Computational Linguistics and Chinese Language Processing
Vol. 26, No. 2, December 2021, pp. 35-48

使用低通時序列語音特徵訓練理想比率遮罩法之語音強化

Employing Low-Pass Filtered Temporal Speech Features for the Training of Ideal Ratio Mask in Speech Enhancement

陳彥同*、洪志偉*

Yan-Tong Chen and Jeih-weih Hung

摘要

在諸多基於深度學習之語音強化法中，遮罩式(masking-based)強化法求取一個遮罩與雜訊語音之時頻圖相乘、藉此使所得乘積之新時頻圖所含雜訊成分降低、以重建相對乾淨的語音訊號。在用以訓練遮罩之深度模型其輸入特徵的選取上，許多長期以來用以語音辨識的特徵、如梅爾倒倒頻譜、振幅調變時頻圖、感知線性估測係數等都是適合的選擇、可使訓練所得的遮罩達到有效的語音強化效果。另外，傳統上若將語音特徵之時序列作低通濾波處理，可以抑制雜訊所帶來的失真，因此，在本研究中，我們嘗試將各種語音特徵時序列，藉由離散小波轉換的方式加以低通濾波，再用它們來訓練語音遮罩的深度模型，探究其是否能使所學習之遮罩能對於原始雜訊語音之時頻圖有更佳的語音強化效果。在我們的初步實驗裡，在人聲雜訊環境中，我們發現上述之低通濾波所得之特徵序列、相較於原始特徵序列而言所學習而得的深度模型，能更有效地提升測試語音之品質與可讀性。

* 國立暨南國際大學電機工程學系
Department of Electrical Engineering, National Chi Nan University
E-mail: s109323508@mail1.ncnu.edu.tw; jwhung@ncnu.edu.tw

Abstract

The masking-based speech enhancement method pursues a multiplicative mask that applies to the spectrogram of input noise-corrupted utterance, and a deep neural network (DNN) is often used to learn the mask. In particular, the features commonly used for automatic speech recognition can serve as the input of the DNN to learn the well-behaved mask that significantly reduce the noise distortion of processed utterances. This study proposes to preprocess the input speech features for the ideal ratio mask (IRM)-based DNN by lowpass filtering in order to alleviate the noise components. In particular, we employ the discrete wavelet transform (DWT) to decompose the temporal speech feature sequence and scale down the detail coefficients, which correspond to the high-pass portion of the sequence. Preliminary experiments conducted on a subset of TIMIT corpus reveal that the proposed method can make the resulting IRM achieve higher speech quality and intelligibility for the babble noise-corrupted signals compared with the original IRM, indicating that the lowpass filtered temporal feature sequence can learn a superior IRM network for speech enhancement.

關鍵詞：語音強化、特徵時序列、低通濾波、理想比例遮罩法、小波轉換

Keywords: Speech Enhancement, Temporal Feature Sequence, Lowpass Filtering, Ideal Ratio Mask, Wavelet Transform

1. 緒論 (Introduction)

深度類神經模型與相關之學習演算法的高度發展，引發許多科技研究的空前突破與創新，過往的許多技術開發，常是基於解釋思維、在多次試錯之後找到一個可行方案，再對此可行方案賦予人們專業的解釋，然而深度學習則普遍基於統計思維、並不著重於方法在解釋上的合理性，而是嘗試將大量觀察（輸入）和對應結果（輸出）的關聯性藉由深度類神經網路加以詮釋，以期對於新的觀察能精準預測出對應的結果。

在語音處理的領域中，近年來基於深度學習所開發出的演算法也琳瑯滿目，且因訓練資料的可取得性越來越高，這些演算法在學習與預測結果的能力也隨之增強。以本研究著重的語音強化法為例，基於深度類神經模型之各式語音強化架構其表現常超越經典且富有高度理論根據的演算法，或是以後者的演算法的原型 (prototype) 出發，但配合深度類神經網路來有效學習該演算法的各項參數，使其語音強化效果更佳。

根據文獻(Wang *et al*., 2014)，許多基於深度學習之語音強化法根據其訓練目標大致可以分為兩大範疇：對映式 (mapping) 與遮罩式 (masking)，前者直接求取一個對映函數，使此對映函數之理想輸出為乾淨語音的呈現式(特徵)，如時域訊號波形、時頻圖 (spectrogram) 或耳蝸時頻譜圖 (cochleagram)，後者是求取一個遮罩 (mask)，用以與原始輸入訊號或特徵呈現作點對點的相乘，使相乘後的訊號呈現式能趨近乾淨時的狀態。簡

單來說對映式所求取的函數，對於輸入訊號特徵的運算可以是任意由所使用之深度學習模型定義的非線性運算，而遮罩式所求取的函數運算，則簡化或限制為對輸入訊號特徵作乘法 (即加權運算)。二者各擅勝場，但近年來似乎是以遮罩式的語音強化更受重視與發展，相關的演算法包括了理想二元遮罩 (ideal binary mask, IBM) (Wang, 2005; Srinivasan *et al*., 2006)、理想比例遮罩 (ideal ratio mask, IRM) (Srinivasan *et al*., 2006)、頻譜強度遮罩 (spectral magnitude mask, SMM) (Wang *et al*., 2014)、複數理想比例遮罩 (complex ideal ratio mask, cIRM) (Williamson *et al*., 2016)、相位敏感型遮罩 (phase-sensitive mask, PSM) (Erdogan *et al*., 2015) 等。

在本研究中，主要是針對上述之遮罩式語音強化法加以改進，我們提出對於訓練遮罩模型的輸入雜訊語音的特徵時序列作簡單的預處理 (pre-processing)，使其包含的雜訊失真較低，以期在之後的訓練遮罩步驟能更加精確。而使用的預處理方法，是透過簡易的一階離散小波轉換 (discrete wavelet transform, DWT) (Mallat, 1999)]，將特徵時序列分為高低兩調變頻帶 (modulation frequency bands)，然後藉由一權重的相乘來降低高調變頻帶之序列的振幅，再將其與原始低調變頻帶序列搭配、透過一階反離散小波轉換 (inverse discrete wavelet transform, IDWT) 重建特徵序列，再使用此相當於透過低通濾波處理後的特徵序列來訓練遮罩模型。

上述低通濾波之處理，主要是基於先前諸多學者所提出的觀察(Kanedera *et al*., 1997; Chen & Bilmes, 2007)：乾淨語音特徵時序列主要分布頻率在 1 Hz 至 16 Hz 之間，以一般的音框取樣率 100 Hz 而言，特徵序列可包含的（調變）頻帶為[0,50 Hz]，因此後半頻帶鮮少包含語音成分，抑制此頻帶不會對語音造成明顯失真，但可有效抑制雜訊的干擾。

另外，基於文獻(Wang *et al*., 2018)所述，使用小波轉換分解語音特徵時序列、消除其細節係數 (detail coefficients，相當於調變高頻成分) 後重建之語音特徵，在雜訊環境下有明顯進步的語音辨識率，我們參照這樣的做法來實現前述之語音特徵序列的低通濾波處理，期許它對應的遮罩深度模型能得到更佳的語音強化效果。

2. 提出的新方法 (Proposed Method)

在本研究中，我們選擇加以研究改進的是理想比例遮罩(ideal ratio mask, IRM)法，此法通常是求取語音之一般時頻圖 (spectrogram) 或耳蝸時頻圖 (cochleagram) 對應的理想遮罩值：

$$M(m,f)=\frac{|s(m,f)|^2}{|s(m,f)|^2+|d(m,f)|^2}, \tag{1}$$

其中，$|s(m,f)|^2$與$|d(m,f)|^2$分別代表了雜訊語音其時頻圖或耳蝸時頻圖在音框時間m與頻率f之時頻單位(time-frequency unit, T-F unit) 所對應的乾淨語音與純雜訊的能量，在人造訓練雜訊語句的準備上，由於事先可得知其乾淨語音及純雜訊的成分，因此可根據式(1)計算其理想比例遮罩的值，作為 IRM 深度模型的訓練目標。

在我們構思的新方法中，嘗試將用以訓練 IRM 深度模型所使用的語音特徵時序列，

加以低通濾波處理、藉此抑制其調變高頻的成分，再使用處理後的語音特徵來求取 IRM 深度模型，預期此 IRM 模型相對於原始特徵對應之 IRM 模型，能求取更佳的遮罩來抑制雜訊對語音時頻圖上的失真。

值得一提的是，我們使用離散小波轉換 (discrete wavelet transform, DWT) (Mallat, 1999; Wang *et al*., 2018)來執行上述的低通濾波處理，部分原因是在 DWT 其分解與重建的濾波器彼此互補，在分解與重建的過程中不會造成序列相位的失真，此相較於一般的低通濾波器而言存在優勢。

以下，我們敘述此新方法的步驟：

訓練階段：

步驟一：將訓練集 (training set) 中的任一雜訊干擾的語音$x[n]$，經音框化 (framing) 與窗化 (windowing) 切割成個別音框訊號$x_m[n]$後（m為音框索引），再將個別音框訊號轉換成語音特徵，如 amplitude modulation spectrogram (簡稱 AMS), relative spectral transformed perceptual linear prediction coefficients (簡稱 RASTA-PLP), mel-frequency cepstral coefficients (簡稱 MFCC) 及 Gammatone filterbank power spectra (簡稱 GF) 等。我們將對應的D維語音特徵向量以$\mathbf{x}_m$表示，$\mathbf{x}_m$為一$D \times 1$ 的行向量，假設該語句共切割成M個音框，則其對應的語音特徵矩陣可表示為：

$$\mathrm{X} = [\mathbf{x}_0 \quad \mathbf{x}_1 \quad \cdots \quad \mathbf{x}_{M-1}] \tag{2}$$

其尺寸為$D \times M$。

步驟二：上述之特徵矩陣X的任一第d個橫列向量

$$[\mathrm{X}_{d,0}\, \mathrm{X}_{d,1} \ldots \mathrm{X}_{d,M-1}] \tag{3}$$

以$\mathrm{X}_d[m]$代表之，其稱作X的第d維特徵時序列，尺寸為$1 \times M$，其中$1 \le d \le D$。

我們將任一維特徵時序列$\mathrm{X}_d[m]$以一階離散小波轉換加以分解如下：

$$[\mathrm{cA}_d[m], \mathrm{cD}_d[m]] = \mathbf{DWT}(\mathrm{X}_d[m]) \tag{4}$$

其中**DWT**(.) 代表離散小波轉換 (discrete wavelet transform)、 $\mathrm{cA}_d[m]$與 $\mathrm{cD}_d[m]$分別為轉換分解而得的近似係數(approximation coefficients)與細節係數(detail coefficients)，其可視為原始序列$\mathrm{X}_d[m]$之低通成分與高通成分，二者頻寬均約等於原始序列頻寬的一半，且點數減半。

步驟三：我們將上一步驟所得的細節係數$\mathrm{cD}_d[m]$]乘上一個小於 1 的權重α，再與原近似係數相組合、經過反離散小波轉換重建第d維特徵時序列，表示如下：

$$\widetilde{\mathrm{X}}_d[m] = \mathbf{IDWT}([\mathrm{cA}_d[m], \alpha \times \mathrm{cD}_d[m]]) \tag{5}$$

其中$\widetilde{\mathrm{X}}_d[m]$為更新的特徵時序列，相較於原始特徵時序列$\mathrm{X}_d[m]$，$\widetilde{\mathrm{X}}_d[m]$包含較低的高通成分，因此應當包含較少雜訊造成的失真。

步驟四：參照一般 IRM 深度模型的訓練法，我們改以新的特徵序列$\{\tilde{\mathrm{X}}_d[m], 1 \le d \le D\}$作為輸入，以理想 IRM 遮罩值為目標輸出，訓練 IRM 深度模型。值得注意的是，若式(4)中的權重$\alpha = 1$，則所訓練的 IRM 模型與原始（即使用原始特徵訓練）IRM 模型完全一致。

測試階段：

將測試之語句如同訓練語句之處理的前三個步驟、求取低通濾波之特徵時序列，將其通過訓練完成的 IRM 模型求取遮罩值，將遮罩值與原設定之對應的時頻圖作點乘積 (dot product)，即可得強化後的時頻圖，經由適當的反轉換重建成強化版的時域訊號。

3. 實驗設置 (Experimental Setup)

參照文獻(Wang *et al*., 2014)所提供的程式碼[1]，我們使用了 TIMIT 資料庫的部分語句 (取樣頻率為 16 kHz) 來實驗評估我們所提出的方法，其中，訓練集包含了 5 位語者、每人 10 句共 50 個語句，而測試集則包含了與訓練集不同的 3 位語者、每人 10 句共 30 個語句。我們將訓練與測試語句摻入 babble 雜訊，訓雜比 (signal-to-noise ratio, SNR) 固定為 -2 dB。在訓練與測試 IRM 之深度模型上，輸入特徵的種類包含了 AMS、RASTA-PLP、MFCC 與 GF 四種，同時，我們將左右相鄰的 5 個音框 (frames) 串接成一個長向量，作為深度模型的輸入單位，深度模型之架構為全連結層(densely connected layers)網路，共包含 4 層隱藏層，每個隱藏層由 1024 個神經元 (neurons) 構成。目標是求取語音之耳蝸時頻圖 (cochleagram) 的遮罩，其每個音框設有 64 維，相當有 64 個通道(channel)。

在我們所提的新 IRM 訓練法上，對於輸入特徵之時序列之細節係數（高頻係數）所給予的權重α，分別設定為 0, 0.25, 0.50, 0.75，藉此觀察細節係數之壓抑程度對於 IRM 效果之影響（原始 IRM 所對應之權重$\alpha = 1$）。

在使用的離散小波轉換與反轉換中，我們使用 db2 小波函數。

在評估效能上，我們使用了 PESQ 分數(Rix *et al*., 2001)作為語音品質 (quality) 的客觀指標、STOI 分數(Taal *et al*., 2011)作為語音可讀性(intelligibility) 的客觀指標，PESQ 分數介於-0.5 與 4.5 之間， STOI 分數介於 0 與 1 之間，分數越高代表語音的品質/可讀性越佳。

4. 實驗結果與討論 (Experimental Results and Discussions)

在我們的評估實驗上，我們將分為三部分來呈現並討論，第一部分是對應於使用所有種類之輸入特徵組合所訓練及測試之 IRM 模型，第二部分是對應於使用單一種類之輸入特徵所訓練及測試之 IRM 模型，我們將在這兩部分中，探究所提新方法之低通濾波特徵時序列對於 IRM 效能的改變，第三部分則是藉由時頻圖的展示，觀察原始與更新之 IRM 所

[1] Matlab toolbox for DNN based speech separation .Retrieved from http://web.cse.ohio-state.edu/pnl/DNN_toolbox/

強化的語音訊號的差異。

4.1 使用所有種類之輸入特徵所得的IRM效能分析 (The IRM Results and Analyses for the Case using all kinds of Features)

首先，表 1 列出了測試雜訊語句在處理前、經由理想 IRM（遮罩直接由乾淨語音與摻雜之雜訊求得）及原始 IRM（使用原始輸入特徵訓練，並可能額外加入差量特徵）處理後所對應的 PESQ 與 STOI 的平均值。從此表中，我們可以看到：

1. 雜訊語句經過理想 IRM 處理後，在 PESQ 與 STOI 都得到了大幅的提升。
2. 原始 IRM 雖然也能帶來顯著的改進，但效果明顯與理想 IRM 有差距，這代表了藉由雜訊語音（特徵）中估測乾淨語音與雜訊成分之精準度仍有很大的進步空間。
3. 差量特徵的有無並未對於訓練而得 IRM 在 STOI 與 PESQ 的表現上有大幅影響，額外使用差量特徵甚至使 IRM 得到較低的 STOI 分數。

表 1.未處理語音與經過理想 IRM、原始 IRM_1（使用原特徵求取）、原始 IRM_2（使用原特徵與其差量特徵求取）處理後對應的 STOI 與 PESQ 平均分數，原特徵由四種特徵 (AMS, RASTA-PLP, MFCC, GF) 排列而得
[Table 1. The PESQ and STOI results for the baseline, oracle IRM, original IRM_1 (using the original combo static features) and IRM_2 (using the original combo static and delta features)]

	未處理語音	理想 IRM	原始 IRM_1	原始 IRM_2
STOI	0.6130	0.9004	0.6763	0.6658
PESQ	1.6081	2.6408	1.7755	1.7748

接下來，我們開始評估所提之新 IRM 訓練法，表 2 列出了在不使用差量特徵時，給定輸入特徵之時序列之高頻係數不同的權重α，經訓練之 IRM 所對應的 STOI 與 PESQ 分數，從此表中，我們有以下的發現：

1. 當使用我們提出之抑制調變高頻的特徵法時，多數α權重設定都得到了更佳的 STOI 與 PESQ 值（$\alpha = 0.25$在 STOI 分數除外，$\alpha = 0.25, 0.50$ 在 PESQ 分數除外），此初步驗證了此方法對於訓練更佳 IRM 模型、以抑制雜訊干擾有更好的效果。
2. 全然移除（設定 $\alpha = 0$）或少量移除（設定 $\alpha = 0.75$）調變高頻成分似乎是較佳選項，二者至少皆可使 PESQ 與 STOI 值提升，$\alpha = 0$得到最佳的 PESQ 值，而$\alpha = 0.75$則使 STOI 進步最大。

表 2. 未處理語音與經過理想 IRM、原始 IRM_1（使用原特徵求取）、不同權重 α 抑制調變高頻之 IRM（未搭配差量特徵）處理後對應的 STOI 與 PESQ 平均分數。原特徵由四種特徵 (AMS, RASTA-PLP, MFCC, GF) 排列而得
[Table 2. The PESQ and STOI results for the baseline, oracle IRM, original IRM_1 (using the original combo static features) and the lowpass-filtered IRM_1 (using the lowpass filtered combo static features with different assignments of parameter α)]

	原始 IRM_1	不同權重α抑制調變高頻之 IRM_1			
		0	0.25	0.50	0.75
STOI	0.6763	0.6767	0.6728	**0.6799**	0.6789
PESQ	1.7755	**1.7844**	1.7612	1.7717	1.7760

其次，表 3 列出了在額外使用差量特徵時，給定輸入特徵之時序列之高頻係數不同的權重α，經訓練之 IRM 所對應的 STOI 與 PESQ 分數，從此表中，我們有以下的發現：

1. 相較於原始 IRM 而言，使用較大權重α (0.75) 在 STOI 與 PESQ 上都有較明顯的改進，其他較小值的α設定值則並未一制性地得到明顯進步的效果，這可能原因是，當使用差量特徵時，差量特徵本身就已經抑制原始特徵的調變高頻成分，因此此時用較大的α值再對原始特徵的調變高頻成分小幅抑制，即可達到預期之進步效果。
2. 若我們將表 2 與表 3 的數據同時比較，發現達到最佳 STOI 值 (0.6799) 的是「不使用差量特徵、使用$\alpha = 0.50$ 之抑制調變高頻」的 IRM 法，而達到最佳 PESQ 值 (1.7996) 的則是「使用差量特徵、使用$\alpha = 0.75$ 之抑制調變高頻」的 IRM 法。

表 3. 未處理語音與經過理想 IRM、原始 IRM_2（使用原特徵與其差量特徵求取）、不同權重 α 抑制調變高頻之 IRM（有搭配差量特徵）處理後對應的 STOI 與 PESQ 平均分數。原特徵由四種特徵 (AMS, RASTA-PLP, MFCC, GF) 排列而得
[Table 3. The PESQ and STOI results for the baseline, oracle IRM, original IRM_2 (using the original combo static and delta features) and the lowpass filtered IRM_2 (using the lowpass filtered combo static and delta features with different assignments of parameter α)]

	原始 IRM_2	不同權重α抑制調變高頻之 IRM_2			
		0	0.25	0.50	0.75
STOI	0.6658	0.6639	0.6671	0.6615	0.6682
PESQ	1.7748	1.7819	1.7916	1.7589	1.7996

4.2 The IRM Results and Analyses for the Case using each Individual Kind of Features

在前一節中，我們已經呈現綜合四類特徵所得之 IRM 的效果，並初步驗證將特徵時序列低通濾波可以進一步強化 IRM。在本節裡，我們想進一步觀察各個類別的特徵（包含 AMS, RASTA-PLP, MFCC, GF）對於 IRM 效能之影響，同時我們也使用低通濾波來處理其序

列、進而比較濾波前與濾波後對於 IRM 效能的影響，表 4 與表 5 分別列出各種不同特徵搭配低通濾波對應之 IRM 所得之測試語句的 STOI 與 PESQ 分數，為了使整體效能優化起見，這裡我們把差量特徵一併加入，同時，我們將前一節四類特徵的組合（以"combo"表示）之結果列在表的最下一列，以供比較。從這兩個表之數據，我們有以下幾點的觀察與討論：

1. 對於語音可讀度指標 STOI 而言，不使用低通濾波之四類特徵中，以 MFCC 表現最佳（0.6740），甚至超越了組合特徵的結果（0.6658），然而，當配合低通濾波時，MFCC 可以達到更佳的 STOI 值，例如當使用$\alpha = 0.25$的權重時，MFCC 對應之 STOI 值可以進一步提升至 0.6772。此外，低通濾波處理並非對每一種特徵都能帶來改進，例如對於 AMS 特徵而言，不使用低通濾波所對應的原始 IRM 表現最好。
2. 對於語音品質指標 PESQ 而言，在不使用低通濾波之四類特徵中，MFCC 仍表現最佳（1.7966），超越了組合特徵（1.7748），而 AMS 特徵表現較不好，只有 1.6721 之 PESQ 值。然而，當配合低通濾波時，各種類特徵皆可以達到更佳的 PESQ 值，例如當使用$\alpha = 0.75$的權重時，MFCC 對應之 PESQ 值可以進一步提升至 1.7977。然而，獲得 PESQ 最佳之特徵是組合特徵配合$\alpha = 0.75$之低通濾波法，可達到 1.7996。

根據以上觀察，四類特徵的組合未必在 STOI 表現上優於單類特徵，而在 PESQ 表現上只能些許超越個別單類特徵，這可能原因在於某類特徵（如 AMS）在表現上與其他特徵差異較大，即使後端的深度模型在學習中理應能淡化這類特徵的負面影響，但是從測試結果上，多類特徵的組合並未發揮顯著的加成性。

表 4. 單一種類特徵的 STOI 分數比較，未處理語音與經過原始 IRM_2（使用原特徵與其差量特徵求取）、不同權重 α 抑制調變高頻之 IRM（有搭配差量特徵）處理後對應的 STOI 平均分數，其中"combo"表示四類特徵之組合

[Table 4. The averaged STOI results for the original IRM_2 (using the original static and delta features of single type) and the lowpass filtered IRM_2 (using the lowpass filtered static and delta features of single type with different assignments of parameter α)]

STOI 分數	原始 IRM_2	不同權重α抑制調變高頻之 IRM_2			
		0	0.25	0.50	0.75
AMS	**0.6472**	0.6430	0.6435	0.6458	0.6466
RASTAPLP	0.6559	0.6600	0.6607	**0.6611**	0.6556
MFCC	0.6740	0.6771	**0.6772**	0.6761	0.6770
GF	0.6695	**0.6698**	0.6667	0.6672	0.6692
combo	0.6658	0.6639	0.6671	0.6615	**0.6682**

表 5. 單一種類特徵的 PESQ 分數比較，未處理語音與經過原始 IRM_2（使用原特徵與其差量特徵求取）、不同權重 α 抑制調變高頻之 IRM（有搭配差量特徵）處理後對應的 PESQ 平均分數，其中"combo"表示四類特徵之組合
[Table 5. The averaged PESQ results for the original IRM_2 (using the original static and delta features of single type) and the lowpass filtered IRM_2 (using the lowpass filtered static and delta features of single type with different assignments of parameter α)]

PESQ 分數	原始 IRM_2	不同權重α抑制調變高頻之 IRM_2			
		0	0.25	0.50	0.75
AMS	1.6721	1.6705	1.6712	**1.6731**	1.6758
RASTA-PLP	1.7463	**1.7634**	**1.7634**	1.7630	1.7426
MFCC	1.7966	1.7870	1.7916	1.7946	**1.7977**
GF	1.7641	**1.7791**	1.7669	1.7635	1.7633
combo	1.7748	1.7819	1.7916	1.7589	**1.7996**

4.3 增加訓練及測試資料且使用單一種類之輸入特徵所得的IRM效能分析 (The IRM Results and Analyses for the Case using a Single Feature with More Training and Test Data)

在前一節中，我們可觀察出在各個類別的特徵中，單獨使用 MFCC 特徵的 IRM 效能明顯優於其他特徵，其同時使用低通濾波與差量特徵處理其序列可得到較佳的 STOI ($\alpha = 0.25$) 與 PESQ ($\alpha = 0.75$) 分數。在本節裡，我們想進一步觀察此表現良好的的 MFCC 特徵，若再增加 1 倍的資料數量 (其中，訓練集包含了 10 位語者、每人 10 句共 100 個語句，而測試集則包含了與訓練集不同的 6 位語者、每人 10 句共 60 個語句) 的情況下，其 IRM 的效能，同時觀察在使用我們所提出的低通濾波法對於 MFCC 特徵在此狀態下之 IRM 效能的影響，這一系列實驗結果分別列在表 6（無差量特徵）與表 7（有差量特徵）。

從表 6 與表 7 我們可以觀察出以下幾點：

1. 把表 6、7 與表 4、5 的數據相比較，我們可以看到增加訓練資料量可以同時使測試資料的 PESQ 與 STOI 的分數都明顯進步，進而驗證訓練資料的增加可以使 IRM 模型在語音強化的效果更好。
2. **當沒有使用差量特徵時**，若增加訓練語料，在 STOI 分數上，原始的 IRM 比使用低通濾波法對應的 IRM 效果較佳，代表此時低通率波處理並未帶來 STOI 分數的進步，然而在 PESQ 分數上，當配合低通濾波時，可以比原始 IRM 達到更佳的結果，例如當使用$\alpha = 0.75$的權重時，MFCC 對應之 PESQ 值可以進一步提升至 1.8192。然而，獲得 PESQ 最佳權重是$\alpha = 0$之低通濾波法，可達到 1.8214。
3. **當使用差量特徵時**，上一點的結果則剛好對調：即若增加訓練語料，在 PESQ 分數上，

原始的 IRM 比使用低通濾波法對應的 IRM 效果較佳，而在 STOI 分數上，當配合低通濾波時，可以比原始 IRM 達到更佳的結果，例如當使用$\alpha = 0.5$的權重時，MFCC 對應之 STOI 值可以進一步提升至 0.6880。然而，獲得 PESQ 最佳權重是$\alpha = 0$之低通濾波法，可達到 1.8214。

4. 當比較表 6 與表 7 的數據，我們可以清楚看到，額外使用差量特徵反而同時使 PESQ 與 STOI 的分數都降低，這結果似乎表明，在訓練資料增加時，差量特徵的參與並未對於 IRM 模型之訓練有正面的影響，這背後原因可能是此時 IRM 模型之複雜度應該進一步提高、以因應額外的差量特徵帶來的資料多樣性。如果在原始 IRM 模型架構的設定下，不使用差量特徵可能是較佳的選擇，同時配合低通濾波處理，可使 PESQ 分數進一步提升。

表 6. 未處理語音與經過原始 IRM_1（使用原 MFCC 特徵求取）、不同權重 α 抑制調變高頻之 IRM_1 處理後對應的 STOI 與 PESQ 平均分數。原特徵由單一特徵 MFCC 而得

[Table 6. The averaged PESQ and STOI results for the original IRM_1 (using the original static MFCC features) and the lowpass filtered IRM_1 (using the lowpass filtered static MFCC features with different assignments of parameter α)]

MFCC 特徵	原始 IRM_1	不同權重α抑制調變高頻之 IRM_1			
		0	0.25	0.50	0.75
STOI	**0.6947**	0.6900	0.6926	0.6918	0.6928
PESQ	1.8182	**1.8214**	1.7996	1.8056	**1.8192**

表 7. 未處理語音與經過原始 IRM_2（使用原 MFCC 特徵與其差量特徵求取）、不同權重 α 抑制調變高頻之 IRM_2（有搭配差量特徵）處理後對應的 STOI 與 PESQ 平均分數。原特徵由單一特徵 MFCC 而得

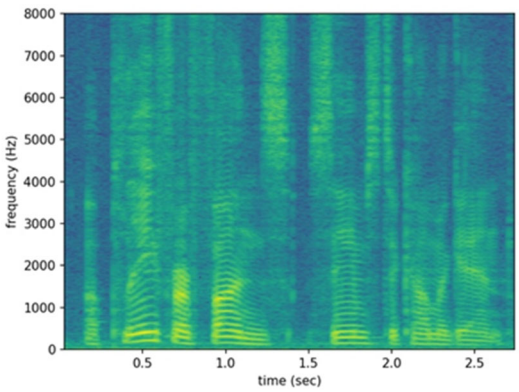
[Table 7. The averaged PESQ and STOI results for the original IRM_2 (using the original static and delta MFCC features) and the lowpass filtered IRM_2 (using the lowpass filtered static and delta MFCC features with different assignments of parameter α)]

MFCC 特徵	原始 IRM_2	不同權重α抑制調變高頻之 IRM_2			
		0	0.25	0.50	0.75
STOI	0.6863	0.6841	0.6840	**0.6880**	0.6837
PESQ	**1.8003**	1.7966	1.7966	1.7853	1.7972

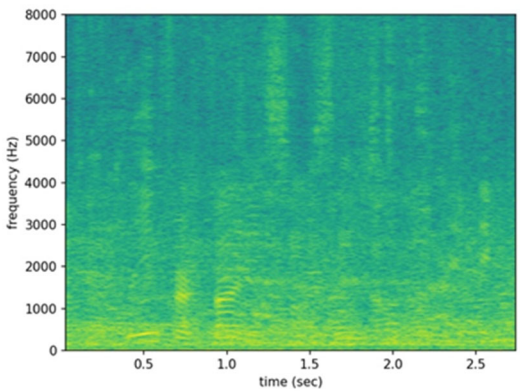
4.4 使用時頻圖演示結果 (Spectrogram Demonstration for Each Method)

最後在這一小節，我們使用語音訊號的強度時頻圖(magnitude spectrogram)，來檢視原始 IRM 與我們提出之低通濾波特徵之 IRM 的強化效能，圖 1(a)-(f) 為一語句在各種狀態下所對應的強度時頻圖，首先，我們比較圖 1(a)與圖 1(b)，發現雜訊對於語音在時頻圖上

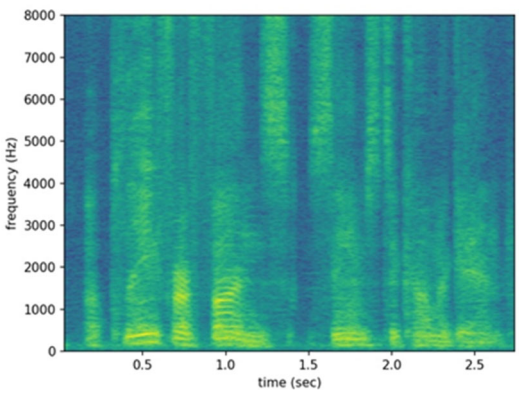
產生顯著的失真，接著，比較圖 1(b)與圖 1(c)可看出，理想的 IRM 可帶來顯著的語音強化效果，最後，觀察原始 IRM 與低通濾波特徵之 IRM 所對應的圖 1(d) 與 圖 2(e) ，相對於圖 1(b)，雜訊所造成的失真明顯降低，但效果並不如理想 IRM 所對應的圖 1(c)，例如在時間 0.1-0.3 秒之間的頻譜強度並未有效重建（在紅色框所標示區域），然而圖 1(e)的在此區域的頻譜重建程度稍優於圖 1(d)，根據此比較結果，我們似乎可看出，低通濾波特徵之 IRM 在此語句的處理上略優於原始 IRM。



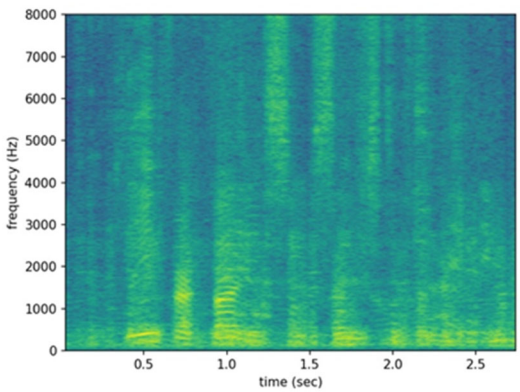
(a) 原始乾淨語音
[a. the original clean utterance]



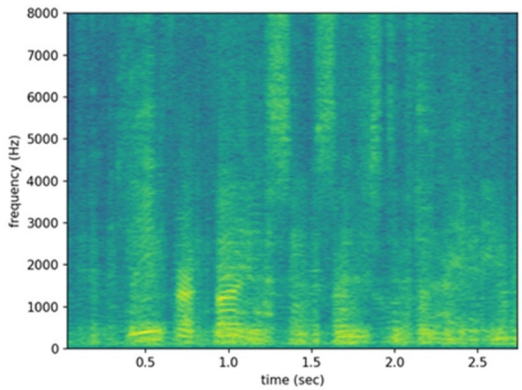
(b) 摻入-2 dB SNR 之 babble 雜訊之語音
[b. the -2 dB SNR utterance with babble noise]



(c) 雜訊語音經由理想 IRM 處理之語音
[c. the oracle-IRM enhanced utterance]



(d) 雜訊語音經由原始 IRM 處理之語音
[d. the original-IRM enhanced utterance]



(e) 雜訊語音經由低通濾波 IRM 處理後之語音
[e. the lowpass-filtered IRM-enhanced utterance]

圖 1. 各種狀態下之語音強度時頻圖
[Figure 1. The magnitude spectrograms of an utterance at different conditions]

5. 結論與未來展望 (Conclusion and future works)

在本研究中，我們提出並初步驗證了當理想比例遮罩(IRM) 之深度模型使用低通濾波之語音特徵時序列來訓練時，相較於使用原特徵時序列訓練，可以得到更佳的語音強化效果。我們使用小波轉換來實現低通濾波的處理，其執行簡易但效果明顯，在未來工作上，我們初步規劃將此低通濾波的時序列處理用在訓練其他種類的語音強化深度模型之特徵上，檢視其是否也能更有效改進該模型的效能、提升語音之品質與可讀性。

致謝 (Acknowledgement)

本論文其部分初階實驗由本校畢業同學林子強 (Mr. Zi-Qiang Lin) 加以執行，特此致謝。

參考文獻 (References)

Chen, C., & Bilmes, J. (2007). MVA processing of speech features. *IEEE Trans. on Audio, Speech, and Language Processing*, *15*(1), 257-270. https://doi.org/10.1109/TASL.2006.876717

Erdogan, H., Hershey, J. R., Watanabe, S., & Le Roux, J. (2015). Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *Proceedings of 40th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, 780-712. https://doi.org/10.1109/ICASSP.2015.7178061

Kanedera, N., Arai, T., Hermansky, H., &. Pavel, M. (1997). On the importance of various modulation frequencies for speech recognition. In *Proceedings of the European Conference on Speech Communication and Technology (Eurospeech)*, *3*, 1079-1082.

Mallat, S. (1999). *A Wavelet Tour of Signal Processing*. (2nd ed.). Academic Press.

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001). Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs. In *Proceedings of of 26th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2001)*, 749-752. https://doi.org/10.1109/ICASSP.2001.941023

Srinivasan, S., Roman, N., & Wang, D. (2006). Binary and ratio time-frequency masks for robust speech recognition. *Speech Communications*, *48*(11), 1486-1501. https://doi.org/10.1016/j.specom.2006.09.003

Taal, C. H., Hendriks, R. C., Heusdens, R., & Jensen, J. (2011). An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech. *IEEE Transactions on Audio, Speech, and Language Processing*, *19*(7), 2125-2136. https://doi.org/10.1109/TASL.2011.2114881

Wang, D. (2005). On ideal binary mask as the computational goal of auditory scene analysis. In: Divenyi P. (eds) *Speech Separation by Humans and Machines*, (pp. 181-197). Springer. https://doi.org/10.1007/0-387-22794-6_12

Wang, S.-S., Lin, P., Tsao, Y., Hung, J.-W., & Su, B. (2018). Suppression by selecting wavelets for feature compression in distributed speech recognition. *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, *26*(3), 564-579. https://doi.org/10.1109/TASLP.2017.2779787

Wang, Y., Narayanan, A., & Wang, D. (2014). On training targets for supervised speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(12), 1849-1858. https://doi.org/10.1109/TASLP.2014.2352935

Williamson, D.S., Wang, Y., & Wang, D. (2016). Complex ratio masking for monaural speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(3), 483-492. https://doi.org/10.1109/TASLP.2015.2512042