Vector Quantized Attention for Speech Enhancement

Advisor :Chung-Hsien Wu

Presenter :Jen-Hung Huang

Outline

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

Speech Enhancement

現實世界中充滿了各式各樣的背景雜音，這些噪音會汙染語音訊號，降低如ASR、助聽器與其他語音任務的正確性。

然而人類在面對這些被背景雜訊干擾的聲音時，可以藉由調整注意力集中處來減低噪音的干擾。
甚至能在理解講者的說話內容時，利用自身具備的語言知識恢復受損的語音訊號。

因此，本次專論研究重點將會著重於如何使用注意力機制與聲學單元來抑制雜訊造成的破壞並重建乾淨的語音。

Methodology

Incorporating Symbolic Sequential Modeling For Speech Enhancement

U-Net

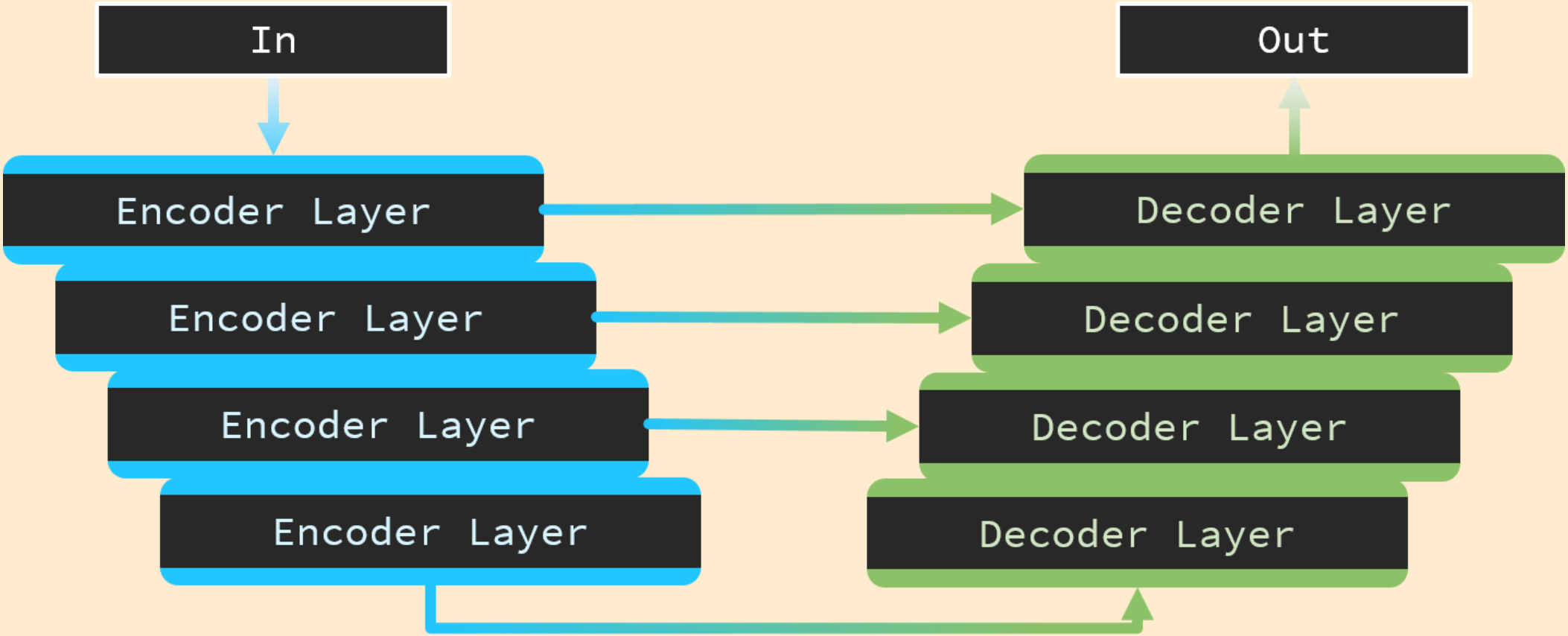
+

VQ-VAE

+

Multi Head Attention

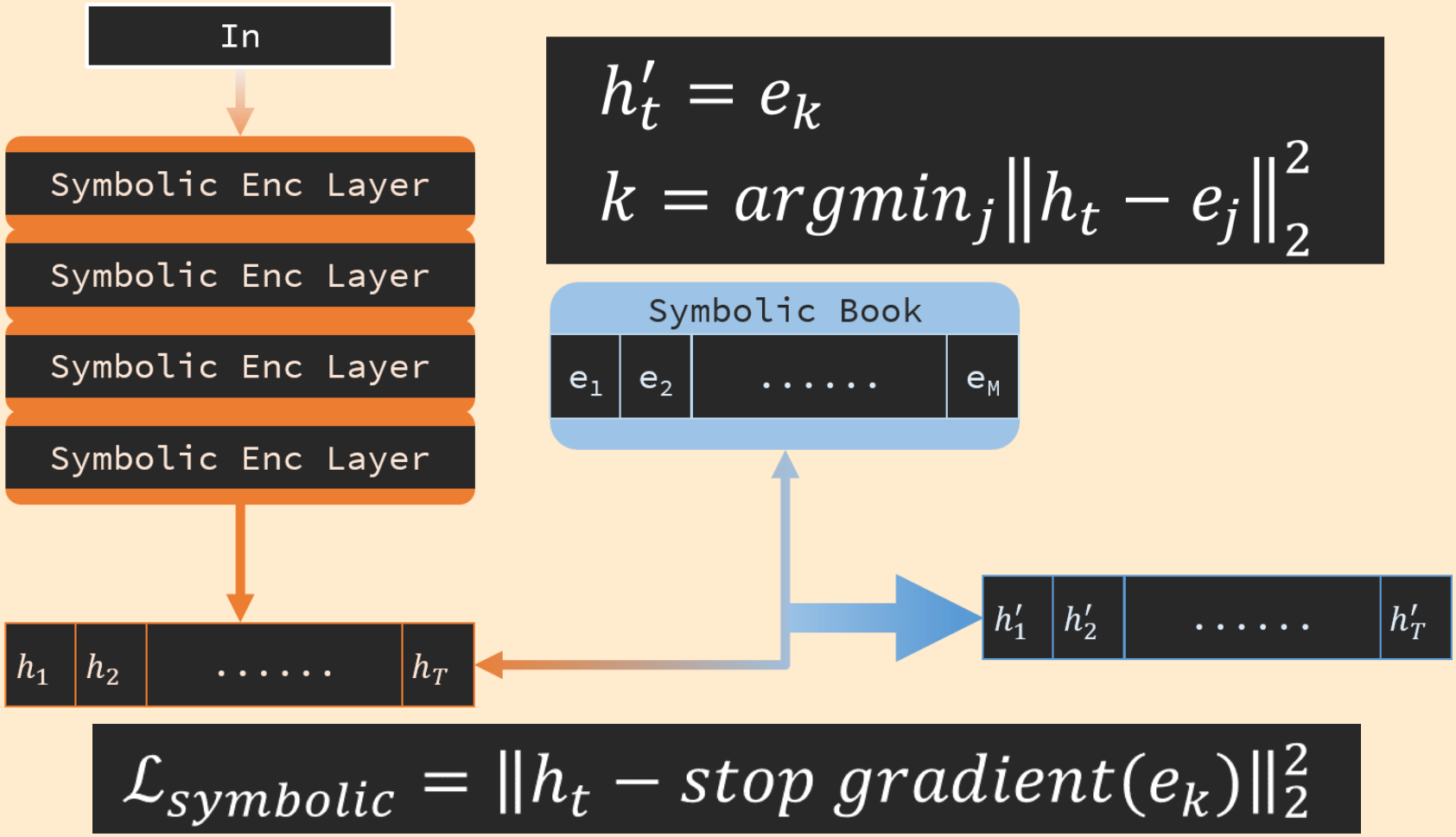
U-Net



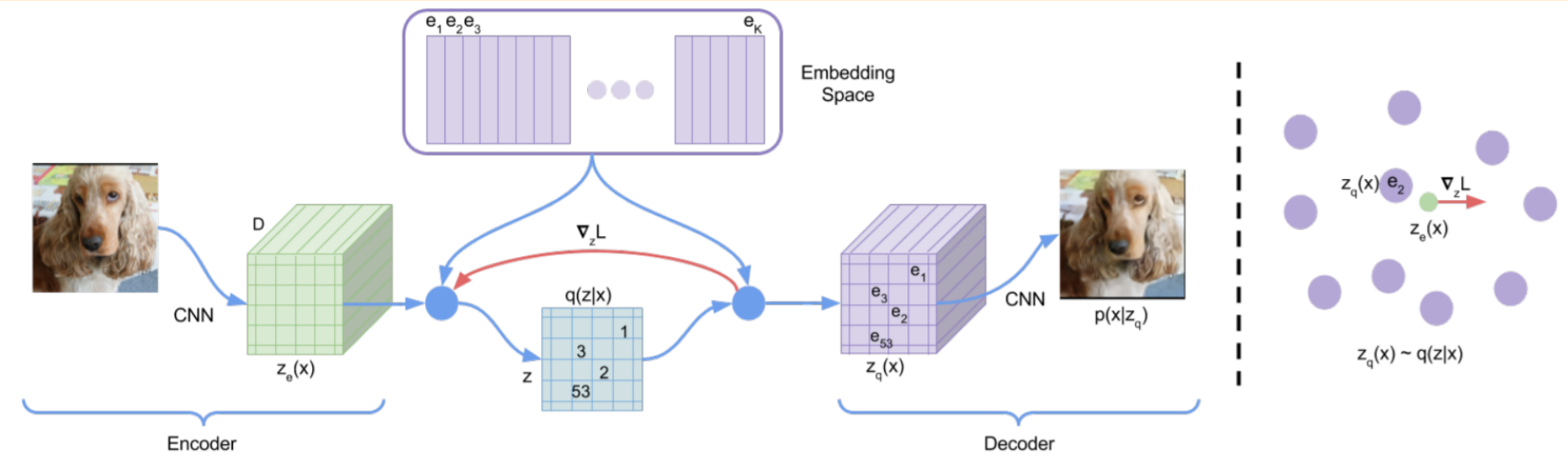
$$\mathcal{L}_{mse} = \frac{1}{N}\sum_{i=1}^{N} \left\| Dec\left(Enc(x_i)\right) - y_i \right\|_2^2$$

Symbolic Encoder

源自於 VQ-VAE 中的 Code Book 概念



VQ-VAE

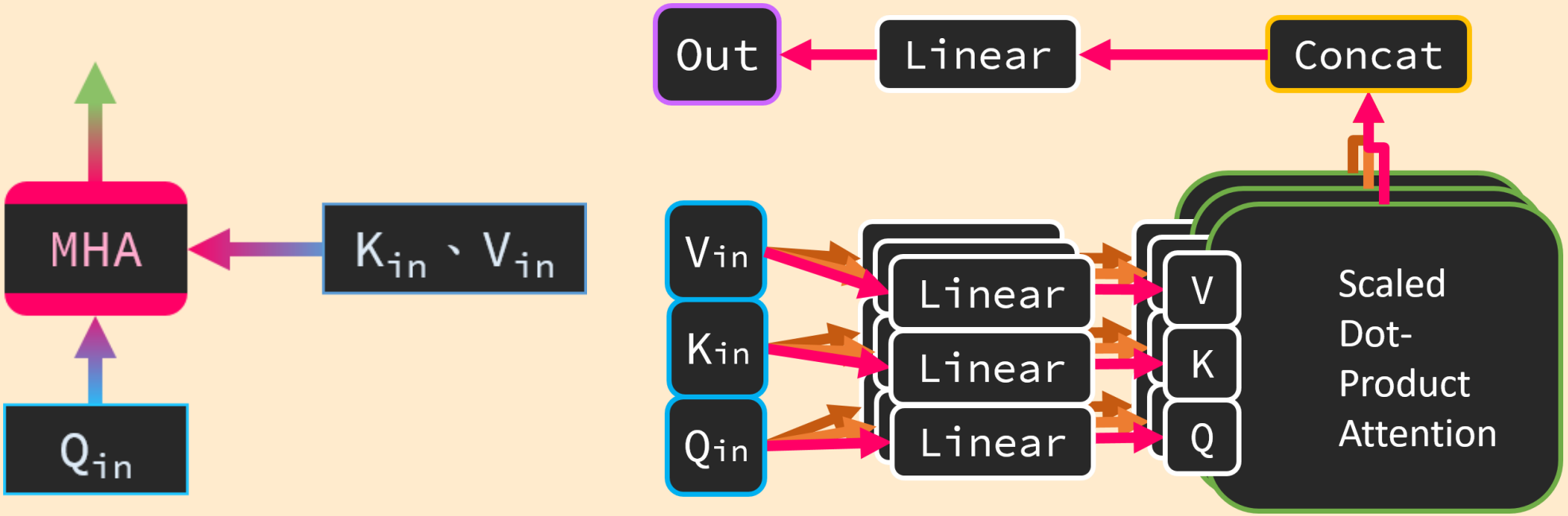


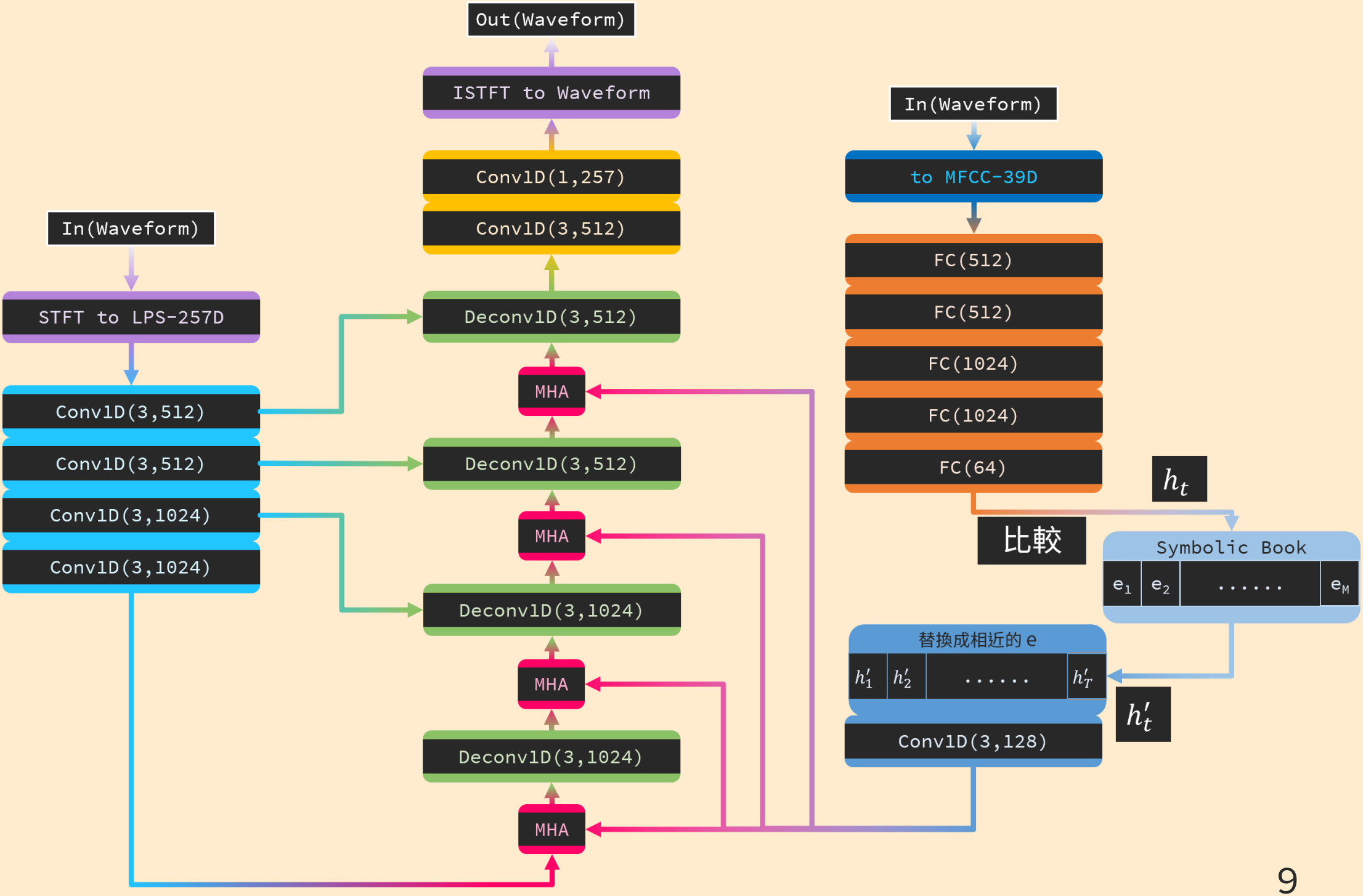
先將 Encoder 輸出的 hidden vector 進行向量量化後才輸入 Decoder 生成

兩步驟訓練：

- 訓練 Encoder-CodeBook-Decoder
- 訓練 Pixel CNN 來生成離散的 hidden variants (上圖的 q(z|x))

Multi Head 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. 在 PESQ 與 STOI 的評估標準上並未到達很理想的結果。
2. 即使是給予了正確聲學資訊的 Oracle 模型，與 Proposed 模型相比也沒有很大的提升。

Possible Solutions

- 更換目前的 Encoder-Decoder：
  即便用上了正確的聲學資訊依然無法帶來顯著提升，
  表示 Encoder-Decoder 可能無法提取重要的資訊。

- 使用 VQ-VAE 2 的多階層(多解析度)，嘗試保留不同細度的聲音特徵。

Possible Solutions

- 更換目前的 Encoder-Decoder：
  即便用上了正確的聲學資訊依然無法帶來顯著提升，
  表示 Encoder-Decoder 可能無法提取重要的資訊。

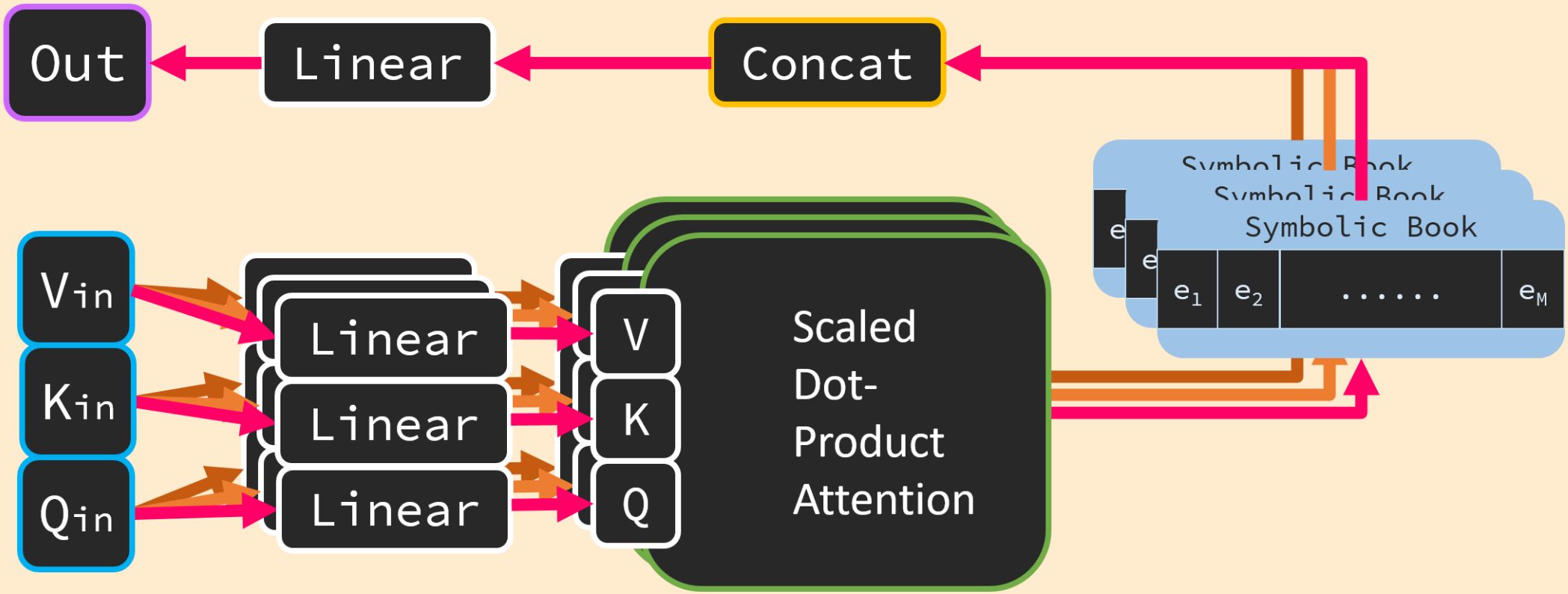
- 使用 VQ-VAE 2 的多階層(多解析度)，嘗試保留不同細度的聲音特徵。

- 挑戰：從 Time Domain 進行計算。

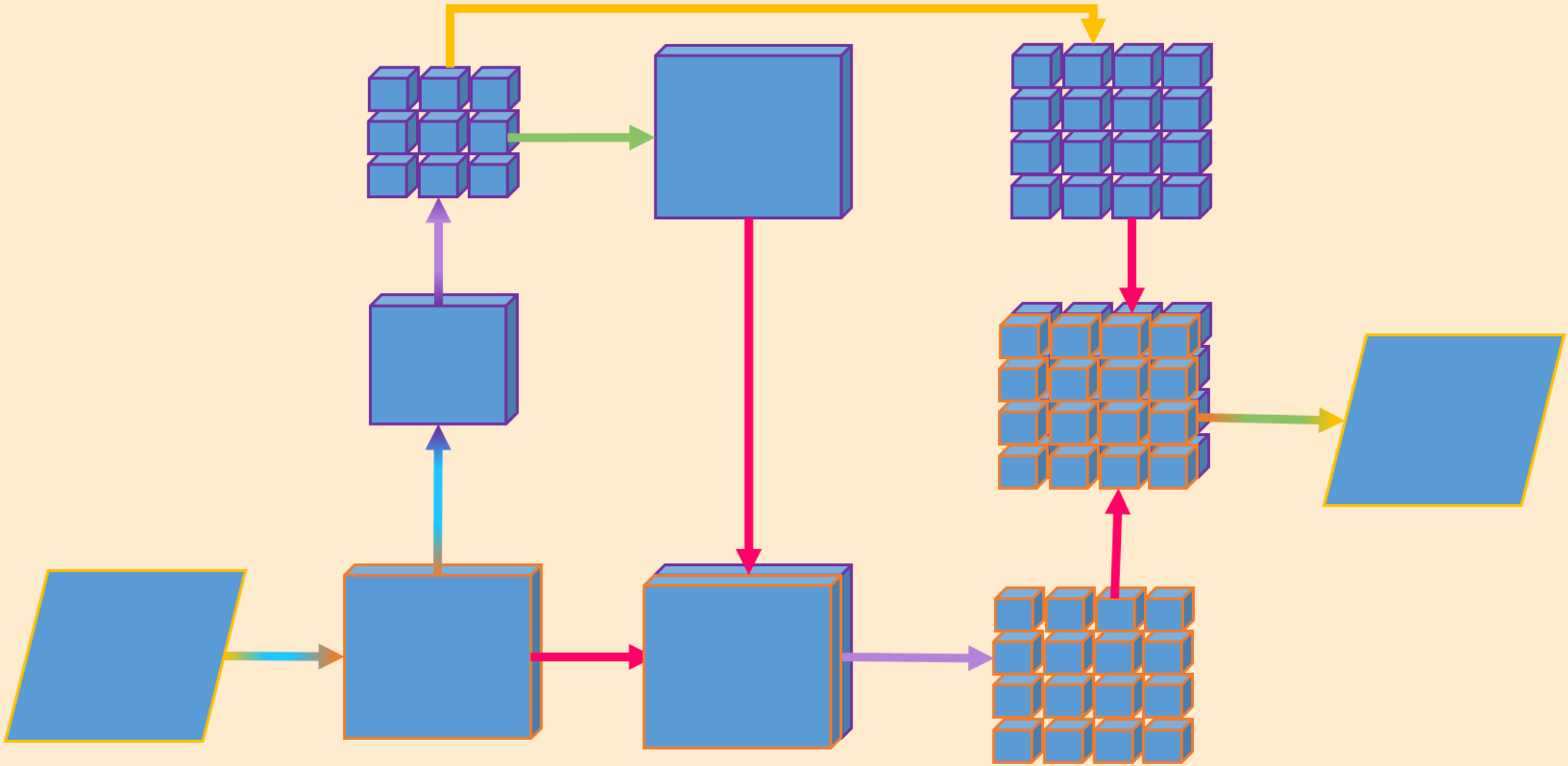
- 挑戰：使用 Phase 的資訊。

Possible Solutions

MVQHA

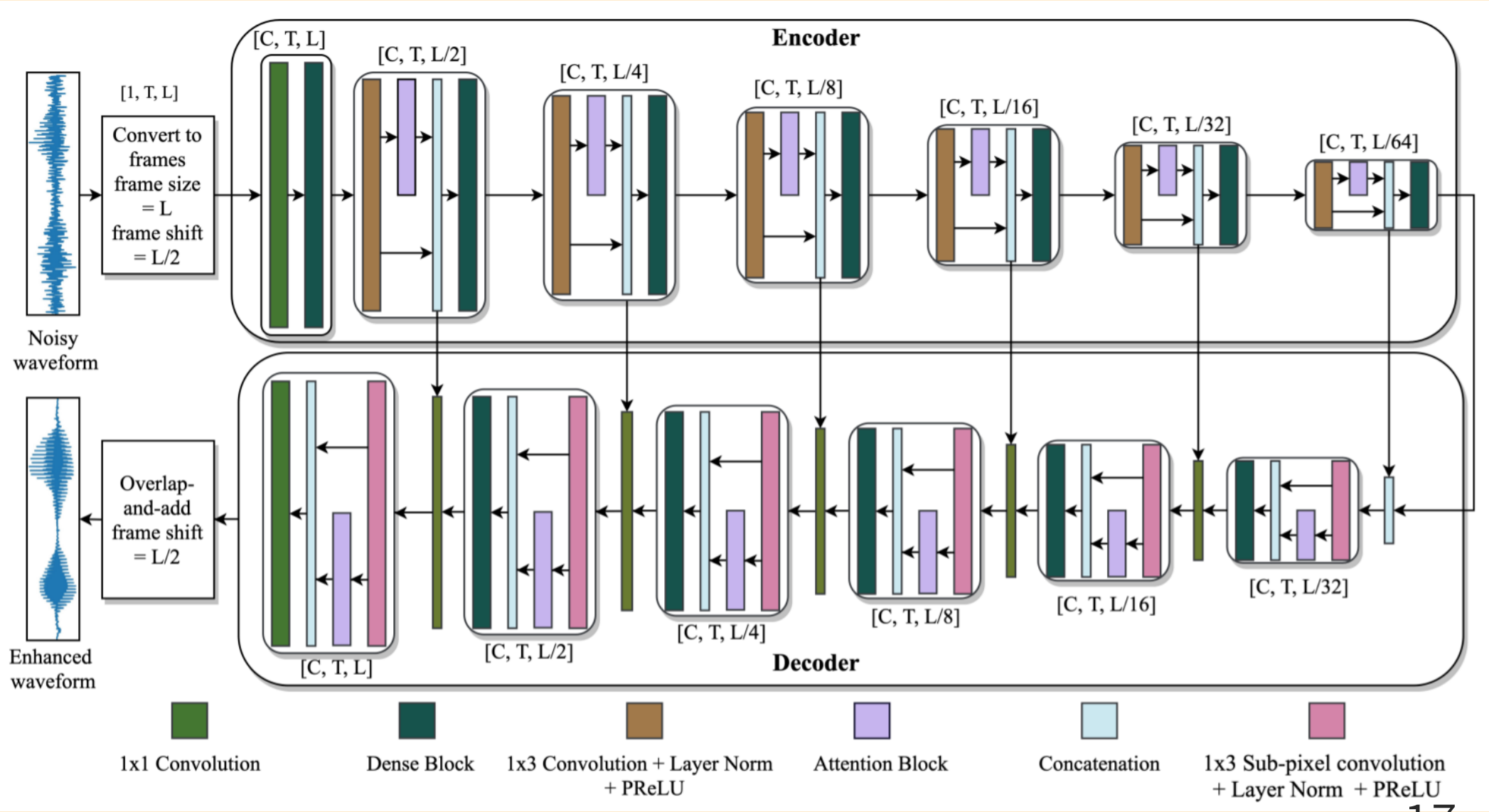


VQ-VAE2



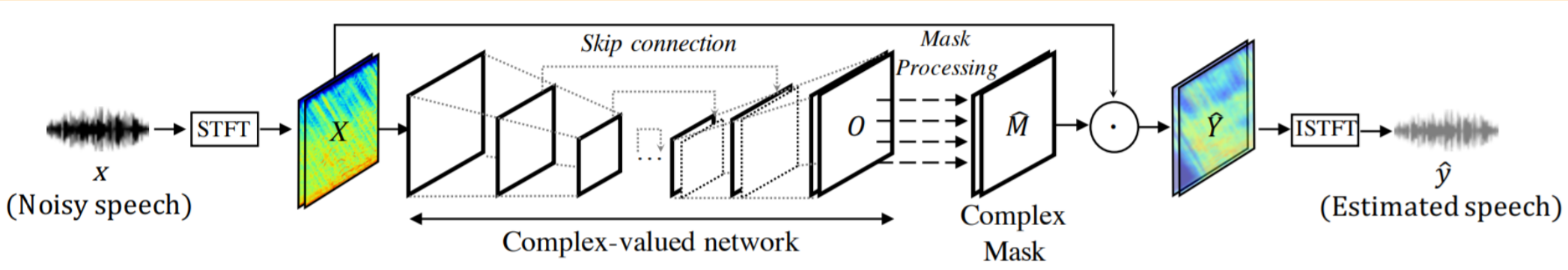
Possible Solutions

Dense CNN

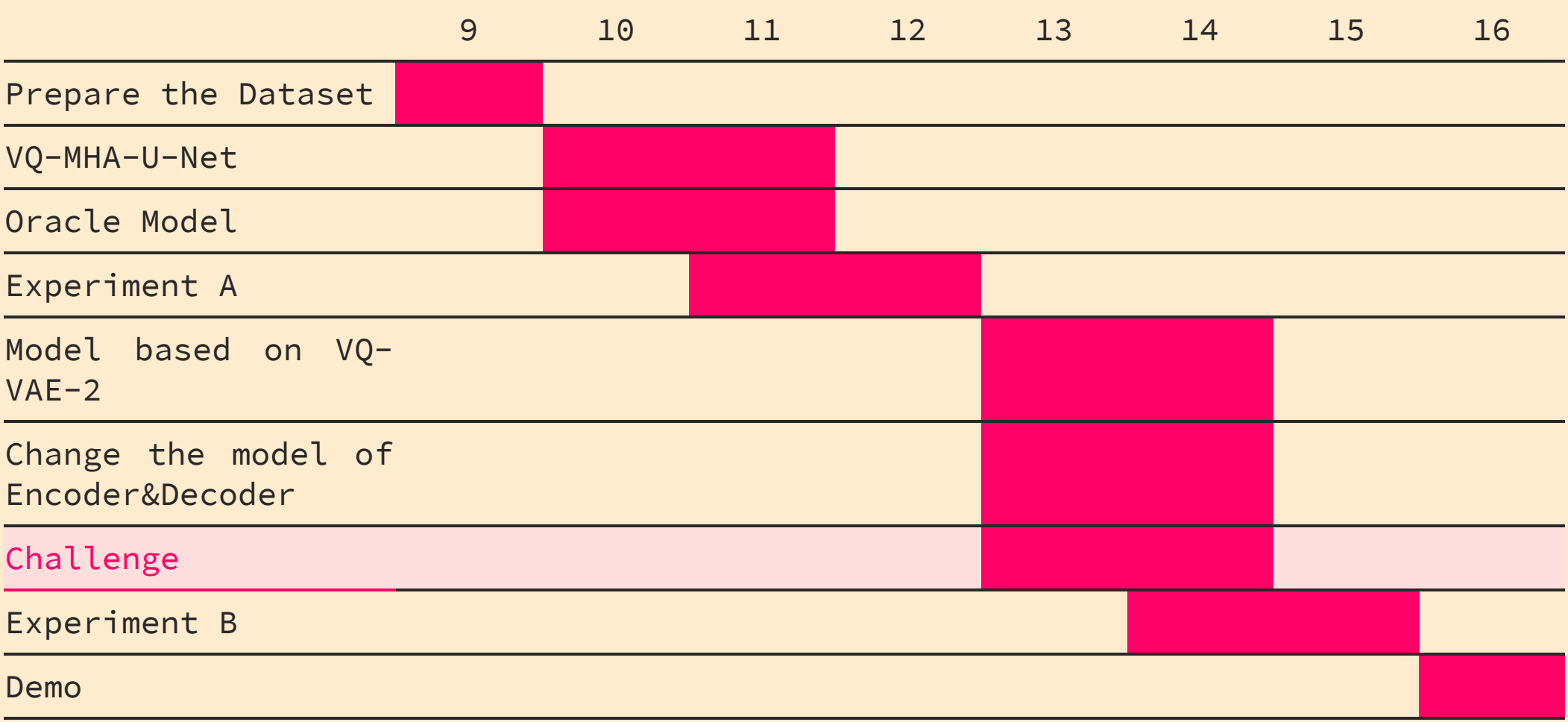


Deep Complex U-Net

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