Title: End-to-End Model for Speech Enhancement by Consistent Spectrogram Masking
Author: Du, Xingjian, Zhu, Mengyao, Shi, Xuan, Zhang, Xinpeng, Zhang, Wen, Chen, Jingdong

Summary

The article proposed a Consistency Spectrogram Masking method to estimate complex spectrogram. The proposed method utilizes a joint real and imaginary mask instead of magnitude and/or phase mask in the STFT domain. Training is achieved with the objective function in the time domain in order to optimize the mask of real and imaginary spectrogram jointly and simultaneously.

Good Things about This Paper

In speech enhancement systems, estimation of masks for complex-valued short-time Fourier transforms (STFTs) is needed to suppress noise and preserve speech. Other models use large amounts of data to train a deep neural network. However, these approaches often neglect an important constraint: STFT consistency. Without STFT consistency, the system's output is not necessarily the STFT of a time-domain signal. This work addresses the spectrogram ″consistency″ by formulating a neural speech enhancement algorithm trained with time-domain loss. Therefore, it makes it possible to improve the phase of enhanced speech. Also, the design of densely connected u-net is reasonable for extracting a multi-scale representation of the spectrogram adaptively.

Major Comments

1. The major weakness of this work is the lack of an "ablation study". Speech enhancement represents an important problem domain with many applications. This work addresses the sub-problem of clean phase estimation which may has an opportunity to improvement over speech enhancement methods that merely retain the noisy phase.

However, the magnitude and phase information of STFT is not independent mutually. The signal can be reconstructed from only the magnitude information with an iterative method [1]. So the phase information of the STFT is not always necessary to enhance the noise-corrupted signals. The proposed method needs to be compared with the methods using only the magnitude information of the STFT.

2. The approach outlined in the paper is technically sound. However, many details were omitted such as the size of the DenseNet and the training procedure. The training corpus is relatively small and no evidence is given regarding its sufficiency. Also, the choice of MSE loss function is not very well motivated (low MSE is sufficient but not necessary to achieve perceptually high-quality enhancement).

3. While algorithmic complexity of the proposed algorithm is unclear in the text, the fact that the proposed algorithm is amenable to ″strong GPU acceleration″ is presented as a positive attribute. Many leading speech enhancement algorithms have quite low computational complexity and do not require a GPU so the complexity of the proposed method may be a negative attribute.

[1]D.W.Griffin et al ″Signal reconstruction from short-time Fourier transform magnitude,″ IEEE Trans. acoust., speech, and signal processing. vol.24, no.3, pp.243-248, June 1976.

Minor Comments

1. Fig. 1: What is the 4 small blocks FCN module? Also, the output signal is not clean but de-noised or enhanced. Besides, is the operator of the JRIM is a multiplication? If not, the plot of this operator needs to be corrected to avoid the confusion.

2. Fig. 2: The x-axis and y-axis need to be labeled if possible. It is better if lines are solid, broken, dotted, because color difference is not visible in the black and white printed paper.