Exploring Feature Enhancement in The Modulation Spectrum Domain via Ideal Ratio Mask for Robust Speech Recognition

Bi-Cheng Yan*†, Meng-Che Wu† and Berlin Chen*
*National Taiwan Normal University, Taipei, Taiwan
E-mail: {80847001s, berlin}@ntnt.edu.tw
†AICS, ASUS, Taipei, Taiwan
E-mail: {bicheng, meng}@asus.com

***Abstract*—Development of robustness techniques is of paramount importance to the success of automatic speech recognition (ASR) systems. In this paper, we present a novel use of the ideal ratio mask (IRM) method to improve ASR robustness. IRM was originally proposed for time-frequency (T-F) masking-based speech enhancement and has shown considerable promise in preserving the intelligibility of a noisy mixture signal. Further, IRM is alternatively used to normalize the intermediate representations of speech feature vector sequences, in a holistic manner, for both training and test utterances. Finally, we instead treat IRM as a data augmentation method, conducted on speech feature vectors of training utterances or their intermediate representations, to generate additional augmented data for increasing the diversity of training data. A series of experiments carried out on the standard Aurora-4 database and task confirm the effectiveness of our methods.**

I. INTRODUCTION

Robustness techniques are normally adopted to alleviate the negative influence caused by environmental distortions so as to make ASR systems retain acceptable performance [1]. To date, pragmatic robustness methods that have been designed and developed can be broadly grouped into three generic families: 1) speech enhancement; 2) feature normalization; and 3) model adaptation. Speech enhancement means to increase intelligibility of a noisy speech through suppression of inherent noise components [2], [3]. Further, feature normalization is dedicated to refining speech features and make them more resistant to noise and channel disturbances [4]. Lastly, the acoustic model of an ASR system can be transformed from the original space (with the clean-condition training setup) to a new space (reflecting the noisy test condition) by using an array of model-based adaptation techniques [5], [6].

More recently, an important trend of robustness research has is to formulate speech enhancement as a supervised learning problem. Notably, this line of research has been shown good generalization ability when given sufficient training data. Furthermore, methods stemming from it generally seek to enhance a speech signal in a frame-by-frame fashion, thereby being more amenable to real-time processing [6], [8]. Among them, the family of time-frequency (T-F) masking-based speech enhancement methods aim to filter noise components out from a noisy mixture via its T-F representation. The simplest instantiation is ideal binary mask (IBM), which boils down the speech enhancement task to a binary classification problem, geared towards the reduction of computational burden [10]. However, such a binary masking operation typically produces musical noise that might hurt the performance of downstream applications. Other more elaborate methods of this family, including ideal ratio mask (IRM), spectral magnitude mask (SMM), complex ideal ratio mask (cIRM), and the like, focus exclusively on estimating a smoothed ratio mask which can yield better intelligibility for an enhanced noisy speech utterance than does IBM [11], [12].

In view of the above, we apply and extend the IRM method to improve ASR robustness, since it has shown superior promise in preserving the intelligibility of a noisy mixture signal [13]. Moreover, in this paper, IRM is used in an alternative manner to normalize the intermediate representations of speech feature vector sequences in a holistic manner for both the training utterances and the test utterances. Such intermediate representations can be embodied in the modulation domain by performing discrete Fourier transform (DFT) along the time-axis of acoustic feature vector sequences [14], [15]. By doing so, the salient linguistic information of noisy utterances can be better manifested, consequently improving ASR robustness. Finally, we exploring treating IRM as a data augmentation method instead, conducted on speech feature vectors of training utterances or their intermediate representations, to generate additional augmented data for increasing the diversity of training data. The remainder of the paper is organized as follows: Section II introduces the notion and formulation of modulation spectrum. The masking-based speech enhancement method leveraged in this paper is briefly reviewed in Section III. After that, the corpus and experimental setup are described in Section IV, followed by a series of experiment and associated discussions in Section V. Finally, Section VI concludes this paper and discusses avenues for future work.

II. THE FORMULATION OF MODULATION SPECTRUM

For a given utterance, we can express each of its speech feature dimensions as an ordered feature sequence $\{\mathbf{y}(n), n =$

$1,2,\dots,T\}$, which contains T frames, and each vector has D dimensions. Processed with the discrete Fourier transform (DFT) on the time trajectory of each feature component sequence $y_d(n)$ of $\mathbf{y}(n)$, the temporal sequence $Y_d(k)$ corresponding to the modulation spectrum of this sequence is expressed as follows:

$$Y_d(k) = \sum_{n=0}^{N-1} y_d(n) e^{\frac{-j2\pi nk}{N}} \quad (1)$$
$$k = 0,\dots,N-1; \quad d = 1,2,\dots,D$$

where k signifies the modulation frequency component index and N is used to designate the DFT sample point number. Put another way, Eq. (1) amounts to treating the acoustic feature component sequence as a signal and rendering its dynamic patterns along the temporal axis. This way, the resulting modulation spectra can be taken as efficient intermediate representations for the purpose of analyzing the dynamic characteristics of speech feature component sequences along the time axis in a holistic manner [14].

Previous research has found that the salient linguistic information of the modulation spectra mostly resides in the range from 2 Hz to 8 Hz, rendering syllabic and phonetic temporal structure of speech, where the most prominent frequency components center around 4 Hz. These characteristics are closely related to human auditory perception. Furthermore, it has also been empirically revealed that in the modulation spectrum, different frequency components have different levels of contributions to the ASR performance [16].

III. Masking-based Speech Enhancement

A standard recipe of masking-based speech enhancement aims to separate a target speech signal from its background interference. Pioneering efforts in the line of research dates back to computational auditory scene analysis (CASA), which devises speech separation algorithms based on perceptual principles of auditory scene analysis and exploits grouping cues such as pitch and onset [17], [18]. T-F masking-based speech enhancement has recently emerged as one of the popular approaches, treating speech separation as a supervised learning problem. To this end, a suitable mask is thus estimated for suppressing noise components while retaining speech components in the T-F representation of a noisy signal. Celebrated instantiations of this approach include, but are not limited to, IBM, IRM and cIRM. Unlike most previous work on them that was devoted mainly to improving the perceptual quality of noisy speech signals, we instead contextualize and extend them for building a more robust ASR system.

Since it has been shown that IRM is considered amenable to real-time implementation and generalizes well to different kinds of T-F representations [13], we thus use it as the cornerstone method. Given a representation of the spectrum of frequencies of a noisy speech signal as it evolves over a time span, the computational loss of IRM is mathematically defined as follows:

$$\mathrm{IRM} = \left(\frac{\mathrm{S}(t,f)^2}{\mathrm{S}(t,f)^2+\mathrm{N}(t,f)^2}\right)^{\beta}. \quad (2)$$

where $\mathrm{S}(t,f)^2$ and $\mathrm{N}(t,f)^2$ denote the speech energy and the noise energy within a time-frequency (T-F) unit, respectively. The tunable parameter β scales the mask, and is commonly chosen to 0.5 in practice. With the square root, IRM preserves the speech energy with each T-F unit, under the assumption that $\mathrm{S}(t,f)$ and $\mathrm{N}(t,f)$ are uncorrelated. This assumption actually holds well in most realistic scenarios. Interestingly, with $\beta = 0.5$, Eq. (2), turns to closely resemble the square-root Wiener filter, which is the optimum estimator of the power spectrum. IRM is typically embodied in the form of a deep neural network whose parameters are estimated with the mean-square error (MSE) criterion. No content with directly applying IRM to obtain a separate speech spectrum at each time frame for robust ASR, we explore three novel extensions in this paper: 1) employing IRM to eliminate the noise effects from the Mel-frequency filter bank (denoted by FBANK) speech feature vector at each time frame; 2) leveraging IRM to normalize the intermediate representation (embodied in the modulation frequency domain) of the noise effects from the FBANK speech feature vector sequence of a noisy utterance in a holistic manner; and 3) treating IRM as a data augmentation method, conducted on the FBANK speech feature vectors of training utterances or their intermediate representations, to generate additional augmented data for increasing the diversity of training data. To the best of our knowledge, this work represents the first exploration of using IRM to normalize the intermediate representations of speech features, as well as treating IRM as a data augmentation method alternatively to expand the training data.

IV. Experimental Setup

A. Corpus and ASR Configuration

Our empirical experiments are conducted on Aurora-4 corpus, which is a subset of Wall Street Journal (WSJ) and designed to evaluate the robustness of ASR systems on a medium to large vocabulary continuous speech recognition task. Aurora-4 was composed of speech utterances recorded in a clean condition, which were further corrupted by different types of noise sources with varying SNR levels, in a range between 5 dB and 15 dB. In the clean-condition training setup, the number of utterances in the training set is 7,138 utterances from 83 speakers recorded using the primary microphone. Furthermore, the multi-condition training set also consists of 7,138 utterances and the same speaker information. One half of the utterances were recorded by the primary Sennheiser microphone and the other half were recorded using one of a number of different secondary microphones. Both of them include a combination of clean speech and speech corrupted by one out of six different noises (street traffic, train station, car, babble, restaurant, airport). The test sets are totally composed of 14 subsets, each of which contains 330 utterances contaminated with various types of environmental noise at different SNR levels. Although both 8 kHz and 16 kHz sampled speech utterances were provided in the Aurora-4 dataset, we only make use of 16-kHz sampled speech utterances for all experiments.

TABLE I.
CONTEXT SPECIFICATION OF TDNN EMPLOYED IN THE MASK PREDICTION NETWORK OF IRM.

Layer	Layer context	Total context
Layer-1	[t-2, t+2]	5
Layer-2	{t-1, t, t+1}	7
Layer-3	{t-1, t, t+1}	9
Layer-4	{t-3, t, t+3}	15
Layer-5	{t-3, t, t+3}	21
Layer-6	{t-6, t-3, t}	24

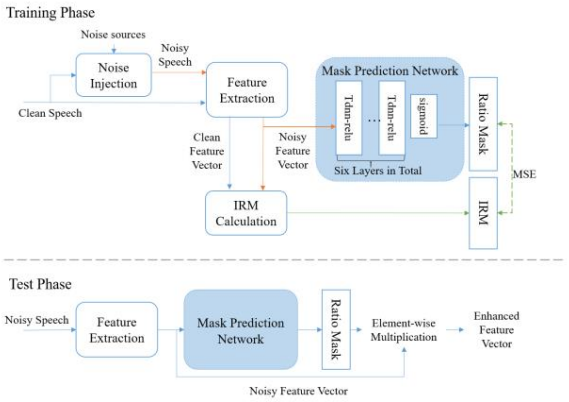

Figure 1: *A schematic depiction of IRM enhancement process.*

The acoustic models were configured with a modeling framework of cascaded deep neural networks and hidden Markov models (denoted by DNN-HMM for short), according to the commonly-used setup suggested in literature [19]. The DNN-HMM framework inherits advantages from the strong representation learning power of DNN and the sequential modeling ability of HMM. Furthermore, the speech feature vectors for ASR merely consist of only the static part of 80 Mel-frequency filter bank (denoted by FBANK) coefficients.

B. Mask Prediction Network

In this paper, IRM employs a mark prediction network to estimate a smoothed ratio mask for each noisy speech feature vector (e.g., FBANK) or the intermediate representation for a sequence of speech feature components for each dimension (i.e., the corresponding modulation spectrum).

In implementation, the network structure is realized with a six-layer time delay neural network (TDNN), where each layer consists of 512 hidden units with the ReLU activation function and the p-norm non-linearity regularization [20], except for that the last hidden layer uses a sigmoid activation function since the value of each element in IRM ranges from 0 to 1. The commonly-used MSE criterion is adopted as the cost function for parameter estimation and during the training procedure the Backstitch optimization method can be used [21].

When the input to IRM is the FBANK speech feature vectors, we further incorporate temporal context for TDNN modeling, rather than splicing together contiguous temporal windows of frames at each layer, which allows gaps between the frames. The TDNN configuration is illustrated in TABLE I. The input (80-dimensional FBANK) features with a frame-length of 25ms. The Layer-1 splices together frames t-2 through t+2 at the input layer (which we could write as context $\{-2, -1, 0, 1, 2\}$, or more compactly as $[t-2, t+2]$); and then the consequently layers have small temporal context centered at the current frame t. For example, at the frame t, the input to Layer 2 is the spliced output of the Layer 1, at frames $t-1$ and $t+1$. The notation $\{t-6, t-3\}$ means that we additionally splice together the input at the current frame minus 3 and the current frame minus 6.

The training utterances for estimating IRM were selected from the clean-condition set of Aurora 4; which were in turn used to generate the corresponding noisy-clean training pairs through random injection of noise made from the MUSAN dataset, which in total consisted of over 900 types of noise and 42 hours of music from various genres, as well as 60 hours of speech from twelve languages [22].

C. Experimental Procedure

A schematic depiction of the IRM enhancement process is diagrammed in Figure 1. In the training phase, we estimate the parameters of the mask prediction network with clean training utterances and their noise-injected counterparts (*cf.* Section 4.2). Specifically, the input to the network is a noisy feature vector and the desired output is the corresponding ideal ratio mask which is expected to faithfully restore it to its clean counterpart. In the test phase, we can use the resulting mask prediction network to obtain a ratio mask for an unseen noisy, or (even clean) speech feature vector, and subsequently apply element-wise multiplication operations on the speech feature vector components with their corresponding mask values. In a similar vein, the above-mentioned procedure can be modified to enhance the intermediate representations of speech features (in the modulation domain).

V. EXPERIMENTAL RESULTS

We report on the empirical results of our proposed IRM-based enhancement and data augmentation methods for ASR in terms of word error rate (WER). In the first set of experiments, we evaluate the performance levels of using IRM to enhance spectral representations (denoted by Spectrum), FBANK feature vectors (denoted FBANK) and intermediate representations of FBANK feature vectors (denoted by Modulation), respectively. Their corresponding results are shown in TABLE II, where the results of a baseline ASR system with a multi-condition training setting are listed for reference. From this table we can make at least two observations. First, IRM-based enhancement conducted on either the spectra or the FBANK feature vectors can, on average, lead to roughly 2.0% relative improvements over the baseline system. Furthermore, IRM-based enhancement conducted on intermediate representations of FBANK feature vector sequences seems to

TABLE II.
THE WORD ERROR RATE (%) RESULTS OF THE BASELINE SYSTEMS AND THE THREE VARIANTS OF THE IRM METHOD.

Enhancement Space	Test sets				Avg.
	Set A	Set B	Set C	Set D	
Baseline (no enhancement)	3.28	8.05	10.09	21.74	10.79
Spectrum	3.25	7.05	8.16	23.71	10.54
FBANK	3.60	7.24	8.92	22.49	10.56
Modulation	3.75	7.95	9.00	19.60	10.08

TABLE III.
THE WORD ERROR RATE (%) RESULTS OF DIFFERENT IRM-BASED DATA AUGMENTATION METHODS.

Augmentation	Test sets				Avg.
	Set A	Set B	Set C	Set D	
Spectrum	3.65	7.33	7.36	17.96	9.08
FBANK	3.78	6.98	7.94	17.34	9.01
Modulation	3.54	6.94	8.40	17.49	9.09

be obviously superior to the two former methods, especially for the test set D which contains both unseen environmental noise and channel distortions, and ultimately yield a relative improvement of 6.6% over the baseline system. This confirms the utility of performing such an enhancement operation in the modulation domain of feature vector sequences.

In the second set of experiments, we treat IRM as a data augmentation method, instead of using it to enhance both training and test speech utterances. This means that IRM is merely used in the training phase to generated enhanced representations of training speech utterances from their noisy counterparts. The enhanced representations then can be treated as augmented training data to be jointly used with the original multi-condition training data for training the acoustic model. Note also that the corresponding feature vectors of test speech utterances are kept unprocessed in the test phase. Again, we can conduct IRM on the spectral representations, the FBANK feature vectors and the intermediate representation of FBANK feature vectors as well. The WER results for these three variants of IRM-based data augmentation are depicted in TABLE III. Inspection of TABLE III reveals two noteworthy points. First, these three IRM-based data augmentation methods tend to perform on par with one another, and all of them also show superiority over the above IRM-based enhancement methods in terms of WER reductions, especially for the test sets C and D. Second, the best among them yields a relative improvement of 16.5% over the baseline system on average.

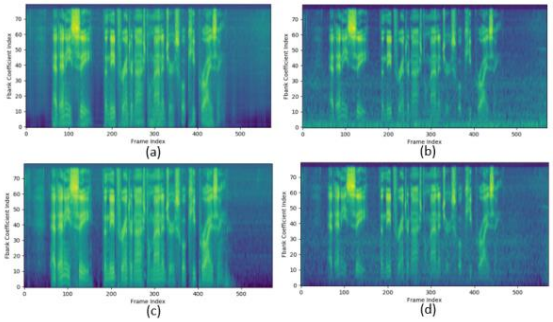

Figure 2: *Visual comparison of the adjoined FBANK feature vectors of (a) a randomly selected clean test utterance, (b) its noisy counterpart, (c) its noisy counterpart enhanced b conducting IRM directly on the FBANK feature vectors, and (d) its noisy counterpart enhanced by conducting IRM in the modulation domain of the FBANK feature vectors.*

Finally, as shown in Figure 2, we conduct visual inspection on the adjoined FBANK feature vectors of (a) a randomly selected clean test utterance, (b) its noisy counterpart, (c) its noisy counterpart enhanced by conducting IRM directly on the FBANK feature vectors, and (d) its noisy counterpart enhanced by conducting IRM in the modulation domain of the FBANK feature vectors. It is observed that the noisy utterance can be restored quite faithfully to its clean counterpart when IRM is directly conducted on the FBANK feature vectors. It instead seems less pronounced for enhancing the noisy utterance by conducting IRM in the modulation domain of the FBANK feature vectors, though such enhancement delivers slightly better ASR performance (*cf.* TABLE III).

VI. CONCLUSIONS

In this paper, we have proposed several novel methods to leverage IRM-based enhancement for robust ASR, which fall roughly into two categories: enhancement and data augmentation. Both these two categories of methods can result in considerable performance improvements over the strong baseline with a multi-condition training setting. As to future work, we plan to explore more sophisticated enhancement methods that can be leveraged to normalize speech feature vectors, or disparate intermediate representations of them, to generate augmented training data for further improving ASR robustness.

VII. EXPERIMENTAL RESULTS

This research is supported in part by ASUS AICS and the Ministry of Science and Technology (MOST), Taiwan, under Grant Number MOST 109-2634-F-008-006- through Pervasive Artificial Intelligence Research (PAIR) Labs, Taiwan, and Grant Numbers MOST 108-2221-E-003-005-MY3 and MOST 109-2221-E-003-020-MY3. Any findings and implications in the paper do not necessarily reflect those of the sponsors.

REFERENCES

[1] J. Li et al., "An overview of noise-robust automatic speech recognition," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(4), pp. 745-777, 2014.

[2] F. Weninger et al., "Speech enhancement with LSTM recurrent neural networks and its application to noise-robust ASR," Proceedings of LVA/ICA, pp. 91-99, 2015.

[3] D. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(10), pp. 1702-1726, 2018.

[4] S. J. Chen et al., "Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline," arXiv preprint arXiv:1803.10109, 2018.

[5] L. Mošner et al., "Improving noise robustness of automatic speech recognition via parallel data and teacher-student learning," Proceedings of ICASSP, pp. 6475-6479, 2019.

[6] P. Ghahremani et al., "Investigation of transfer learning for ASR using LF-MMI trained neural networks," Proceedings of ASRU, pp. 279-286, 2017.

[7] J. Chen et al., "Large-scale training to increase speech intelligibility for hearing-impaired listeners in novel noises," The Journal of the Acoustical Society of America, 139(5), pp. 2604-2612, 2016.

[8] A. Pandey and D. Wang, "A new framework for CNN-based speech enhancement in the time domain," IEEE/ACM Transactions on Audio, Speech and Language Processing, 27(7), pp. 1179-1188, 2019.

[9] N. Shah et al., "Time-frequency mask-based speech enhancement using convolutional generative adversarial network," Proceedings of APSIPA, pp. 1246-125, 2018.

[10] S. Srinivasan et al., "Binary and ratio time-frequency masks for robust speech recognition," Speech Communication, 48(11), pp. 1486-1501, 2006.

[11] Y. Wang et al., "On training targets for supervised speech separation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 22(12), pp. 1849-1858, 2014.

[12] D. S. Williamson et al., "Complex ratio masking for monaural speech separation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(3), pp. 483-492, 2016.

[13] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," Proceedings of ICASSP, pp. 7092–7096, 2013.

[14] J.-W. Hung et al., "Robust speech recognition via enhancing the complex-valued acoustic spectrum in modulation domain," IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(2), pp. 236-251, 2016.

[15] B.-C. Yan et al., "Exploring low-dimensional structures of modulation spectra for robust speech recognition," Proceedings of Interspeech, pp. 3637-3641, 2017.

[16] H. Hermansky, "Modulation spectrum in speech processing," Signal Analysis and Prediction, pp. 395-406, 1998.

[17] D. Wang and G. J. Brown, Computational Auditory Scene Analysis: Principles, Algorithms, and Applications, Wiley-IEEE Press, 2006.

[18] G. Hu and D. Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," IEEE Transactions on Audio, Speech, and Language Processing, 18(8), pp. 2067-2079, 2010.

[19] M. L. Seltzer et al., "An investigation of deep neural networks for noise robust speech recognition," Proceedings of ICASSP, 7398-7402, 2013.

[20] X. Zhang et al., "Improving deep neural network acoustic models using generalized maxout networks," Proceedings of ICASSP, pp. 215-219, 2014.

[21] Y. Wang et al., "Backstitch: counteracting finite-sample bias via negative steps," Proceedings of Interspeech, pp. 1631-1635, 2017.

[22] D. Snyder et al., "MUSAN: A music, speech, and noise corpus," arXiv preprint arXiv:1510.08484, 2015.

[23] D. S. Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," arXiv preprint arXiv:1904.08779, 2019