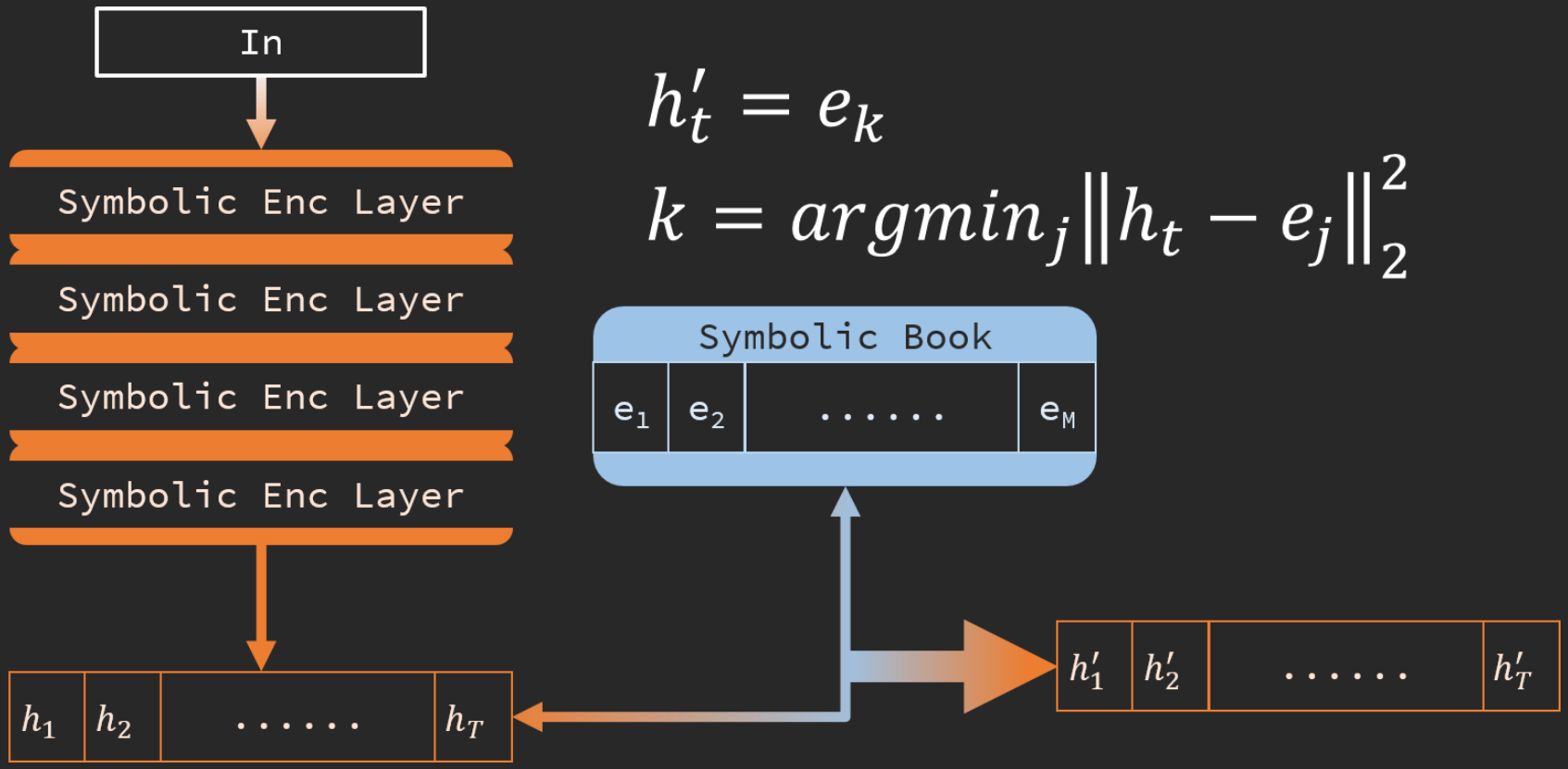
Multi Head Attention

U-Net



$$\mathcal{L}_{mse} = \frac{1}{N}\sum_{i=1}^{N} \left\| Dec\left(Enc(x_i)\right) - y_i \right\|_2^2$$

Symbolic Encoder

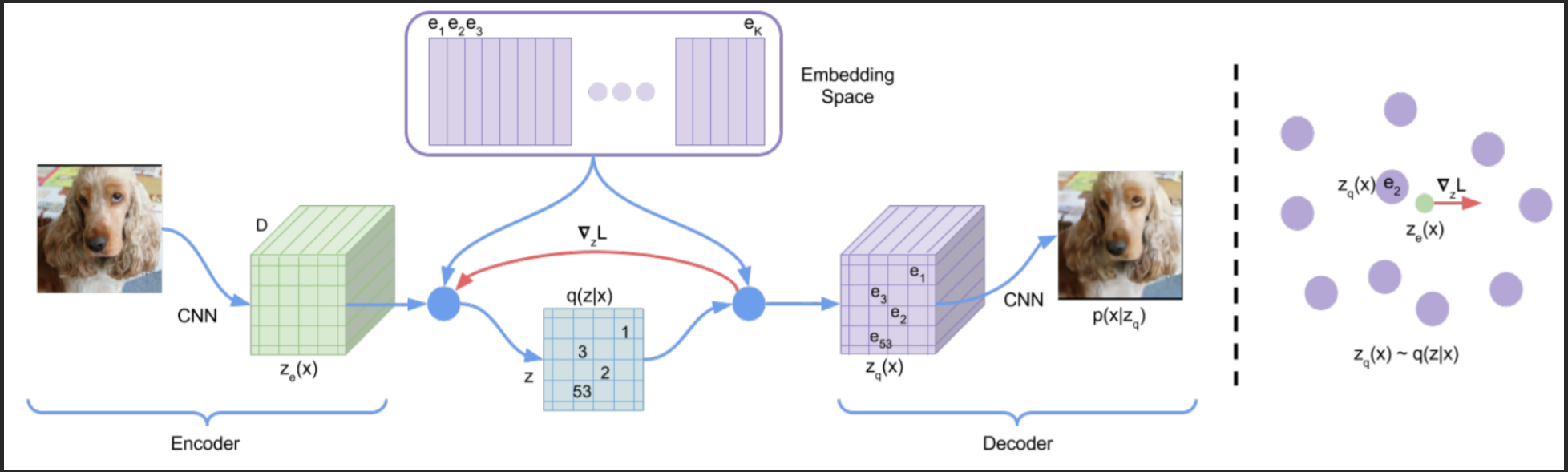


$$h'_t = e_k$$

$$k = argmin_j \|h_t - e_j\|_2^2$$

$$\mathcal{L}_{symbolic} = \|h_t - stop\ gradient(e_k)\|_2^2$$

VQ-VAE

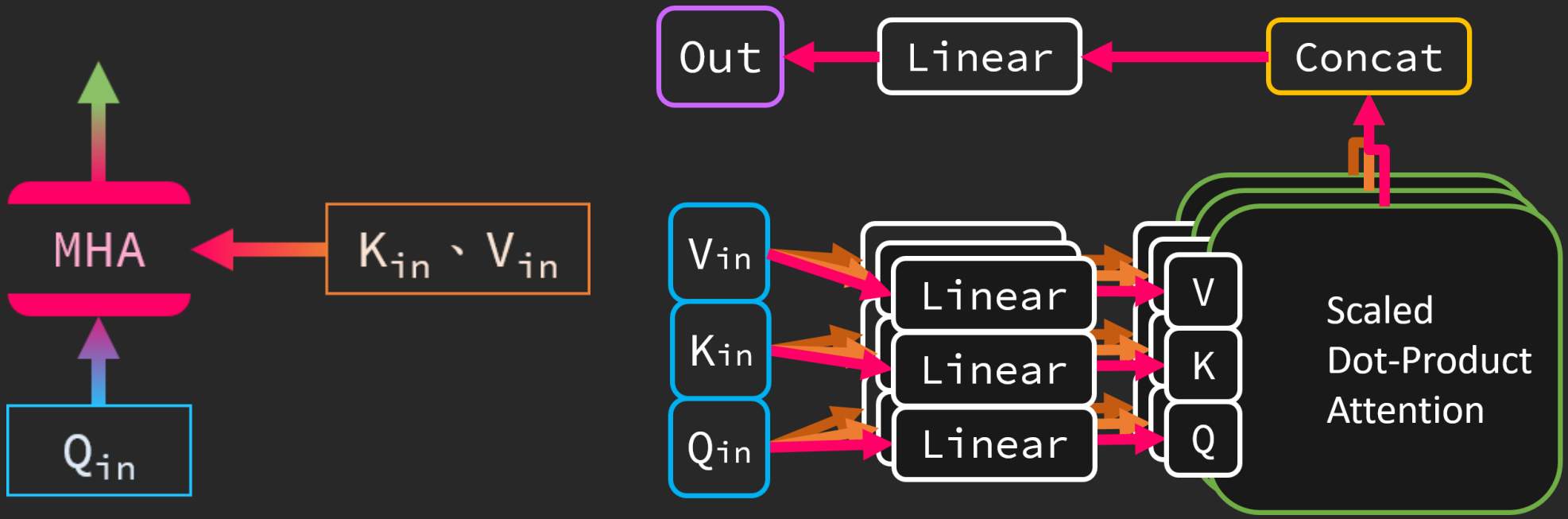


The hidden vector output by the Encoder is vector quantized before being input to the Decoder for generation.

Two-step training:

- Train Encoder-CodeBook-Decoder.
- Train Pixel CNN to generate discrete hidden variants. (Q(z|x) in the figure above)

Multi Head Attention



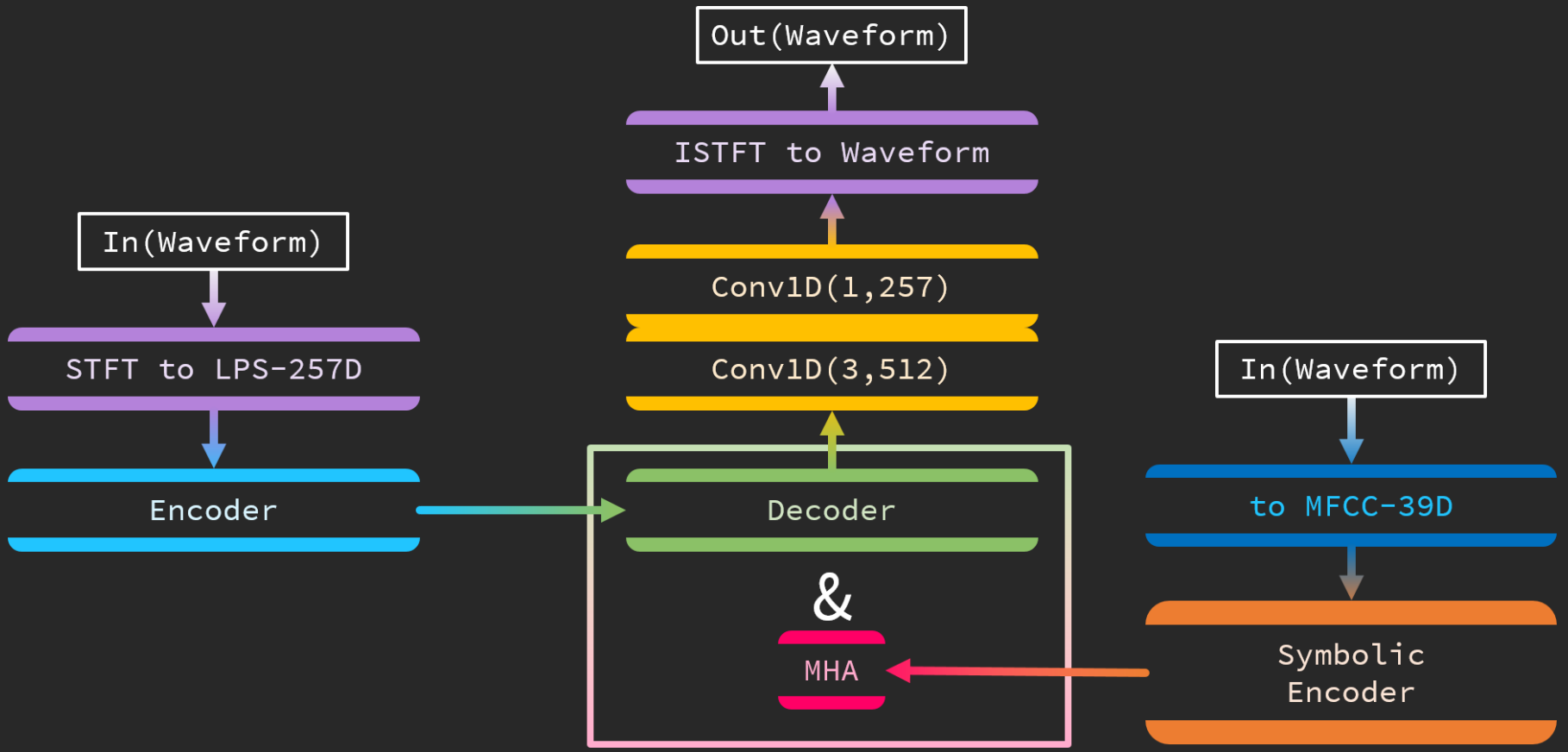
Loss

$$\mathcal{L}_{mse} = \frac{1}{N}\sum_{i=1}^{N} \left\| Dec\left(Enc(x_i)\right) - y_i \right\|_2^2$$

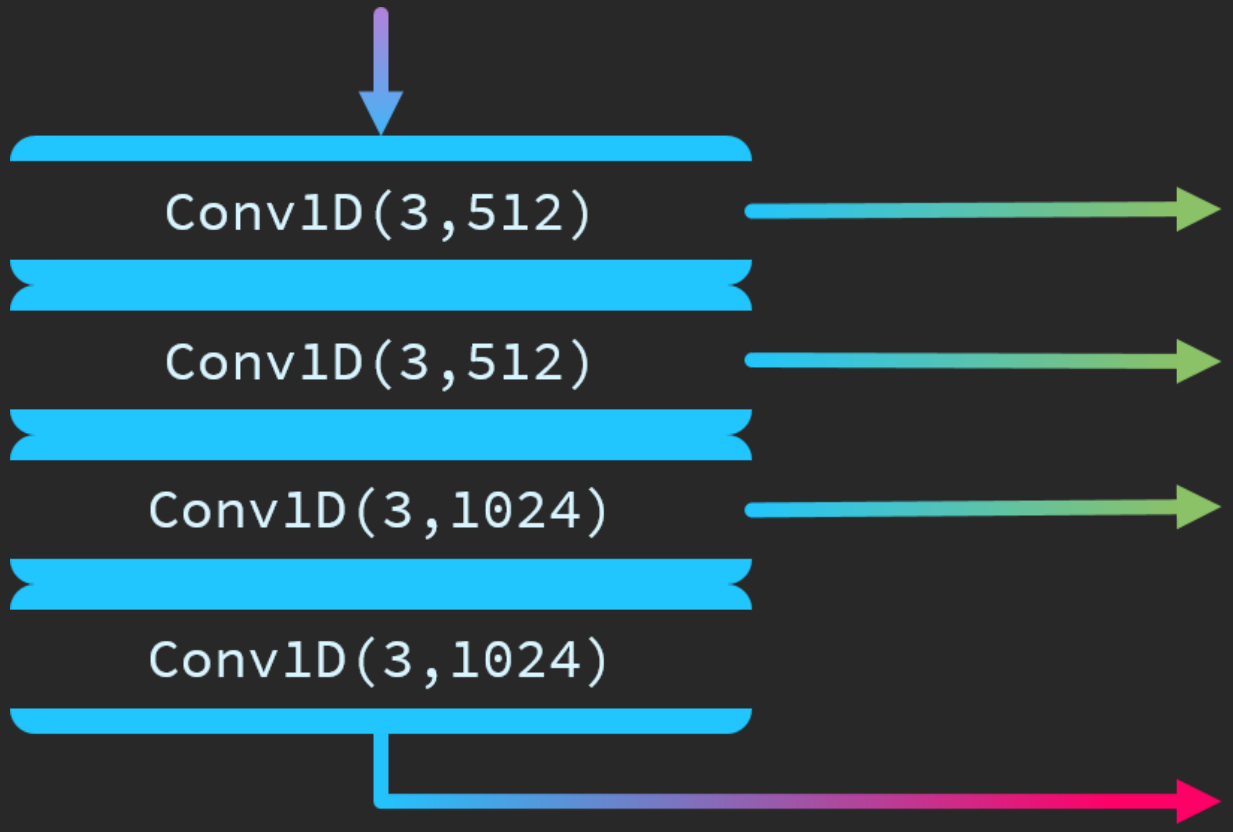
$$\mathcal{L}_{symbolic} = \| h_t - stop\ gradient(e_k) \|_2^2$$

$$\mathcal{L}_{total} = \mathcal{L}_{mse} + \lambda \cdot \mathcal{L}_{symbolic}$$

Architecture

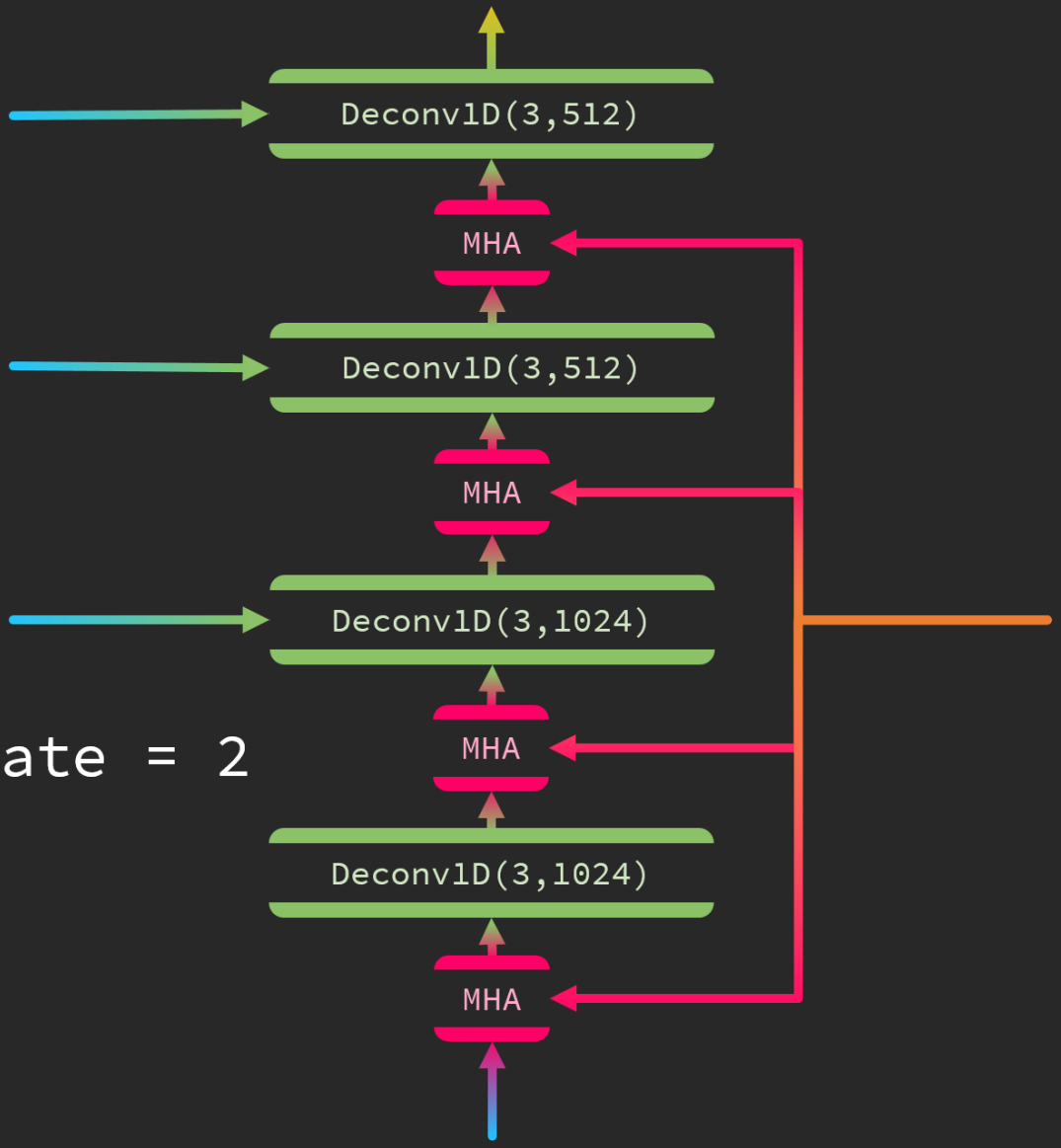


Encoder



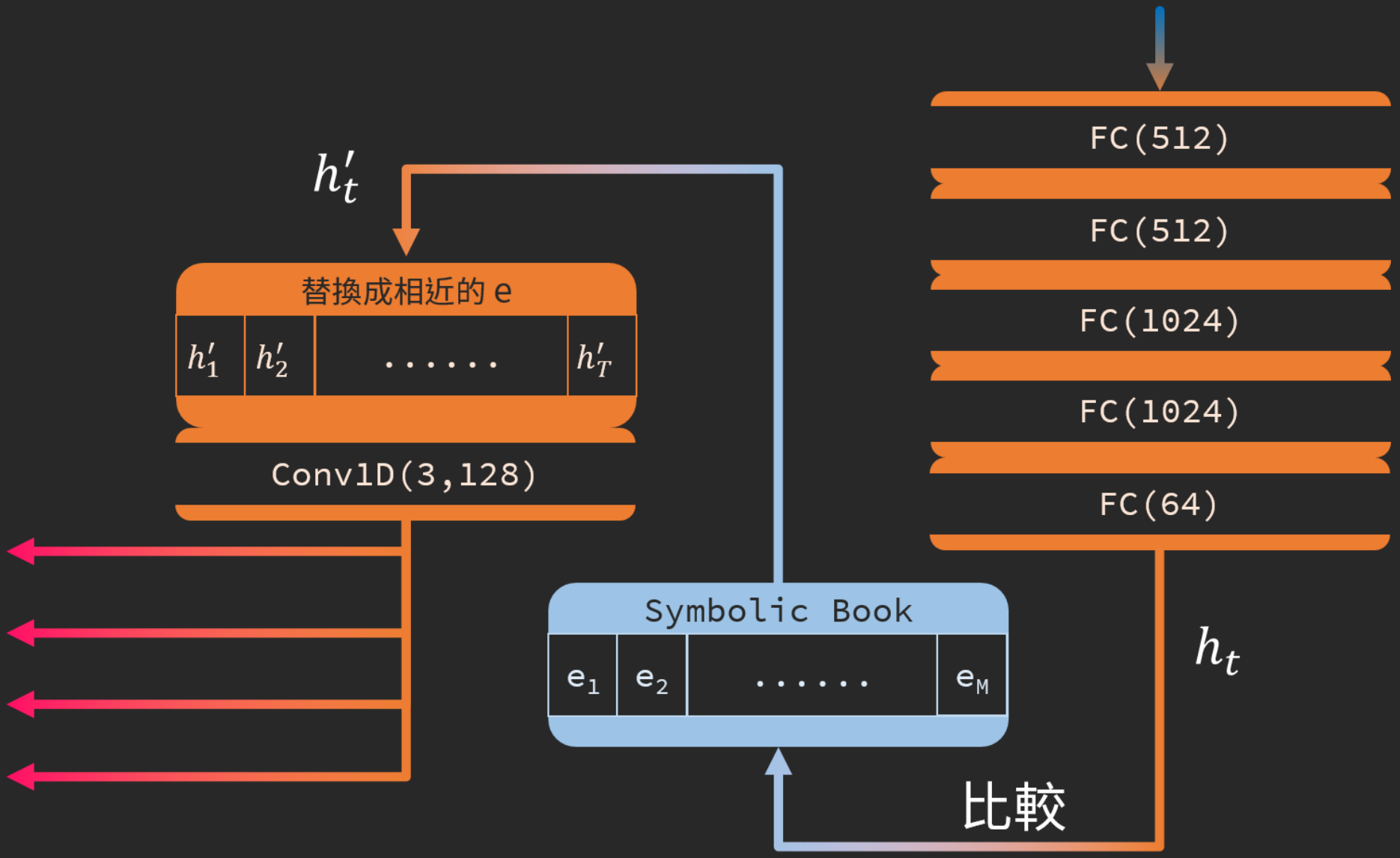
Downsample Rate = 2

Decoder

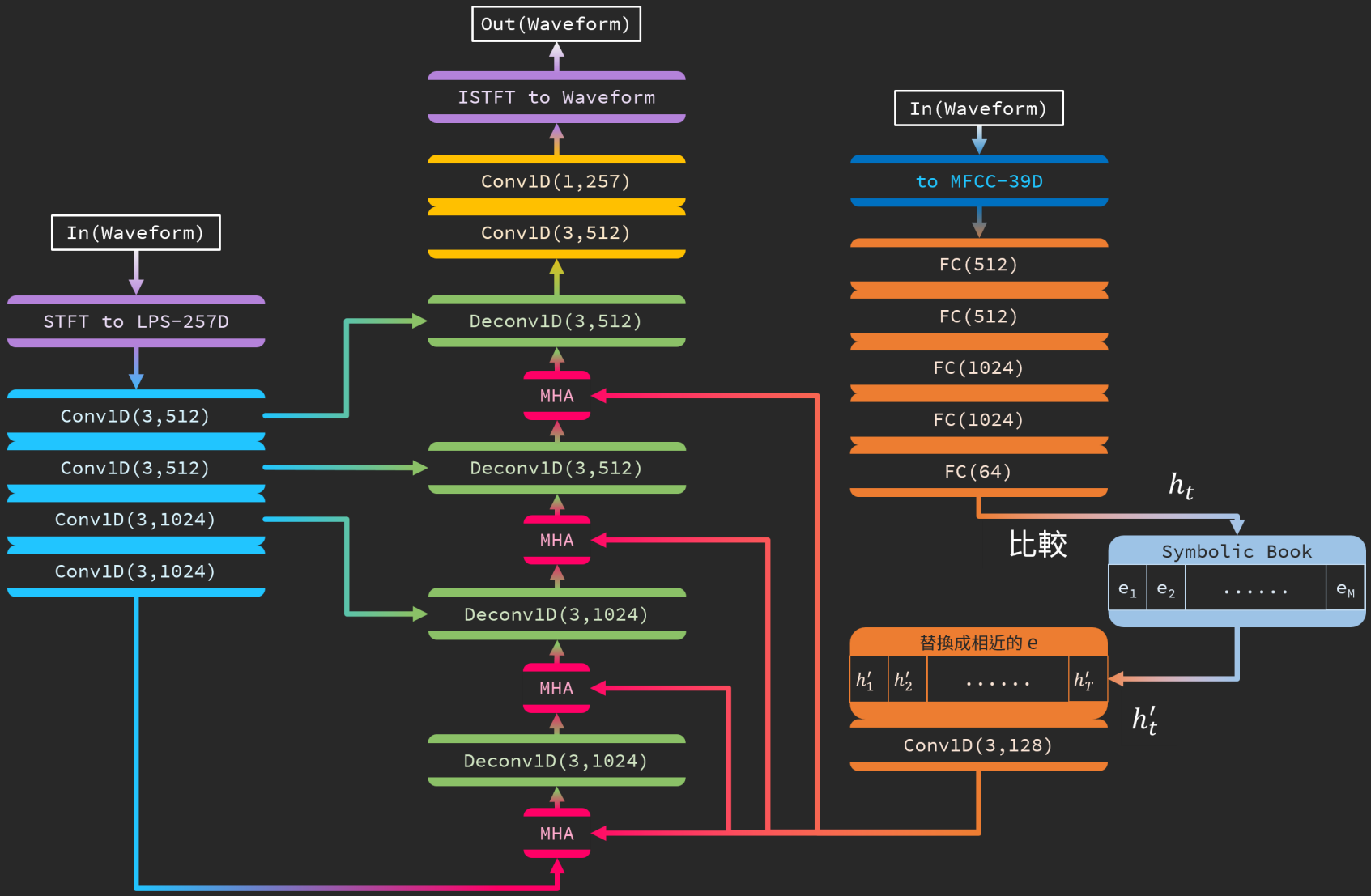


Upsample Rate = 2

Symbolic Encoder



Complete Model



Experiments

- Sample rate：16kHz
- Hamming window
  - size：512
  - stride：256
- Input frame：64
- Optimizer：Adam
  - learning rate：0.0001
  - beta 1：0.5
  - beta 9：0.9

Data Set

- Speech：TIMIT
- Noise：PNL 100 Nonspeech Sounds

The voice and noise are mixed with different signal-to-noise ratio (SNR) as input, and the clean voice is used as ground truth.

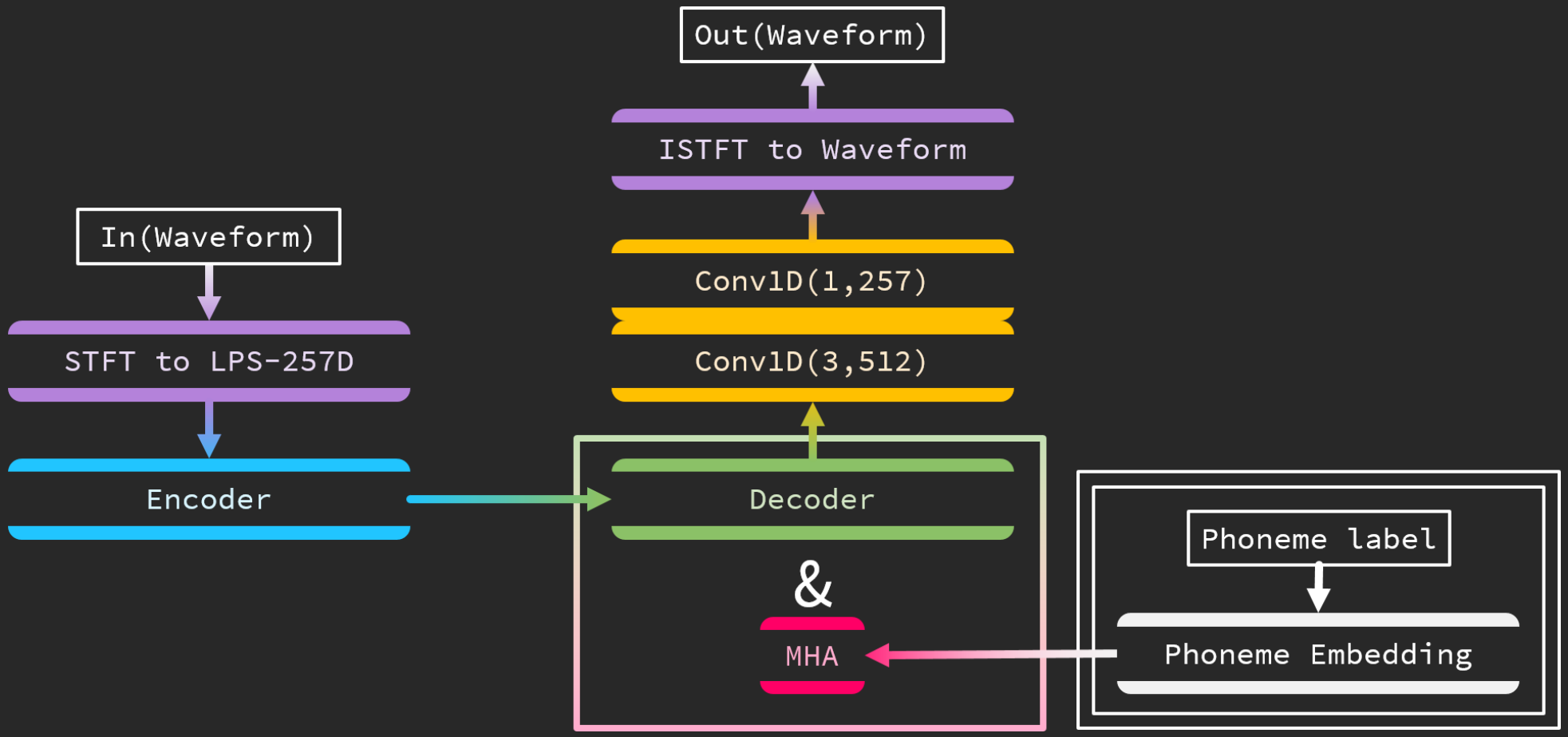
Baseline Model

1. U-Net
   Just a U-Net.

2. U-Net-MOL
   using multi objective learning.

3. Oracle
   Replace Symbolic Encoder with phoneme (phoneme) embedding of input speech.

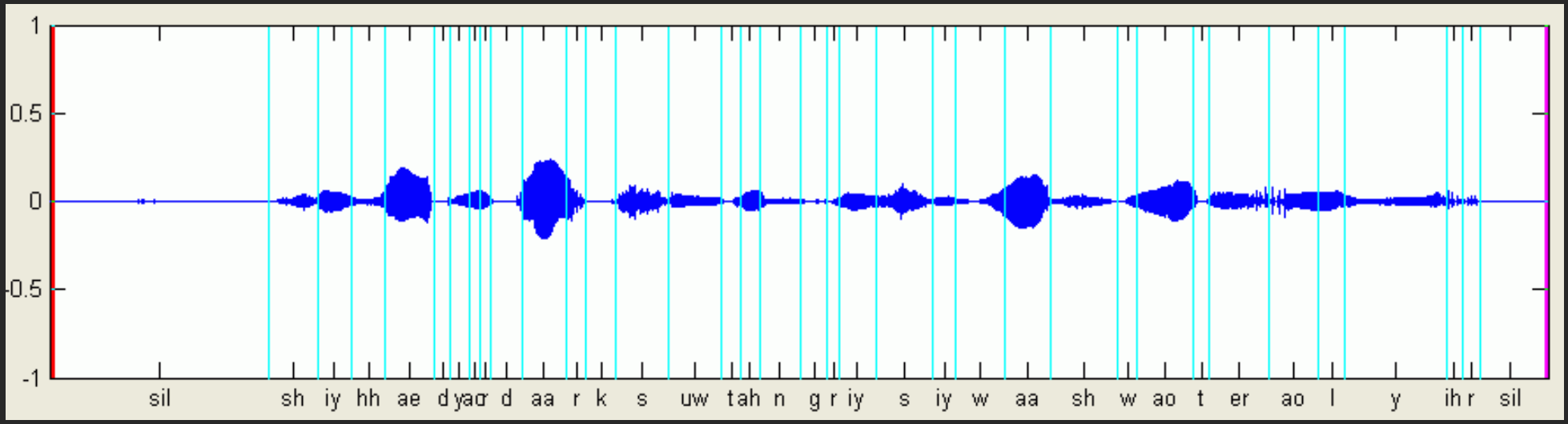
Oracle


Out(Waveform)
ISTFT to Waveform
Conv1D(1,257)
Conv1D(3,512)
In(Waveform)
STFT to LPS-257D
Encoder
Decoder
&
MHA
Phoneme label
Phoneme Embedding

Oracle

In the TIMIT data set, in addition to speech, the phonemes at each time point are also marked.

Oracle embeds the phonemes as the input of MHA's Key and value.



Phoneme example (source)

Experiments

SNR	Noisy		U-Net		U-Net-MOL		Proposed (64)		Oracle	
	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI	PESQ	STOI
-6	1.213	0.532	1.685	0.602	1.800	0.619	**1.828**	**0.624**	1.961	0.703
-3	1.353	0.598	1.880	0.669	1.974	0.681	**2.045**	**0.693**	2.140	0.741
0	1.517	0.669	2.071	0.725	2.140	0.736	**2.240**	**0.750**	2.306	0.776
3	1.702	0.739	2.237	0.770	2.290	0.779	**2.416**	**0.794**	2.456	0.806
6	1.902	0.823	2.387	0.805	2.424	0.813	**2.581**	**0.830**	2.592	0.831
Avg.	1.537	0.669	2.052	0.714	2.126	0.725	**2.222**	**0.738**	2.291	0.771

1. Models using Symbolic Encoder and MHA have better results than Baseline Models other than Oracle.

2. Oracle proved that adding phoneme information to the Speech Enhancement question can improve the performance of the model.

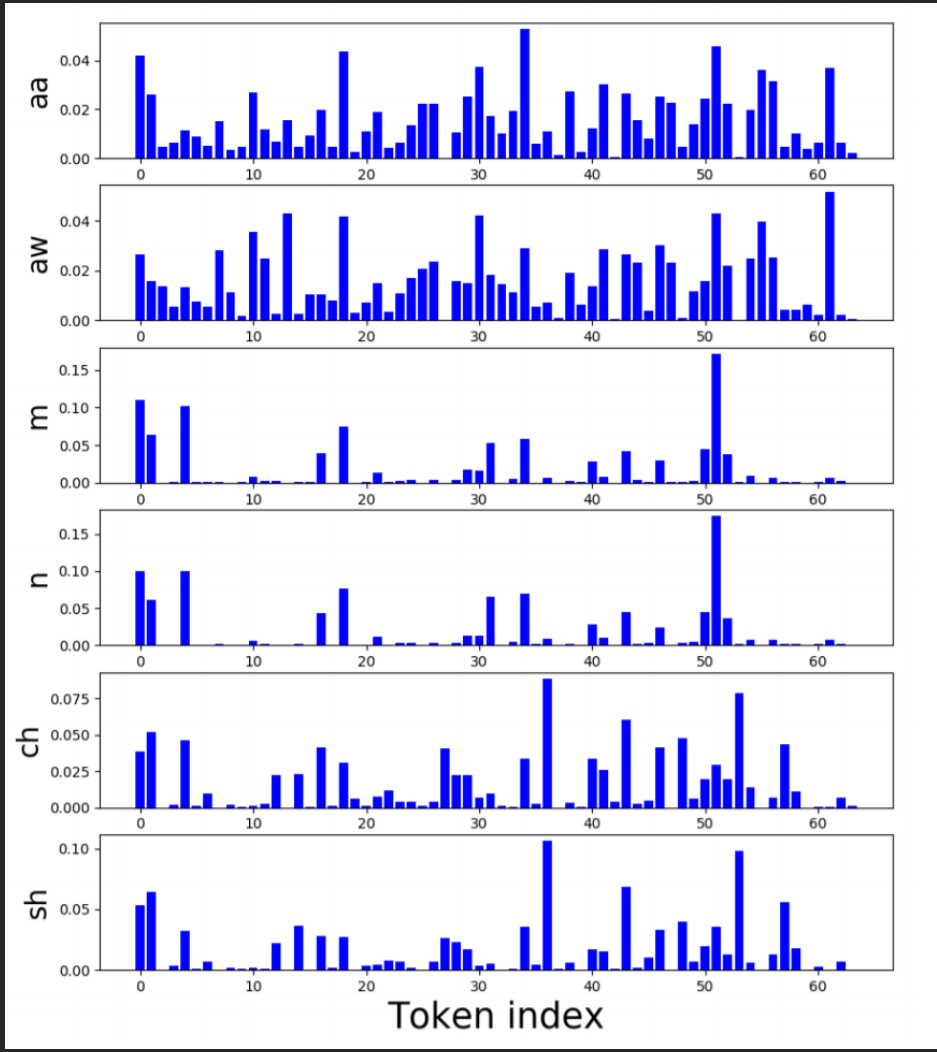
Experiments

Book size M	PESQ	STOI
39	2.061	0.711
64	**2.108**	**0.713**
128	2.027	0.712
256	2.041	0.711

"Index collapse" will appear when the Book Size is too large.

As a result, part of the symbolic vector will not be used.

Experiments

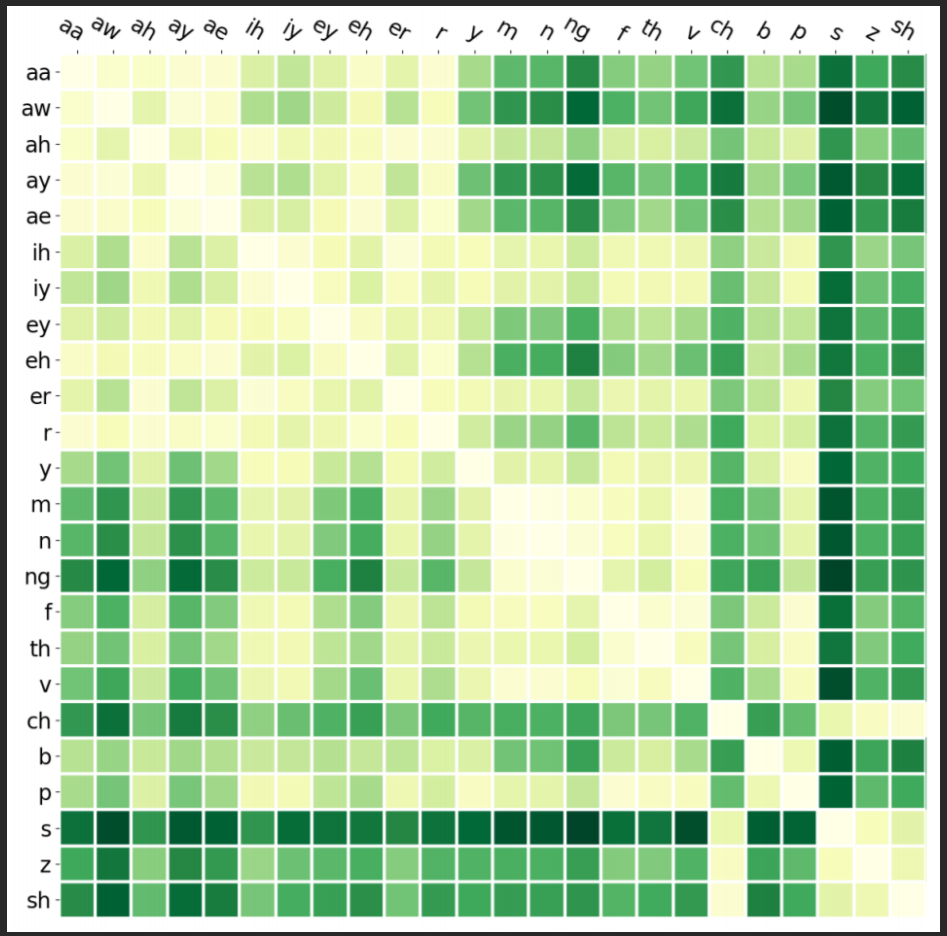


The horizontal axis represents Symbolic index.

And each value above represents the number of times the pronunciation corresponds to a certain Symbolic.

Similar pronunciations have similar distributions.

Experiments



Calculate the similarity of pronunciation distribution.

Some different pronunciations will give similar results because the input contains noise.

Conclusion

Use Symbolic Encoder to first encode speech into high-level phoneme-like content, and extract features through Multi Head Attention, which can achieve better results than before.

Conclusion

Review：

- The structure of Symbolic Encoder + Multi Head Attention is to establish a basic skeleton for speech (speaking content).
- U-Net's shortcut can assist in adding intonation information (tonal fluctuations, timbre, etc.) to the basic skeleton.
- In the second generation of VQ-VAE, you can try to use a multi-layer (multi-resolution) method to retain more acoustic features for the model