2024 AMD人工智慧終端節點運算創意競賽

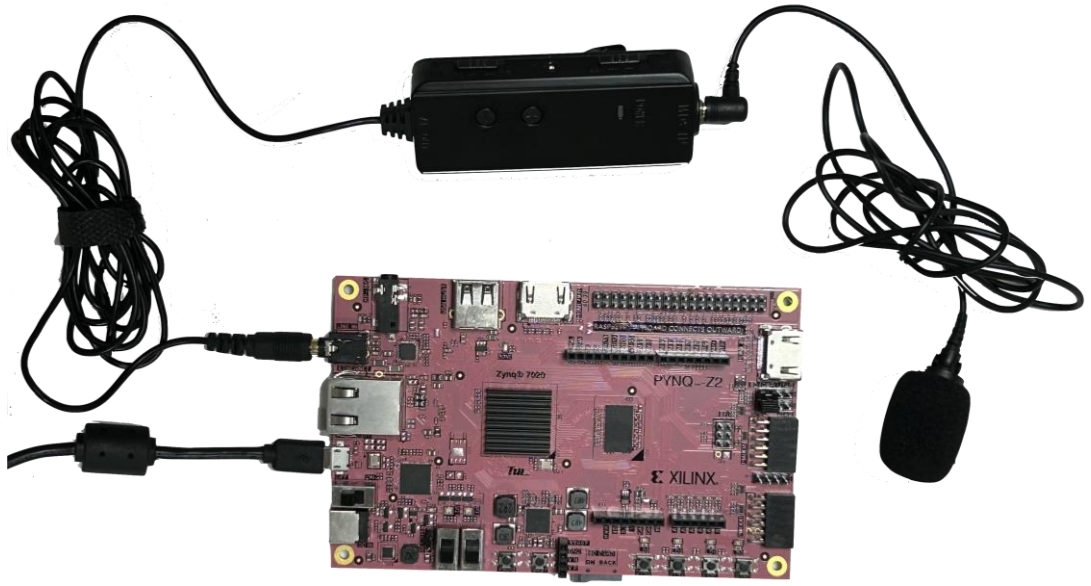
基於 ShuffleNet 之道路事件聲音識別輕量化邊緣智慧系統

中央大學電機系 我會出手
指導老師：陳聿廣 教授
學生： 吳育丞、戴仕庭、鄭凱方、呂翊銓

Outline

於FPGA上實作聲音辨識偵測系統

Board: PYNQ-Z2



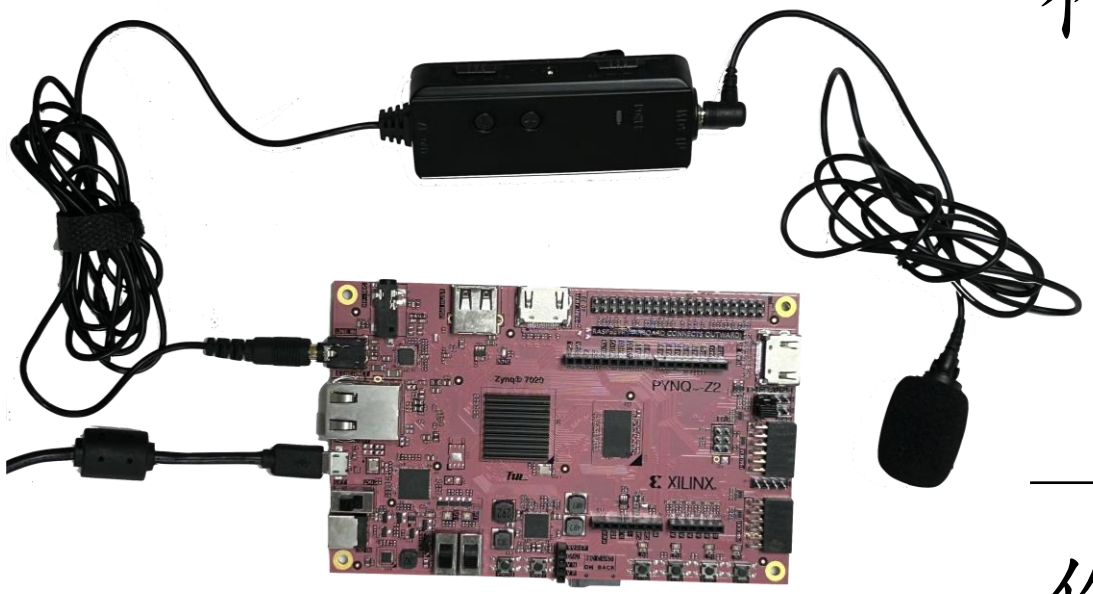
聲音辨識

車禍偵測

Mic: AT9901

Outline

PYNQ-Z2搭載AT9901



前言

開發背景、應用場域、開發目標

開發背景

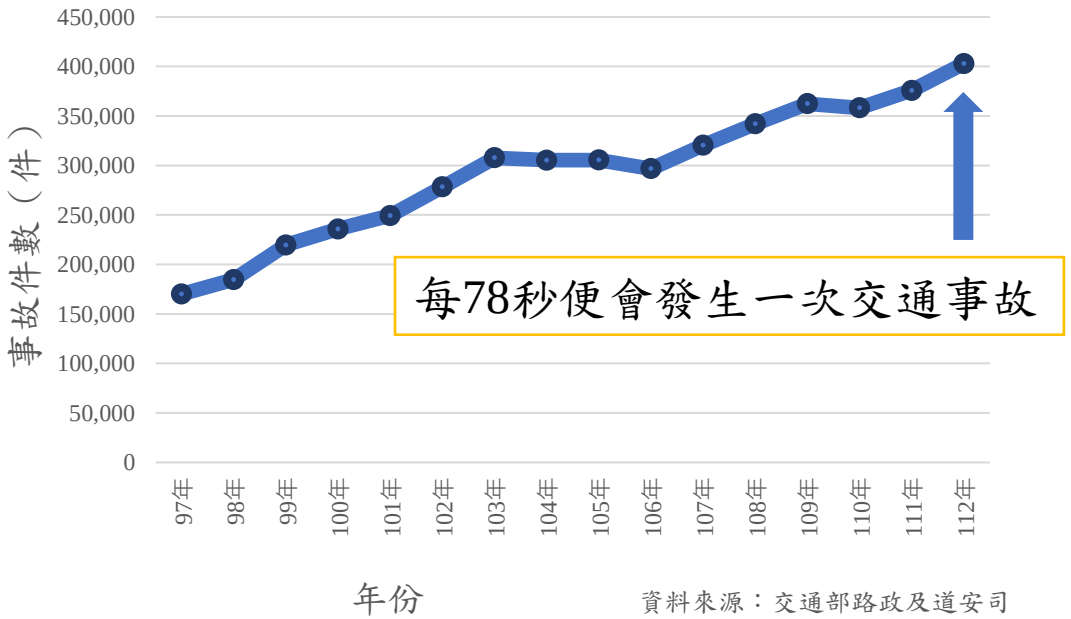

台灣歷年交通事故統計
450,000
400,000
350,000
300,000
250,000
200,000
150,000
100,000
50,000
0
事故件數（件）
97年
98年
99年
100年
101年
102年
103年
104年
105年
106年
107年
108年
109年
110年
111年
112年
年份
170,127件
402,926件
2.35倍!
資料來源：交通部路政及道安司

開發背景

通報流程
繁瑣

自動通報

台灣歷年交通事故統計



資料來源：交通部路政及道安司

交通事故通報流程

目擊事件

人為通報

需人為介入

等[illegible]抵達

事[illegible]

延誤送醫！

開發背景

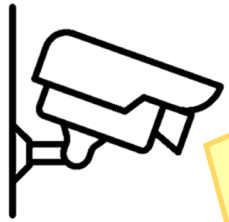
通報流程
繁瑣

自動通報

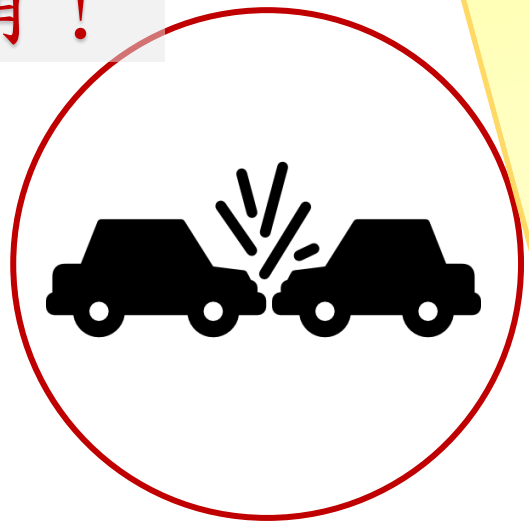
監視器
死角問題

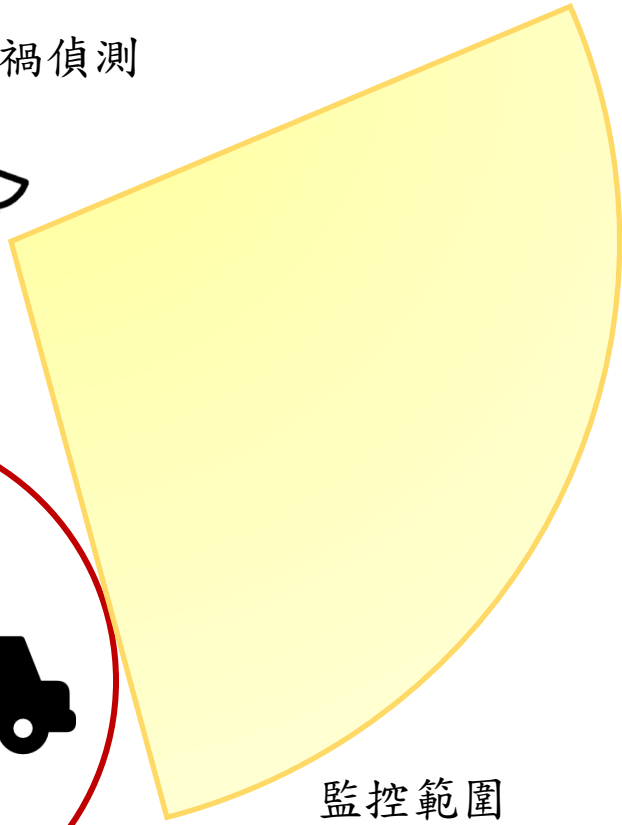
音訊偵測

視覺辨識車禍偵測



視覺死角！





監控範圍

開發背景

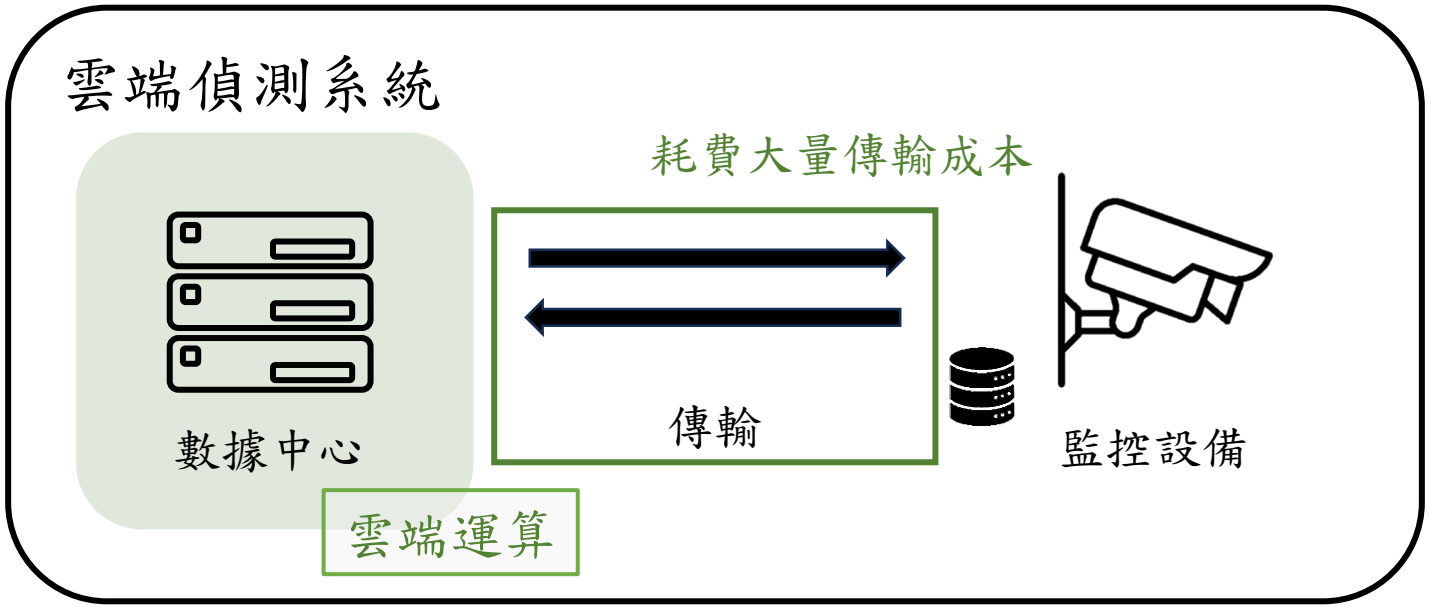
通報流程繁瑣

自動通報

監視器死角問題

音訊偵測

邊緣運算重要性

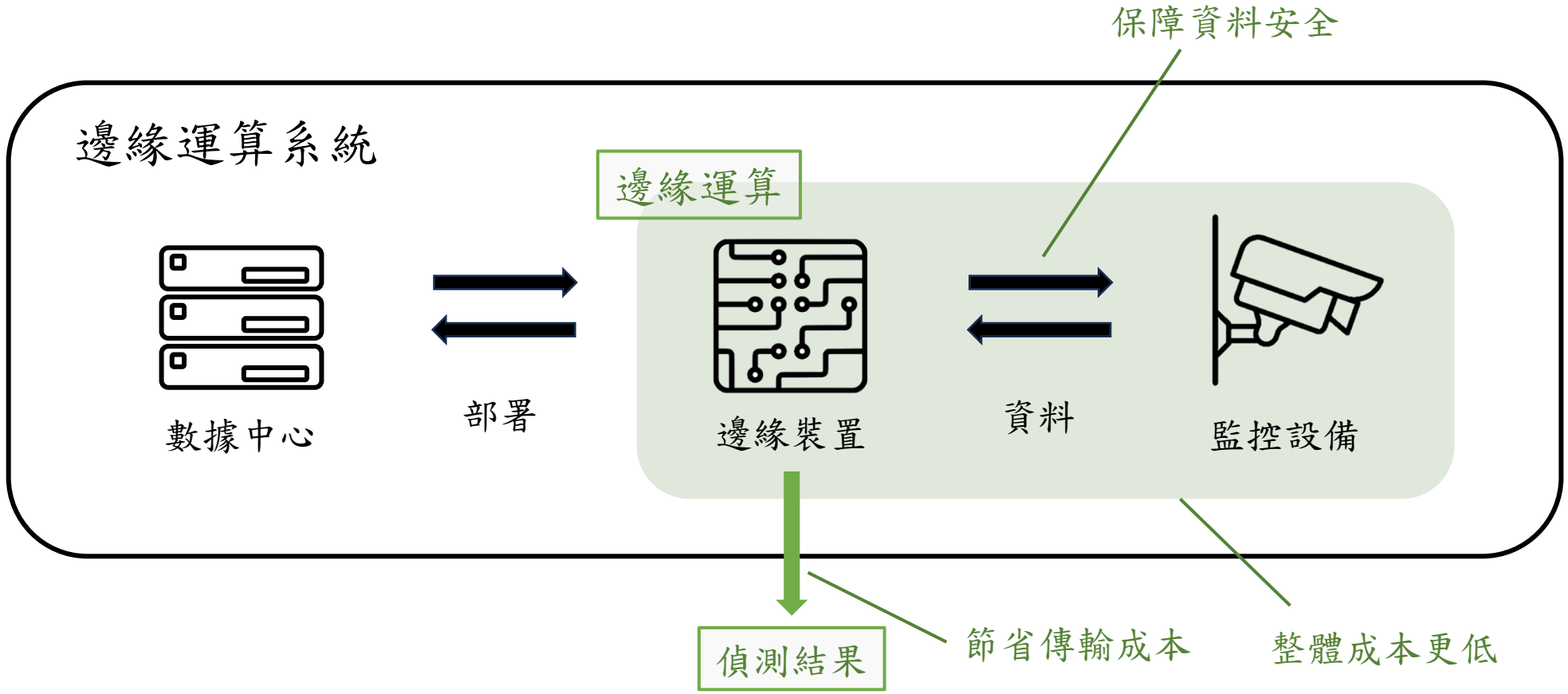


開發背景

通報流程繁瑣 → 自動通報

監視器死角問題 → 音訊偵測

邊緣運算重要性 → 邊緣運算



開發背景

基於 ShuffleNet 之
道路事件聲音識別輕量化
邊緣智慧系統

音訊偵測

自動通報

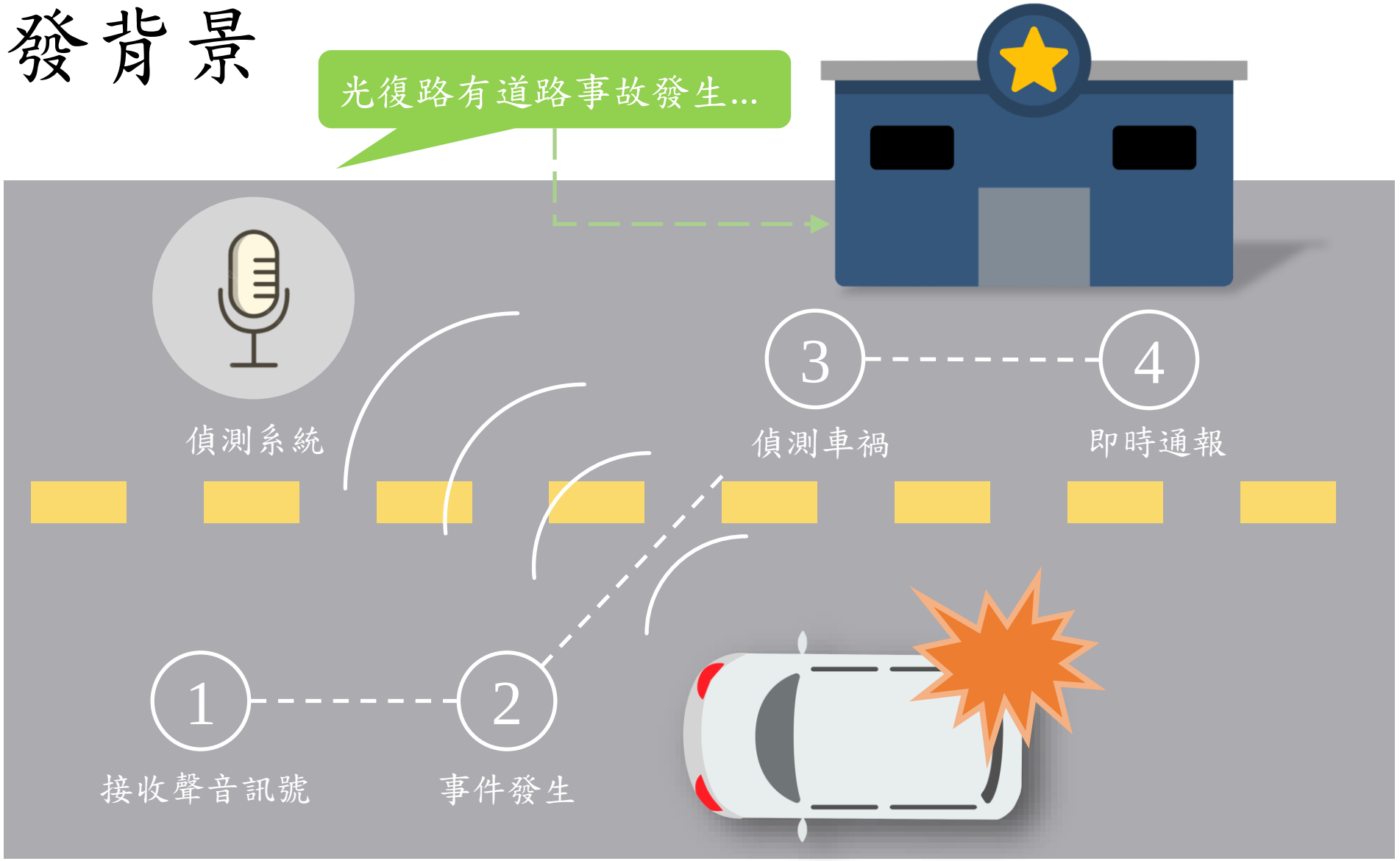
邊緣運算

通報流程
繁瑣

監視器
死角問題

邊緣運算
重要性

開發背景

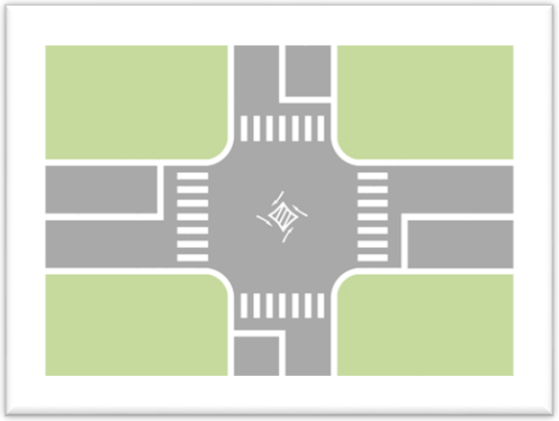


▲ 系統運作流程

應用場域

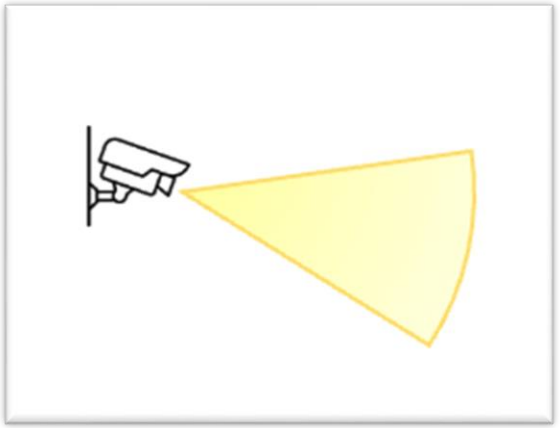
偏遠
人煙稀少

- 降低人為依賴



易發生
道路事件

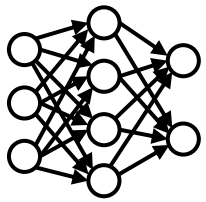
- 加速通報流程



已裝設
傳統監視器

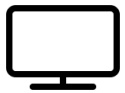
- 消除視線死角

開發目標



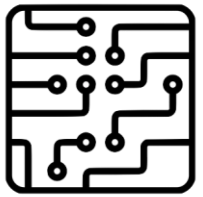
準確度高

類比聲音訊號



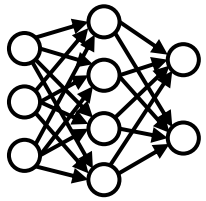
辨識結果

即時辨識



硬體輕量化

開發目標

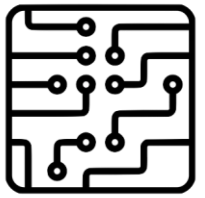


準確度高

類比聲音訊號

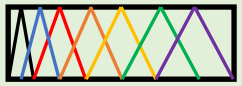
辨識結果

即時辨識

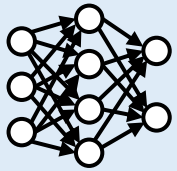


硬體輕量化

類比聲音訊號



MFSC
音訊預處理



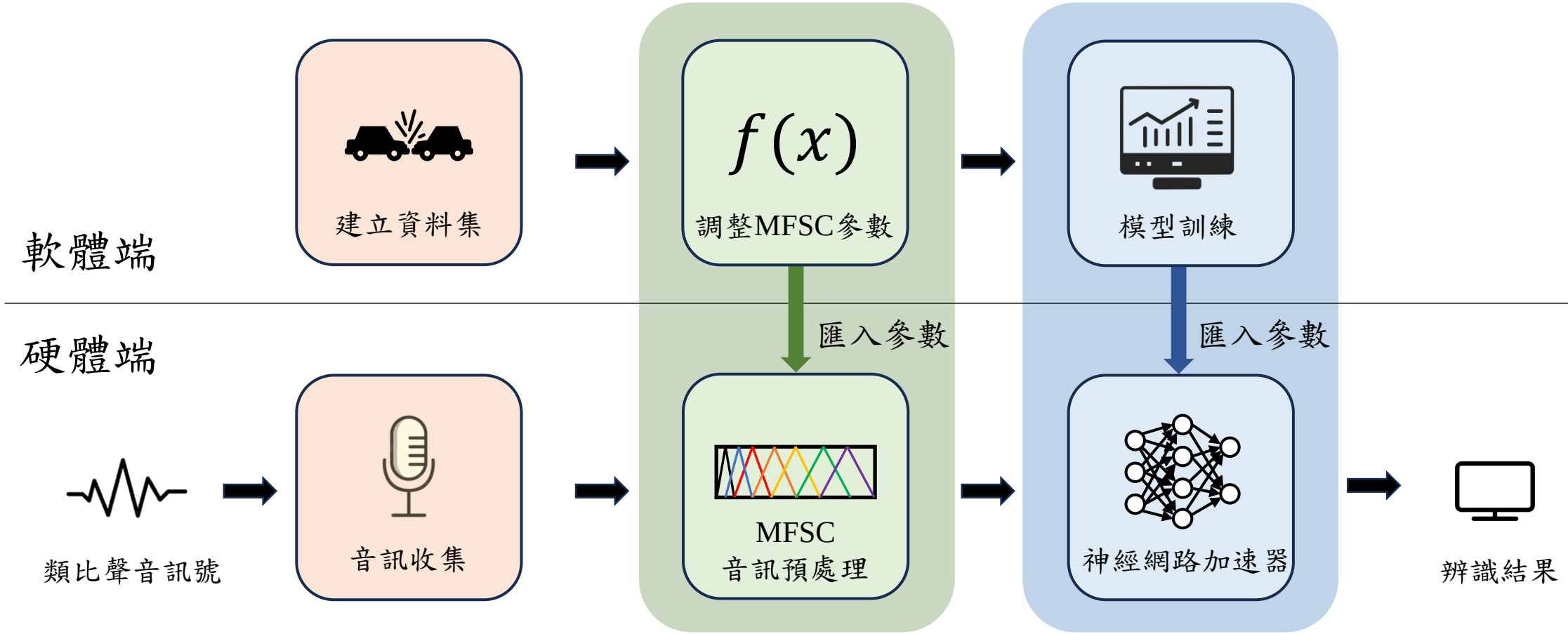
ShuffleNet

辨識結果

音訊預處理流程

輕量化模型

開發流程


軟體端
硬體端
建立資料集
$f(x)$
調整MFSC參數
模型訓練
匯入參數
匯入參數
類比聲音訊號
音訊收集
MFSC
音訊預處理
神經網路加速器
辨識結果

相關技術文獻探討

MFSC音訊預處理技術、ShuffleNet介紹

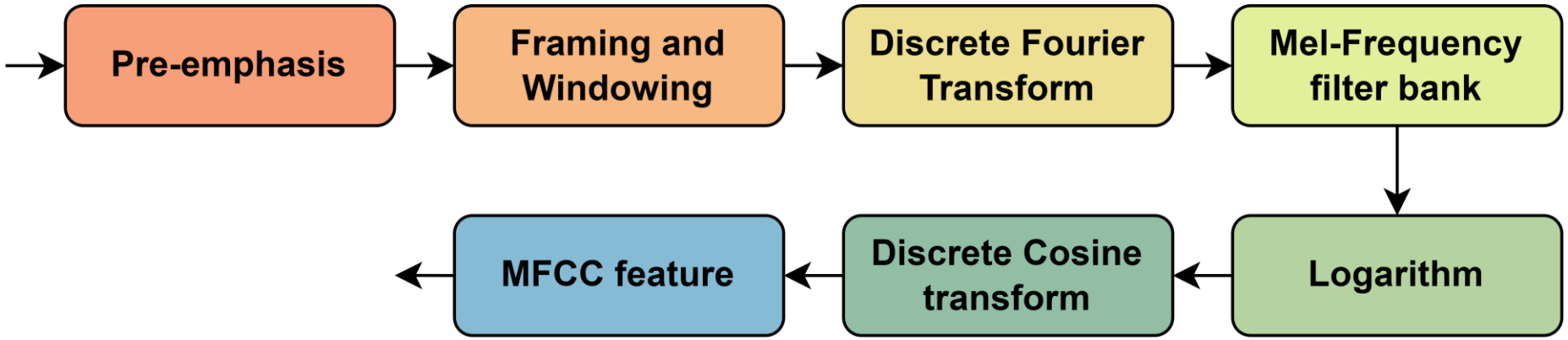
文獻探討 — MFCC 梅爾頻率倒譜係數

- 頻域分析比時域分析更能有效過濾雜音、保留關鍵聲音特徵

  Nossier, S. A., Wall, J., Moniri, M., Glackin, C., & Cannings, N. (2020, July). A comparative study of time and frequency domain approaches to deep learning based speech enhancement. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

- MFCC (Mel-Frequency Cepstral Coefficients) 能夠快速提取高質量的低維特徵，在實時檢測系統中尤為重要

  Wang, Feifan, and Xizhong Shen. "Research on Speech Emotion Recognition Based on Teager Energy Operator Coefficients and Inverted MFCC Feature Fusion." Electronics 12.17 (2023): 3599.



文獻探討 — MFSC 梅爾頻率係數

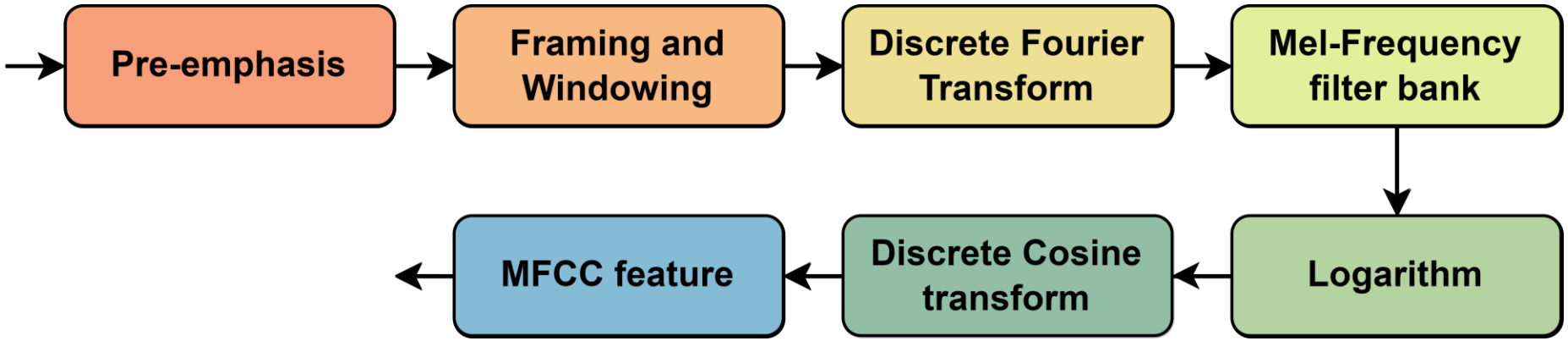
- MFSC (Mel-Frequency Spectral Coefficients) 使用更少的運算步驟

(省略DCT)、可節省硬體資源

- MFSC更能保留資料的局部特性，避免持續累積資料儲存量

Abdel-Hamid, Ossama, et al. "Convolutional neural networks for speech recognition." IEEE/ACM Transactions on audio, speech, and language processing 22.10 (2014): 1533-1545.

- 本實驗設計採MFSC音訊預處理



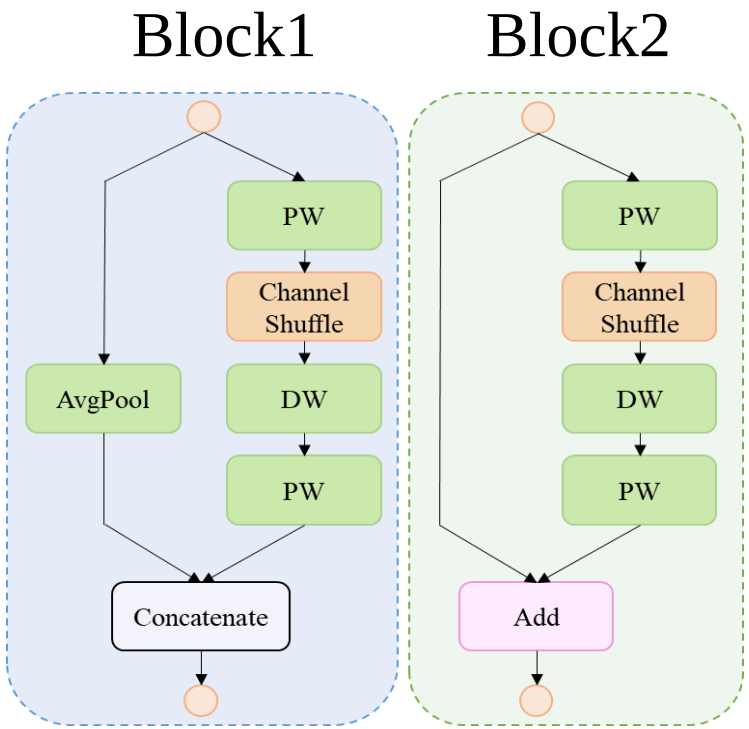
節省硬體運算、儲存資源

文獻探討 — ShuffleNet

- 神經網路架構

O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional Neural Networks for Speech Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 10, pp. 1533-1545, Oct. 2014.

Layer		Stride	Output Size	Output Channel
Input		-	64×64	1
3×3 Global Conv		2	32×32	16
MaxPool		2	16×16	16
Stage1	Block1	2	8×8	32
	Block2	1	8×8	32
Stage2	Block1	2	4×4	64
	Block2	1	4×4	64
	Block2	1	4×4	64
	Block2	1	4×4	64
Global AvgPool		-	1×1	64
Fully Connected		-	2	1



Pointwise Conv.

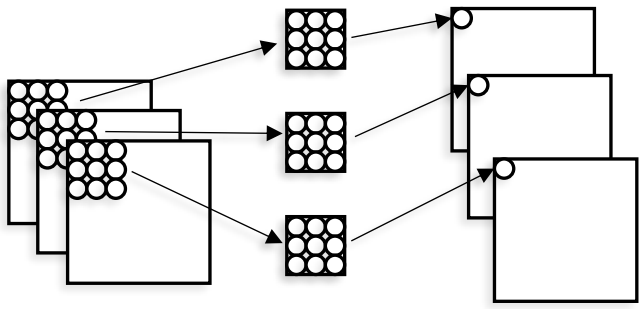
Channel Shuffle

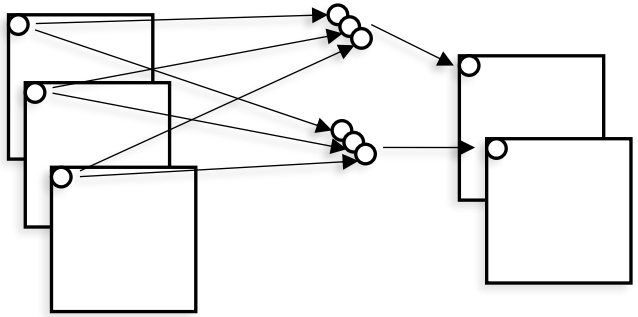
Depthwise Conv.

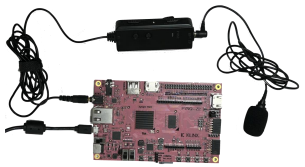
Pointwise Conv.

文獻探討 — ShuffleNet

- 核心技術
  - Depthwise & Pointwise Convolution

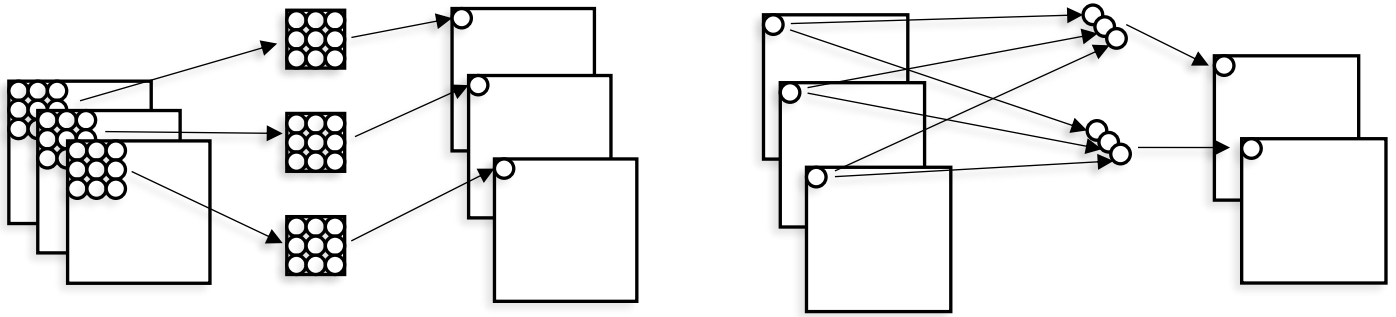




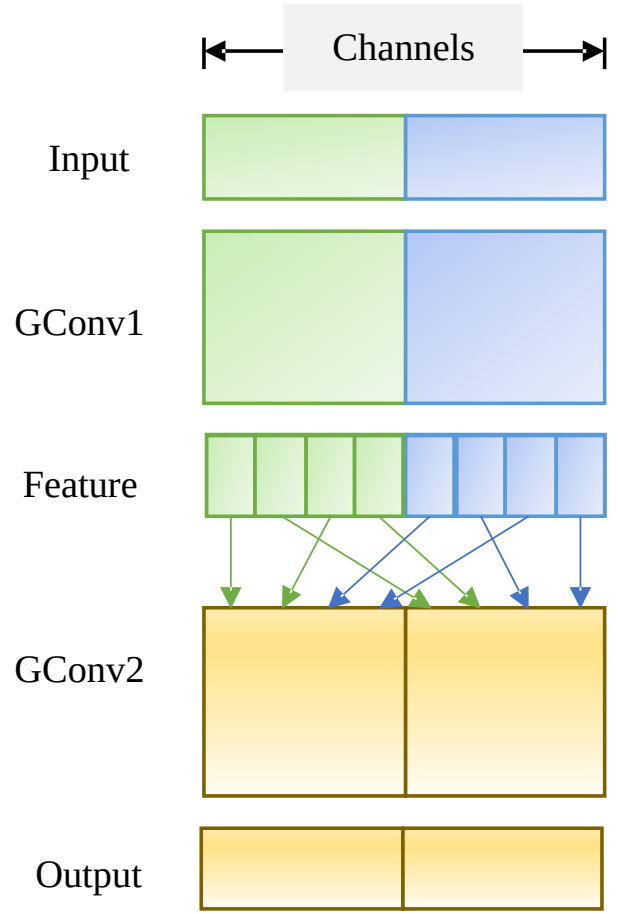


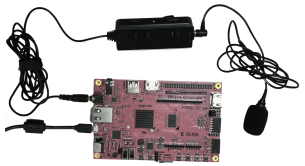
文獻探討 — ShuffleNet

- 核心技術
  - Depthwise & Pointwise Convolution



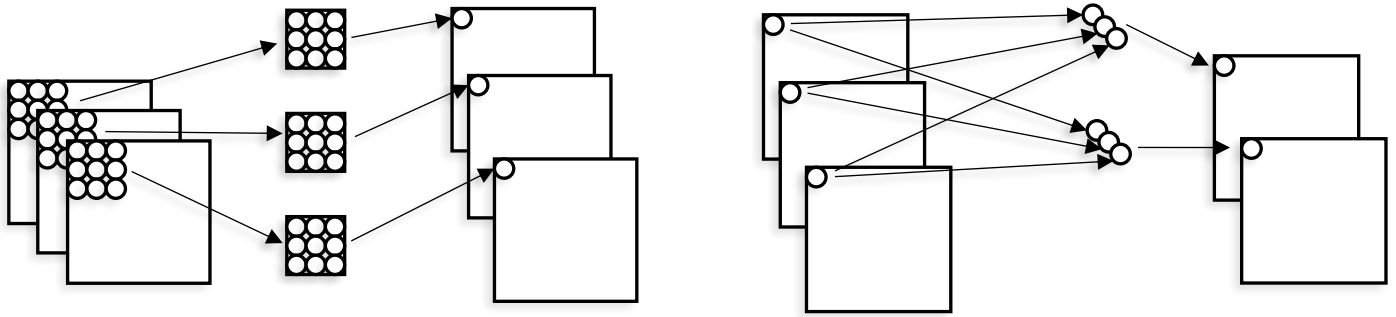
- Channel Shuffle





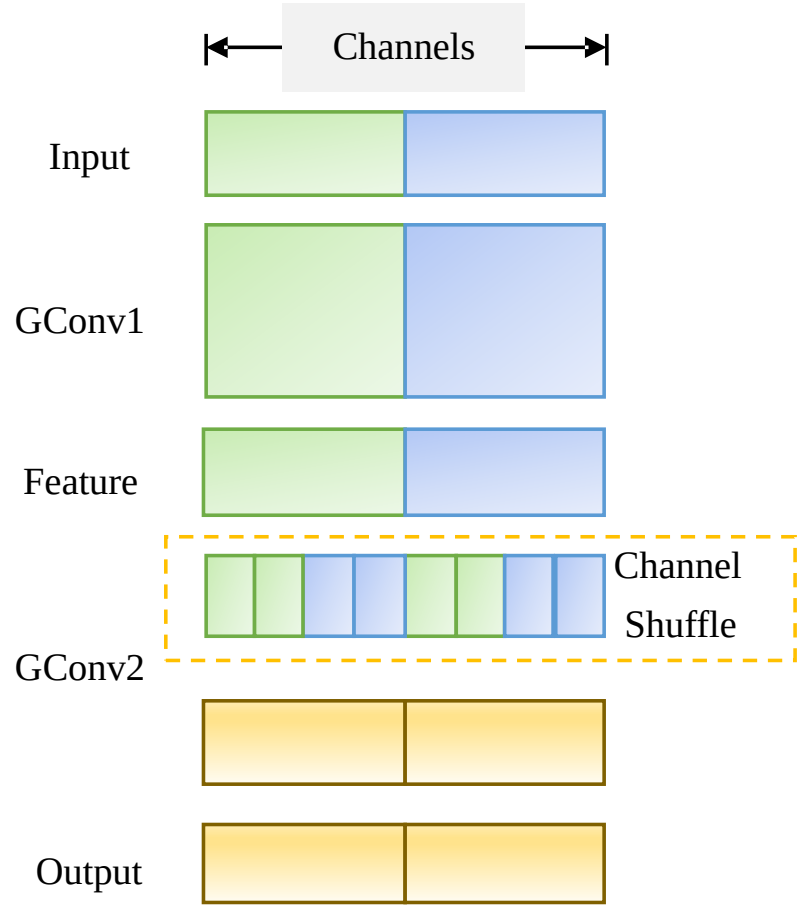
文獻探討 — ShuffleNet

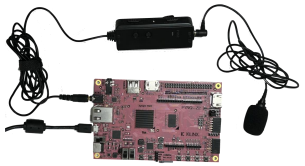
- 核心技術
  - Depthwise & Pointwise Convolution



優點：減少運算量、參數儲存量，學習更複雜特徵

- Channel Shuffle

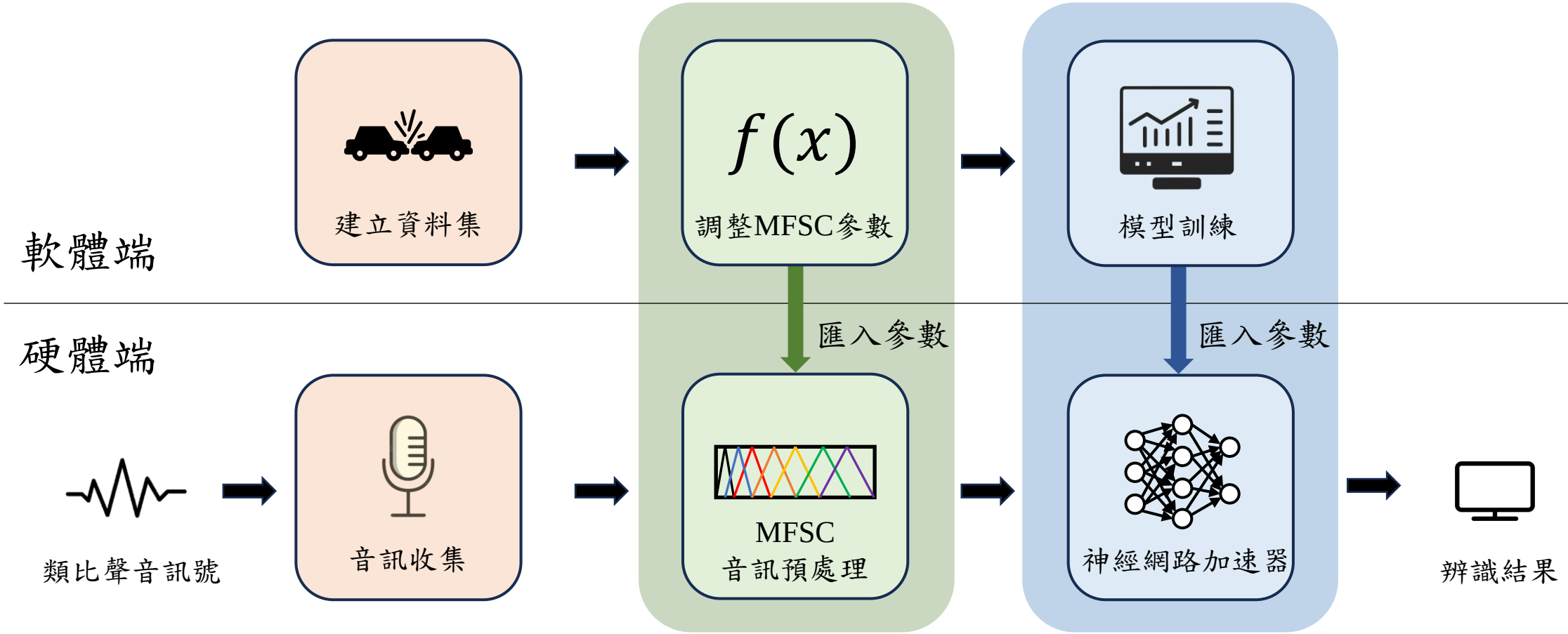




作品設計概念

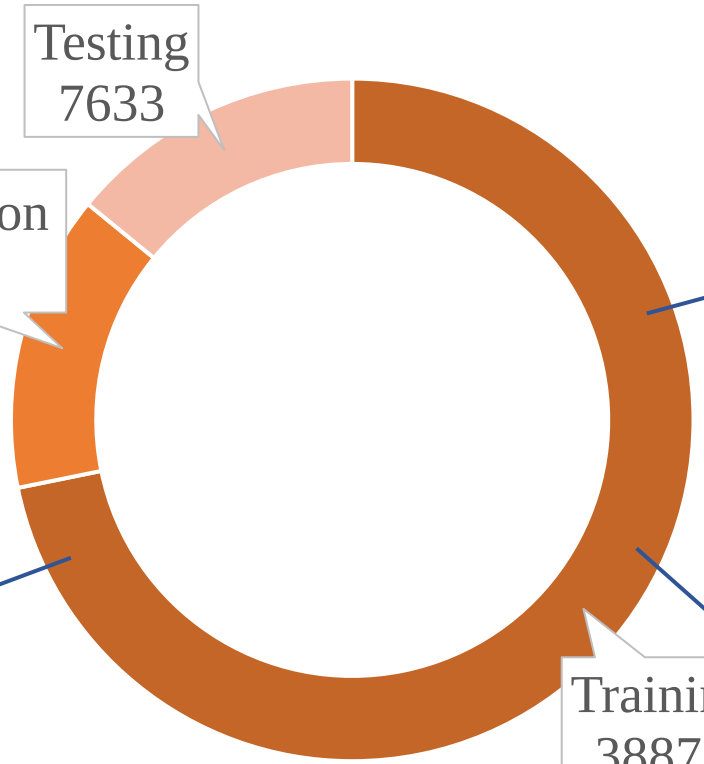
軟硬體整合流程、軟體端神經網路訓練、硬體開發

軟硬體整合流程

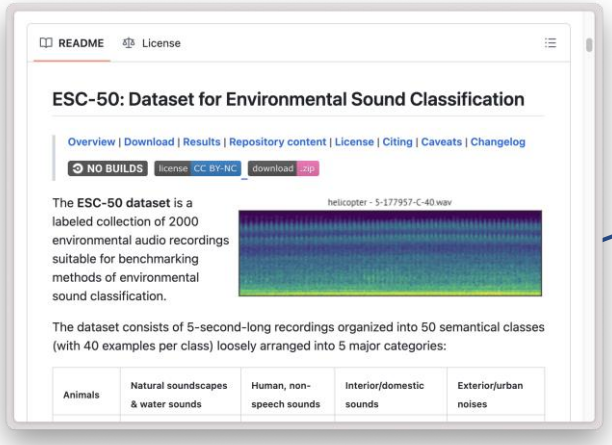

軟體端
硬體端
建立資料集
$f(x)$
調整MFSC參數
模型訓練
匯入參數
匯入參數
類比聲音訊號
音訊收集
MFSC
音訊預處理
神經網路加速器
辨識結果

軟體端神經網路訓練

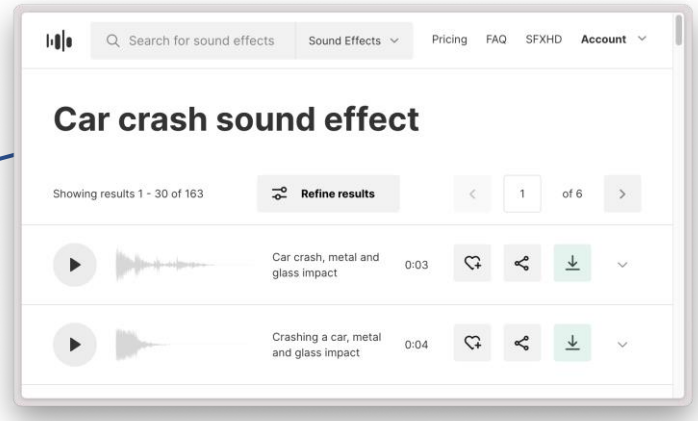
- 車禍聲音資料集



- ESC-50 開源資料庫



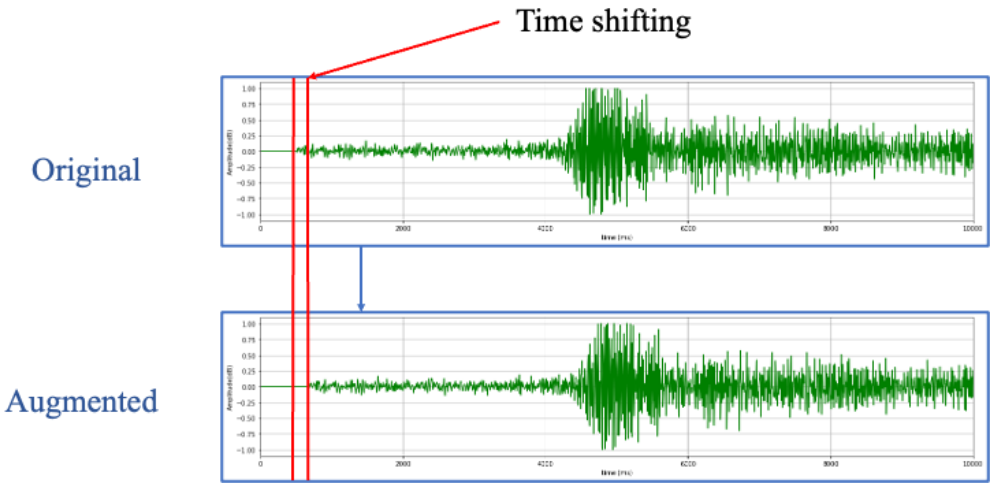
- Soundsnap 專業資料庫

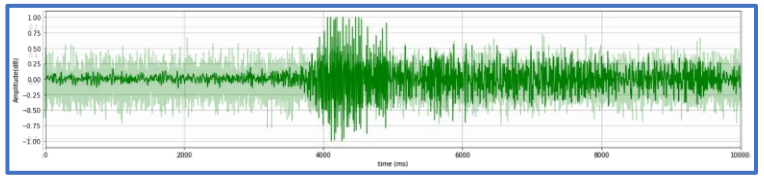


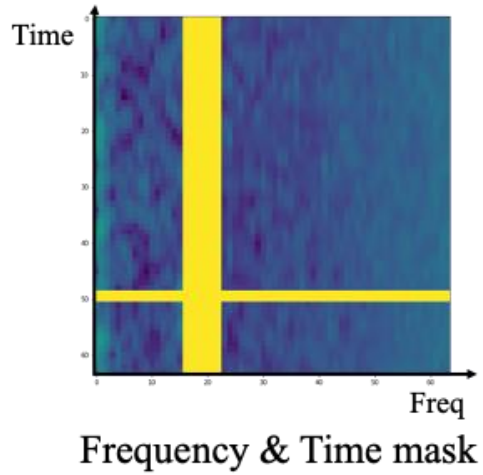
- 環境雜音

軟體端神經網路訓練

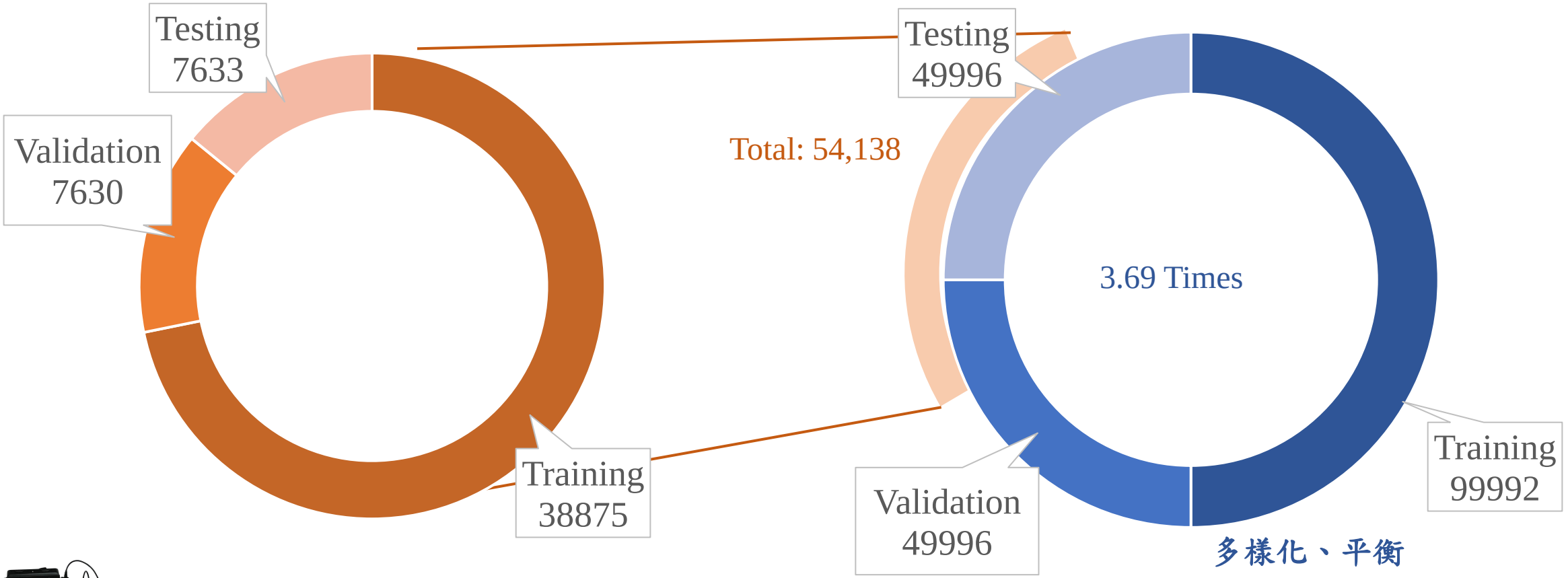
- Data augmentation
  - 音訊時序平移 — 提升適應不同時序音訊資料的能力
  - 加入環境噪音 — 提升對抗環境雜音的能力
  - 加入遮罩(mask) —避免模型訓練overfitting

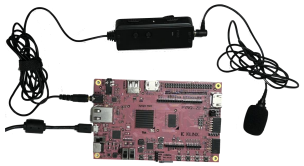






軟體端神經網路訓練



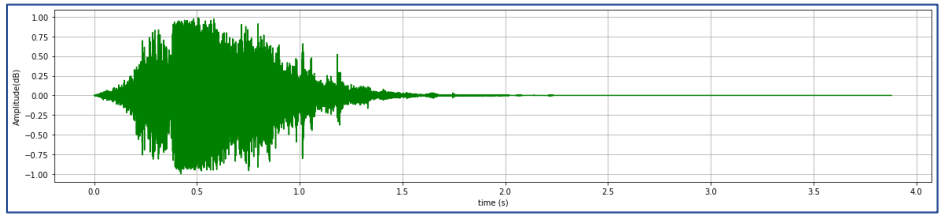


- 原資料集
- 擴充後

軟體端神經網路訓練

- 資料集預處理

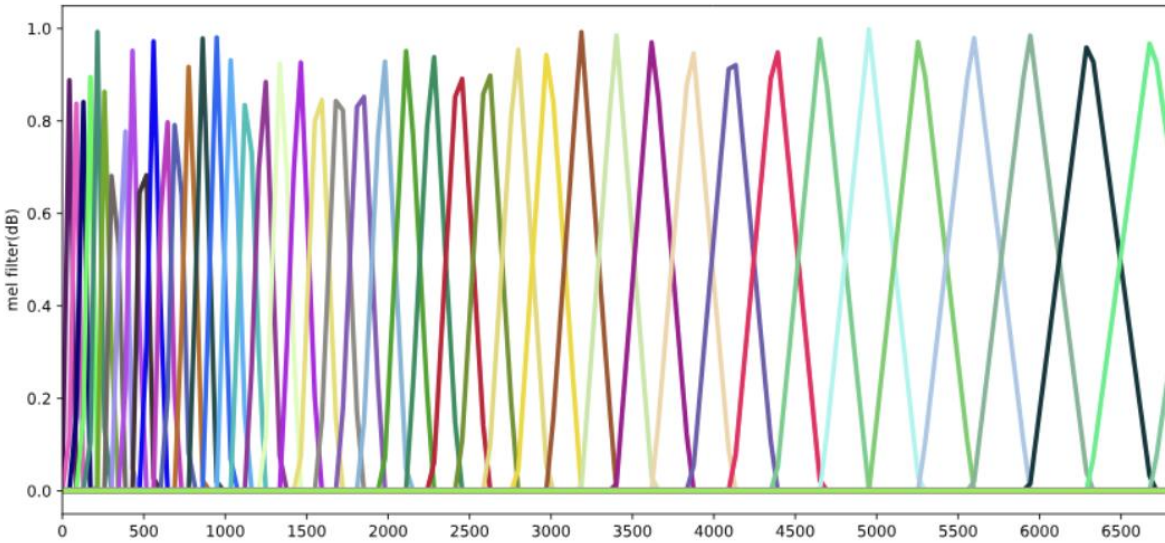
正規化



- 聲音訊號

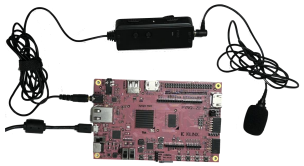
濾波

調整參數



- 梅爾濾波器

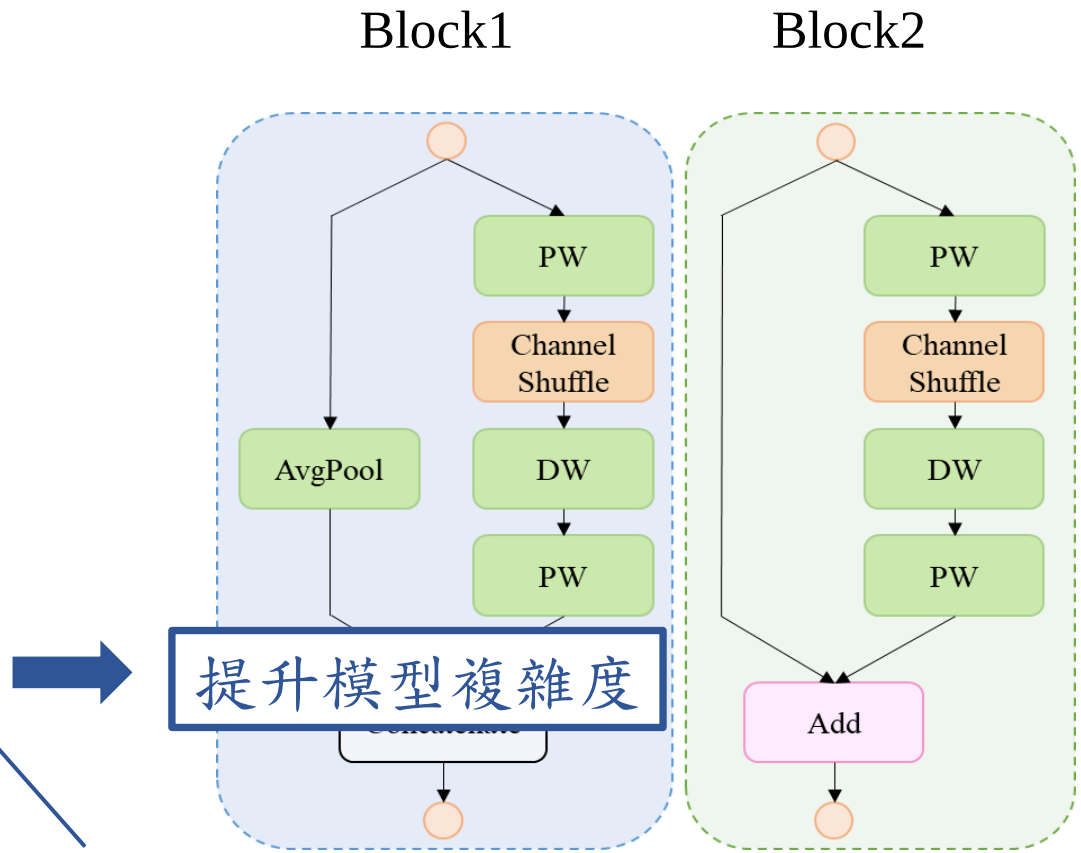
雜音過濾、特徵提取



軟體端神經網路訓練

- ShuffleNet 神經網路架構

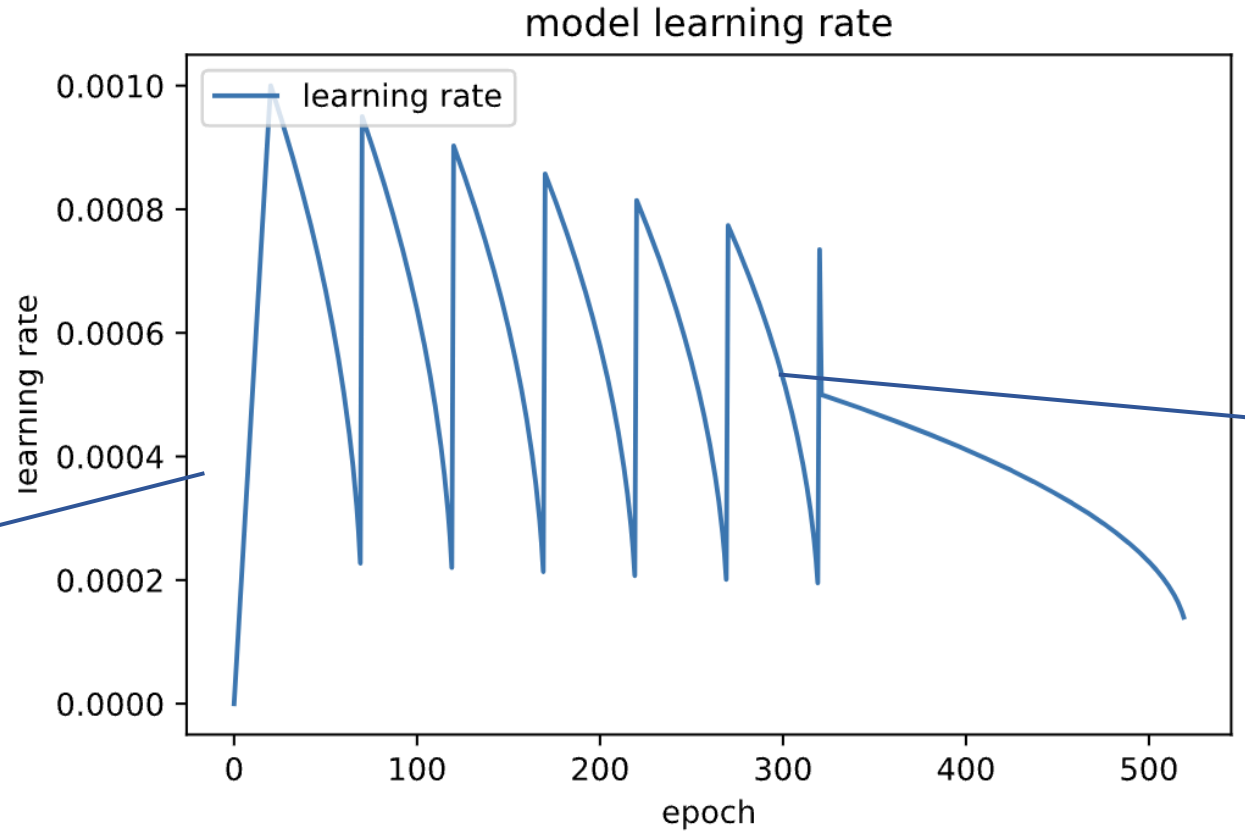
Layer		Stride	Output Size	Output Channel
Input		-	64×64	1
3×3 Global Conv		2	32×32	16
MaxPool		2	16×16	16
Stage1	Block1	2	8×8	32
	Block2	1	8×8	32
Stage2	Block1	2	4×4	64
	Block2	1	4×4	64
	Block2	1	4×4	64
	Block2	1	4×4	64
Global AvgPool		-	1×1	64
Fully Connected		-	2	1



增加Block2數量

軟體端神經網路訓練

- ShuffleNet模型訓練

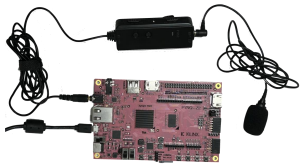


Warm up

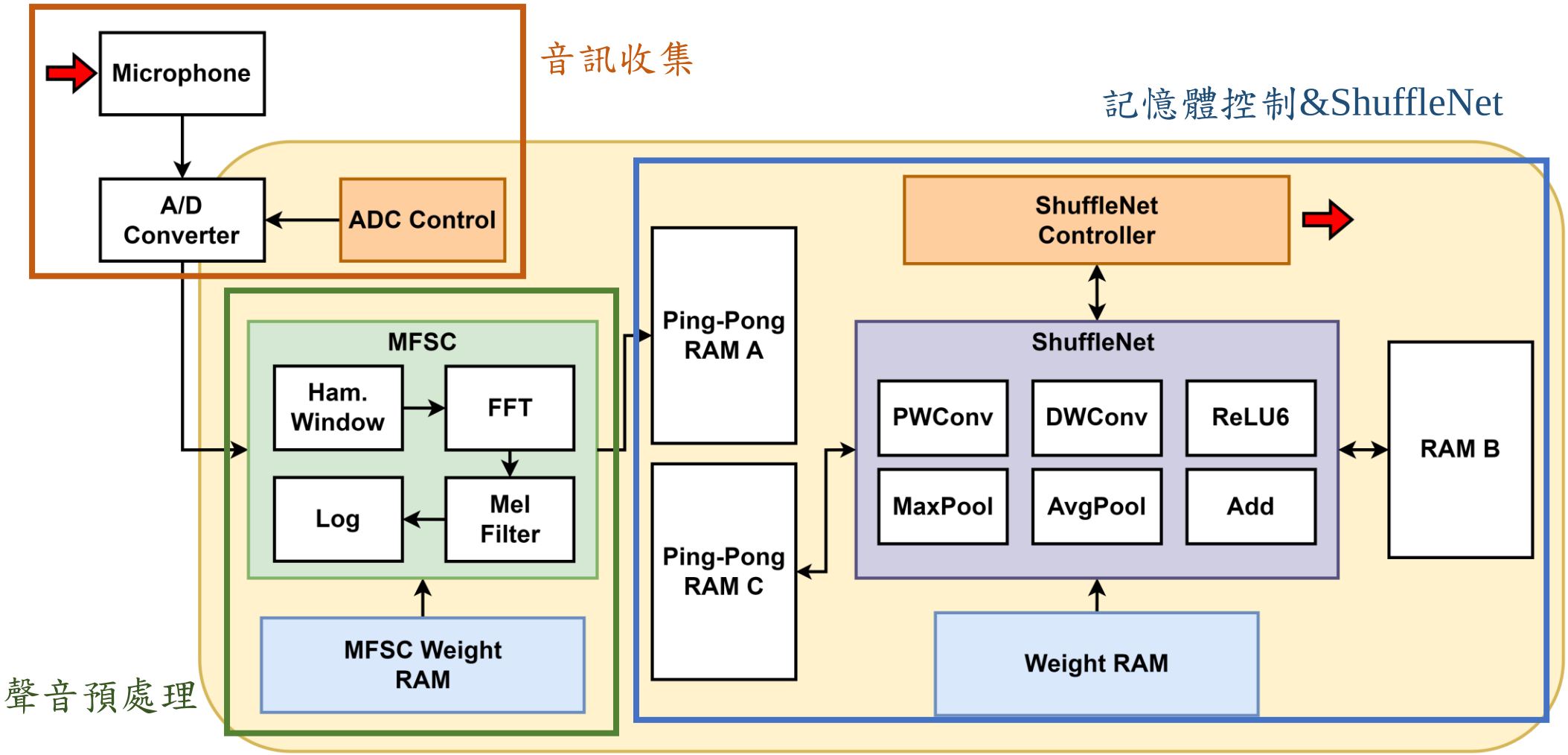
前期訓練更加穩定

Polynomial Decay

跳出區域最佳解

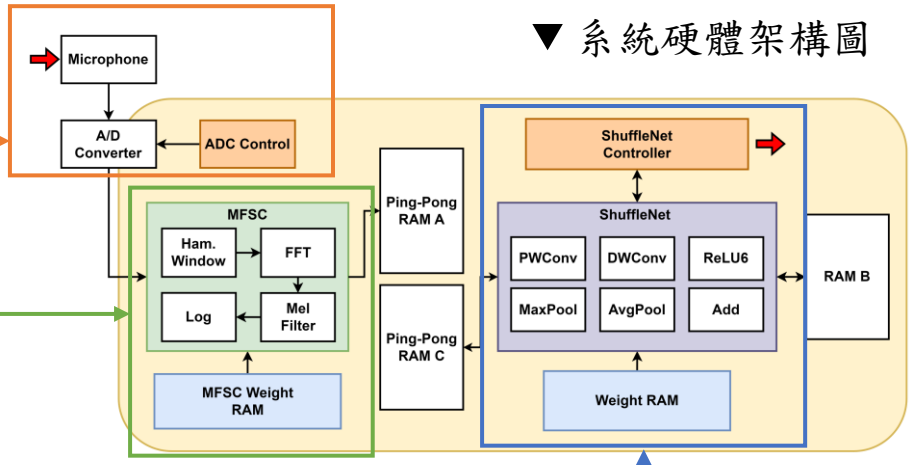


軟硬體整合流程

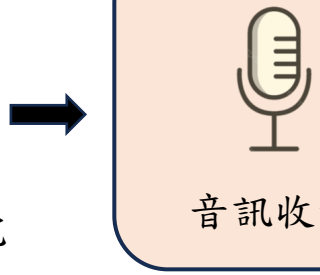

音訊收集
Microphone
A/D Converter
ADC Control
記憶體控制&ShuffleNet
ShuffleNet Controller
MFSC
Ham. Window
FFT
Log
Mel Filter
MFSC Weight RAM
Ping-Pong RAM A
Ping-Pong RAM C
ShuffleNet
PWConv
DWConv
ReLU6
MaxPool
AvgPool
Add
RAM B
Weight RAM
聲音預處理

軟硬體整合流程

▼ 系統硬體架構圖

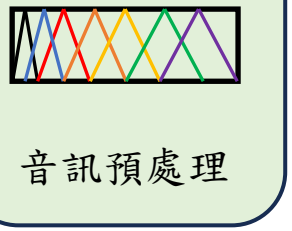


類比聲音訊號

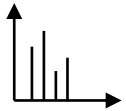


音訊收集

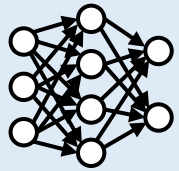
數位訊號



音訊預處理



頻域訊號



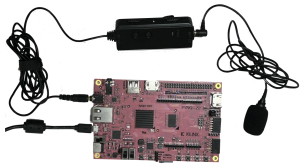
ShuffleNet

辨識結果

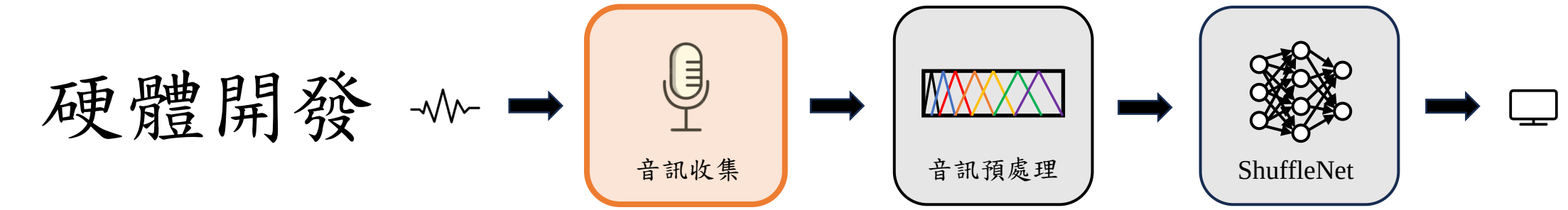
- 麥克風收音
- ADC 類比轉數位

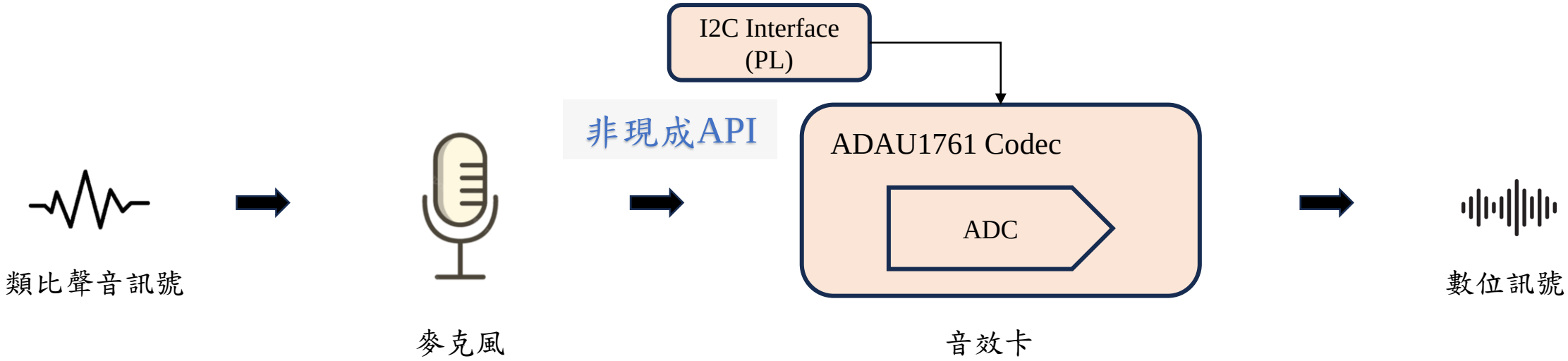
- 分幀加窗
- FFT & Modulo
- 梅爾濾波器
- 取絕對值&對數

- 記憶體調控
- 深度可分離卷積



硬體開發





- 克服難點
  - 編寫純PL端音效卡I2C驅動 (**並未使用現成API**)
  - **避免PS端硬體成本與功耗**

硬體開發

音訊收集 → 音訊預處理 → ShuffleNet

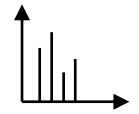
梅爾頻譜係數 (MFSC)

不利於辨識

時域訊號

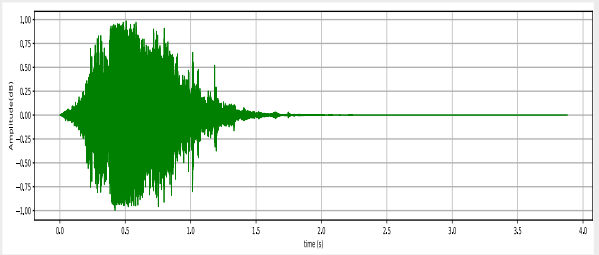
- 分幀加窗
- FFT & Modulo
- 梅爾濾波器

易提取特徵

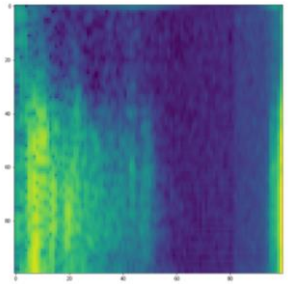


頻域訊號

- 克服難點
  - **各功能IP整合**
  - 梅爾濾波器參數調整
  - **兼顧硬體成本與特徵提取能力**

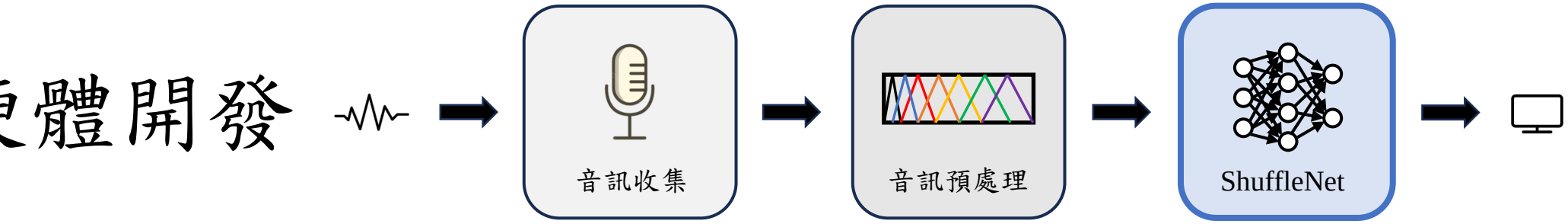


時域訊號



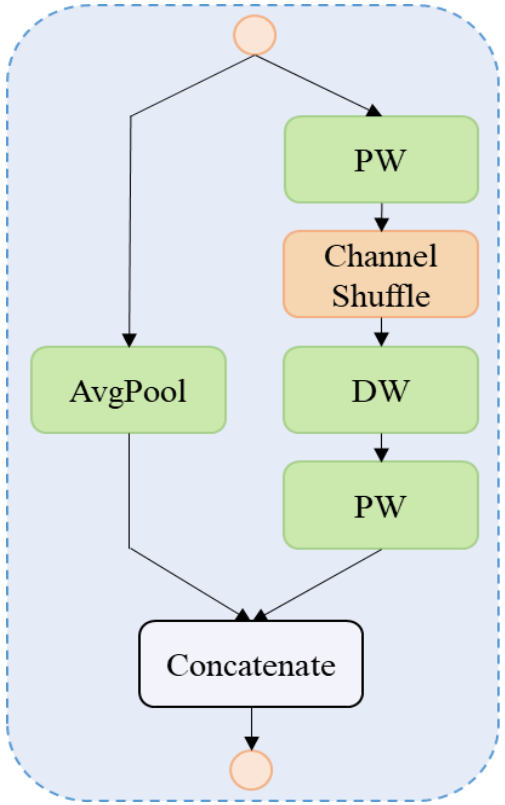
頻譜圖

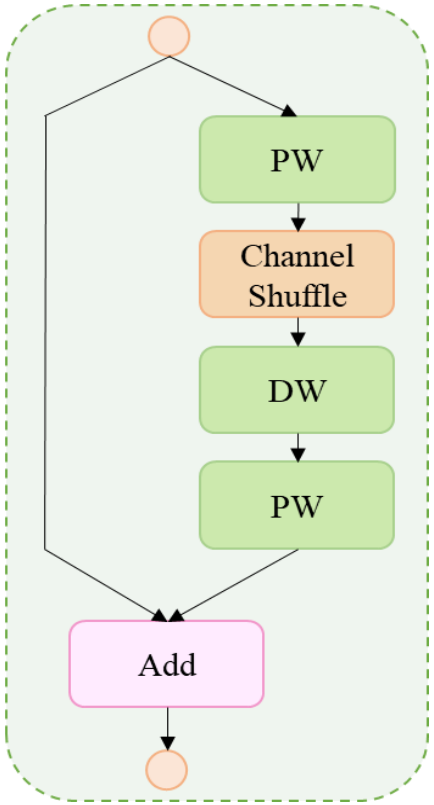
硬體開發



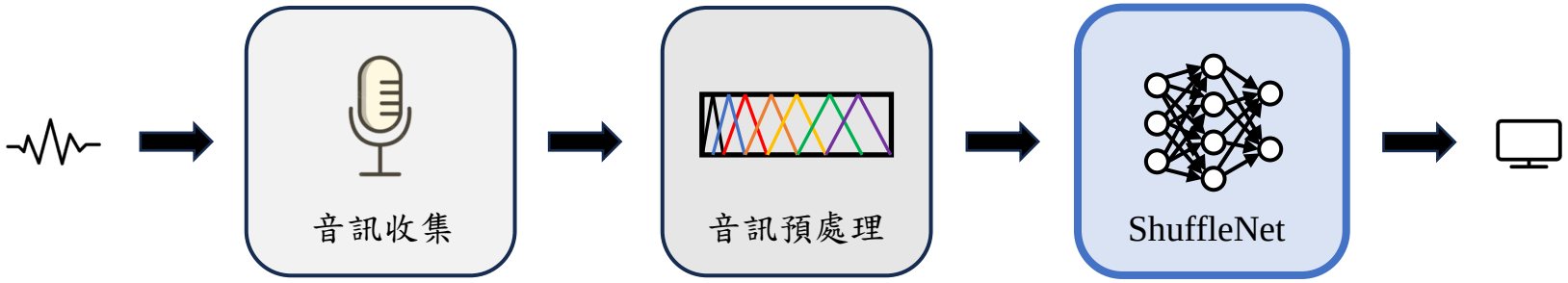
- ShuffleNet 神經網路架構

Layer		Stride	Output Size	Output Channel
Input		-	64x64	1
3x3 Global Conv		2	32x32	16
MaxPool		2	16x16	16
Stage1	Block1	2	8x8	32
	Block2	1	8x8	32
Stage2	Block1	2	4x4	64
	Block2	1	4x4	64
	Block2	1	4x4	64
	Block2	1	4x4	64
Global AvgPool		-	1x1	64
Fully Connected		-	2	1





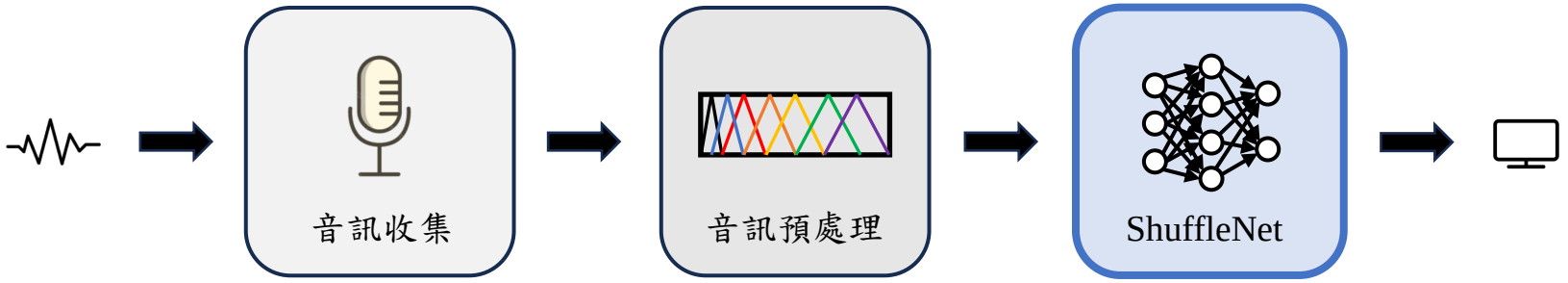
硬體開發



- ShuffleNet 計算步驟類別

Max Pool	Channel Shuffle	Add	
Depthwise Conv	Avg Pool	Pointwise Conv	Fully Connected

硬體開發



- 最大池化 Max Pool

  與全域卷積步驟合併做Pipeline處理

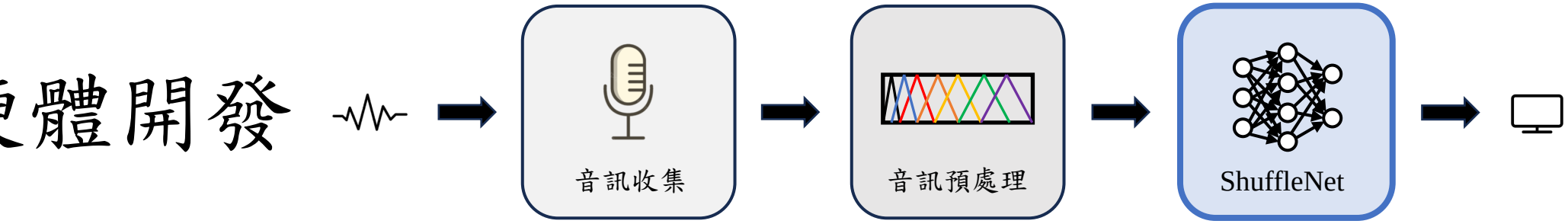
Layer		Stride	Output Size	Output Channel
Input		-	64×64	1
3×3 Global Conv		2	32×32	16
MaxPool		2	16×16	16
Stage1	Block1	2	8×8	32
	Block2	1	8×8	32
Stage2	Block1	2	4×4	64
	Block2	1	4×4	64
	Block2	1	4×4	64
	Block2	1	4×4	64
Global AvgPool		-	1×1	64
Fully Connected		-	2	1

Max Pool

Channel Shuffle

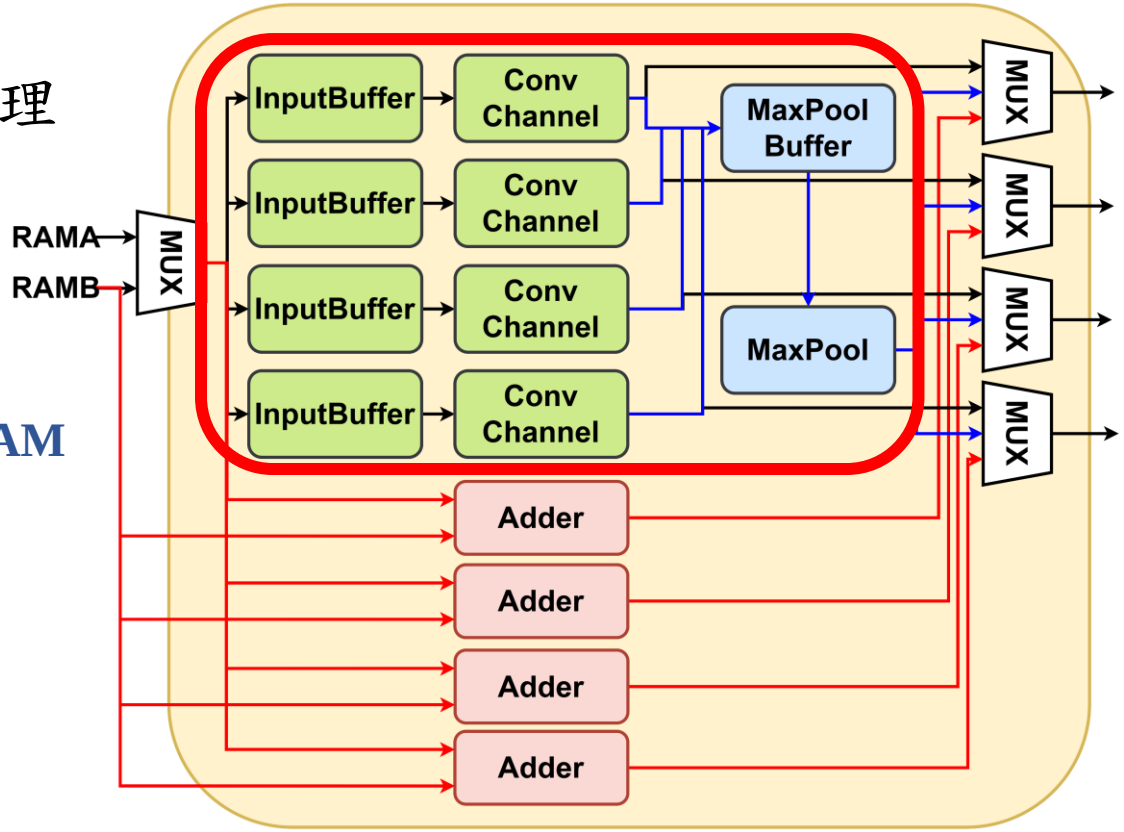
Add

硬體開發



- 最大池化 Max Pool

  與全域卷積步驟合併做Pipeline處理



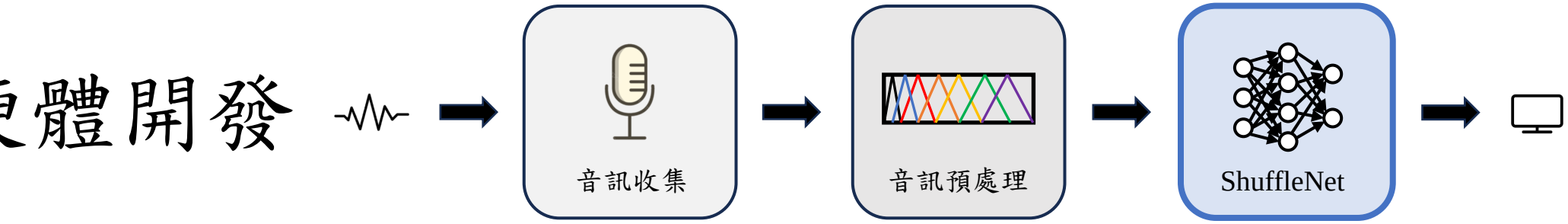
節省80%RAM

Max Pool

Channel Shuffle

Add

硬體開發

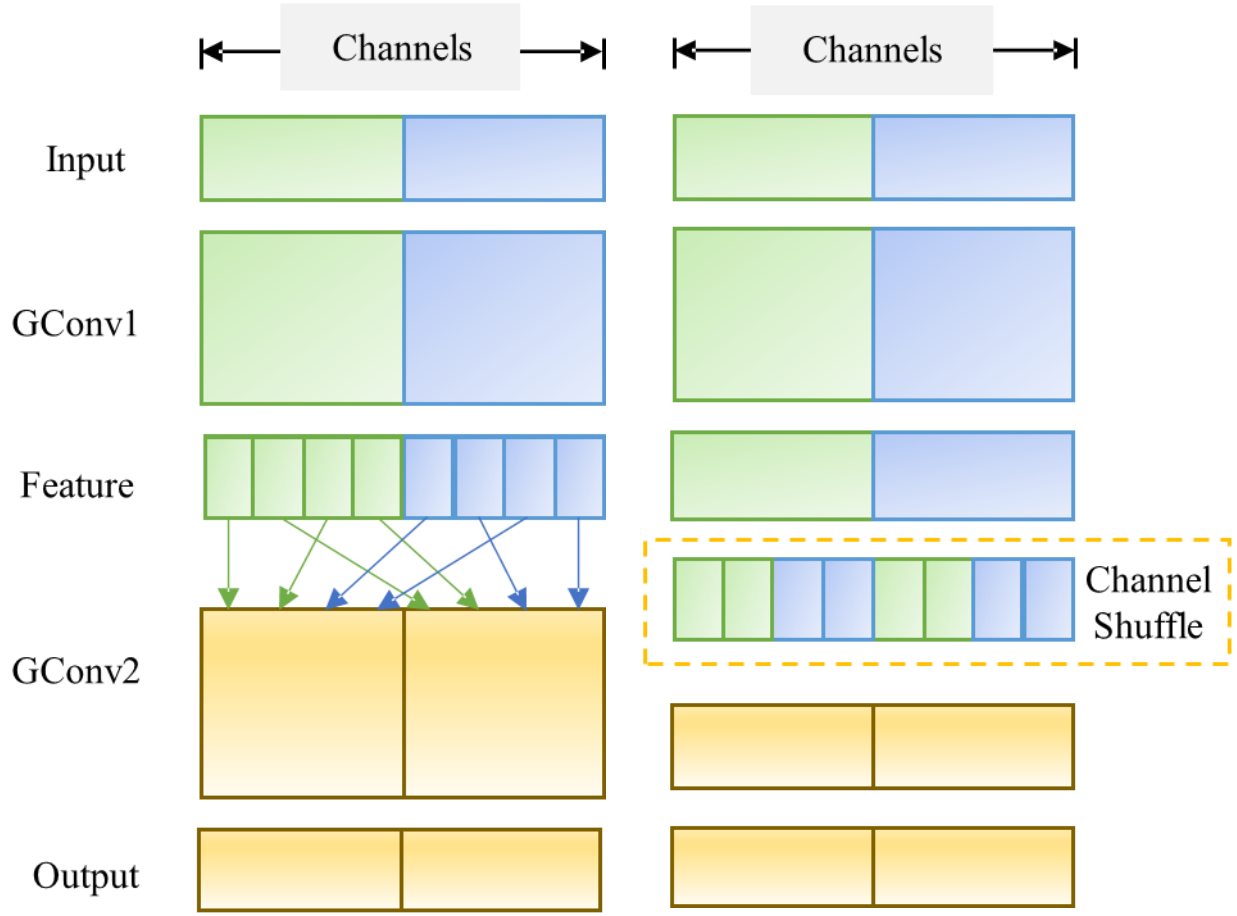


- 最大池化 Max Pool

  與全域卷積步驟合併做Pipeline處理

- 通道洗牌 Channel Shuffle

  直接調整記憶體儲存位址

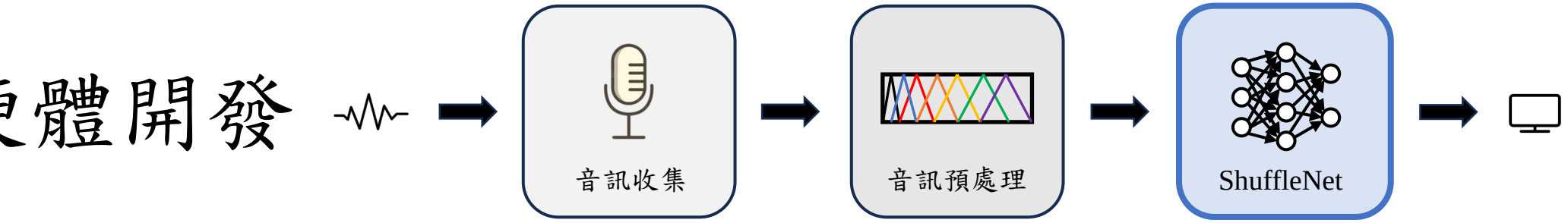


Max Pool

Channel Shuffle

Add

硬體開發



- 最大池化 Max Pool

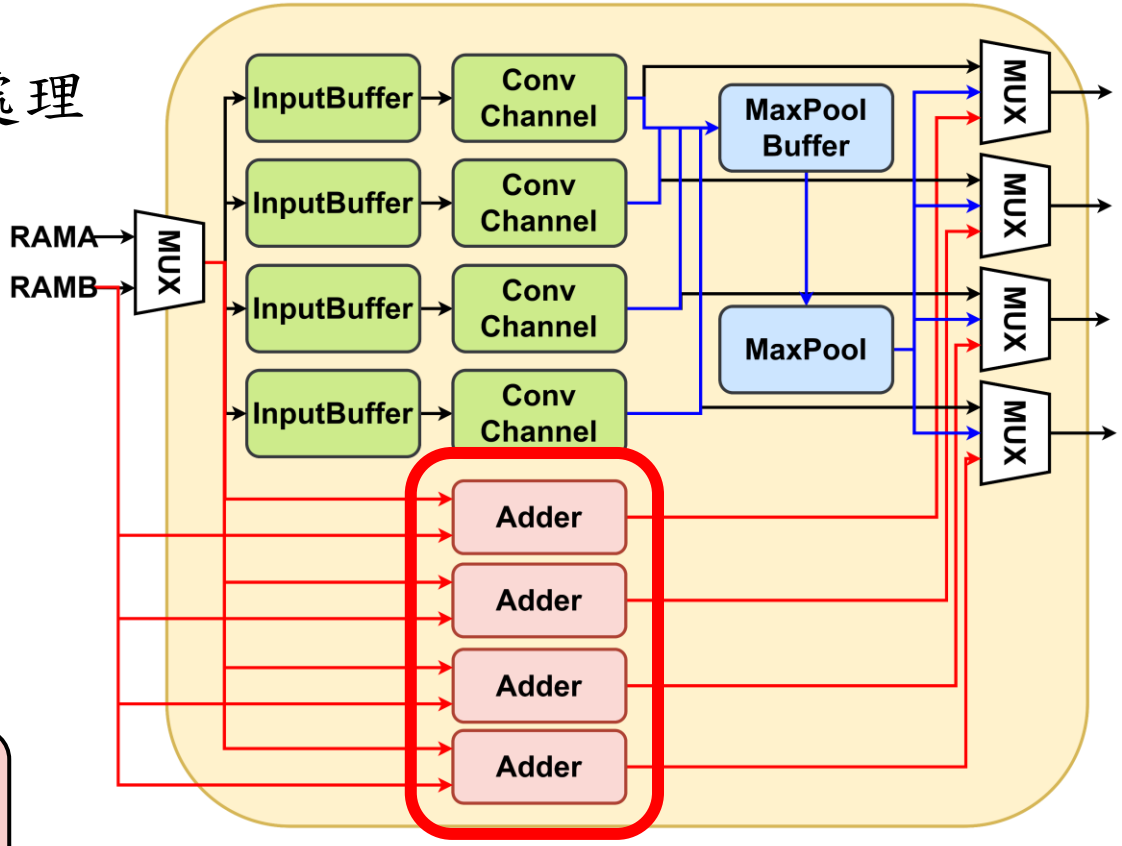
  與全域卷積步驟合併做Pipeline處理

- 通道洗牌 Channel Shuffle

  直接調整記憶體儲存位址

- 特徵圖疊加 Add

  額外設計加法器



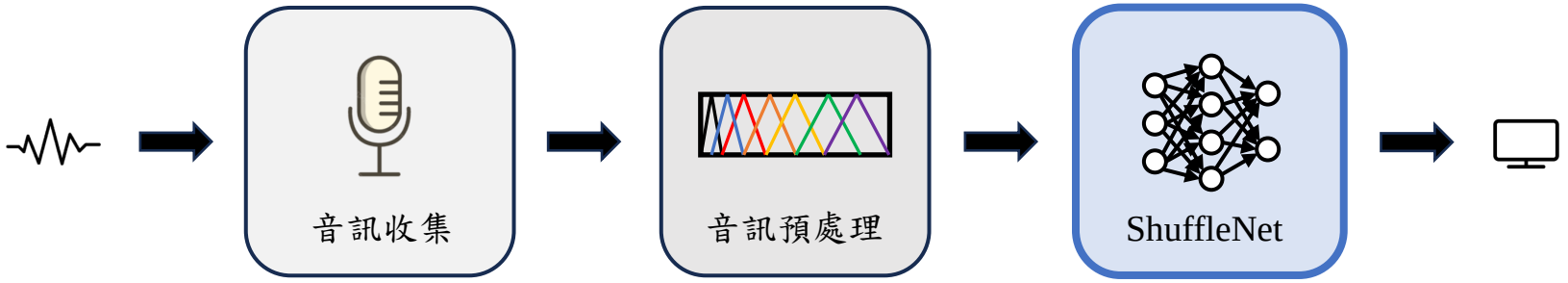
使用四組運算單元進行平行化運算

Max Pool

Channel Shuffle

Add

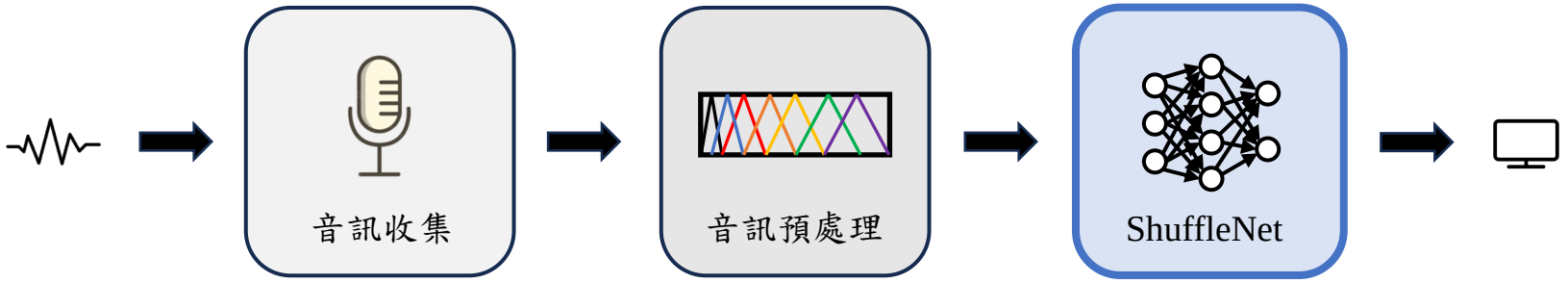
硬體開發



- 乘法-加法計算

Depthwise Conv	Avg Pool	Pointwise Conv	Fully Connected

硬體開發



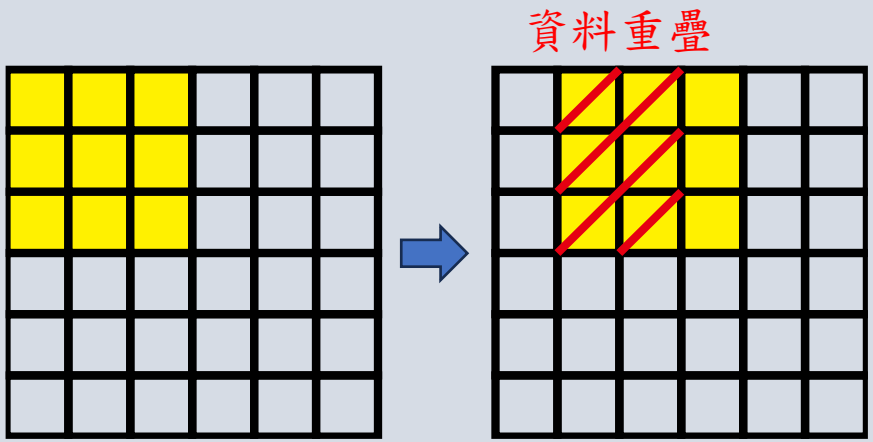
- 乘法-加法計算

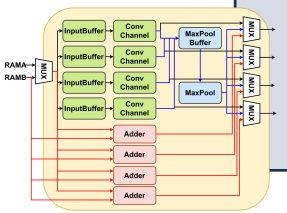
Depthwise類

計算時存在資料重疊的現象

Depthwise Conv

3*3 深度卷積





Pointwise類

計算時不存在資料重疊的現象

Avg Pool

平均池化 (2*2 與 4*4)

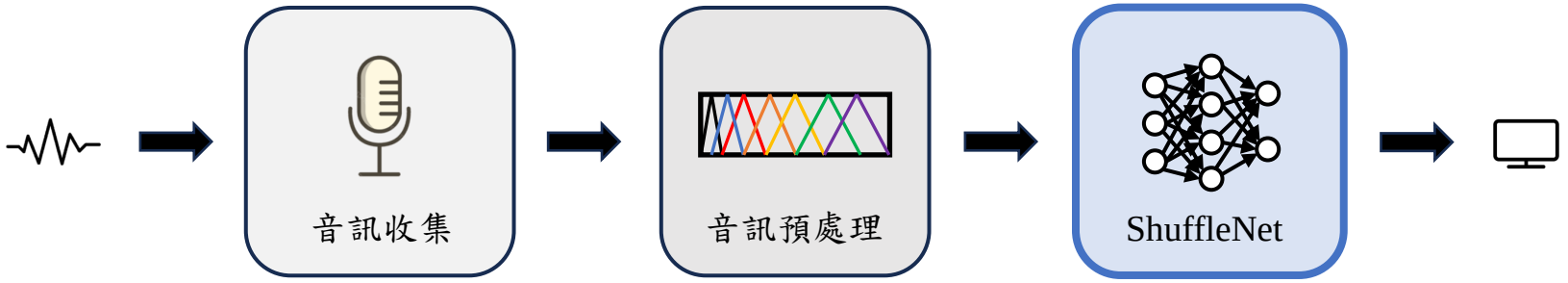
Pointwise Conv

逐點卷積

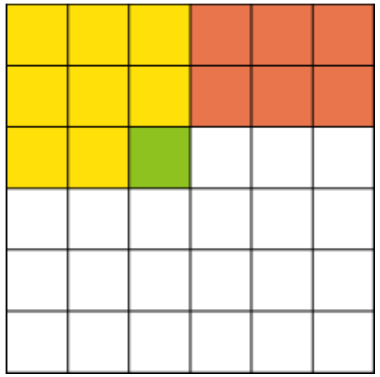
Fully Connected

全連接層

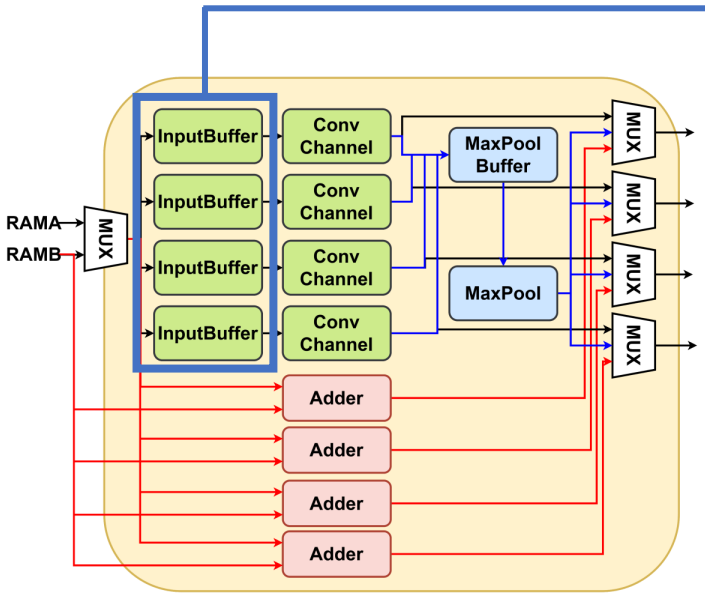
硬體開發

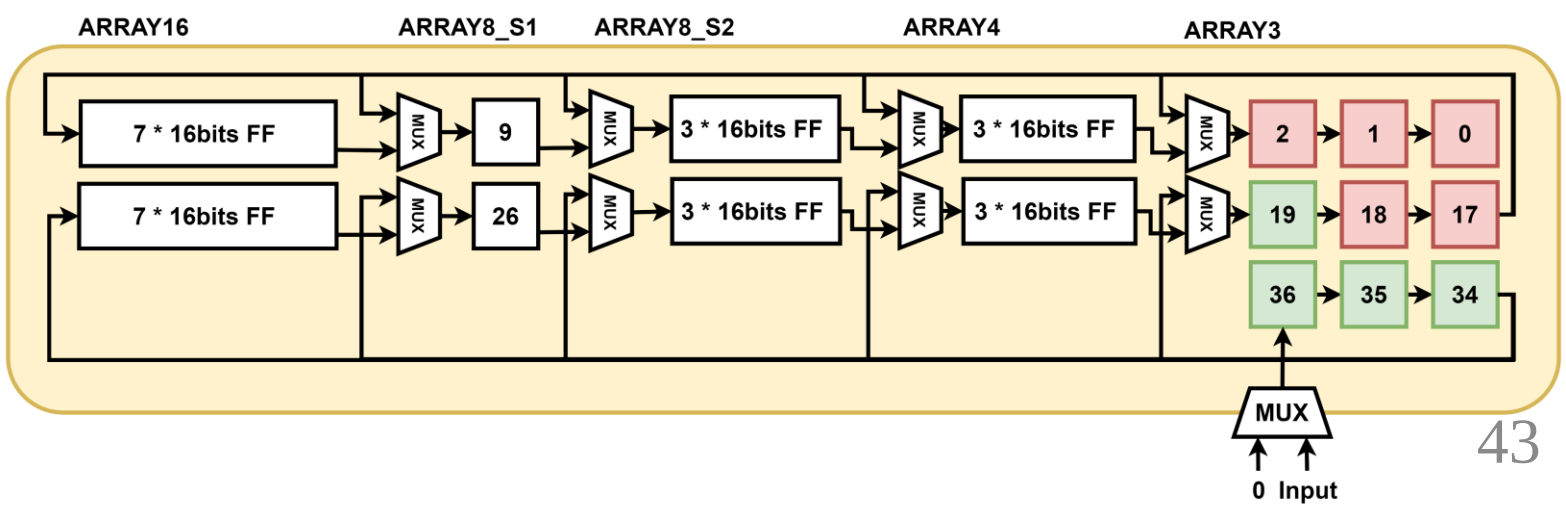


- 輸入緩衝
  - 消除RAM輸出位寬限制
  - Tapped delay line 容許緩衝多條線路移位

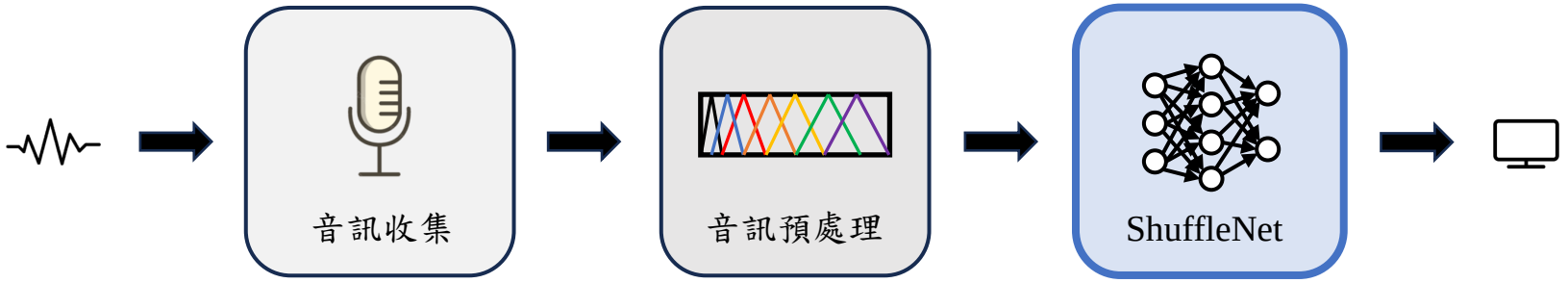


特徵圖存取

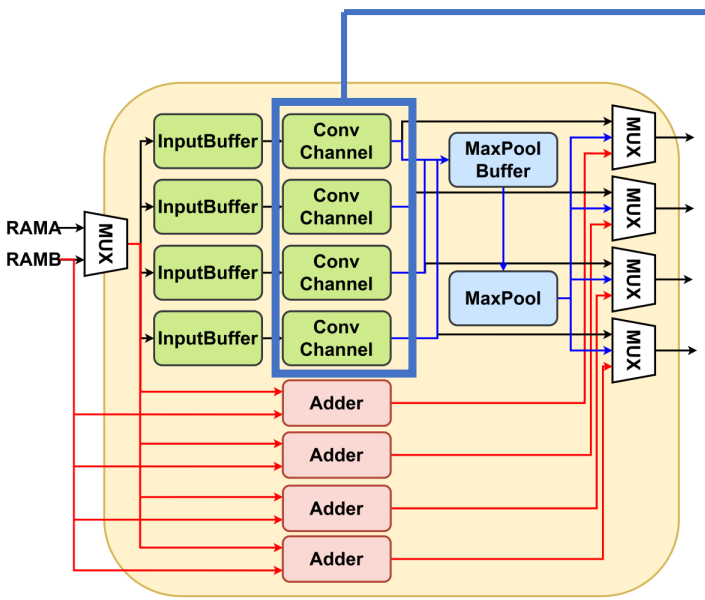


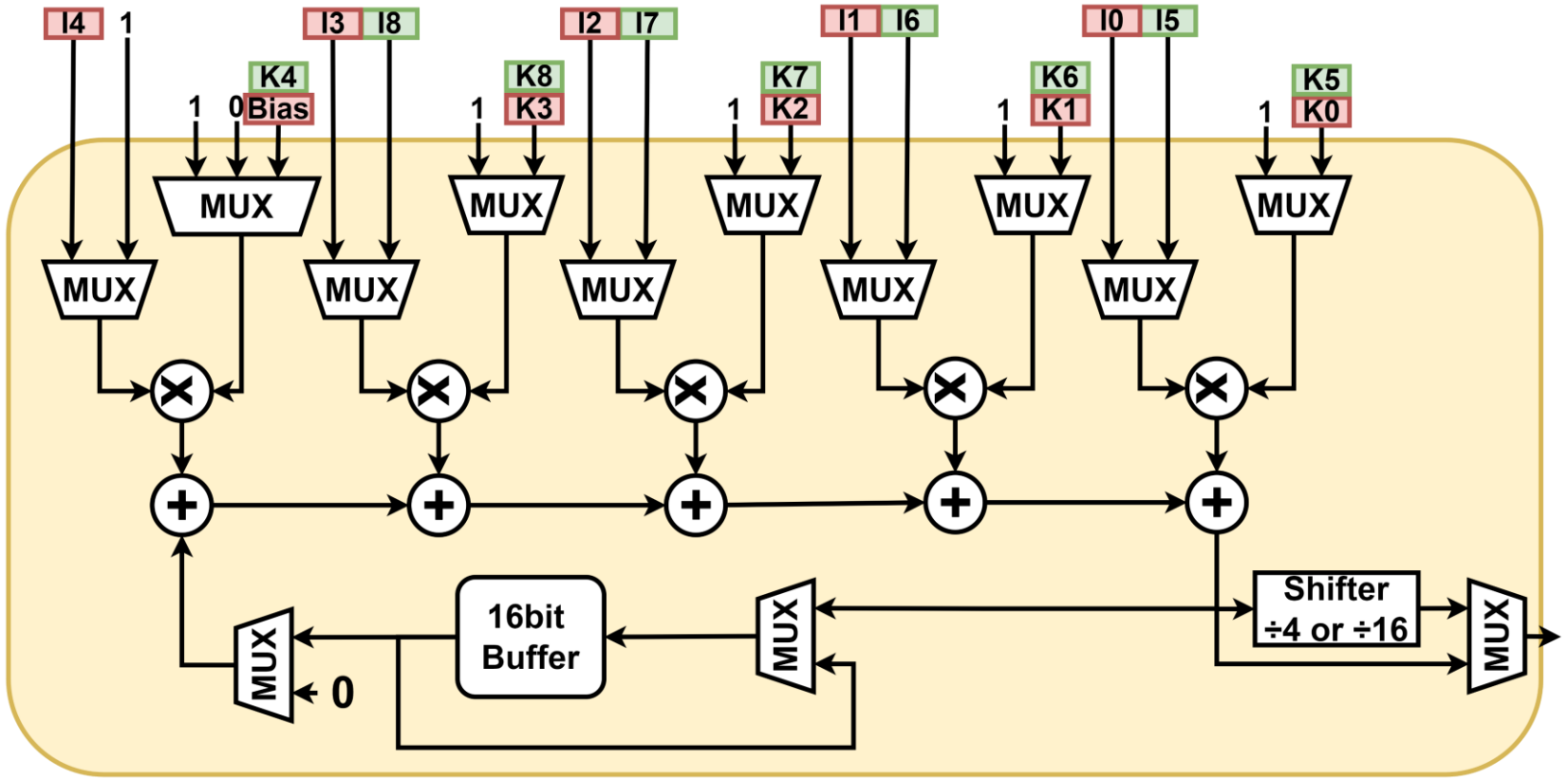


硬體開發

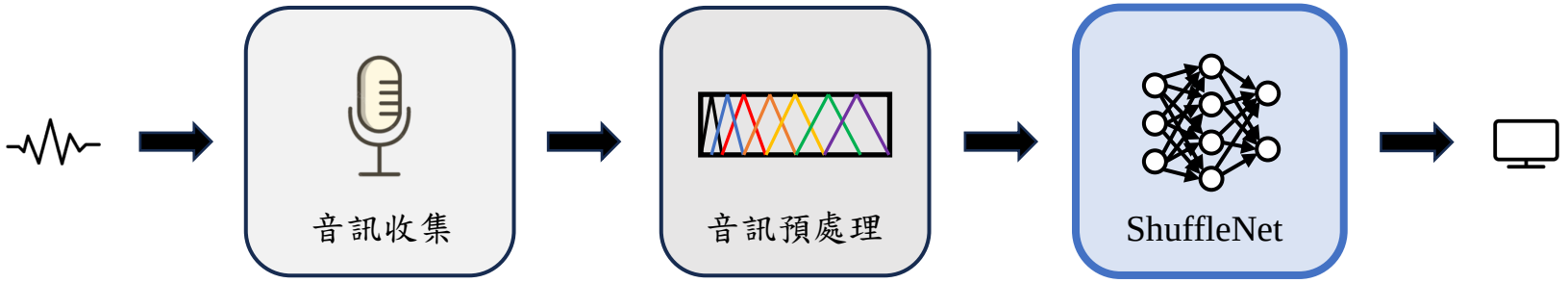


- 卷積通道
  - 3*3卷積
  - 多通道卷積

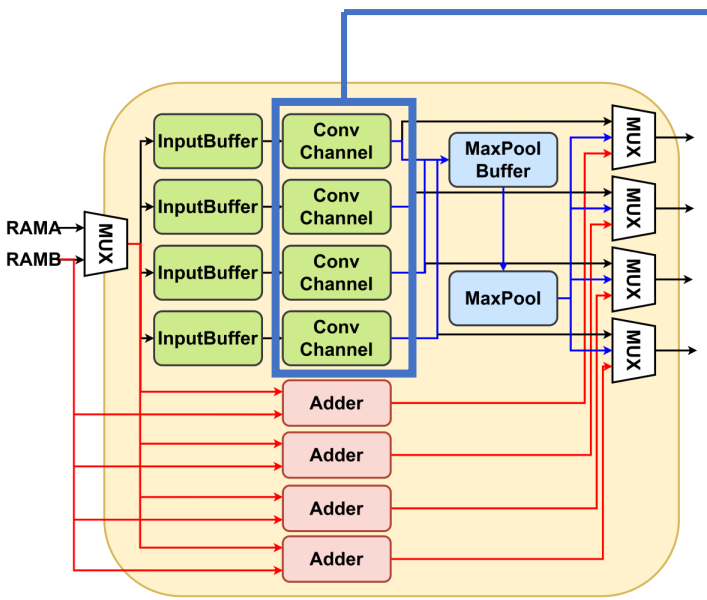


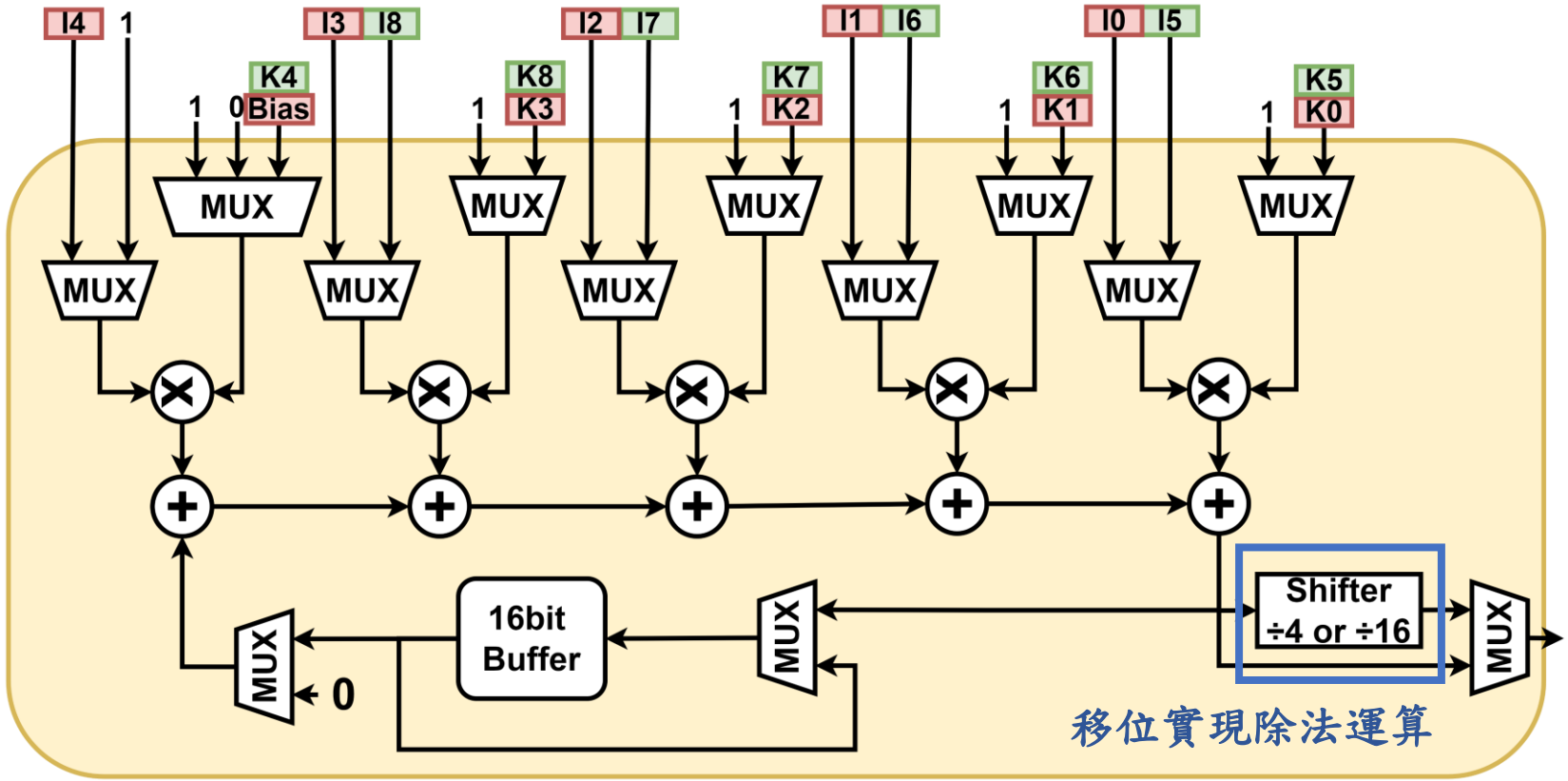


硬體開發

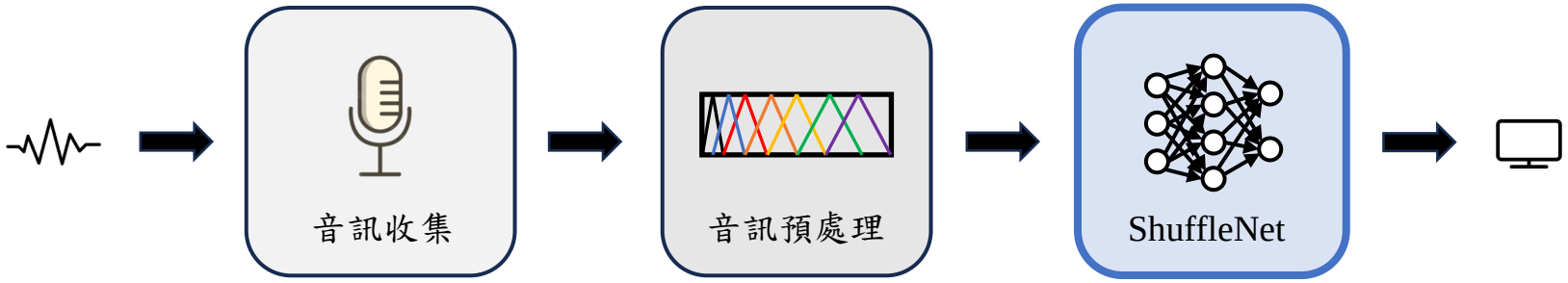


- 卷積通道
  - 3*3卷積
  - 多通道卷積
  - 平均池化





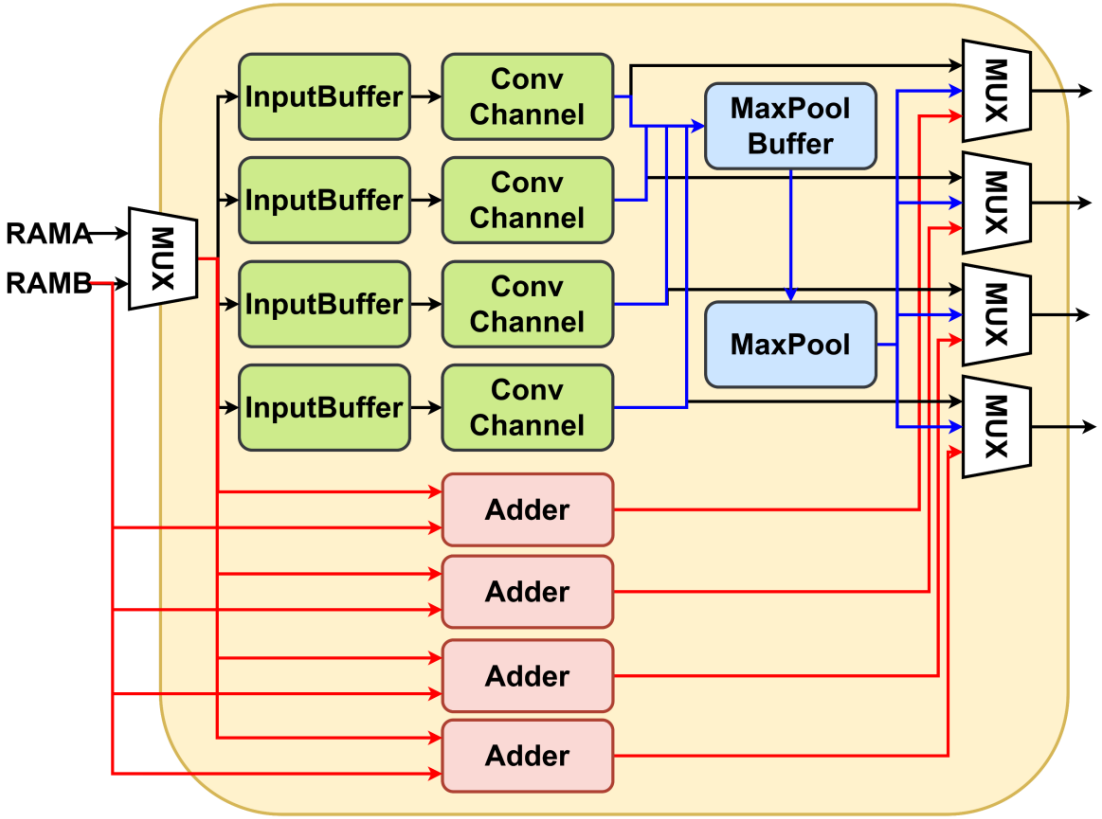
硬體開發



平行運算

流水線處理

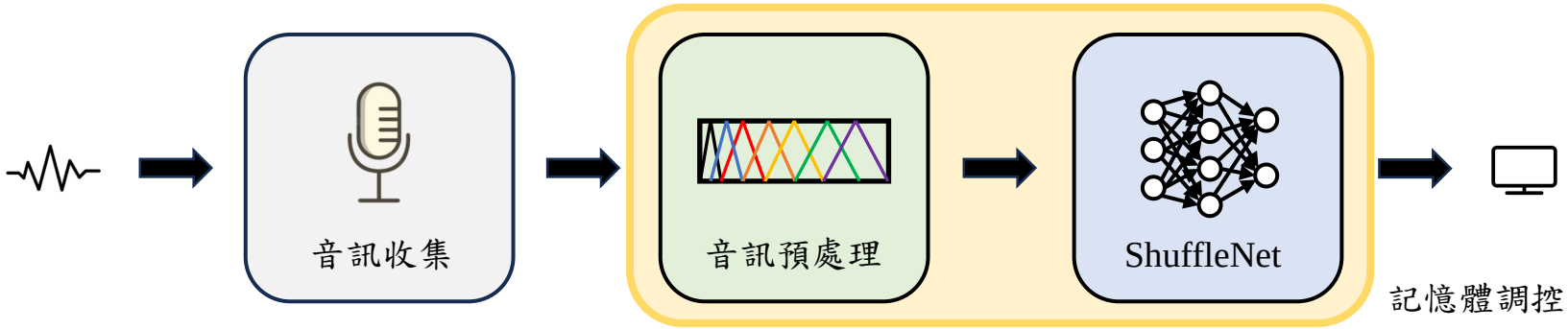
硬體複用



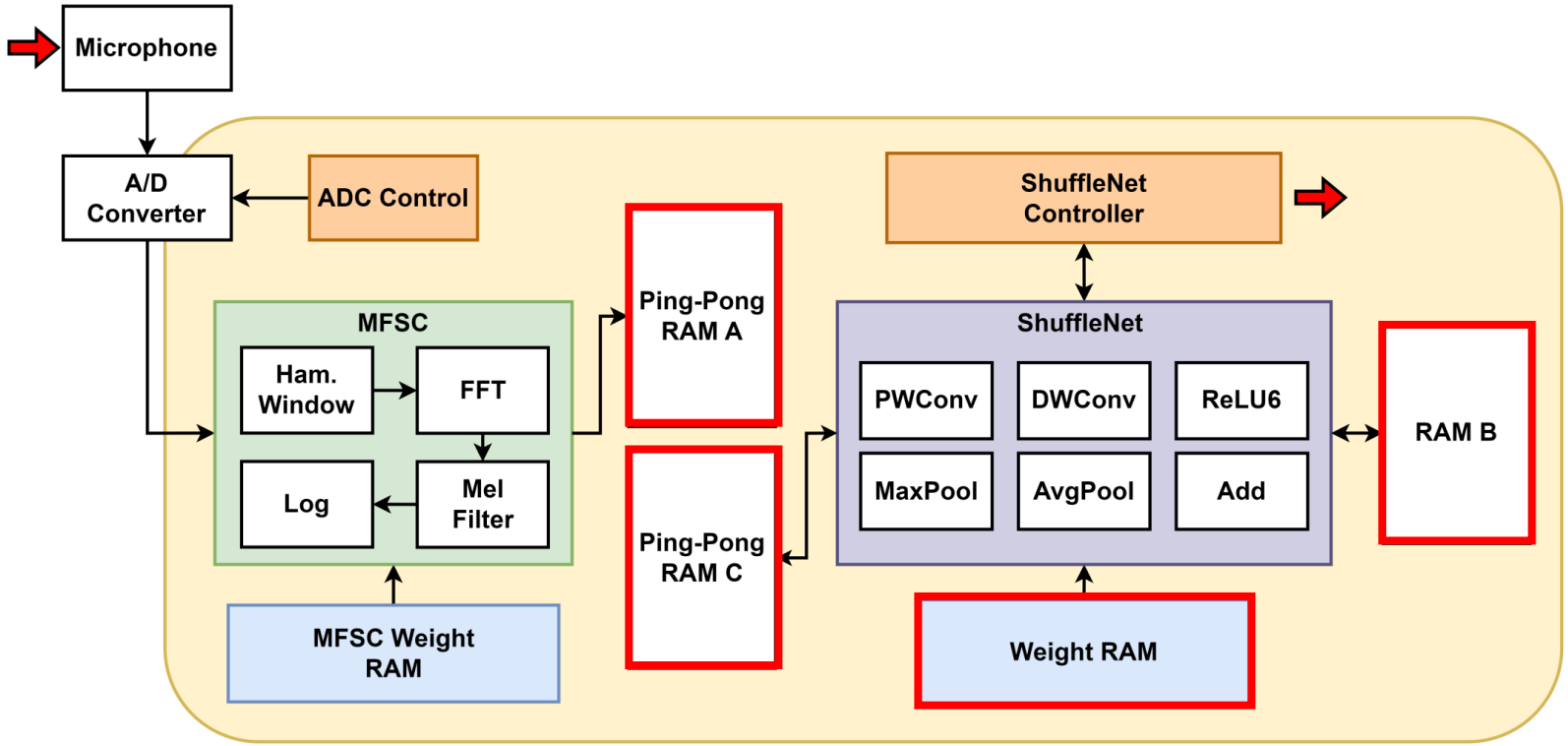
低硬體資源使用

高運算效率

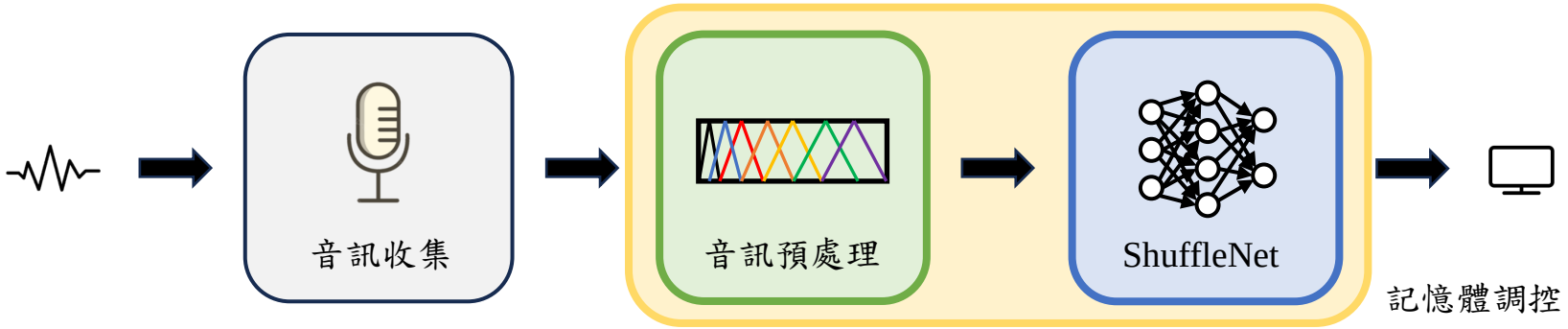
硬體開發



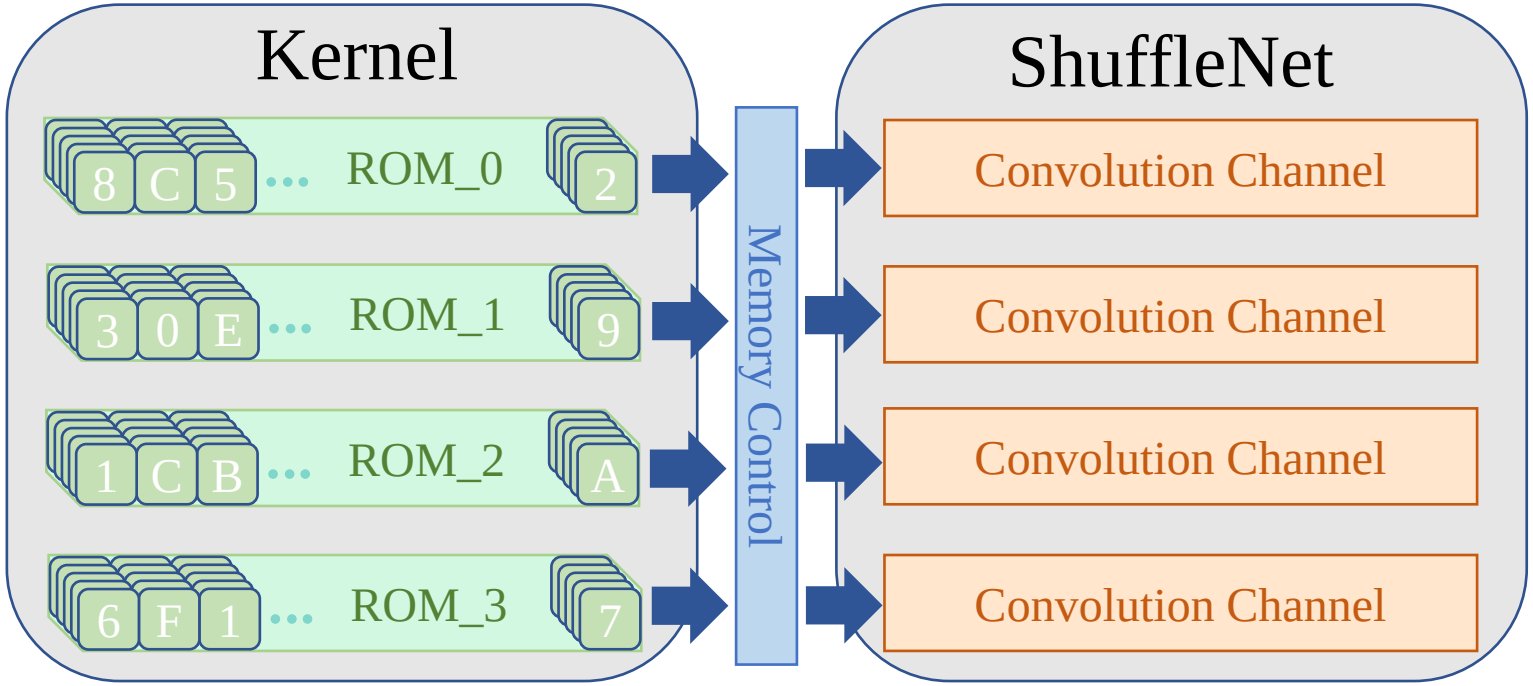
- 將MFSC所產生的頻譜資料依序在RAM中建立出頻譜圖
- 搭配ShuffleNet運算單元同步進行儲存與寫入



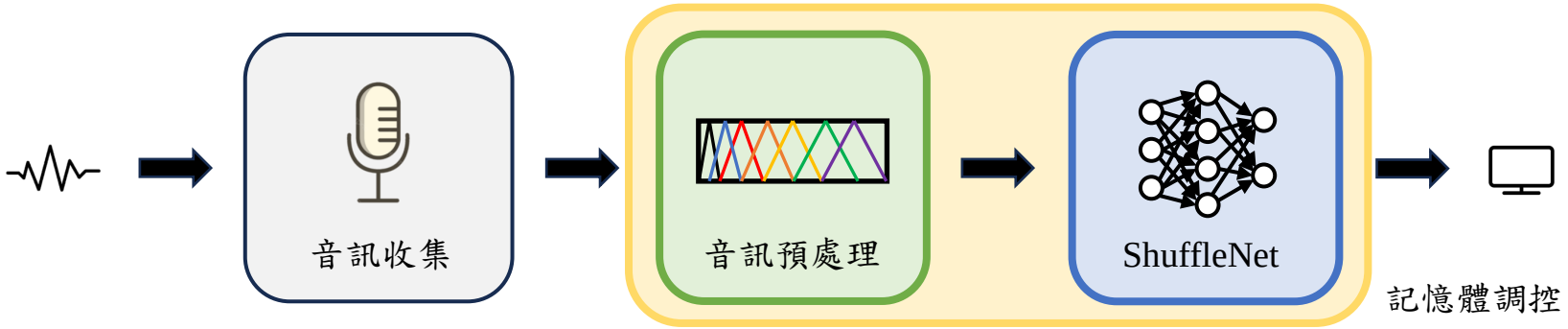
硬體開發



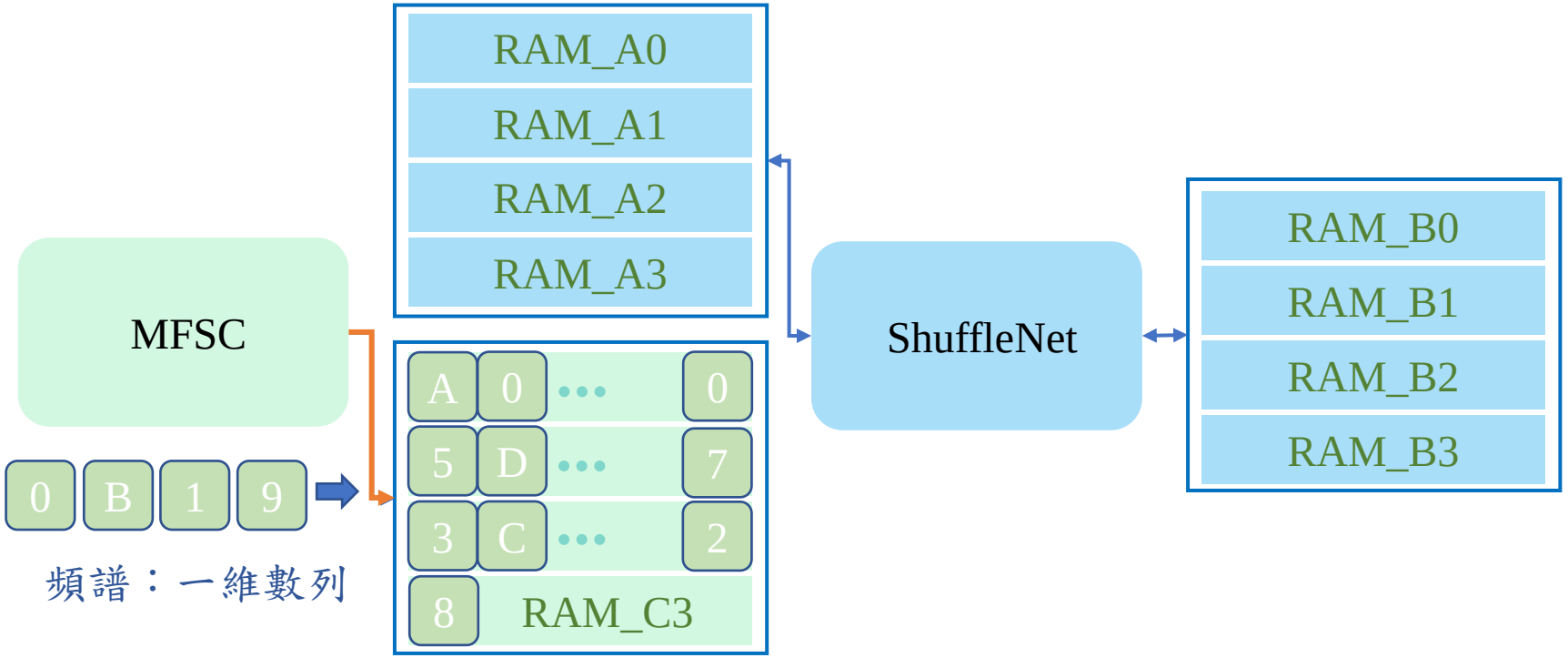
- ROM儲存ShuffleNet的Kernel參數
- 4個ROM為一組對應ShuffleNet的4個Convolution channel



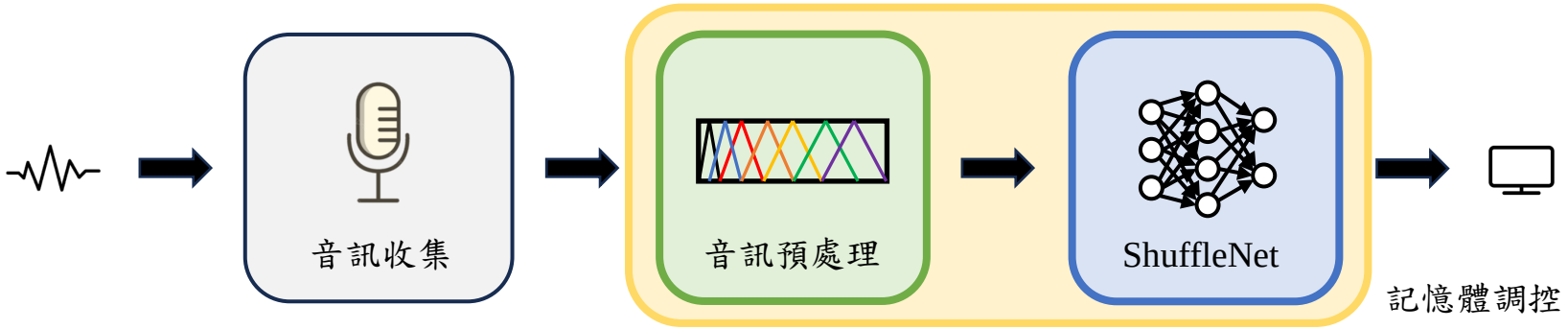
硬體開發



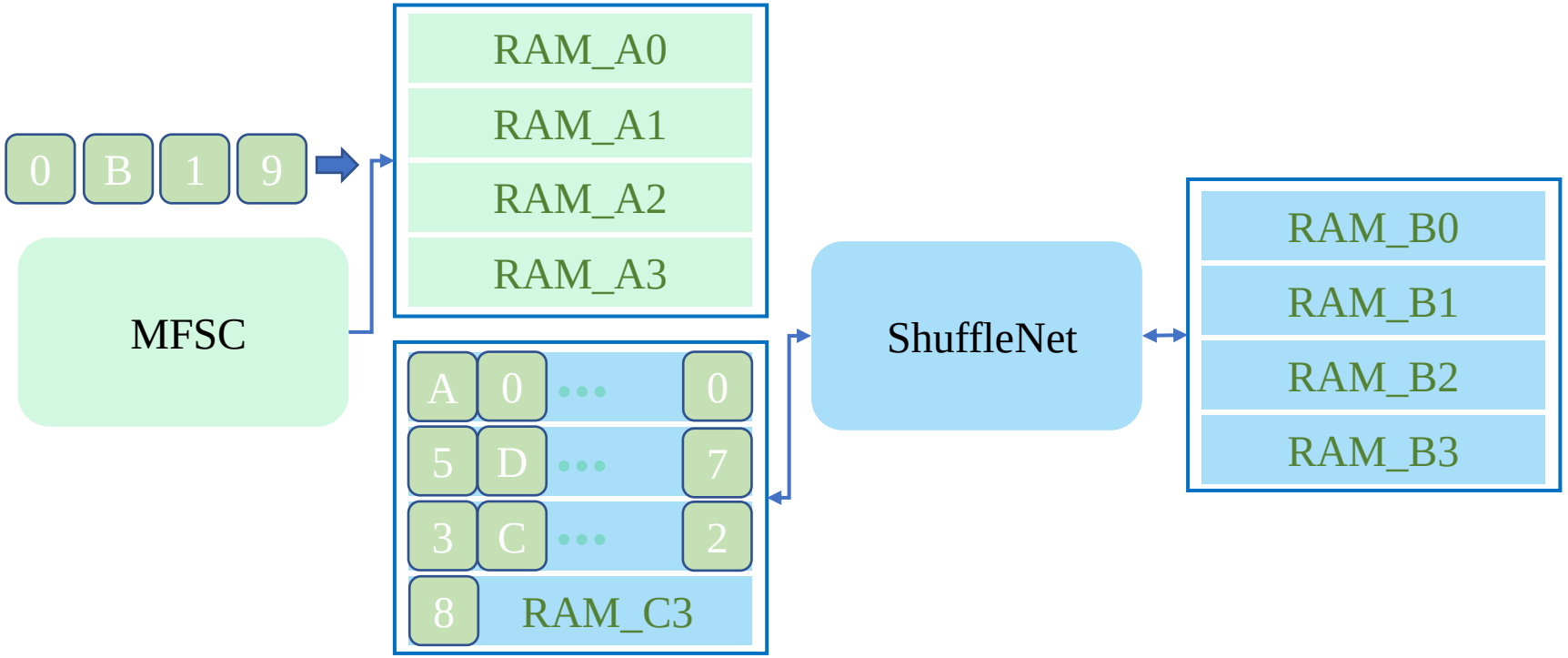
- MFSC與ShuffleNet之間為Ping-Pong-RAM設計
- ShuffleNet計算會同時運用2組RAM分別進行讀寫



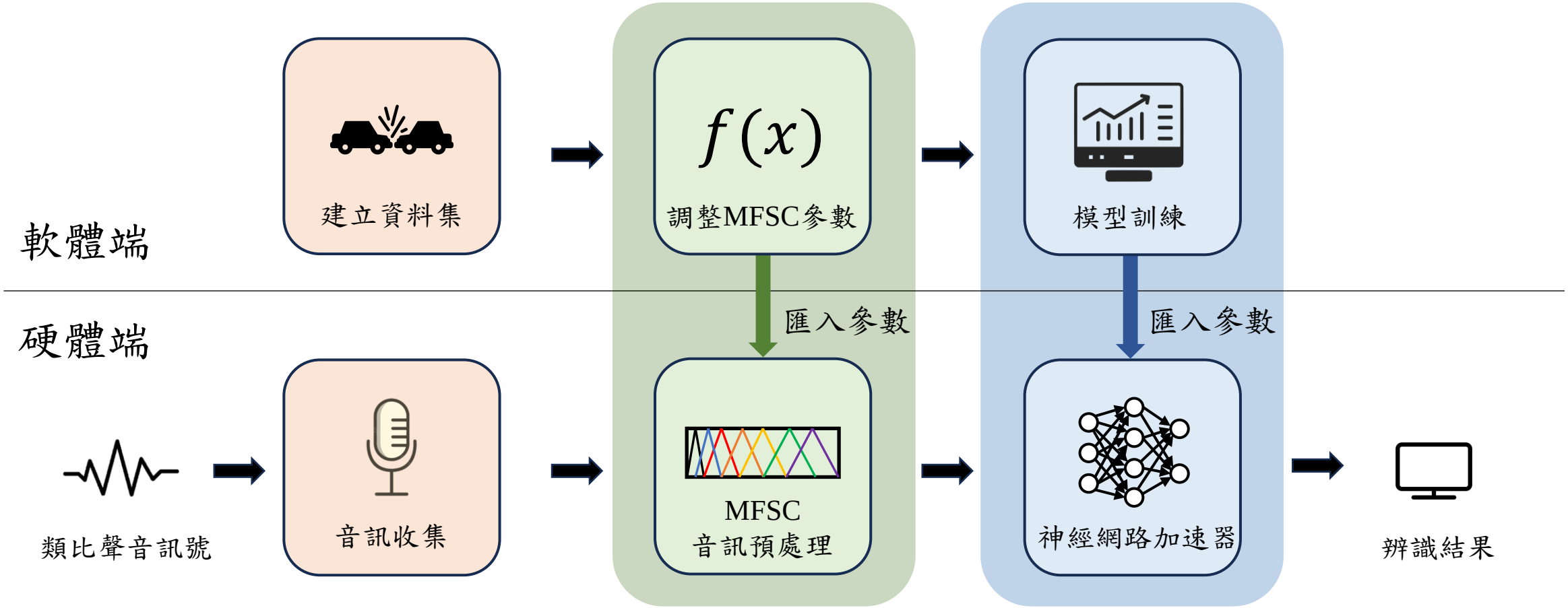
硬體開發



- MFSC與ShuffleNet之間為Ping-Pong-RAM設計
- ShuffleNet計算會同時運用2組RAM分別進行讀寫



軟硬體整合流程

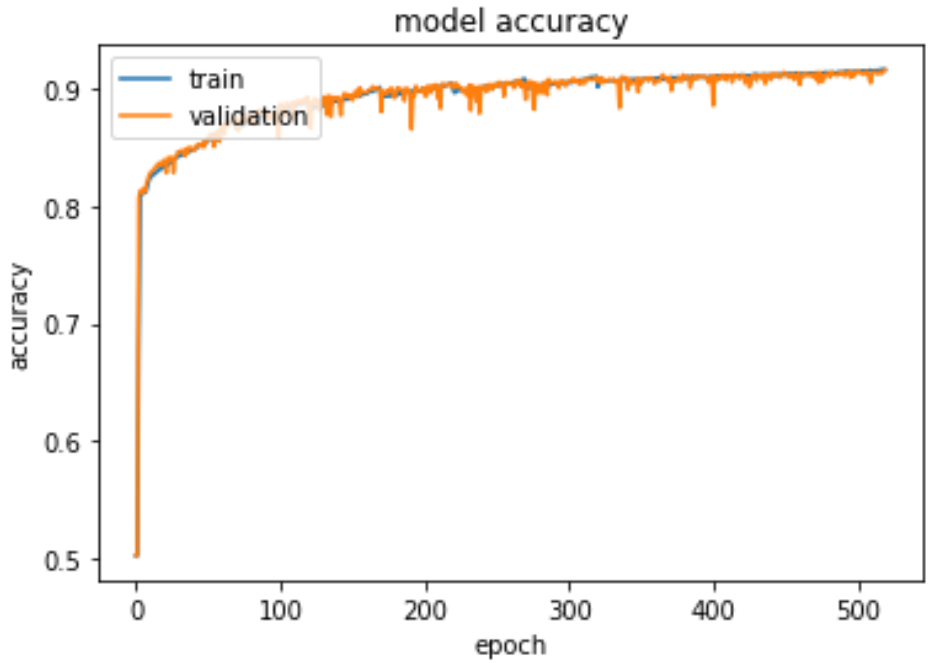

軟體端
硬體端
建立資料集
$f(x)$
調整MFSC參數
模型訓練
匯入參數
匯入參數
類比聲音訊號
音訊收集
MFSC
音訊預處理
神經網路加速器
辨識結果

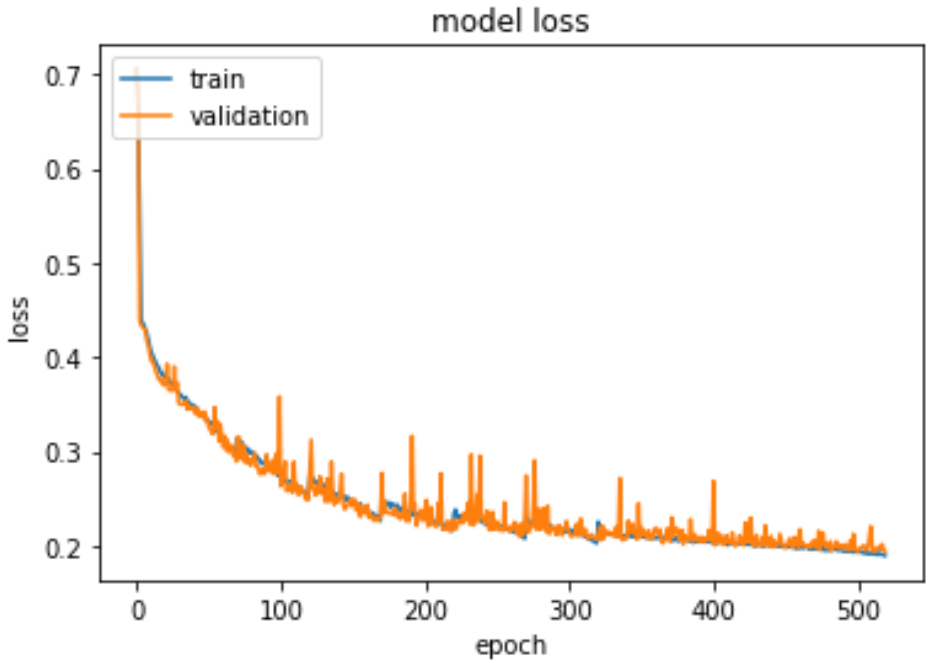
作品成果與分析

模型訓練成果、開發板資源使用量

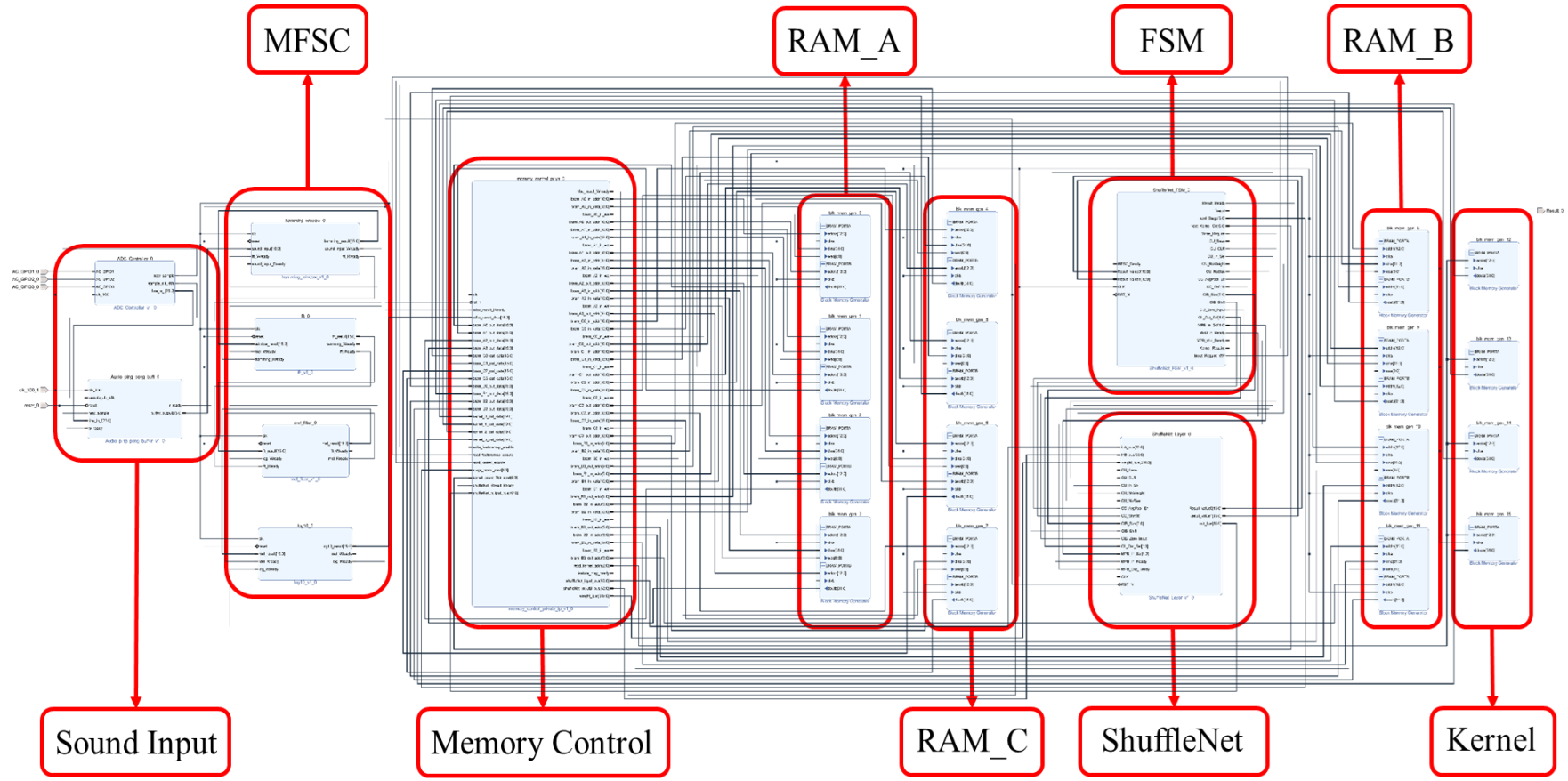
結果與分析

- 模型訓練成果：模型辨識準確度高達92.8%，Loss為0.166。





結果與分析



硬體系統方塊圖

結果與分析

- 開發板資源使用量：硬體使用率30%、降低85%記憶體用量，達成硬體輕量化。

	LUT	Flip-Flop	RAMB18	RAMB36	DSP
Sound Input + MFSC	10935	23171	8	15	3
Memory Control + ShuffleNet	5059	4504	8	12	20
Total	**15,994**	**27790**	**16**	**27**	**23**

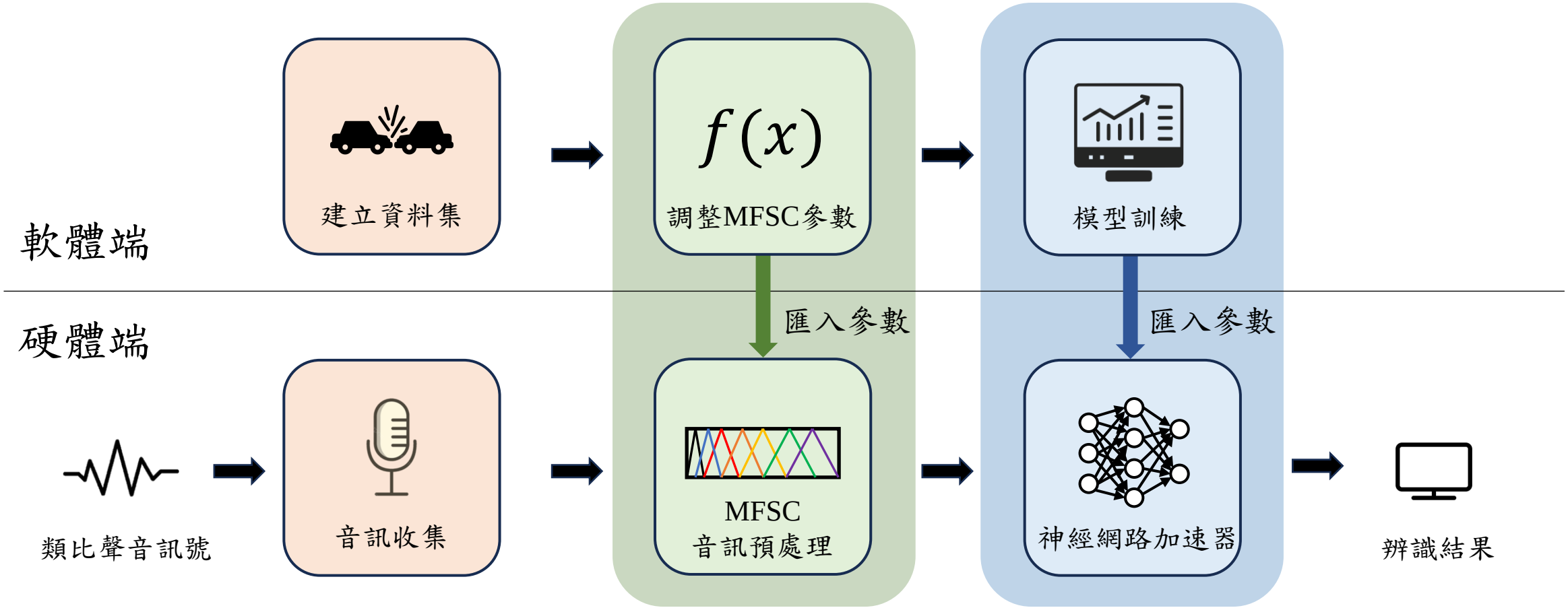
Demo


1. Busy street ambience
2. Car crash 1
3. Traffic cars pass by
4. Car crash 2
5. Human announcement
若中間綠燈亮起，
代表偵測到車禍
PYNQ-Z2
XILINX

結論

開發流程、主要貢獻

結論

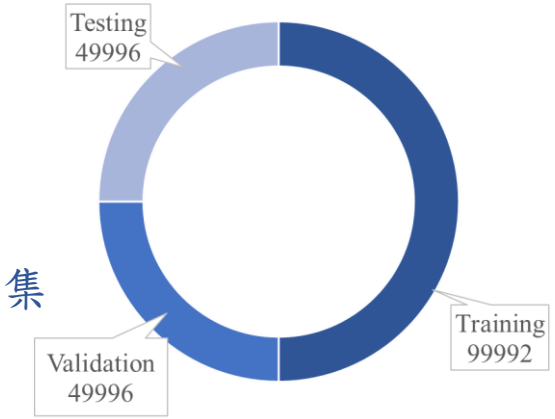

軟體端
硬體端
建立資料集
$f(x)$
調整MFSC參數
模型訓練
匯入參數
匯入參數
類比聲音訊號
音訊收集
MFSC
音訊預處理
神經網路加速器
辨識結果

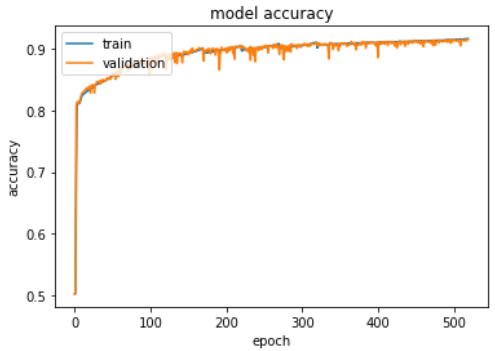
結論

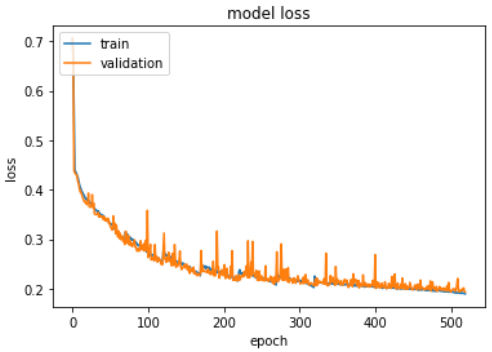
- **準確率高**：自行建立資料集訓練高準確率模型。

對抗雜音、防止誤報

自行建立資料集

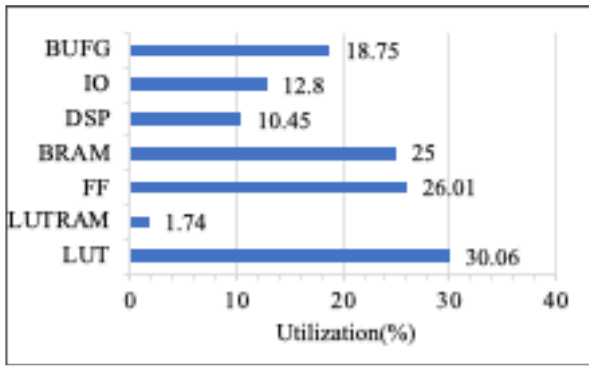


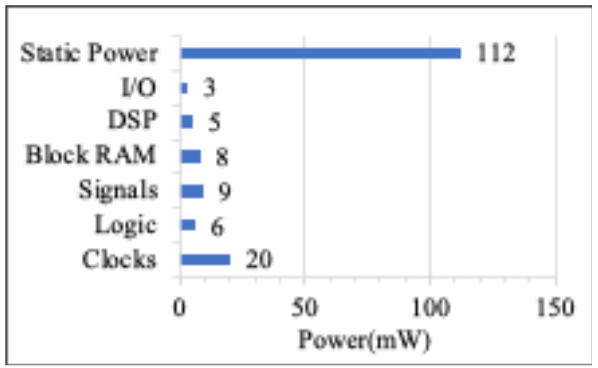




- **硬體輕量化**：自行設計開發輕量化硬體電路 & 神經網路加速器。

降低硬體用量



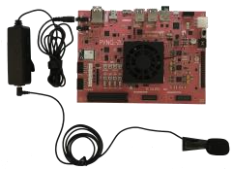


Thank You

- 圖片來源

https://www.vecteezy.com/free-vector/cctv-icon

https://www.vecteezy.com/free-vector/police-station



Why Frequency Domain?

- 有效過濾雜音、保留關鍵聲音特徵。

S. A. Nossier, J. Wall, M. Moniri, C. Glackin, and N. Cannings, "A Comparative Study of Time and Frequency Domain Approaches to Deep Learning based Speech Enhancement," presented at the University of East London, London, UK.

A Comparative Study of Time and Frequency Domain Approaches to Deep Learning based Speech Enhancement

Soha A. Nossier
Dept. of Engineering and Computing
University of East London
London, UK
soha.abdallah.nossier@gmail.com

Julie Wall
Dept. of Engineering and Computing
University of East London
London, UK
j.wall@uel.ac.uk

Mansour Moniri
Dept. of Engineering and Computing
University of East London
London, UK
m.moniri@uel.ac.uk

Cornelius Glackin
Intelligent Voice Ltd
London, UK
neil.glackin@intelligentvoice.com

Nigel Cannings
Intelligent Voice Ltd
London, UK
nigel.cannings@intelligentvoice.com

***Abstract*—Deep learning has recently made a breakthrough in the speech enhancement process. Some architectures are based on a time domain representation, while others operate in the frequency domain; however, the study and comparison of different networks working in time and frequency is not reported in the literature. In this paper, this comparison between time and frequency domain learning for five Deep Neural Network (DNN) based speech enhancement architectures is presented. The comparison covers the evaluation of the output speech using four objective evaluation metrics: PESQ, STOI, LSD, and SSNR increase. Furthermore, the complexity of the five networks**

There are many applications for speech enhancement, for example, it is an essential process in hearing aids, mobile communication systems, Automatic Speech Recognition, head phones, and VoIP communication [1].

Researchers have been developing speech enhancement techniques for decades, which predict clean speech based on statistical assumptions about the relationship between speech and noise [2]. Recently, a new era of speech enhancement has emerged with the introduction of deep learning based

Why MFSC ?

- mfsc使用更少的運算步驟 (DCT)、可節省硬體資源。
- mfsc更能保留資料的局部特性，避免持續累積資料儲存量。

O. Abdel-Hamid, A. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional Neural Networks for Speech Recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 10, pp. 1533-1545, Oct. 2014.

Convolutional Neural Networks for Speech Recognition

Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu

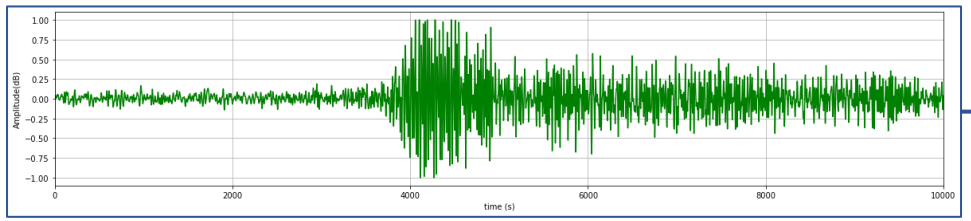
***Abstract*—Recently, the hybrid deep neural network (DNN)-hidden Markov model (HMM) has been shown to significantly improve speech recognition performance over the conventional Gaussian mixture model (GMM)-HMM. The performance improvement is partially attributed to the ability of the DNN to model complex correlations in speech features. In this paper, we show that further error rate reduction can be obtained by using convolutional neural networks (CNNs). We first present a concise description of the basic CNN and explain how it can be used for speech recognition. We further propose a limited-weight-sharing scheme that can better model speech features. The special structure such as local connectivity, weight sharing, and pooling in CNNs exhibits some degree of invariance to small shifts of speech**

for estimating the probabilistic distribution of speech signals associated with each of these HMM states. Meanwhile, the generative training methods of GMM-HMMs have been well developed for ASR based on the popular expectation maximization (EM) algorithm. In addition, a plethora of discriminative training methods, as reviewed in [1], [2], [3], are typically employed to further improve HMMs to yield the state-of-the-art ASR systems.

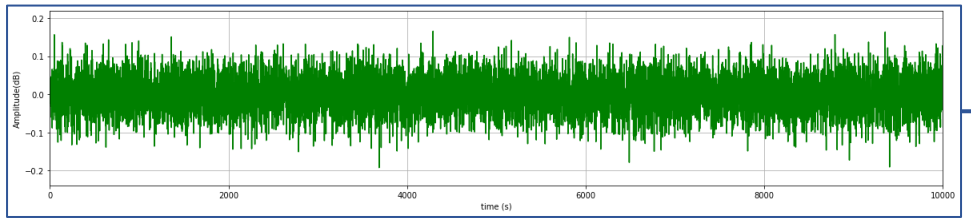
Very recently, HMM models that use artificial neural networks (ANNs) instead of GMMs have witnessed a significant resurgence of research interest [4], [5], [6], [7], [8], [9], ini-

如何增加準確率、降低 loss？

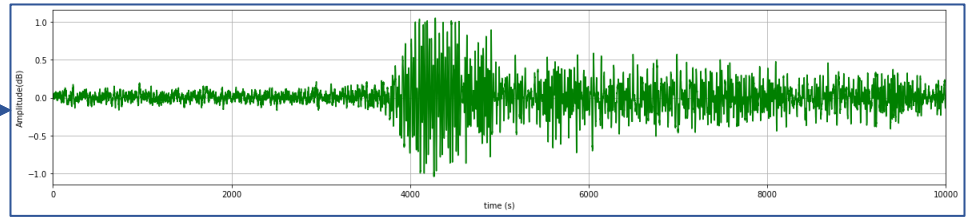
- 我們可以將資料集做data augmentation增加資料集數量
- 提高運算精度（Trade-Off: 硬體成本增加）



Original



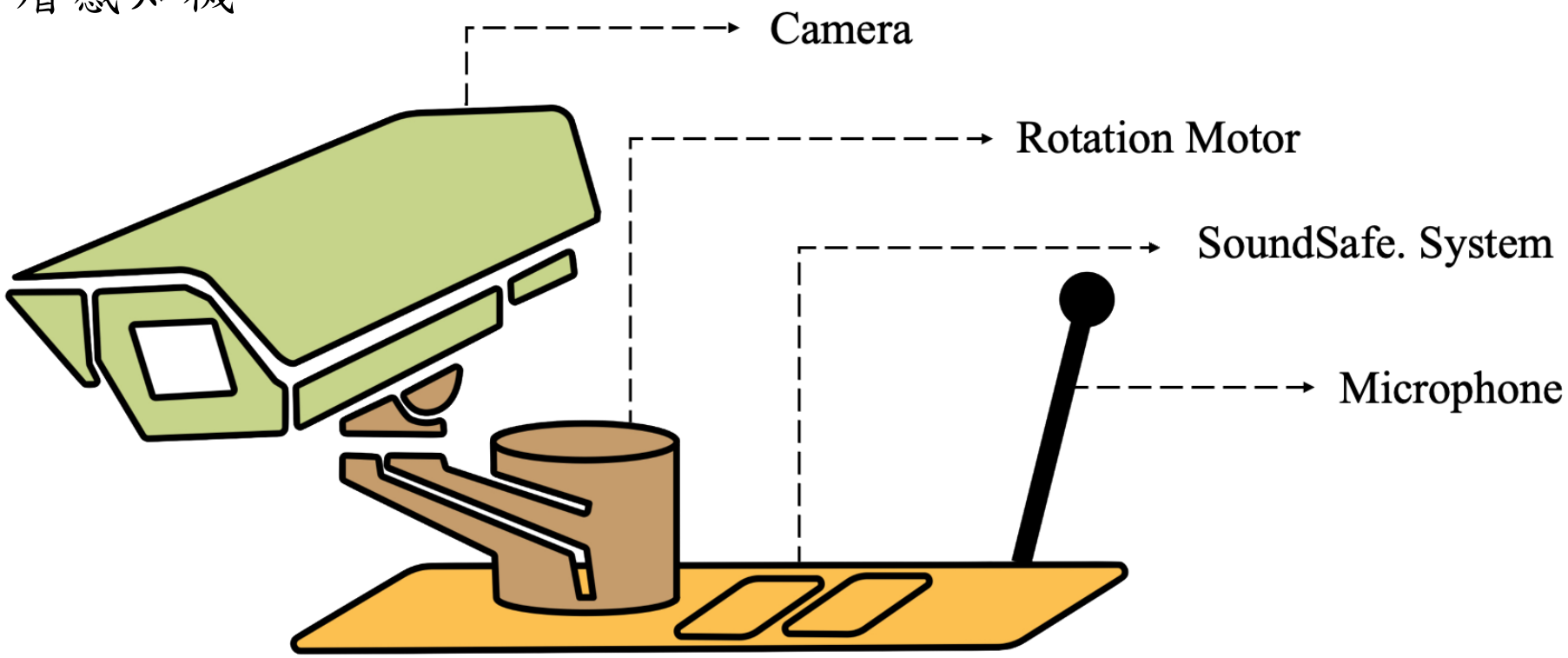
Noise



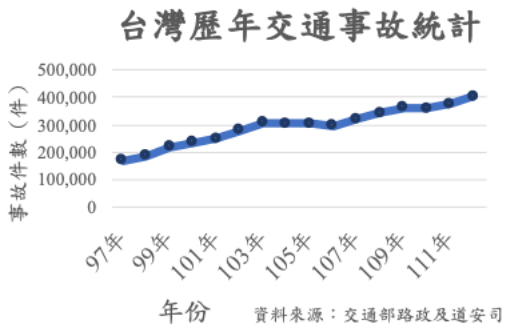
Augmented

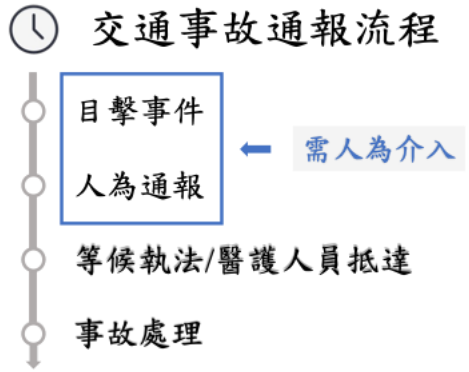
如何提高實用性？

- 結合影像辨識
- 多層感知機



開發動機



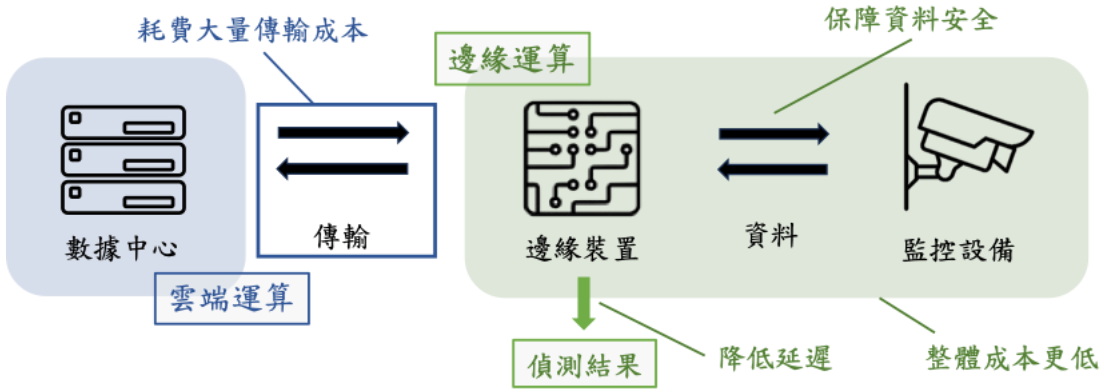


通報流程繁瑣

- 車禍數量逐年增加
- 車禍通報流程繁瑣+**人為依賴度高**

監視器死角問題

- 現有視覺辨識**車禍偵測**
- 傳統監視器存在**視線死角問題**



邊緣運算重要性

- **節省傳輸時間、成本**
- 保障資料隱私與安全性
- 降低能耗與系統搭建總成本

文獻探討—ShuffleNet

- 神經網路架構

P. Peng et al., "Design of an Efficient CNN-Based Cough Detection System on Lightweight FPGA," in IEEE Transactions on Biomedical Circuits and Systems, vol. 17, no. 1, pp. 116-128, Feb. 2023, doi: 10.1109/TBCAS.2023.3236976.
keywords: {Field programmable gate arrays;Convolutional neural networks;Hidden Markov models;Feature extraction;Deep learning;Hardware acceleration;Cough Detection;CNN;deep learning;FPGA;hardware acceleration},

Model	Parameters	No of Layers	Training time (s)	Inference time
1D ShuffleNet	9,510	42*	6,901.89	12.39
1D CNN	79,088	23*	6,733.52	16.40

* Including MaxPooling layer

Performance comparison between 1D ShuffleNet & 1D