Vector Quantized Attention for Speech Enhancement

Advisor :Chung-Hsien Wu

Presenter :Jen-Hung Huang

Outline

- Speech Enhancement
- Methodology
- Problems
- Possible Solutions
- Schedule

Speech Enhancement

The real world is full of various background noises. These noises can pollute the speech signal and reduce the accuracy of ASR, hearing aids and other speech tasks.

However, when humans face these sounds disturbed by background noise, they can reduce the noise interference by adjusting their focus.

Speech Enhancement

When they understand the speaker's content, they can even use their knowledge of the language to recover damaged speech signals.

Therefore, the research focus of this monograph will focus on how to use the attention mechanism and acoustic units to suppress the damage caused by noise and reconstruct clean speech.

Methodology

Incorporating Symbolic Sequential Modeling For Speech Enhancement

U-Net

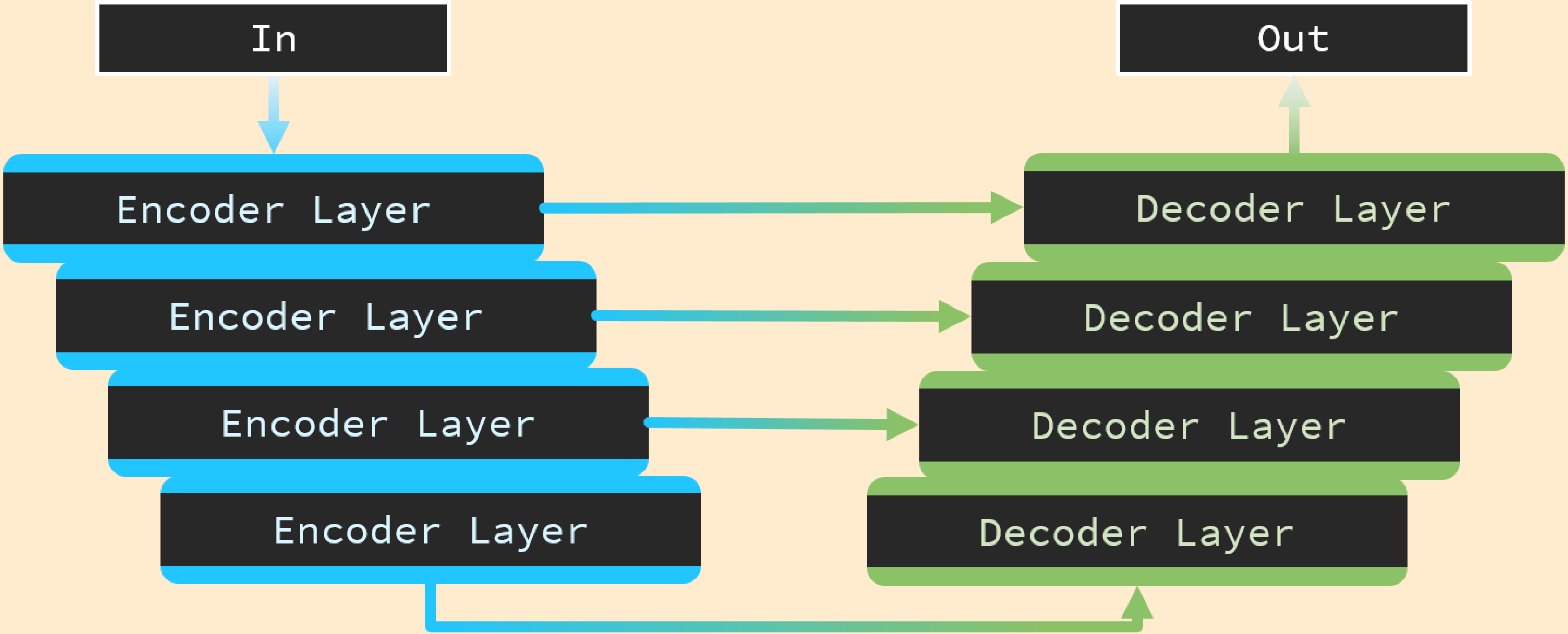
+

VQ-VAE

+

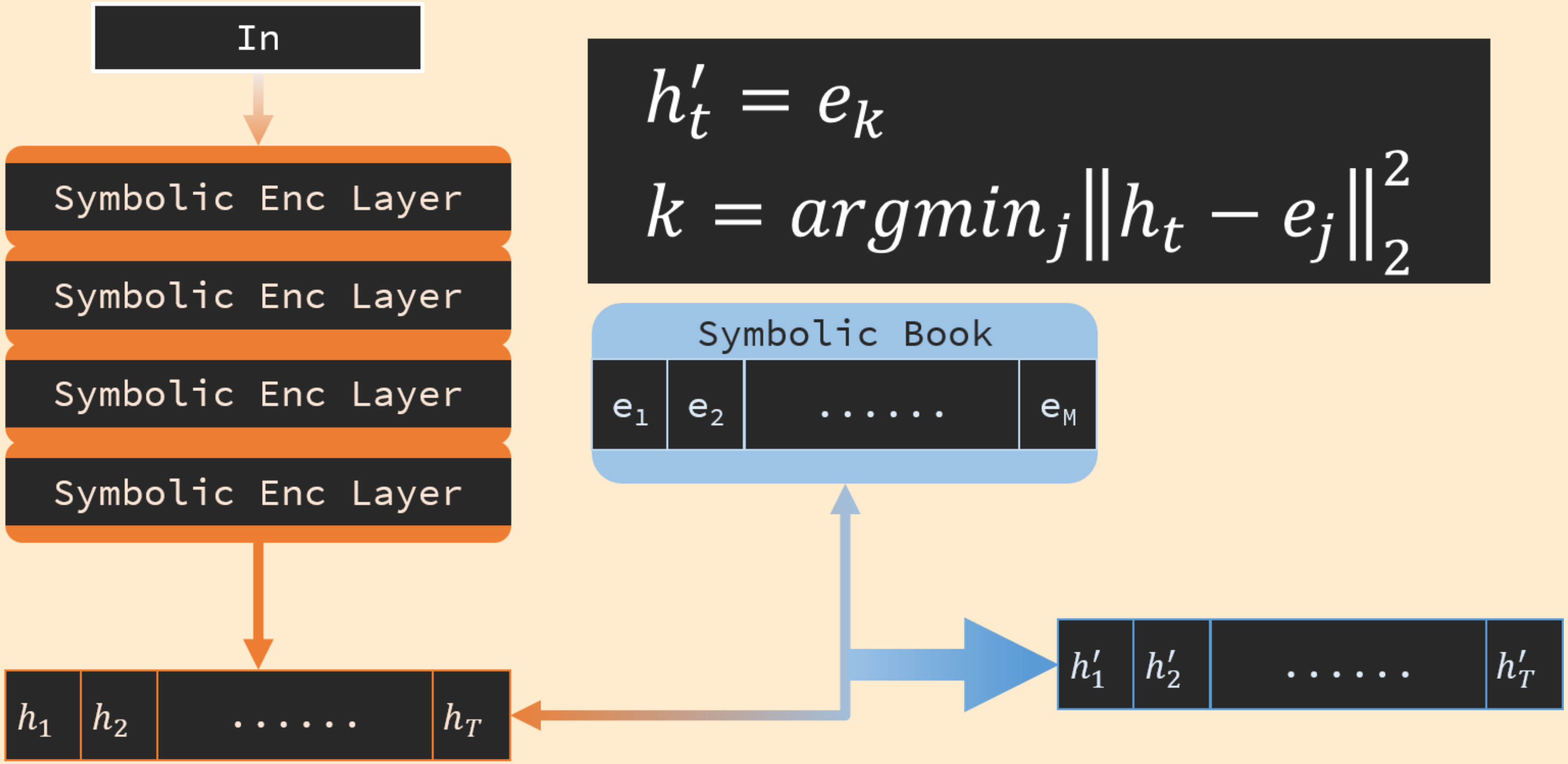
Multi Head Attention

U-Net



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

Symbolic Encoder

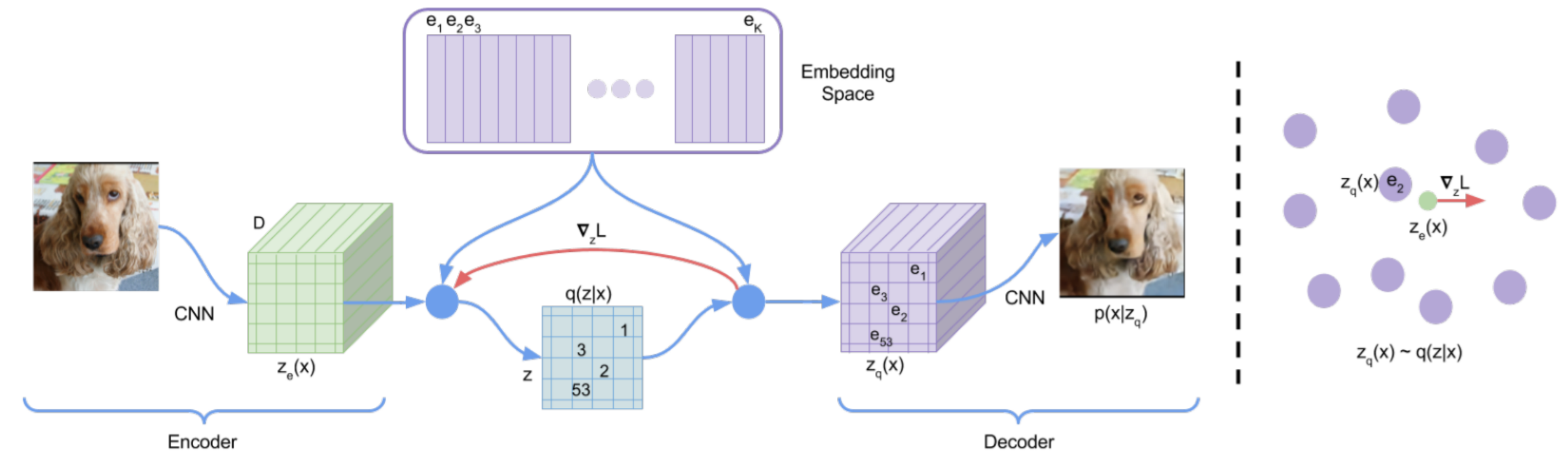


$$h'_t = e_k$$

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

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

VQ-VAE

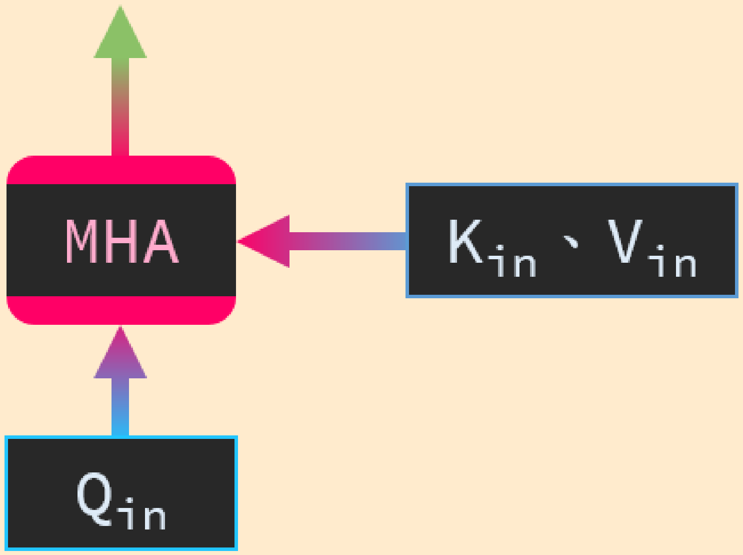


The hidden vector output by the Encoder is vector quantized before being input to the Decoder for generation.

Two-step training：

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

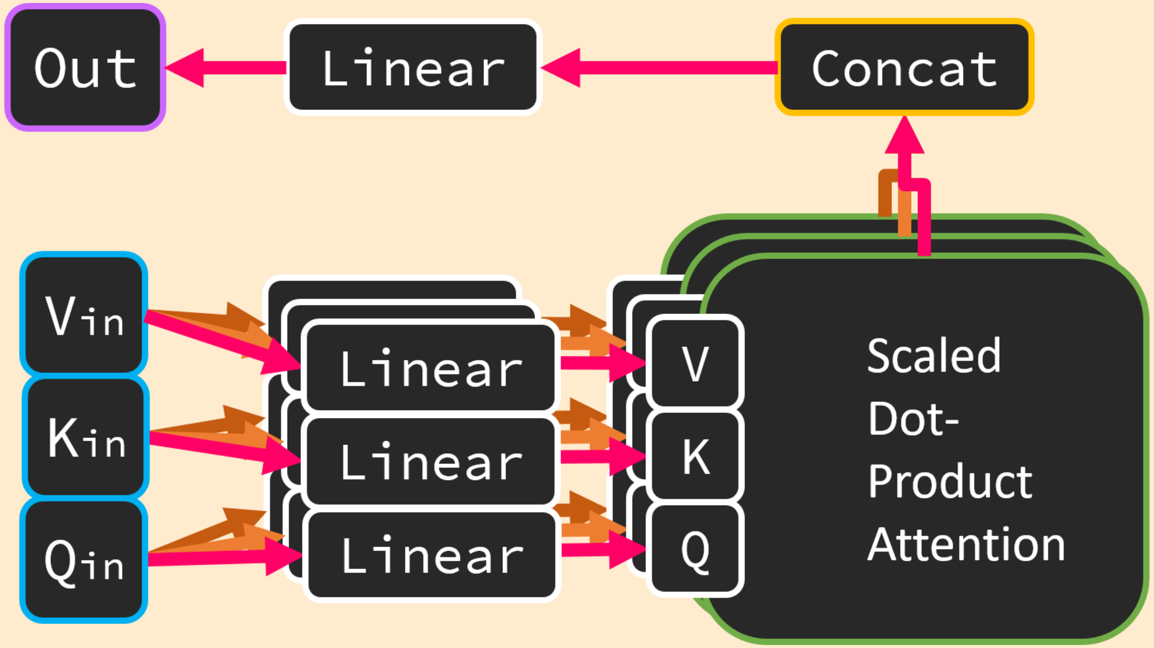
Multi Head Attention



MHA

K_{in}、V_{in}

Q_{in}



Out

Linear

Concat

V_{in}

K_{in}

Q_{in}

Linear

Linear

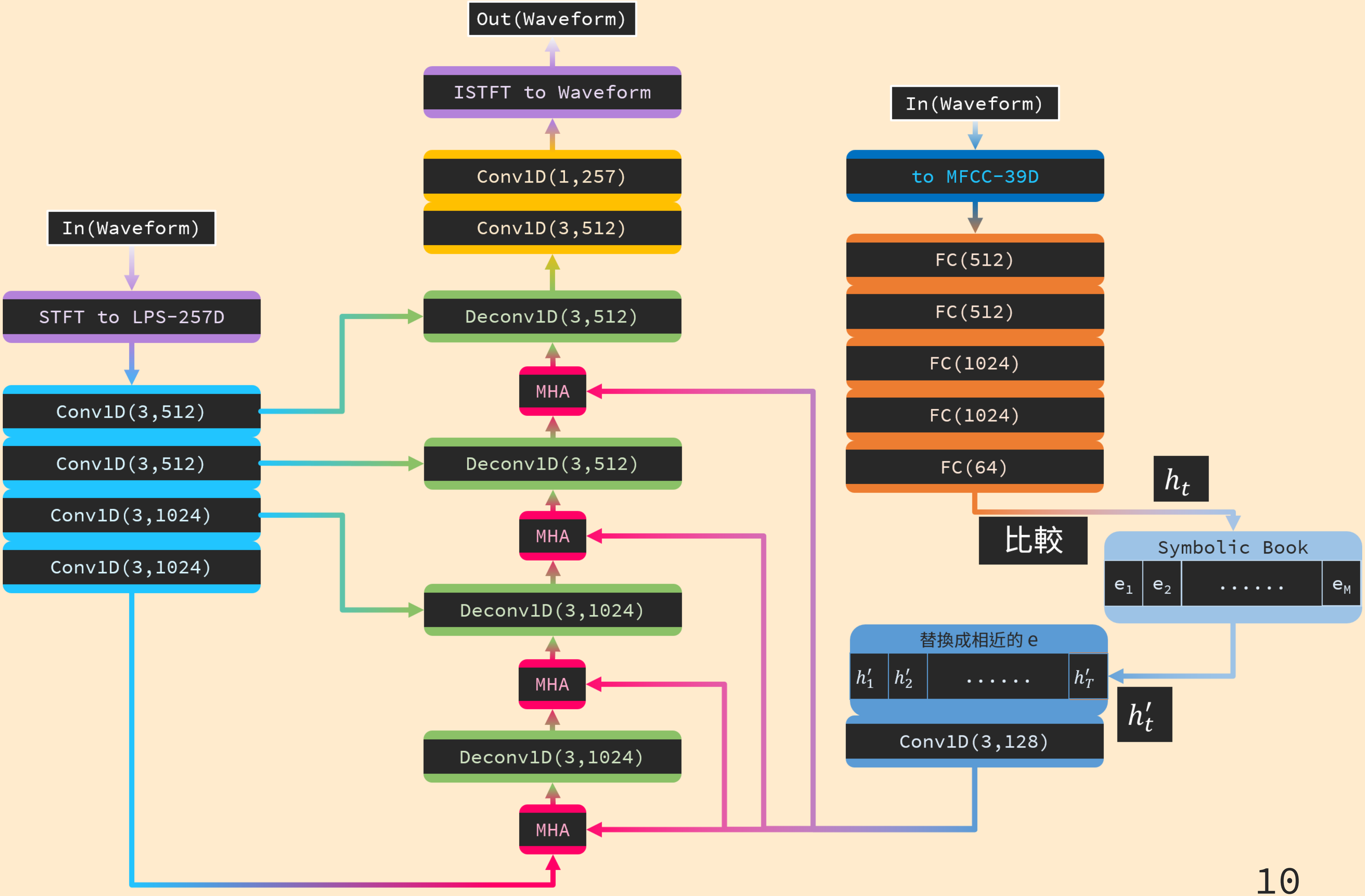
Linear

V

K

Q

Scaled Dot-Product Attention



Out(Waveform)

ISTFT to Waveform

Conv1D(1,257)

Conv1D(3,512)

Deconv1D(3,512)

MHA

Deconv1D(3,512)

MHA

Deconv1D(3,1024)

MHA

Deconv1D(3,1024)

MHA

In(Waveform)

STFT to LPS-257D

Conv1D(3,512)

Conv1D(3,512)

Conv1D(3,1024)

Conv1D(3,1024)

In(Waveform)

to MFCC-39D

FC(512)

FC(512)

FC(1024)

FC(1024)

FC(64)

h_t

比較

Symbolic Book

e_1 e_2 ...... e_M

替換成相近的 e

h'_1 h'_2 ...... h'_T

h'_t

Conv1D(3,128)

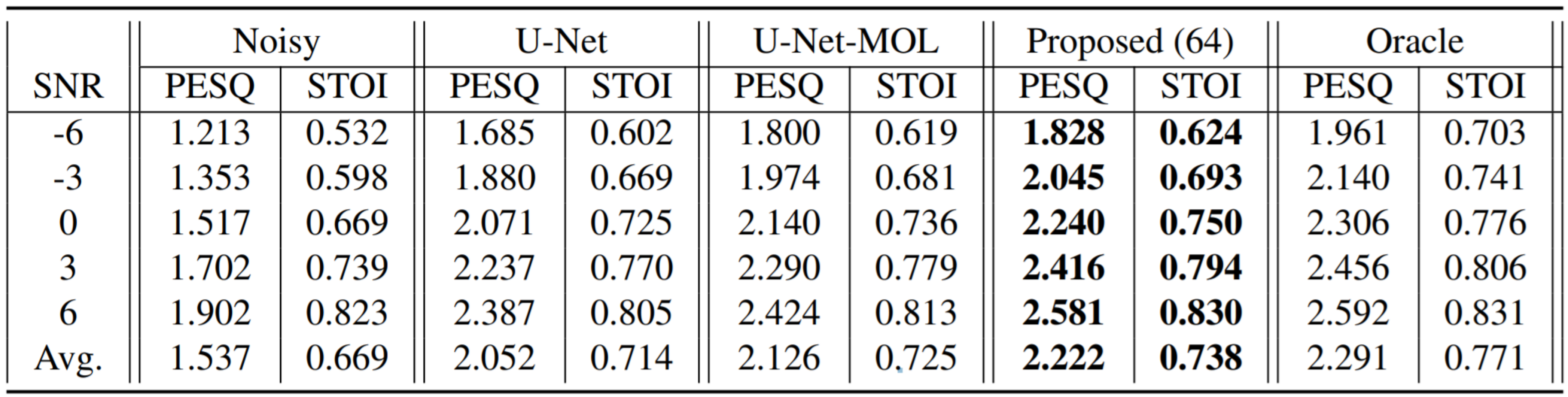
Loss

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

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

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

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

Problems

1. The experiment did not reach the desired result on the evaluation criteria of PESQ and STOI.

2. Even if the correct acoustic information is given to the Oracle model, there is not much improvement compared with the Proposed model.

Possible Solutions

- Change the current Encoder-Decoder :

  Even if the correct acoustic information is used, it cannot bring significant improvement.

  Indicates that the current Encoder-Decoder may not be able to extract important information.

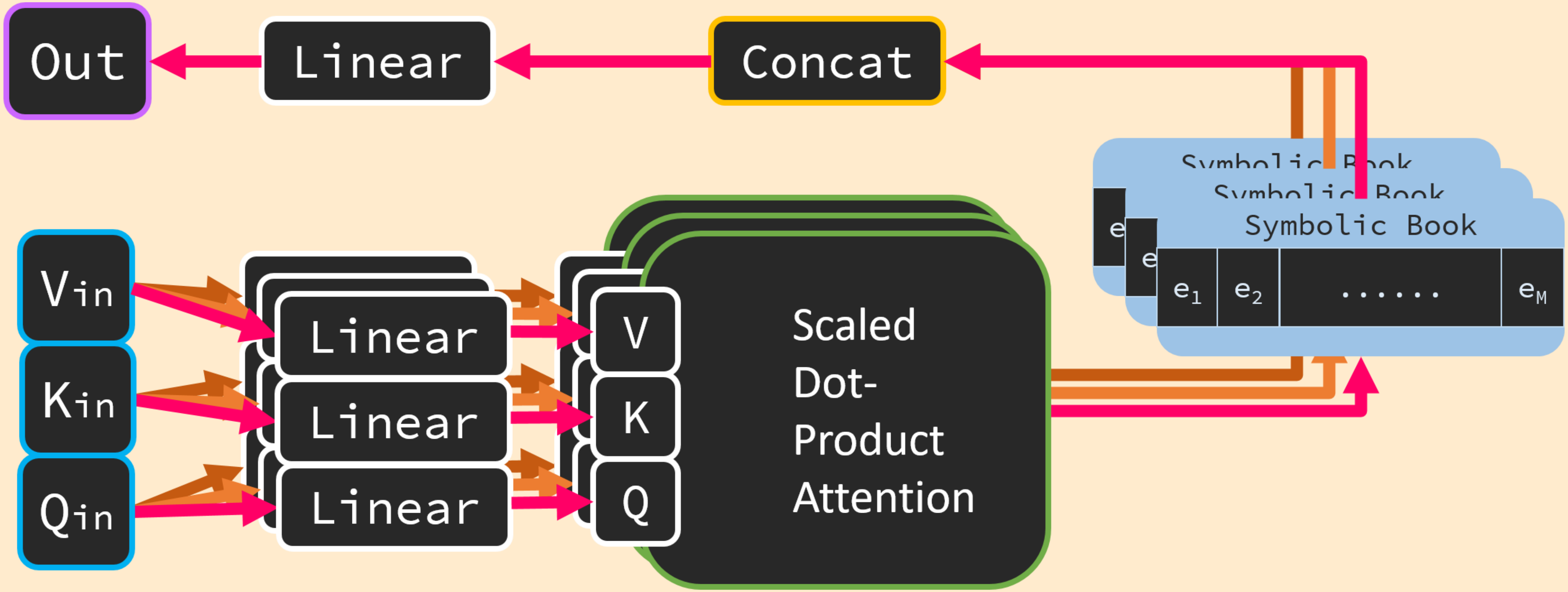
Possible Solutions

- Use the multi-layer (multi-resolution) of VQ-VAE 2 to try to preserve the sound characteristics of different fineness.

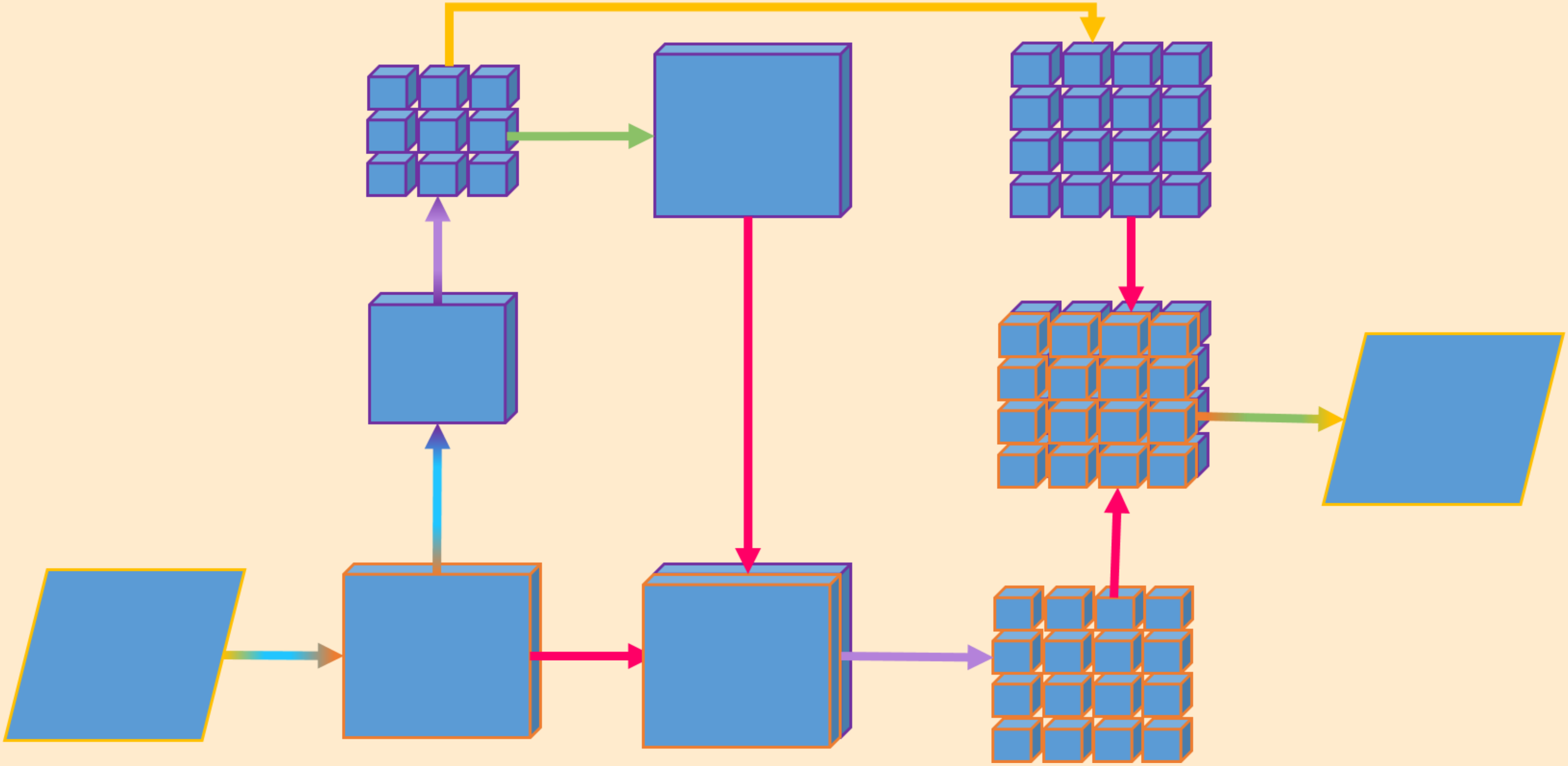
Challenge

- Calculate from Time Domain.
- Use Phase information.

MVQHA

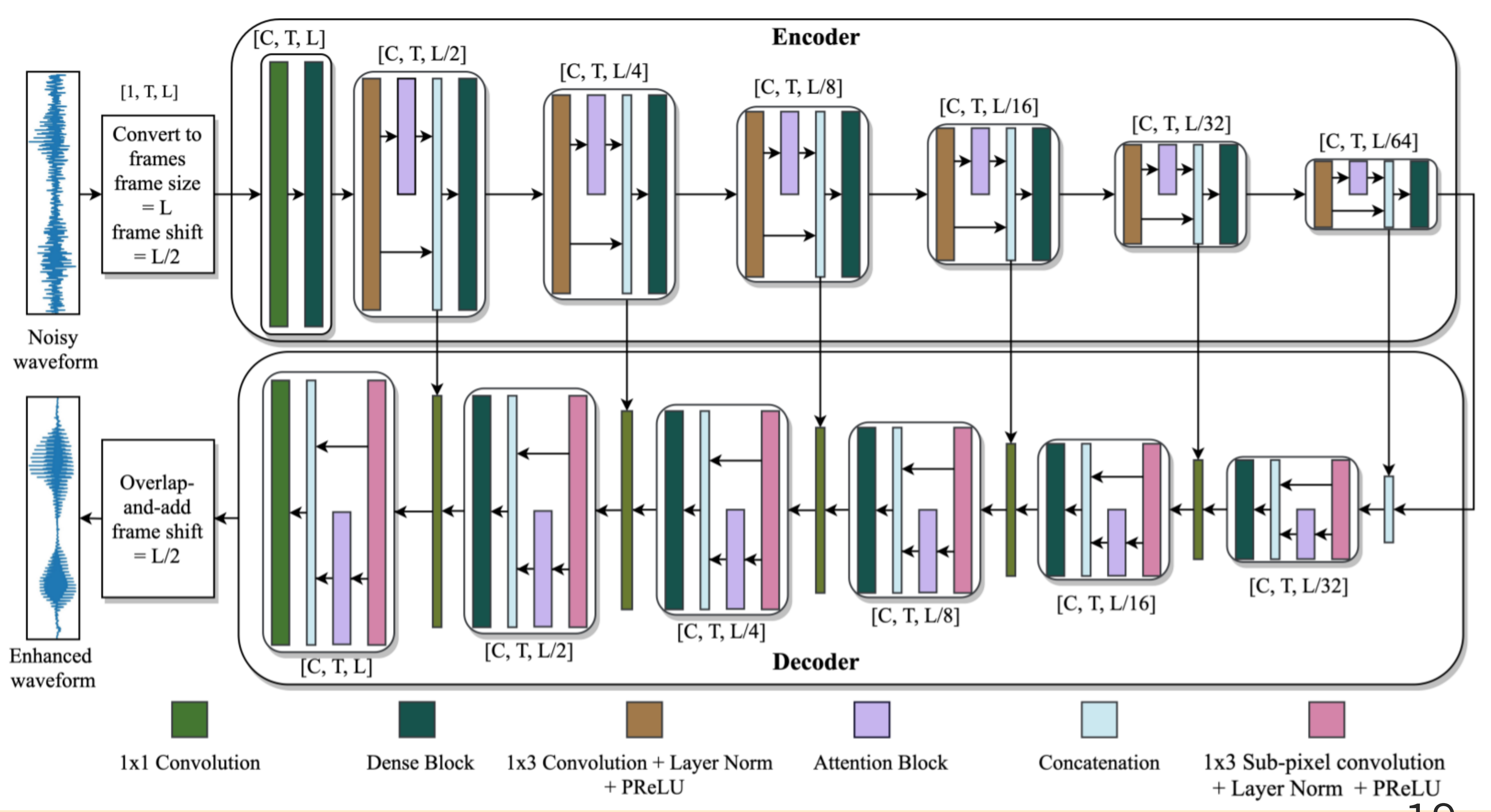

Out
Linear
Concat
Symbolic Book
Symbolic Book
Symbolic Book
e
e
e_1
e_2
....
e_M
Vin
Kin
Qin
Linear
Linear
Linear
V
K
Q
Scaled
Dot-
Product
Attention

VQ-VAE2



Possible Solutions

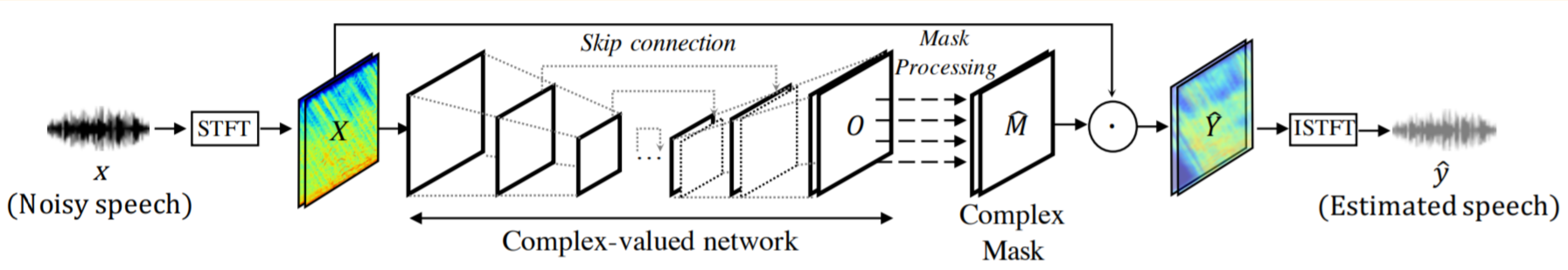
Dense CNN



Possible Solutions

Deep Complex U-Net

Phase-Aware Speech Enhancement with Deep Complex U-Net



Schedule

	9	10	11	12	13	14	15	16
Prepare the Dataset	■							
VQ-MHA-U-Net		■	■					
Oracle Model		■	■					
Experiment A			■	■				
Model based on VQ-VAE-2					■	■		
Change the model of Encoder&Decoder					■	■		
Challenge					■	■		
Experiment B						■	■	
Demo								■