Studies on time-frequency transform for DNN-based speech enhancement

DNN 音声強調のための時間周波数変換に関する研究

5118E009-1　竹内 大起　　指導教員　及川 靖広　教授
TAKEUCHI Daiki　　Prof. OIKAWA Yasuhiro

概要： 音声強調は様々なタスクの前処理として応用される重要な音響信号処理である．近年，深層ニューラルネットワーク (DNN) を用いた時間周波数マスキングが音声強調に適用されており，この手法ではマスクを推定する DNN と時間周波数変換の設計によって性能が決まる．本研究では，DNN 音声強調のための完全再構成フィルタバンクの設計を提案する．提案手法では，DNN の学習に利用する誤差関数の性質に基づいて完全再構成フィルタバンクを設計することで，より学習が安定する時間周波数変換を設計する．提案手法を用いた DNN 音声強調と従来の DNN 音声強調の性能を比較する数値実験を行い，提案手法が従来手法と比べてより高い性能を実現することを確認した．

キーワード：深層学習，時間周波数解析，完全再構成フィルタバンク，warped filterbank frame
Keywords: deep learning, time-frequency analysis, perfect reconstruction filterbank, warped filterbank frame

1. Introduction

Speech enhancement aims to recover the target speech from a noisy observed signal. For the single-channel case, time-frequency (T-F) masking is used as the standard method. Since a mask is multiplied in the T-F domain, the quality of an enhanced signal is determined by *both* T-F mask estimator and T-F transform. While T-F mask estimation is an active research field in deep learning, there is less research on T-F transform *from the viewpoint of speech enhancement* because it has been investigated in the context of speech analysis. Recently, some enhancement methods based on deep neural network (DNN) have demonstrated that particular T-F transforms can improve the quality of enhancement, and thus T-F transform should be a worth-investigating topic by itself.

In order to investigate the optimal T-F transform, I propose the method of designing T-F transform for improving DNN-based speech enhancement [1]. My strategy is to design a T-F transform from a dataset before training the DNN for T-F mask estimation so that the training becomes easier. To do so, the warped filterbank frame (WFBF) [2,3] is utilized as T-F transform, and its frequency-warping function is adapted to the database.

2. Speech enhancement based on T-F masking

The problem of speech enhancement is to recover a target signal s_t degraded by noise n_t from an observed monaural signal x_t, $x_t = s_t + n_t$, where t is the time index. It can be rewritten in T-F domain as $X_{\omega,k} = S_{\omega,k} + N_{\omega,k}$, where X is the T-F representation of x, and k and ω denote the indices of time frame and frequency, respectively. As T-F transform, STFT is often used and, it can be defined as

$$X_{\omega,k} = \sum_{l=0}^{L-1} x_{l+ak}\, \overline{g_l\, e^{2\pi j\omega l/M}}, \tag{1}$$

where $j = \sqrt{-1}$, $\overline{z}$ is complex conjugate of z, g is an analysis window, a is the time-shifting step, L is the length of the signal, and M is the number of frequency channels (or DFT length). In this definition of STFT, the number of time frames is $N = L/a$.

In T-F masking, the estimated target signal $\hat{S}_{\omega,k}$ is acquired by the element-wise multiplication of a T-F mask $G_{\omega,k}$ to the observation $X_{\omega,k}$: $\hat{S}_{\omega,k} = G_{\omega,k}\, X_{\omega,k}$. Then, the result is transformed back to the time domain by the inverse transform. The T-F mask $G_{\omega,k}$ must be estimated solely from $X_{\omega,k}$, which is the difficult part.

Many methods have applied DNN to estimate the T-F mask. In deep learning approach, a T-F mask $G_{\omega,k}$ is estimated as $\hat{G}_{\omega,k} = \mathcal{M}_\theta(\Psi)_{\omega,k}$ where $\mathcal{M}_\theta$ is a regression function implemented by DNN, θ is the set of its parameters, and $\Psi = \Psi(X)$ is the input acoustic feature.

2.1 Phase sensitive approximation (PSA)

Typically, a T-F mask G is chosen to be real-valued. The truncated phase sensitive mask (PSM) G^{PSM} is one of the real-valued T-F masks which minimizes MSE between $\hat{S}_{\omega,k}$ and $S_{\omega,k}$ on the complex plane [4]:

$$G^{\mathrm{PSM}}_{\omega,k} = \mathcal{T}_{[0,1]}\left[\frac{|S_{\omega,k}|}{|X_{\omega,k}|}\cos(\phi_{S_{\omega,k}} - \phi_{X_{\omega,k}})\right], \tag{2}$$

where $\mathcal{T}_{[a,b]}[z] = \min(\max(z,a),b)$ is the truncation operator, and $\phi_{S_{\omega,k}}$ and $\phi_{X_{\omega,k}}$ are phase angles of $S_{\omega,k}$ and $X_{\omega,k}$, respectively. For approximating this mask by DNN $\mathcal{M}_\theta$, its parameters θ are trained to minimize the following MSE for all data in a dataset:

$$\mathcal{J}_{\mathrm{PSA}}(\theta) = \sum_{\omega=1}^{\Omega}\sum_{k=1}^{K} \left|\mathcal{M}_\theta(\Psi)_{\omega,k}\, X_{\omega,k} - S_{\omega,k}\right|^2. \tag{3}$$

2.2 Assumption mismatch of MSE

The above cost function, MSE, assumes that the error between the clean and masked T-F bin has the uniform variance for all bins. However, this assumption cannot be met in reality because both target source and noise have non-uniform spectral distribution in practical situations. Such assumption mismatch is problematic since it underestimates the error in the frequency range having small power. That is, higher frequency range, which contains less power for practical sounds (see Fig. 1), is difficult to train than the lower range.

3. Proposed method

To resolve assumption mismatch, we propose to modify the T-F transform instead of modifying the cost function.

3.1 Warped filterbank frame (WFBF)

To normalize frequency-wise error in MSE, we propose to use a frequency-warped PRFB so that the error for each frequency band has the same power. For warping the frequency axis as desired, the WFBF [2, 3] is considered in this paper.

The WFBF is a PRFB whose frequency scale can be defined by a user. The WFBF can be written as the following form:

$$X^{\mathscr{F}}_{\omega,k} = \sum_{l=0}^{L-1} x_{l+a_\omega k}\, \overline{g_{\omega,l}\, e^{2\pi j\Phi^{-1}(\omega)l}}, \tag{4}$$

where Φ is a frequency-warping function which is defined so that the resulting WFBF has the desired frequency scale (see [2, 3] for the regularity required for Φ such as $\mathcal{C}^1$-diffeomorphism and having positive derivative), and the parameters with the subscript ω may be different for each frequency band. Since Eq. (4) is the downsampled convolution between the signal and $g_\omega[l]\, e^{2\pi j\Phi^{-1}(\omega)l}$ which can be computed efficiently via the fast Fourier transform (FFT), WFBF is a collection of bandpass filters whose center frequencies are decided by the warping function, and the window functions are automatically derived according to the design requirement [2, 3].

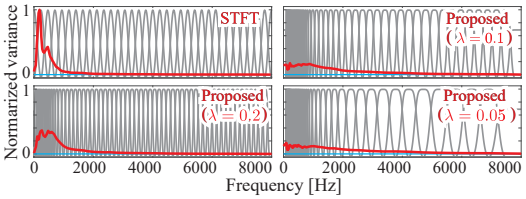

Fig. 1 The variance of the oracle PSM masking error ε (red) calculated in the STFT domain and the learned WFBF domains.

3. 2 Proposed speech enhancement system

T-F masking is applied in WFBF domain instead of STFT domain as $S^{\mathscr{F}}_{\omega,k} = \mathcal{M}^{\mathscr{F}}_{\theta}(\Psi_{\mathscr{F}})_{\omega,k}\, X^{\mathscr{F}}_{\omega,k}$, where $S^{\mathscr{F}}_{\omega,k}$ is the estimated target signal, $\mathcal{M}^{\mathscr{F}}_{\theta}$ is a DNN-based regression function, and $\Psi_{\mathscr{F}}$ is the input feature. After masking, the estimated time-domain signal is recovered from $S^{\mathscr{F}}_{\omega,k}$ thanks to the perfect reconstruction property of WFBF.

In order to design the warping function, we propose to use the frequency-wise energy of a training dataset. Since MSE assumes that the error of all frequency has uniform variance, T-F transform should be designed to fulfill this assumption. To do so, the error of T-F masking by the oracle PSM is collected from the training dataset of $\mathcal{M}^{\mathscr{F}}_{\theta}$ because that is what the training tries to minimize. Then, its power spectral density (PSD) is estimated from the corrected error, which should be normalized to met the assumption of MSE. For the normalization, we propose to obtain the warping function Φ by

$$\Phi_{\mathrm{Prop}} = \mathrm{cumsum}(\boldsymbol{\sigma} + \lambda), \tag{5}$$

where $\boldsymbol{\sigma}$ is a vector of the error PSD, $\lambda \geqq 0$ is a small regularization parameter, and cumsum is the cumulative sum taken from lower to higher frequency. The role of λ is to avoid an excessively wide frequency band, where a larger λ makes WFBF closer to STFT.

Using the proposed PRFB based on WFBF, MSE can be appropriately used as the cost function without the normalization weight. The assumption mismatch is removed by training a DNN based on this cost function because the energy is normalized for each frequency by the filterbank, and its gradient is stably balanced as the weight is removed. The proposed PRFB can be viewed as the pre-emphasis which emphasize the important frequency range and de-emphasize unimportant range that are learned from the training data.

4. Experiment

In order to confirm the correctness of the proposed method, the performance of speech enhancement is investigated by comparing the STFT-domain PSM [4] and the proposed method.

Before that, as a preliminary experiment, frequency-wise variance of the error ε was calculated to see how the proposed warping function Φ_{Prop} in Eq. (5) works. By using the training dataset explained below, the error of the oracle PSM was collected, and its PSD was estimated by the Welch estimator. The variances of the error ε for each frequency are shown in Fig. 1. From the figure, it can be seen that the proposed filterbank obtained more balanced distribution of the masking error. While smaller λ results in more balanced variance distribution, wider frequency band appears in high-frequency range which is not favorable. We chose $\lambda = 0.1$ to balance this trade-off.

4. 1 Experimental conditions

The Wall Street Journal (WSJ-0) corpus and noise dataset CHiME-3 were used as the training datasets. The noisy signals for the training dataset were formed by mixing clean speech utterances with the noise at signal-to-noise ratio (SNR) levels of -6 to 12 dB. As the test datasets, 500 utterances randomly selected from the TIMIT corpus were used for the target-source dataset, and four types of ambient noise *factory1*, *factory 2*, *f16*, and *babble* from the NOISEX92 dataset were used as the noise dataset. All files were recorded at sampling rate of 16 kHz.

Table 1 Experimental result

Input SNR	Dim.	TF transform	cost func.	SDR
6 dB	64	STFT	MSE	10.55
		WFBF	MSE	**11.39**
	257	STFT	MSE	10.93
0 dB	64	STFT	MSE	6.70
		WFBF	MSE	**7.52**
	257	STFT	MSE	7.05
-6 dB	64	STFT	MSE	2.91
		WFBF	MSE	**3.64**
	257	STFT	MSE	3.28

The performances of speech enhancement in the proposed WFBF domain and the conventional STFT domain were compared on the network with the two bidirectional long short-term memory (BLSTM) consisting of 512 cells. The sigmoid function was used at the output layer for limiting the values within the range 0 to 1. In the proposed method, log-amplitude of $X^{\mathscr{F}}_{\omega,k}$ is used asn input acoustic feature, and the number of frequency bins (and thus input dimension) was set to 64. In the conventional STFT, the number of frequency bins was set to 512 (i.e., input dimension was 257), and the window is shifted by 256 samples. To match the input dimensions between the conventional and proposed methods, 64 dimensional log-mel transform matrix and its pseudo-inverse were applied to STFT. The conventional method without log-mel transform is also considered as a baseline. They were trained 200 epochs, where each epoch contained 1000 utterances, and mini-batch size was 5. The learning and dropout rates were decreased linearly.

4. 2 Experimental result

The performances of speech enhancement were measured by the signal-to-distortion ratio (SDR). The experimental results are summarized in Table 1. Bold font indicates the best score within the same condition. For all cases, the proposed WFBF achieved the highest scores. The scores of the conventional STFT tends to decrease as the input dimension decreases, while those of the proposed WFBF did not. It was also confirmed that weighted MSE in STFT domain obtained less performances than the usual MSE, which should be because of the difficulty of the optimization. These results indicate that using the learned T-F transform instead of the ordinary STFT is more efficient for speech enhancement.

5. Conclusions

In this paper, a data-driven design method of T-F transform using WFBF is proposed . By considering WFBF, the learning problem of T-F representation was reduced to calculation of the one-dimensional frequency-warping function, which is obtained through PSD of oracle masking error. Since the calculation of the proposed warping is cheap, it can be easily adapted to different dataset in contrast to a fully DNN-based learning of T-F-like representation.

References

[1] D. Takeuchi, K. Yatabe, Y. Koizumi, Y. Oikawa, and N. Harada, "Data-driven design of perfect reconstruction filterbank for DNN-based sound source enhancement," in *2019 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2019, pp. 596–600.

[2] N. Holighaus, C. Wiesmeyr, and Z. Průša, "A class of warped filter bank frames tailored to non-linear frequency scales," *arXiv preprint arXiv:1409.7203*, 2014.

[3] N. Holighaus, Z. Průša, and C. Wiesmeyr, "Designing tight filter bank frames for nonlinear frequency scales," in *Int. Conf. Sampl. Theory Appl.*, 2015.

[4] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *2015 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 2015, pp. 708–712.