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

Speech Enhancement(語音增強，SE) 的目的是強化在吵雜環境中的語音訊號，相等於去噪的效果。

如果 SE 做得好，還可以作為 ASR(Automatic Speech Recognition，語音識別) 的前處理系統，以提升 ASR 的效果。

而本篇論文嘗試使用 VQ-VAE 的 Symbolic Book 與 Transformer 的 Multi Head Attention(MHA) 來提取聲學單元，幫助提升 SE 的效果。

Methodology

U-Net

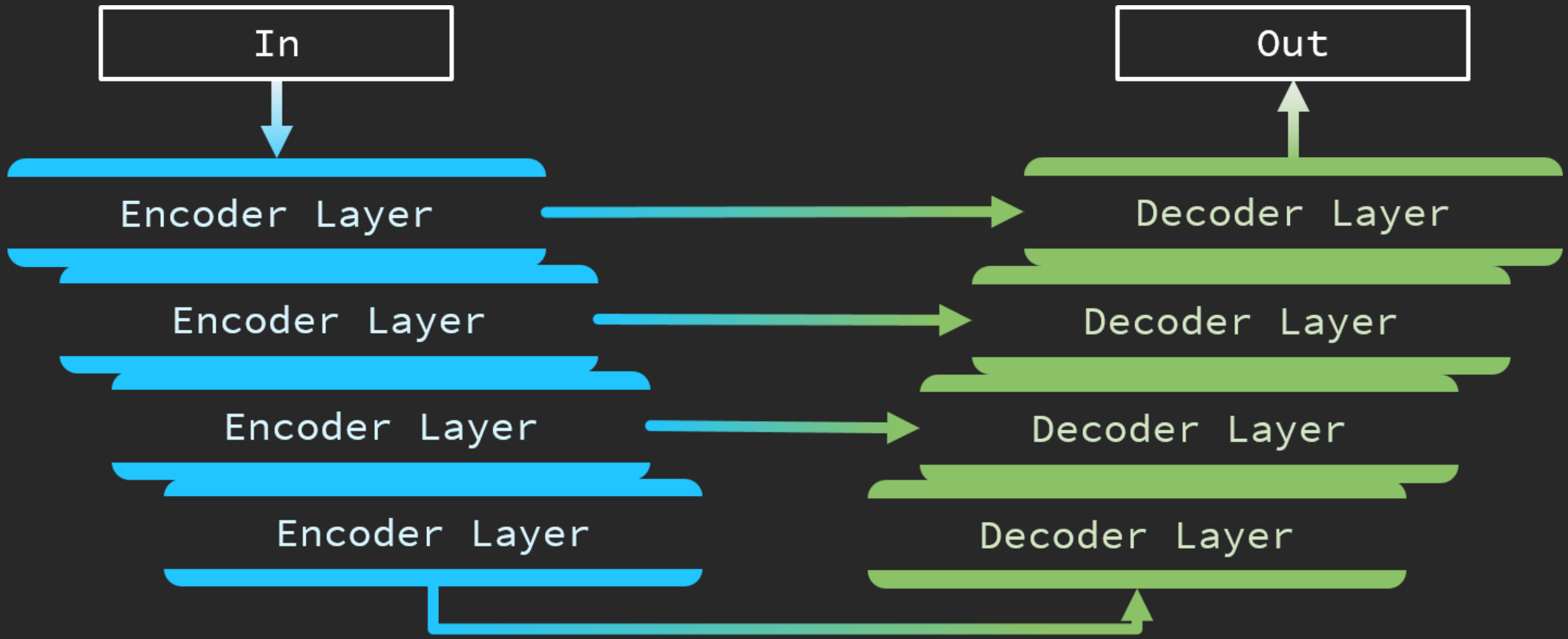
+

VQ-VAE

+

Multi Head Attention

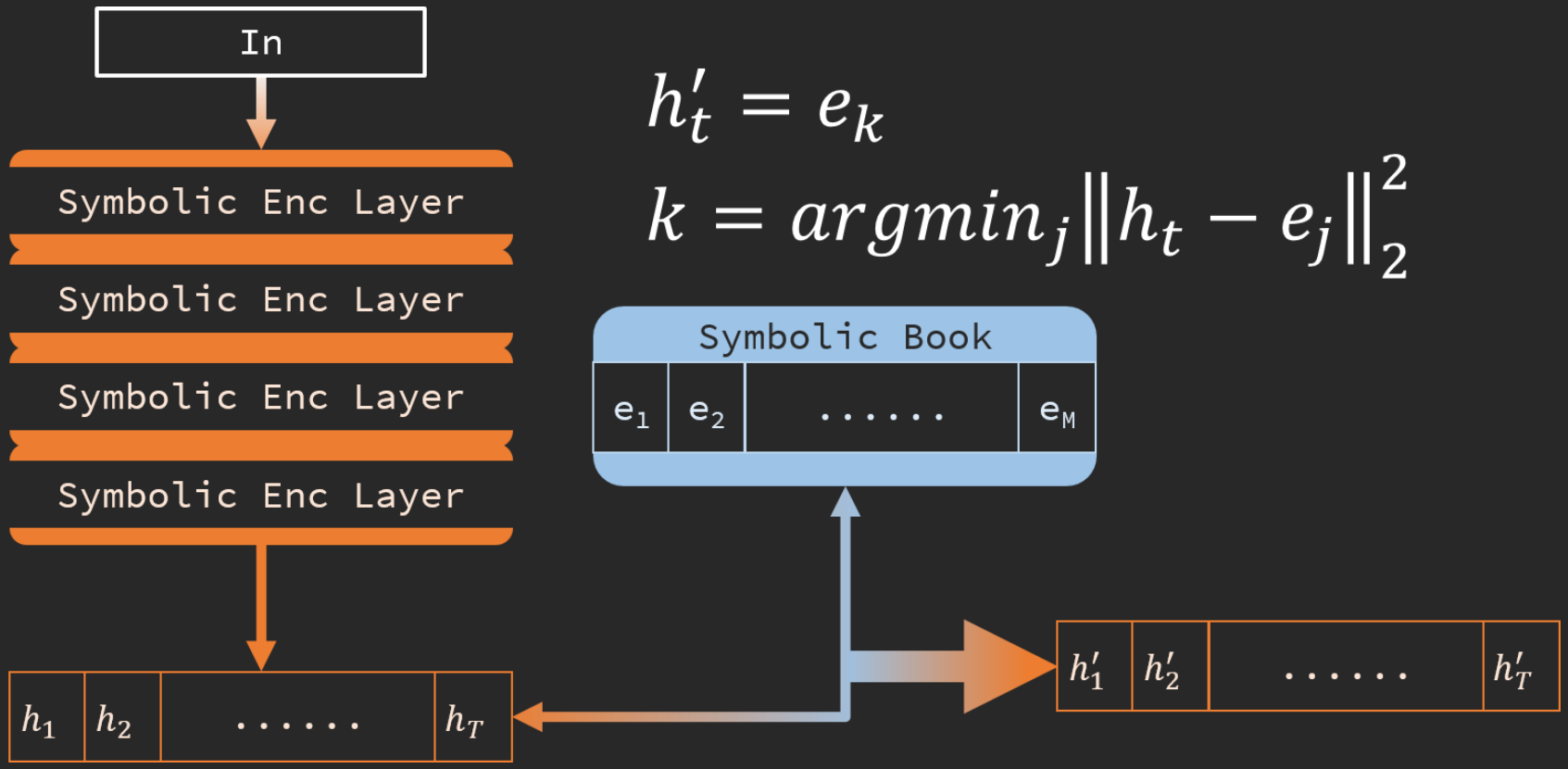
U-Net



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

Symbolic Encoder

源自於 VQ-VAE 中的 Code Book 概念

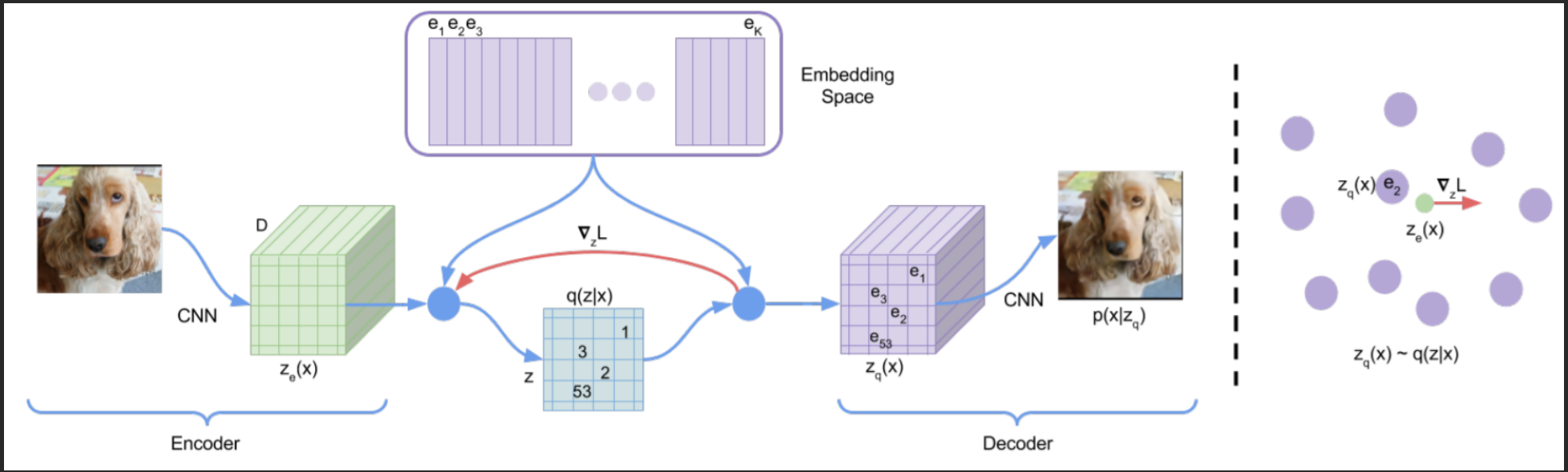


$$h'_t = e_k$$

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

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

VQ-VAE

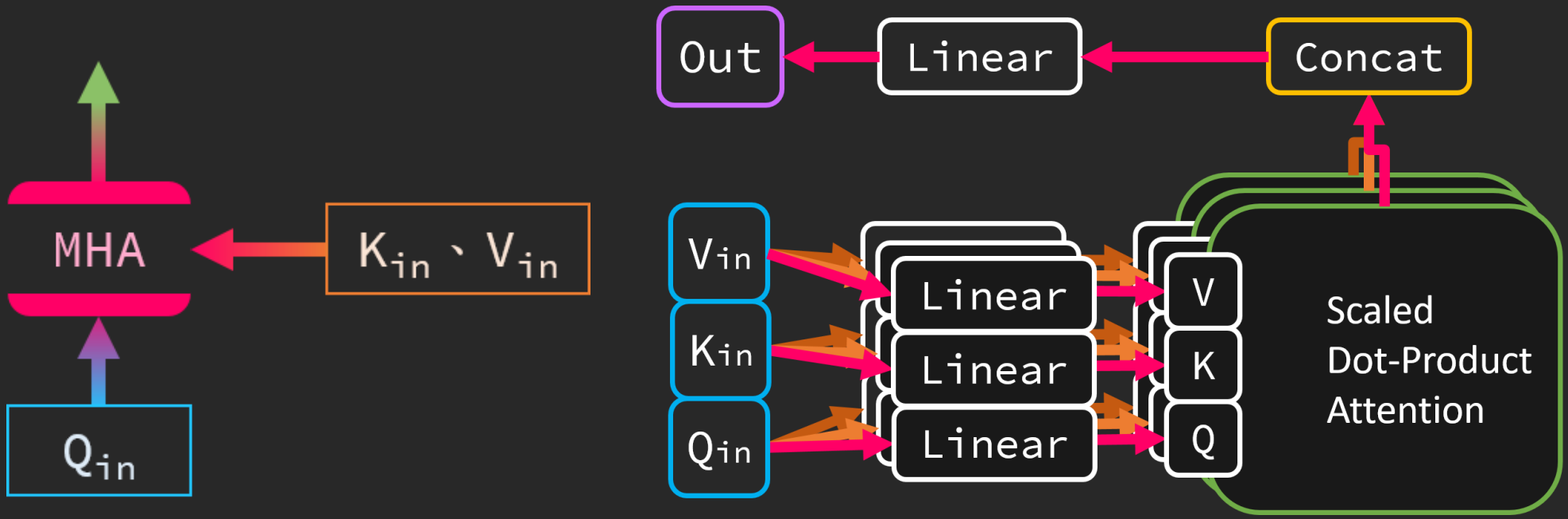


先將 Encoder 輸出的 hidden vector 進行量化後才輸入 Decoder 生成

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



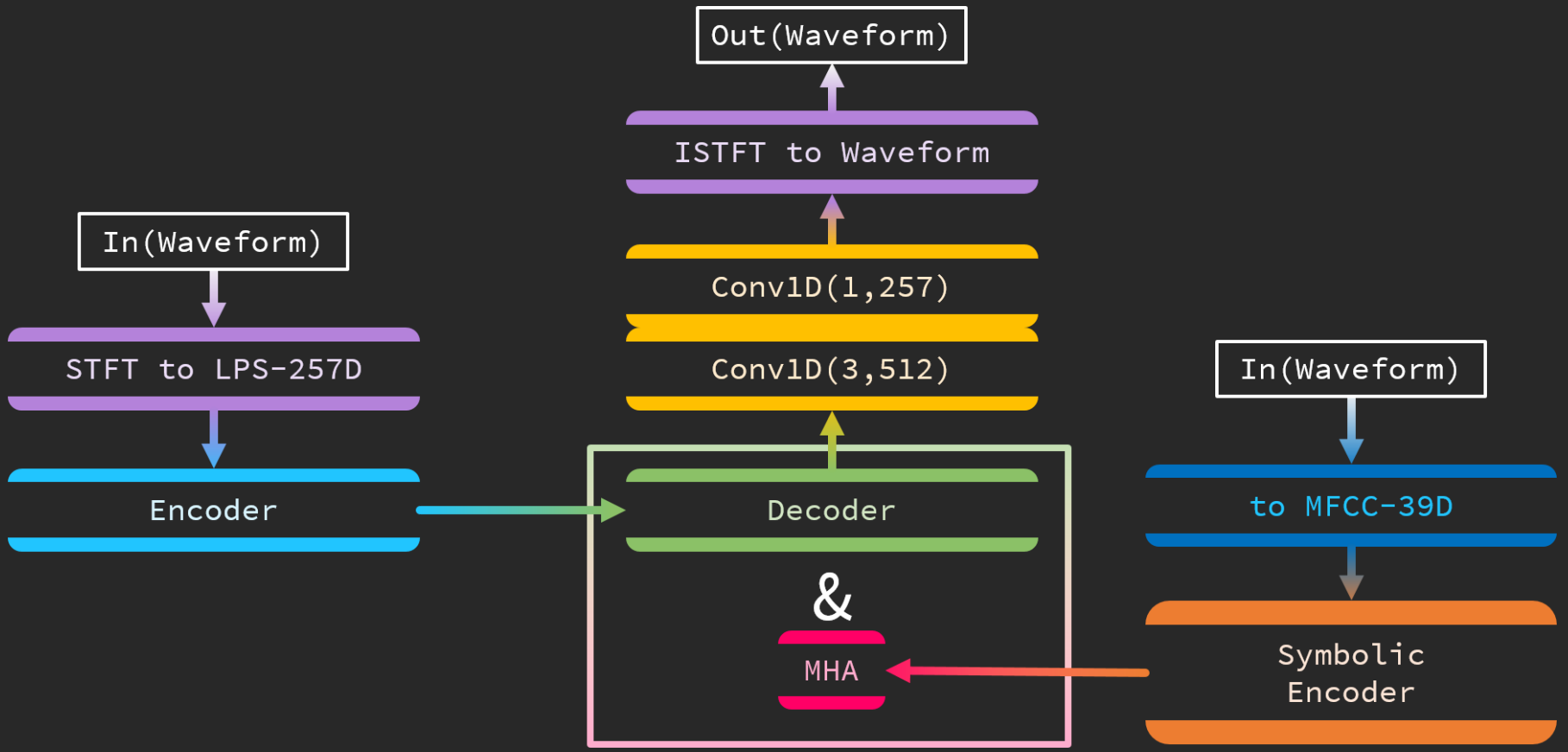
Loss

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

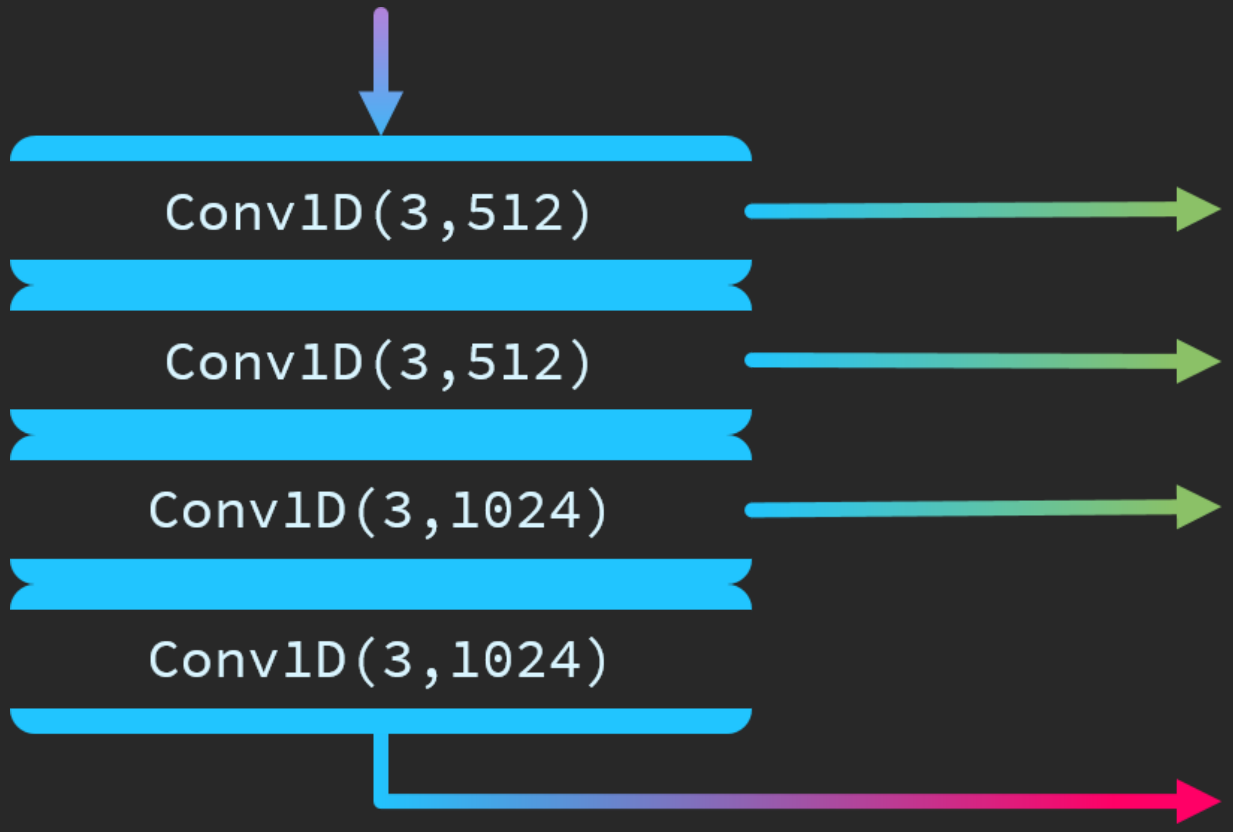
$$\mathcal{L}_{symbolic} = \| h_t - stop\ gradient(e_k) \|_2^2$$

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

Architecture

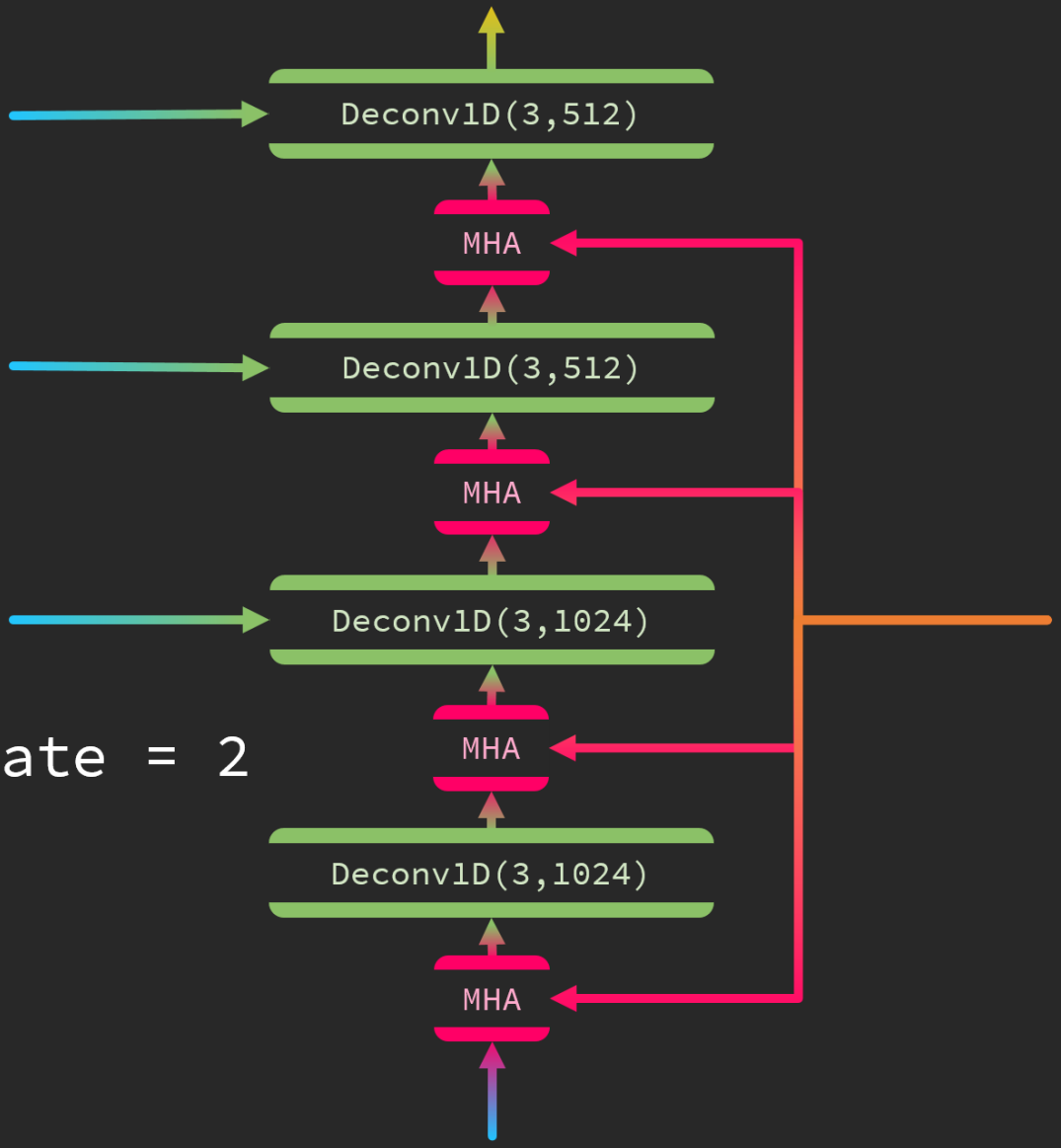


Encoder



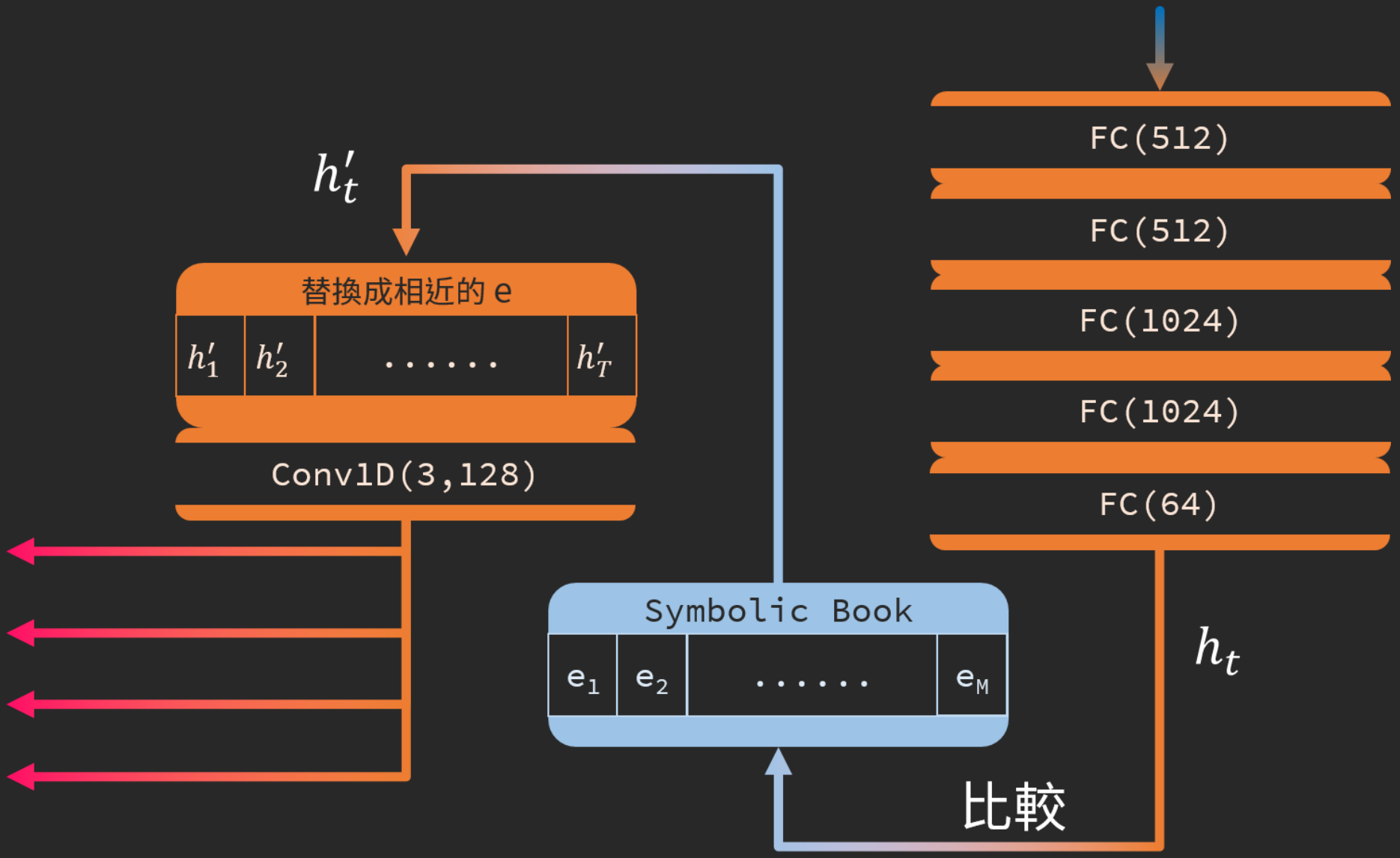
Downsample Rate = 2

Decoder

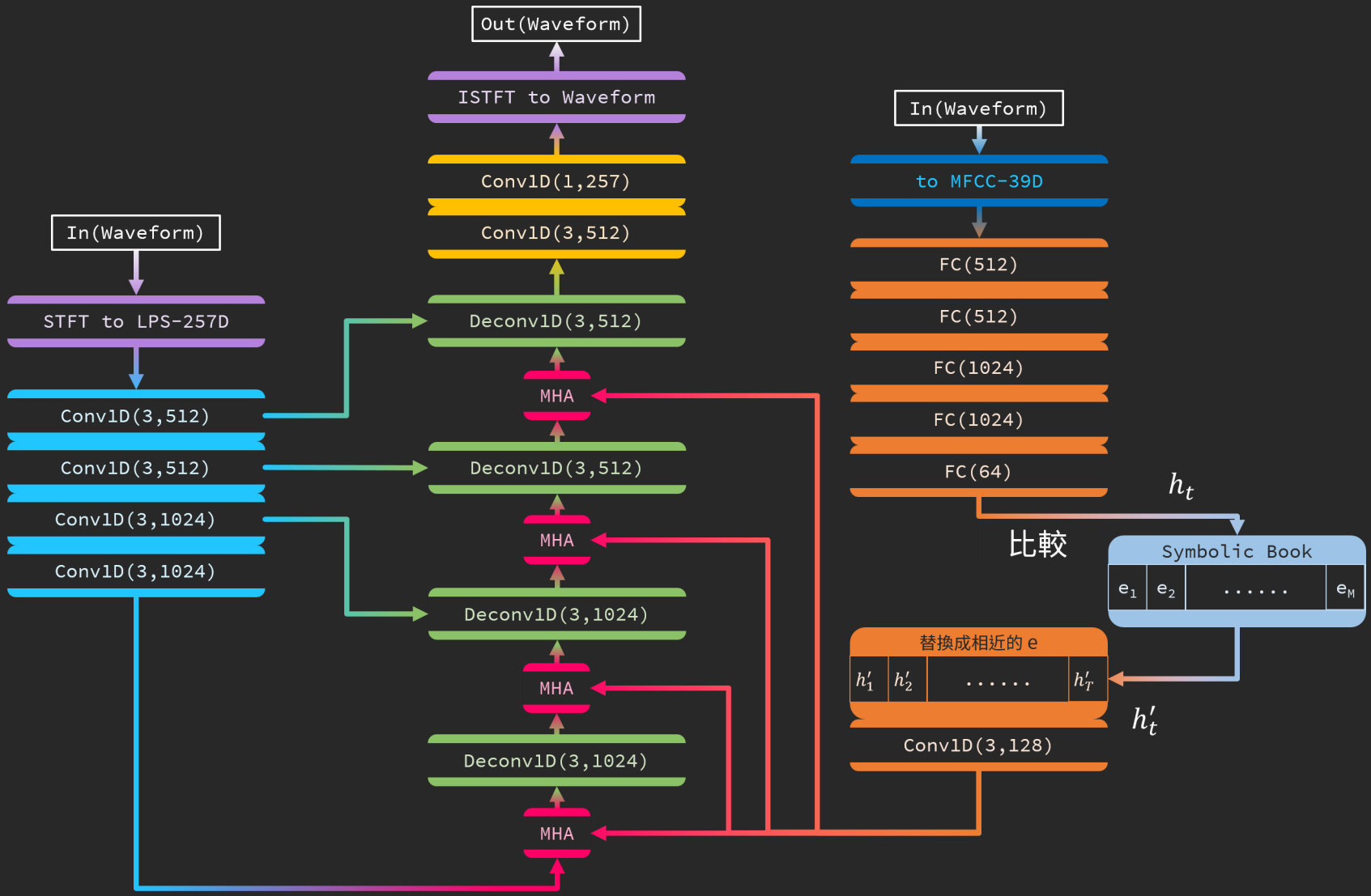


Upsample Rate = 2

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

- 語音：TIMIT
- 噪音：PNL 100 Nonspeech Sounds

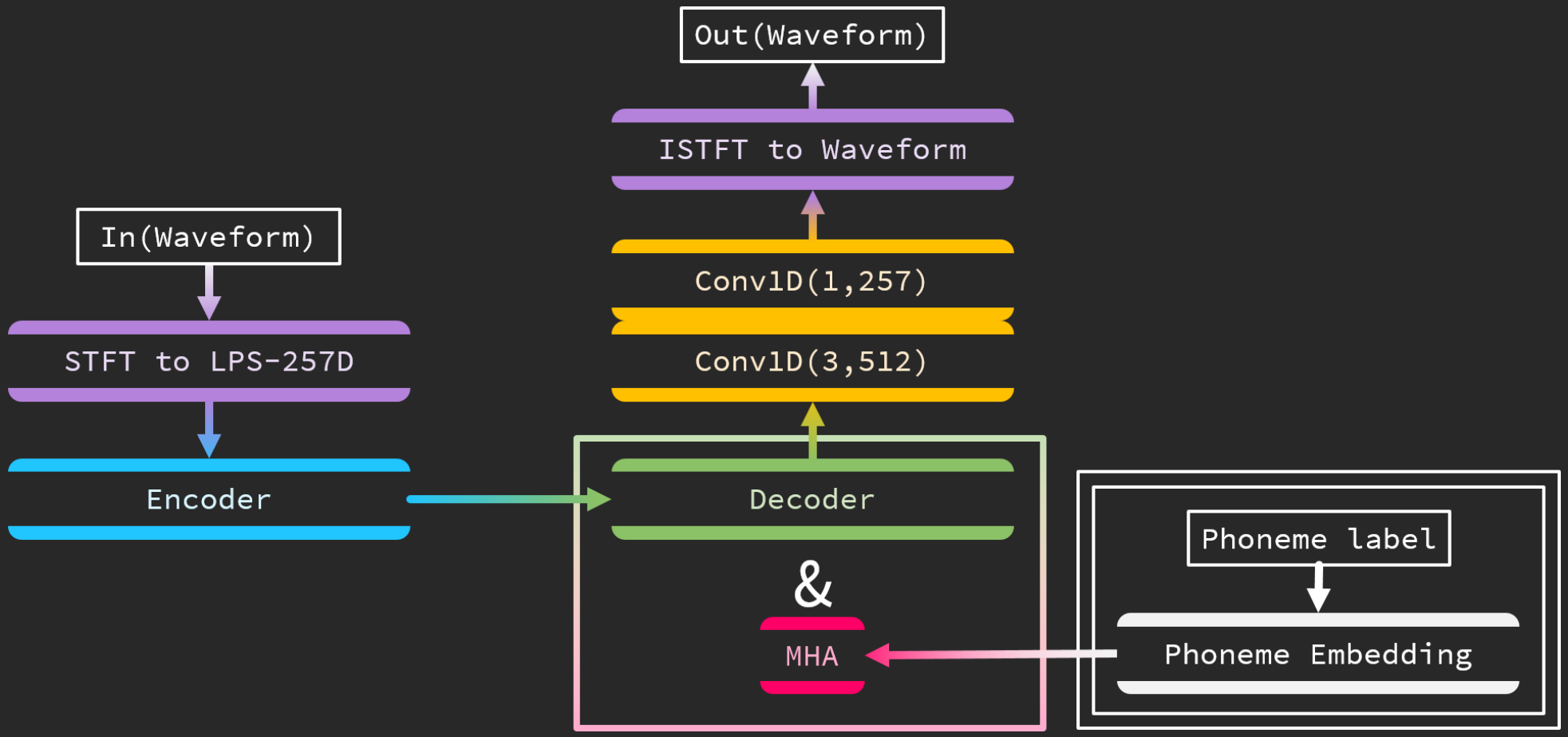
將語音與噪音以不同的訊號雜訊比 (SNR, Signal-to-noise ratio) 混和後作為輸入，並將乾淨的語音作為 Ground Truth

Baseline Model

論文中提出的方法與另外三組模型進行比較
分別是 U-Net、U-Net-MOL 與 Oracle：

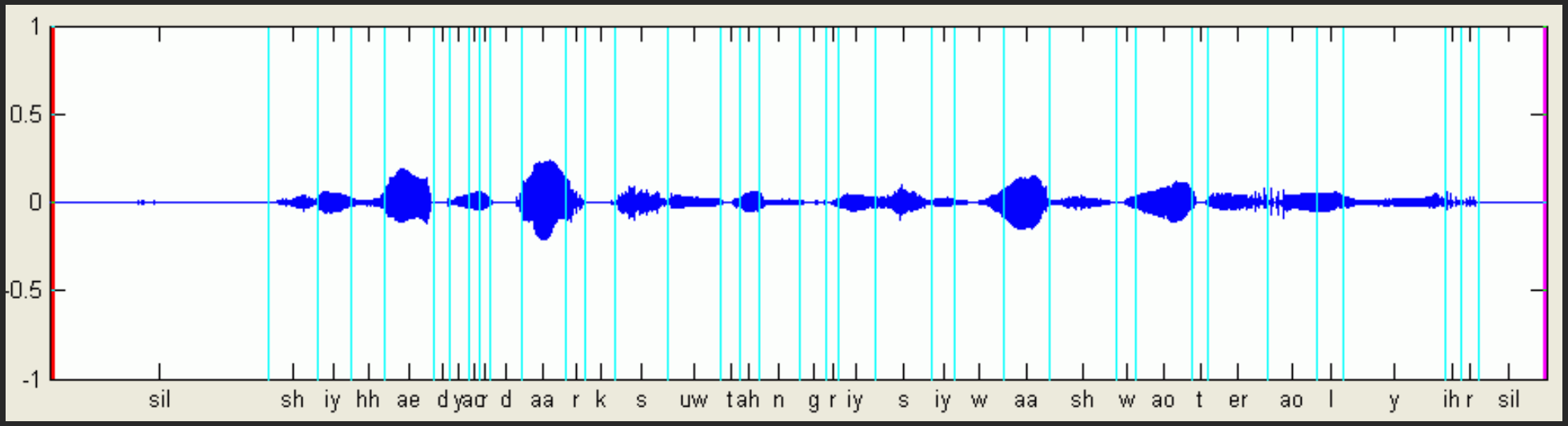
1. U-Net
   單純的 U-Net
2. U-Net-MOL
   使用了 multi objective learning 的 U-Net
3. Oracle
   將原本的 Symbolic Encoder 換成輸入聲音的 phoneme(音素) embedding

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

TIMIT 資料集內，除了語音以外也將每個時間點的音素都標記好了
而 Oracle 就是將音素進行 Embedding 後作為 K_{in}、V_{in} 使用



音素範例（來源）

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. 使用了 Symbolic Encoder 與 MHA 的模型，效果比 Oracle 以外的 Baseline Model 更好

2. Oracle 證明了在 Speech Enhancement 的問題中加入音素資訊能夠提升模型的表現

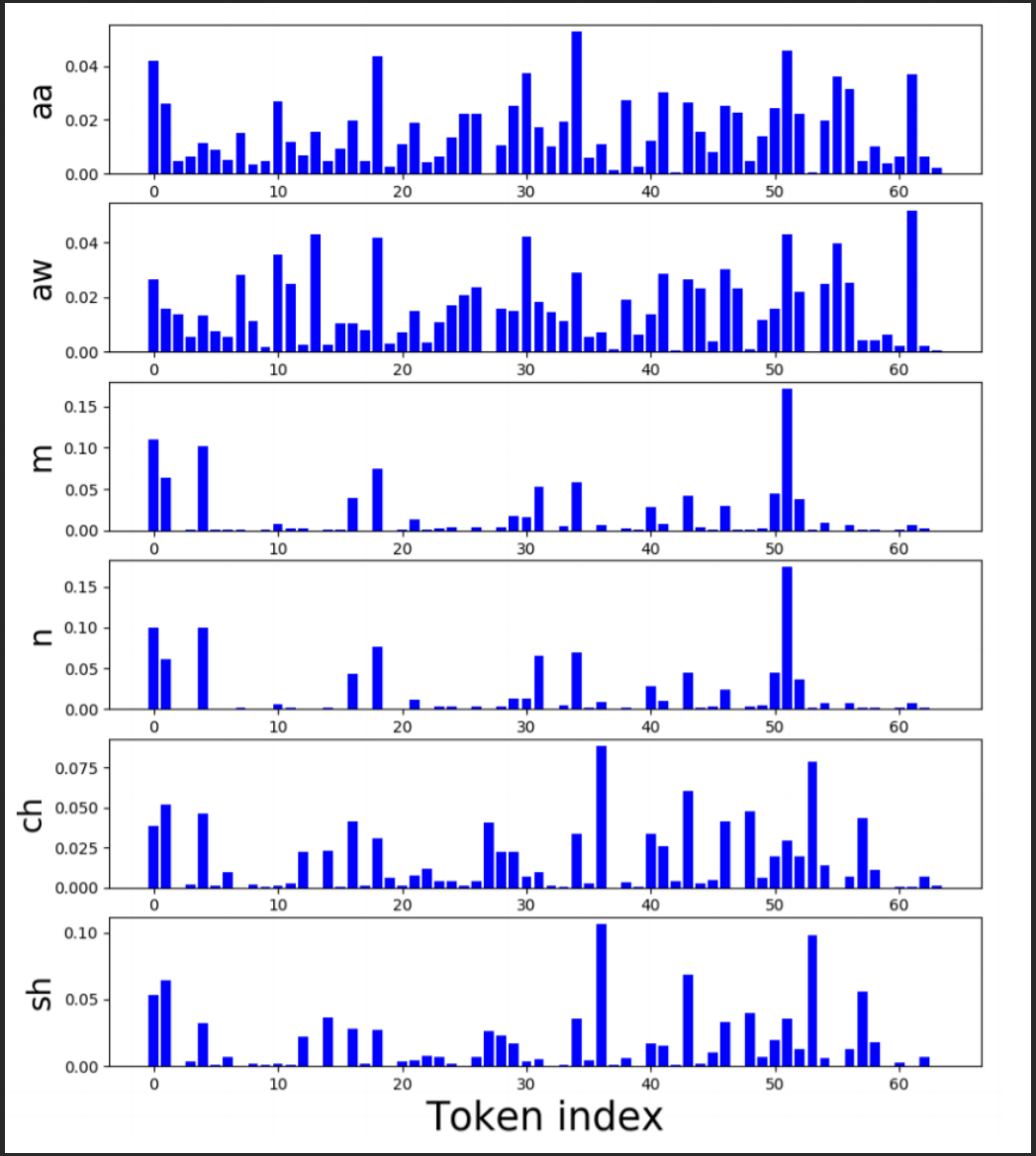
Experiments

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

當 Book Size 過大時會出現「index collapse」

導致有部分的 symbolic vector 不會被使用到

Experiments

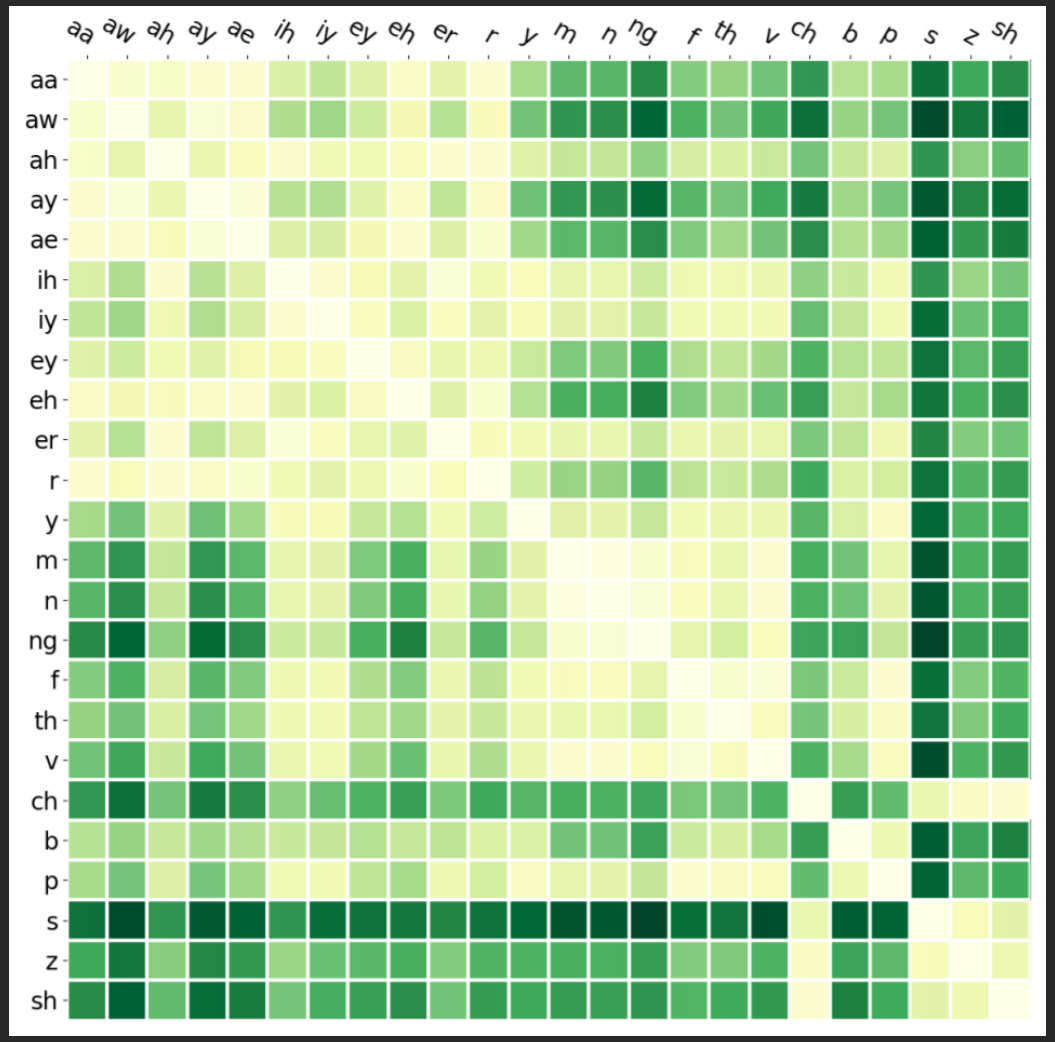


橫軸是 Symbolic index

而上面每個值代表某個發音作為輸入時，會被 mapping 到某個 Symbolic 的次數

當發音越相似，其分佈就越接近

Experiments



將各種發音的分佈進行相似度計算

有些不同的發音會得出相似的結果
是因為輸入含有噪音的緣故

Conclusion

利用 Symbolic Encoder 先將語音編碼成聲音單元，在通過 Multi Head Attention 與提取特徵，能取得比原先更好的結果

讀者想法：

- Symbolic Encoder + Multi Head Attention 的結構就是在為語音建立基礎結構(說話內容)
- 而利用 U-Net 的 shortcut 便能幫助基礎結構補上語調的資訊(聲調起伏、音色等等)
- 在 VQ-VAE 都出第 2 代的現在，可以試試用多階層(多解析度)的方式，來為模型保留住更多的聲音特徵