Employing the Exponentiated Magnitude Spectrogram in the Deep Learning-Based Mask Estimation for Speech Enhancement

Jeih-Weih Hung*, Chi-En Dai, Ping-Chen Wu, and Che-Wei Liao

***Abstract*—The objective of speech enhancement (SE) is to alleviate various types of distortion (noise, channel effect, reverberation, etc.) in received speech signals to improve the corresponding perceptual quality and intelligibility. SE techniques are essential in speech signal-related online education and learning applications and devices.**

Thanks to the rapid development of deep neural network (DNN) techniques, various SE methods based on DNN have been proposed. They usually outperform the conventional statistics-based SE methods in non-stationary environments. These DNN-based SE methods can be further divided into mapping-based and masking-based. In particular, masking-based methods have attracted more attention in recent years.

This study focuses on improving a well-known masking-based method: the ideal ratio mask (IRM). We propose to revise the spectrogram for the input utterances in the learning of the IRM network to improve its speech enhancement performance. For each utterance, the magnitude spectrogram is raised to a particular power (exponentiated) first and then used to create various speech features, including Mel-frequency cepstral coefficients (MFCC) and gammatone-frequency features (GF). We feed these features to the deep network for IRM estimation. The exponentiation operation for the magnitude spectrogram is believed to highlight the speech portion of an utterance. Thus the exponentiated spectrogram probably benefits the following speech feature representation employed to learn the deep neural network for IRM.

We conduct a series of evaluation experiments on a subset of the TIMIT database. The utterances in the training and test sets are corrupted by factory noise at a signal-to-noise ratio (SNR) of -2 dB. We use the Perceptual Evaluation of Speech Quality (PESQ) and Short-Time Objective Intelligibility (STOI) as the speech enhancement metrics.

The preliminary results reveal that, compared with the IRM from the original spectrogram, the new IRM created with the exponentiated spectrogram provides the test utterances with superior perceptual quality and intelligibility scores.

***Index Terms*—Speech enhancement, exponentiated spectrogram, ideal ratio mask**

I. Introduction

People use mobile devices like smartphones and tablets for various activities, including communication, online interactive learning, and education. In particular, voice/speech-wise applications, such as sound recording, music playing, and speech recognition, are crucial in these mobile devices. There is no doubt that the quality or readability of received acoustic signals is in high demand. However, various sources of distortion deteriorate speech signals during transmission, thus sabotaging the capability of the functions mentioned above and their applications. These distortion sources include additive noise, channel distortion, and reverberation. Various speech enhancement (SE) techniques have been developed in recent decades to solve or alleviate the distortion issues. Most of the novel SE methods exploit a deep neural network (DNN) to learn the relationship between the clean noise-free speech and their distorted counterparts. Compared with the conventional SE methods primarily based on statistical modeling of speech or noise, the DNN-based SE methods behave superior, especially in non-stationary noise scenarios.

According to [1], the DNN-based SE methods can be roughly divided into two categories according to their training objectives: mapping and masking. The mapping-wise SE methods directly pursue a mapping function from the input distorted signal to the perfectly noise-free signal or its various representation, such as a time-domain signal waveform, a time-frequency diagram (spectrogram), or a cochleagram. Comparatively, the masking-wise SE methods search for a multiplicative mask to perform point-to-point multiplication with the original input signal or feature representation. The resulting product can approach its clean noise-free state. The mapping-wise methods possess a bigger hypothesis space for the mapping solution, while the masking-wise methods restrict its hypothesis space to be simply multiplicative masking. In recent years, masking-wise SE methods have attracted more attention and gained diversity and development. The training target of these masking-wise methods include ideal binary mask (IBM) [2, 3], ideal ratio Mask (ideal ratio mask, IRM) [3], spectral intensity mask (spectral magnitude mask, SMM) [1], complex ideal ratio mask (complex ideal ratio mask, cIRM) [4], phase sensitive mask (phase-sensitive mask, PSM) [5], etc.

This study attempts to revise the process to learn a well-known target for masking-wise methods: the ideal ratio mask (IRM). Unlike a lot of other algorithms that focus on updating the DMM structure employed in IRM to learn a more effective and generalizable mask for speech enhancement, we propose to deal with the very front-end module in IRM by revising the input utterances in their spectrogram. We exponentiate the magnitude spectrogram with a power value larger than 1 with the purpose of enlarging the discrepancy between speech and noise portions in the utterance to enlarge it signal-to-noise ratio (SNR). Preliminary experiments conducted on a subset of TIMIT

Manuscript received November 15, 2022; revised November 29, 2022; accepted December 30, 2022.

Jeih-Weih Hung, Chi-En Dai, Ping-Chen Wu, and Che-Wei Liao are with the Department of Electrical engineering, National Chi Nan University, Taiwan.

*correspondence: jwhung@ncnu.edu.tw (J.W.H.)

database with the script provided in [6] show that the presented method can promote the SE performance of the resulting IRM network.

The rest of the paper is organized as follows: Section II introduces the background and the procedures of the newly presented method. The experimental setup is given in Section III, and Section IV covers the experimental results and the corresponding analyses. Finally, a brief concluding remark is provided n Section V.

II. Presented Method

In this study, we focus on improving the ideal ratio mask (IRM) method, which generally pursues the approximate mask values for the spectrogram or cochleagram corresponding to an arbitrary input utterance, and one of the multiple choices of the desired mask is from the concept of Wiener filtering:

$$M(m,f) = \frac{|S(m,f)|^2}{|S(m,f)|^2 + |D(m,f)|^2} \quad (1)$$

where $|S(m,f)|^2$ and $|D(m,f)|^2$ represent the energy of clean speech and noise with respect to the time-frequency unit (T-F unit) of the spectrogram or cochleagram at frame m and frequency f respectively. As for preparing the artificial noisy data in the training set, both clean speech and noise components are pre-known, and thus we can accurately obtain the mask values described in Eq. (1) and use them as the desired output for the IRM network to be trained. Furthermore, the input utterances for the learning and inferencing of the IRM network are often converted to speech feature representations, such as mel-frequency cepstral coefficients (MFCC), gammatone features (GF) and relative spectral-perceptual linear predictive features (RASTA-PLP), which serve as an excellent encoding to the subsequent mask estimation network.

In this work, we propose pre-processing the input utterances in their spectrogram before feeding them into the IRM network for training and testing. We extract the magnitude part of the spectrogram for each utterance, raise it to a particular power, integrate it with the initial phase part, and convert it back to the time-domain utterance. The idea behind this method is to emphasize the relatively high-energy portions of the utterance in order to highlight the clean-speech component. The resulting utterances are supposed to possess a higher signal-to-noise ratio (SNR) and thus behave better in the IRM estimation.

We describe the steps of this new method in the following:

Step 1: create the spectrogram

For each time-domain utterance $x[n]$ in the training and test sets, we employ the short-time Fourier transform (STFT) to create its spectrogram $\{X[m,k], 0 \le m \le L-1, 0 \le k \le K-1\}$, where m and k are the indices of frame and acoustic frequency, and L and K are the total numbers of frames and acoustic frequency points, respectively.

Step 2: exponentiate the magnitude spectrogram

The spectrogram $X[m,k]$ is complex-valued and can be presented in polar form $X[m,k] = A[m,k]\exp(j\phi[m,k])$, where $A[m,k]$ and $\phi[m,k]$ are the magnitude and phase components, respectively. We extract its magnitude part $A[m,k]$ and raise it to a power larger than 1. Therefore, the new spectrogram can be formulated by

$$\tilde{X}[m,k] = (A[m,k])^r \exp(j\phi[m,k]), \quad (2)$$

where r is the pre-set power value, and $r > 1$.

Step 3: create the new time-domain utterance

We apply the inverse STFT on the new spectrogram $\tilde{X}[m,k]$ in Eq. (2) to obtain the updated version of the time-domain utterance $\tilde{x}[n]$.

Step 4: Training and testing the IRM network

The updated utterances $\tilde{x}[n]$ in either of the training and test sets are converted to speech features (MFCC, GF, RASTA-PLP, etc.) and used to learn or evaluate the IRM network as the usual IRM preparation procedures.

Some of the characteristics of the presented method include the following:

1. The main component of this method is Step 2, which updates the magnitude part of the input spectrogram, hoping to highlight the clean-signal portion in an utterance. The updated spectrogram is converted back to the time-domain utterance as in Step 3. As such, the presented method can be exploited as a pre-processing procedure of all of the other speech enhancement techniques, exhibiting its flexibility.
2. The presented method focuses on enhancing the magnitude part while leaving the phase part unchanged. One underlying reason is that correcting the phase part is a more challenging task and the phase estimation has an ambiguity/discontinuity issue. We will improve this part by either enhancing the real and imaginary parts individually or processing the real-valued spectrogram created by discrete cosine transform (DCT).

III. Experimental Setup

Referring to the MATLAB toolbox for the speech separation task provided in [6], we use a subset of the TIMIT database to evaluate the presented method in learning an IRM speech enhancement network. The training set contains 100 utterances evenly produced by 10 speakers, and the test set includes 60 utterances from 6 speakers different from those for the training set. The utterances in the training and test sets are corrupted with factory noise at −2 dB signal-to-noise ratio (SNR). Following the procedures stated in Section III, each time-domain utterance is transformed to the STFT-wise spectrogram, the magnitude part being raised to a power r, and converted back to the time domain via inverse STFT.

As for the STFT procedure, the frame size is 20 ms and the hop size is 10 ms, and the hamming window function is used.

The updated utterances in the training set are converted to speech feature representations to train the IRM network as described in [6]. Two speech representations are selected: Mel-frequency cepstral coefficients (MFCC) and gammatone features (GF). The learned IRM network comprises densely connected layers, with 4 hidden layers, each having 1024 neurons. The five adjacent frames are concatenated into a long vector to be the IRM network's input unit. The learning

objective is to obtain the mask of the cochleagram of speech, each frame having 64 dimensions (channels). To evaluate the performance, we use the metrics of perceptual evaluation of speech quality (PESQ) [7] and Short-Time Objective Intelligibility (STOI) [8] as objective indicators of speech quality and intelligibility, respectively. The PESQ score is between -0.5 and 4.5, and the STOI score is between 0 and 1, with higher scores representing better speech quality/intelligibility.

IV. Experimental Results and Discussions

First, Table I reports the PESQ and STOI scores from the unprocessed baseline, the oracle IRM, and the learned IRM using different speech features. From this table, we have the following findings:

1. Noise causes a severe degradation in the quality and intelligibility of speech signals. The respective STOI and PESQ are 0.725 and 1.433, respectively, while their upper bounds are 1.0 and 4.5.
2. The oracle IRM, which employs the prior knowledge of clean speech and noise in noisy utterances of the test set, significantly improves the two metric indices. The respective scores are the upper bound of all the IRM networks discussed later.
3. Compared with the unprocessed baseline, the IRM learned from different feature representations achieves higher PESQ and STOI scores, indicating that the IRM network can be learned well to fulfill a speech enhancement task.
4. MFCC alone behaves better than GF alone for PESQ, while the situation is converse for STOI. When added with delta features, only GF benefits PESQ, while the STOI scores worsen for both MFCC and GF.

TABLE I: The PESQ and STOI Scores Obtained from Various Situations: Unprocessed Baseline, Oracle IRM, Learned IRM Using Different Speech Feature Sets

		STOI	PESQ
unprocessed		0.625	1.433
oracle IRM		0.906	2.737
Learned IRM	MFCC	0.722	1.937
	GF	0.728	1.903
	MFCC + Delta MFCC	0.718	1.932
	GF + Delta GF	0.723	1.927

Next, Tables II and III list the STOI and PESQ scores for the IRM with the new exponentiated spectrogram at different exponent values for MFCC and GF features. We have had several discussions about the results shown in these two tables:

1. Using the presented methods with the exponent $r > 1$ always provides higher STOI and PESQ scores than the original IRM when MFCC features are used. As for the case of GF features, the improvements in PESQ and STOI are similar, except for the situations of $r = 2$ and $r = 2.5$, where STOI is lightly decreased.
2. Setting $r = 2$ in the presented method gives optimal PESQ and STOI scores in almost all cases. Further increasing r to 2.5 causes PESQ degradation compared with the case $r = 2$. The probable explanation is that setting $r = 2.5$ over-amplifies the magnitude spectrogram and brings extra distortion.
3. To further validate if the presented exponentiation operation with $r > 1$ that potentially enlarges the magnitude can improve IRM, we set $r = 0.5$ as a contrary test. From these two tables, we find the setting $r = 0.5$ worsens both PESQ and STOI scores in almost all cases, which agrees with our proposition.

TABLE II: The PESQ and STOI Scores Obtained from Various Situations: Learned IRM using **MFCC Features** with Respect to Exponentiated Spectrogram with Exponent r (the case $r = 1$ corresponds to the Original IRM

	exponent r				
	0.5	1.0 (original)	1.5	2.0	2.5
STOI	0.714	0.722	**0.724**	**0.725**	**0.725**
PESQ	1.914	1.937	**1.943**	**1.946**	**1.944**

TABLE III: The PESQ and STOI Scores Obtained from Various Situations: Learned IRM Using **GF Features** with Respect to Exponentiated Spectrogram with Exponent r (the case $r = 1$ Corresponds to the Original IRM

	exponent r				
	0.5	1.0 (original)	1.5	2.0	2.5
STOI	0.721	0.728	**0.729**	0.726	0.723
PESQ	1.093	1.903	**1.914**	**1.932**	**1.927**

Finally, Tables IV and V list the evaluation scores for the IRM with the new exponentiated spectrogram at different exponent values with respect to MFCC and GF *plus their delta features*. Similar to Tables II and III, the presented methods with the exponent $r > 1$ exhibit higher STOI and PESQ scores in most cases (except for the case with GF and delta GF features).

To briefly sum up, the optimal PESQ is 1.947, which occurs in the case with $r = 2$ using GF and delta GF features, and the optimal STOI is 0.729, which is achieved by setting $r = 1.5$ and using GF features. Moreover, MFCC outperforms GF in the original IRM, while the presented method with $r > 1$ makes GF outperform MFCC in the revised IRM.

TABLE IV: The PESQ and STOI Scores Obtained from Various Situations: Learned IRM using **MFCC and Delta MFCC Features** with Respect to Exponentiated Spectrogram with Exponent r (the case $r = 1$ Corresponds to the Original IRM

	exponent r				
	0.5	1.0 (original)	1.5	2.0	2.5
STOI	0.712	0.718	**0.720**	**0.721**	**0.719**
PESQ	1.900	1.932	**1.945**	**1.945**	**1.933**

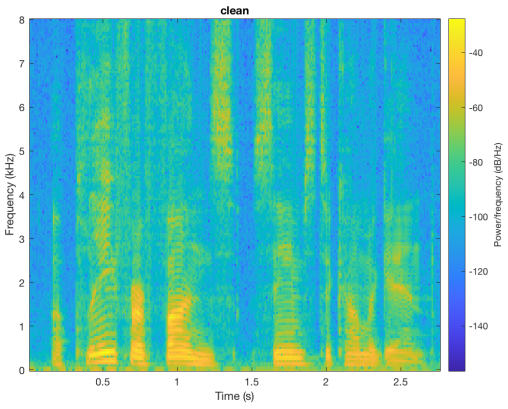
TABLE V: The PESQ and STOI Scores Obtained from Various Situations: Learned IRM Using **GF and Delta GF Features** with Respect to Exponentiated Spectrogram with Exponent r (the case $r = 1$ Corresponds to the Original IRM

	exponent r				
	0.5	1.0 (original)	1.5	2.0	2.5
STOI	0.718	0.723	**0.726**	**0.726**	**0.725**
PESQ	1.914	1.927	1.920	**1.947**	**1.937**

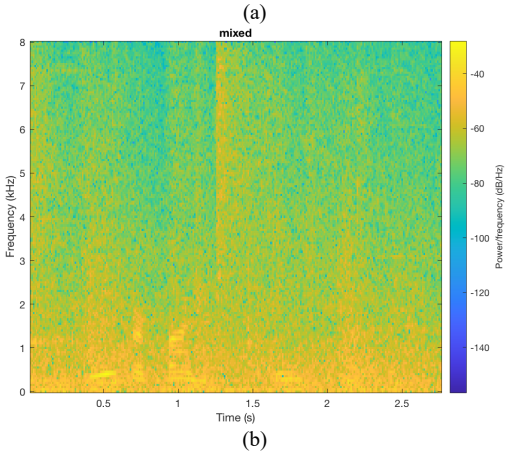
In addition to the quantitative metric comparison, here we pick an utterance in the test set and present its magnitude spectrograms at different conditions as a demonstration: clean, mixed, enhanced with an oracle IRM, the original

learned IRM with GF features using exponent $r = 1$, and the learned IRM with GF features using exponent $r = 2, 0.5$. They are depicted in Fig. 1(a–f). We have two observations from these figures:

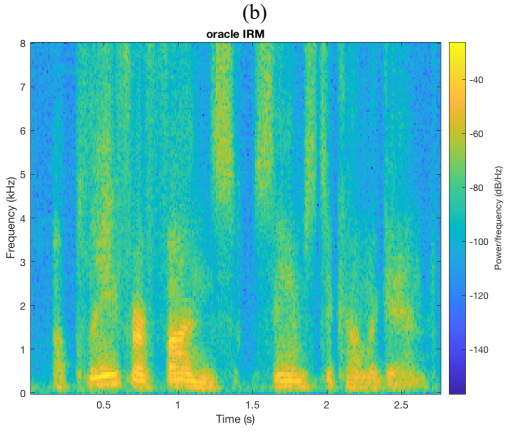
1. Comparing Fig. 1(a, b), the apparent mismatch in the spectrogram reveals that noise significantly disturbs speech. However, the oracle IRM reduces the mismatch greatly (from Fig. 1(b, c) and nearly reconstructs the original clean speech (from Figs. 1(a)(c)).
2. When employing the IRM learned from the training set either with or without an exponentiated spectrogram (as in Fig. 1(d–f)), the distortion caused by noise is moderately reduced, even though the corresponding denoising effect is not as good as the oracle IRM. Moreover, the differences among Fig. 1(d–f) are not significant, which agrees with the PESQ and STOI results provided earlier that the improvement or degradation relative to the original IRM is moderate.



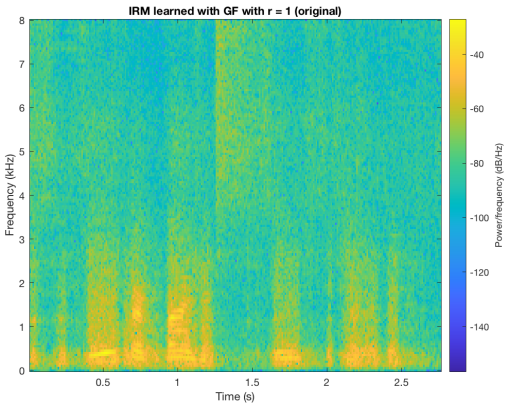
(a)



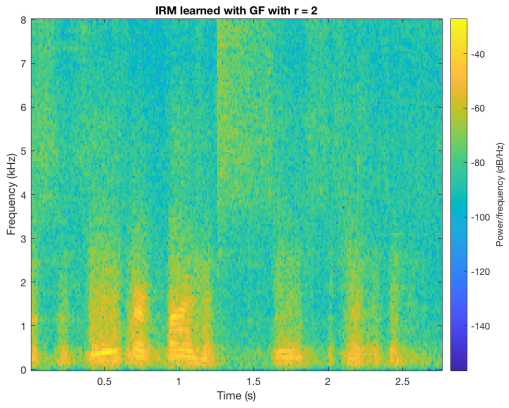
(b)



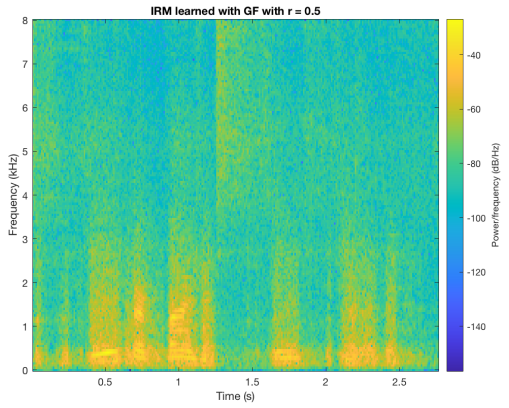
(c)



(d)



(e)



(f)

Fig. 1. The spectrogram of an utterance in the test set at different situations: (a) clean (b) mixed with factory noise at -2 dB SNR, and noise mixed and enhanced by (c) the oracle IRM (d) the GF-feature learned IRM (e) the GF-feature learned IRM with exponent $r = 2$, (f) the GF-feature learned IRM with exponent $r = 0.5$.

V. Conclusion and Future Work

This study investigates whether the pre-emphasis of the high-magnitude time-spatial units in the spectrogram of the utterances in the training set would benefit the backward IRM deep neural network. We present using the exponentiation operation to perform the pre-emphasis, and the preliminary experimental results reveal that the presented method moderately improves the SE behavior of the IRM network. As for the future avenue, we plan to embed the exponent term in the presented method into the neural network to make it learnable to fit the training dataset.

Conflict of Interest

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

J-W. Hung and C-E. Dai conducted the research and designed the initial evaluation experiments; P-C. Wu and C-W. Liao extended the evaluation experiments and analyzed the results; J-w. Hung and C-E. Dai wrote the paper; all authors had approved the final version.

REFERENCES

[1] Y. Wang, A. Narayanan and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2014.
[2] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," *Speech Separation by Humans and Machines; Springer*, 2005.
[3] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communications*, 2006
[4] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016
[5] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks," in *Proc. ICASSP*, 2015
[6] Matlab toolbox for DNN based speech separation. [Online]. Available: http://web.cse.ohio-state.edu/pnl/DNN_toolbox/
[7] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ) – a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*, 2001.
[8] C. H. Taal, R. C. Hendrks, R. Heusdens and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, 2011