INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Ideal ratio mask estimation using deep neural networks for monaural speech segregation in noisy reverberant conditions

Xu Li[1,2]*, Junfeng Li*[1,2] *and Yonghong Yan*[1,2,3]

[1]Key Laboratory of Speech Acoustics and Content Understanding,
Institute of Acoustics, Chinese Academy of Sciences, Beijing 1001090, China
[2]University of Chinese Academy of Sciences, China
[3]Xinjiang Laboratory of Minority Speech and Language Information Processing

Abstract

Monaural speech segregation is an important problem in robust speech processing and has been formulated as a supervised learning problem. In supervised learning methods, the ideal binary mask (IBM) is usually used as the target because of its simplicity and large speech intelligibility gains. Recently, the ideal ratio mask (IRM) has been found to improve the speech quality over the IBM. However, the IRM was originally defined in anechoic conditions and did not consider the effect of reverberation. In this paper, the IRM is extended to reverberant conditions where the direct sound and early reflections of target speech are regarded as the desired signal. Deep neural networks (DNNs) is employed to estimate the extended IRM in the noisy reverberant conditions. The estimated IRM is then applied to the noisy reverberant mixture for speech segregation. Experimental results show that the estimated IRM provides substantial improvements in speech intelligibility and speech quality over the unprocessed mixture signals under various noisy and reverberant conditions.

Index Terms: speech segregation, deep neural networks, ideal ratio mask

1. Introduction

Monaural speech segregation has attracted much attention due to its widely potential applications such as automatic speech recognition (ASR), hearing aids design, and speech communication. It attempts to segregate the speech signal from a noisy mixture using only one microphone.

Recently, monaural speech segregation has been formulated as a supervised learning problem. In its simplest form, supervised speech segregation learns a mapping function from a mixture signal to a time-frequency (T-F) mask, and then uses the estimated mask to segregate the mixture signal. In many studies [1, 2], the T-F mask is typically set to the ideal binary mask (IBM). For the IBM [3, 4], a T-F unit is assigned 1, if the target signal is stronger than noise by a certain local criterion (LC). Otherwise, it is assigned 0. Motivated by the success of deep neural networks (DNNs) with more than one hidden layer, Wang $et\ al.$ [5] first introduced DNN to perform binary classification for speech segregation. Their DNN-based method significantly outperforms earlier segregation methods. Healy $et\ al.$ [6] has shown the DNN-based IBM estimator has demonstrated to improve intelligibility of noisy speech by hearing-impaired listeners.

The IBMs used and estimated in the aforementioned studies were all defined in anechoic conditions. In addition to noise, reverberation produces a temporal smearing of the speech signal and tends to fill in the spectral valleys, which drastically degrades speech intelligibility [7, 8]. As suggested by studies in room acoustics [9], reverberation is generally considered to consist of three parts: direct sound, early reflections, and late reverberation. While late reflections are detrimental to speech perception, it has shown that early reflections are beneficial for speech intelligibility [10]. Roman and Woodruff [11] proposed a novel approach for computing the IBM in reverberant conditions by regarding the direct path and early reflections of target signal as the desired signal, and showed that this new IBM processing yielded improvements over unsegregated signals. In the computation of this IBM, a fixed threshold of 50ms was used to distinguish early and late reflections [11], which was mainly motivated by the finding in [10] where the early reflections of up to 50ms have been found to benefit speech perception for normal-hearing and hearing-impaired listeners. Furthermore, Li $et\ al.$ [12] conducted a lot of experiments to find how the division between early and late reflections impacts on the intelligibility of the IBM-processed noisy reverberant speech. Results showed the the IBMs with different divisions between early and late reflections provided substantial improvements in speech intelligibility over the unprocessed mixture signals in all conditions tested, and there were small, but statistically significant, differences in speech intelligibility between the different IBMs in some conditions tested.

A recent study [13] on training targets for supervised speech segregation shows that the ideal ratio mask (IRM) performs better than the IBM in terms of speech quality and intelligibility in anechoic conditions. For the IRM [14, 15, 16], a T-F unit is assigned some ratio of target energy and mixture energy. Therefore, in this study the IRM is extend to reverberant conditions where the direct sound and early reflections of target speech are regarded as the desired signal. DNNs is employed to estimate the extended IRM in the noisy reverberant conditions. The estimated IRM is then applied to the noisy reverberant mixture for speech segregation. Experiments are conducted to examine the effect of noise and reverberation on the segregation results. Experimental results show that the estimated IRM provides substantial improvements in speech intelligibility and speech quality over the unprocessed mixture signals in all test conditions.

2. System description

An overview of the proposed system is illustrated in Fig.1. In the training stage, a complementary feature set is first extracted from the speech mixture, the feature set is then used to train a DNN. In the test stage, the complementary feature set extract-

http://dx.doi.org/10.21437/Interspeech.2017-549

ed from the test data is fed into the well-trained DNN to obtain an estimate of the target IRM. The estimated IRM is then employed for signal reconstruction. The detailed implementation is described in follows.

2.1. The IRM in reverberant conditions

The computing procedure of this extended IRM in reverberant conditions is briefly summarized below.

Similar to the computation of the IBM in reverberant conditions [11], the IRM in reverberant conditions is also computed from the cochleagram representation of target and noise signals. Specifically, the cochleagram is calculated from the outputs of a 64-channel gammatone filterbank with the center frequencies from 50 to 8000 Hz equally spaced on the equivalent rectangular bandwidth scale. The output of each filter in the filterbank is divided using 20-ms rectangular frames with 10-ms overlap into a set of time-frequency units, and the cochleagram corresponds to a two-dimensional response energy computed across all the time-frequency units. Suppose that $D(k,l)$ and $R(k,l)$ be the energy of the desired and residual signals in the kth frequency channel and the lth time frame, the IRM is then defined as

$$\mathrm{IRM}(k,l) = \sqrt{\frac{D(k,l)}{D(k,l)+R(k,l)}} \tag{1}$$

where the desired signal $D(k,l)$ consists of both the direct path and early reflections of the target signal, which is derived by convolving the target signal with the direct plus early impulse responses. The residual signal $R(k,l)$ is obtained by subtracting the desired signal $D(k,l)$ from the noisy reverberant mixture signal. It is clear that the residual signal $R(k,l)$ includes both the late reverberant target signal and the reverberant noise signal. In the computation of the IRM, a fixed threshold of 50ms is used to distinguish early and late reflections [11].

2.2. Features

In the IRM estimation, a complementary feature set is extracted from the mixture signal; these features are normalized to zero mean and unit variance in every dimension. Specifically, the complementary feature set consists of the 15-dimensional amplitude modulation spectrogram (AMS), 13-dimensional relative spectral transformed perceptual linear prediction (RASTA-PLP), 31-dimensional mel-frequency cepstral coefficients (MFCC), 64-dimensional gammatone frequency features (GF) extracted from each frame of the noisy reverberant mixture signal, and their delta (123 dimensions) components. Therefore, the feature dimension for each time frame is 246. These features have been employed previously in machine speech segregation, and their descriptions can be found in [17].

2.3. DNN training

In the training stage, a DNN is employed to learn the functions that maps the extracted features to the extended IRM. Specifically, the DNN has four hidden layers and 1024 rectified linear units (ReLU)[18] in each hidden layer. Since the target is in the range [0,1], the sigmoid function is used in the output layer. The DNN is trained using backpropagation algorithm and mean square error (MSE) as the loss function. A minibatch size of 1024 and dropout ratio of 0.2 are used. The Adagrad algorithm [19] is used to adjust the learning rate. The maximum epoch is set to 300. To incorporate context, five frames of features (two to each side of the current frame) are used to simultaneously predict five frames of the IRM. By incrementing frame-by-frame, each frame of the IRM is estimated five times and the average is taken as the final estimate for each frame [13].

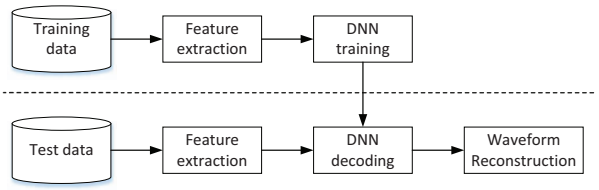

Figure 1: *The proposed system. Above the dash line is the training stage. Below the dash line is the test stage.*

2.4. Waveform reconstruction

In the test stage, the complementary feature set extracted from the test data is fed into the trained DNN to obtain the estimated IRM. Finally, the estimated IRM is applied to the mixture cochleagrams in a synthesis step to generate the segregated target speech in the time domain [4].

3. Experiments and Results Analysis

3.1. Experimental setup

In the speech segregation experiments, the TIMIT corpus [20] were adopted as the speech material. 600 utterances randomly chosen from the TIMIT training set were used as the training utterances. 100 utterances randomly chosen from the TIMIT training set (without overlapping sentences in the training utterances) were used as the validation utterances.

To generate reverberant conditions, room impulse responses (RIRs) were selected from the database provided by Jeub *et al.* [21] in which the RIRs were measured in four typical rooms with different dimensions and acoustic properties. Two RIRs corresponding to the farthest and the nearest distances between the sound source and the recording microphone in two rooms (meeting and lecture) were adopted to generate the training data. One RIR (not the same used in the training set) in each of the two rooms (meeting and lecture) was adopted to generate the validation data. All the RIRs were stored with a sampling frequency of 48 kHz. Consequently, there were 600×2(Rooms)×2(RIRs) = 2400 reverberant utterances in the training set, 100×2(Rooms)×1(RIR) = 200 reverberant utterances in the validation set.

Two types of noises (babble and the speech-shaped noise(SSN)) were employed to obtained the training set and the validation set, with babble noise being nonstationary and SSN stationary. The SSN was obtained by passing white noise through the average spectrum of the speech database. Both noises last about 4 min, and were devided into two parts: the first 3 min was used for training and validation and the remaining noise was used for testing. Thus there was no noise overlap between training/validation and test data. Both speech and noise signal were first upsampled to 48 kHz and then convolved with each RIR, which was followed by being downsampled to 16 kHz. The long-term mean square level of reverberant target speech in the absence of noise was fixed across all sentences in all conditions tested. The reverberant noise signal was scaled accordingly to reach the desired signal-to-noise (SNR) level and then added to the reverberant target signal at the SNRs

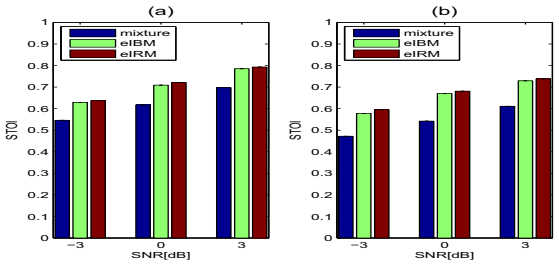

Figure 2: *Mean STOI of the masked speech in the speech-shaped noise with two reverberant conditions: (a) meeting room; (b) lecture room.*

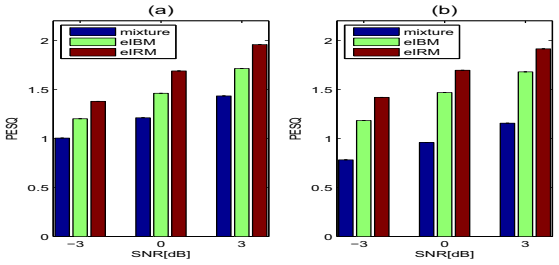

Figure 3: *Mean PESQ of the masked speech in the speech-shaped noise with two reverberant conditions: (a) meeting room; (b) lecture room.*

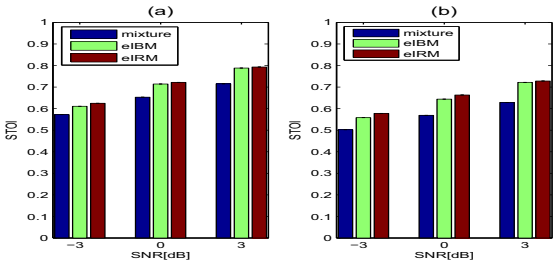

Figure 4: *Mean STOI of the masked speech in the babble noise with two reverberant conditions: (a) meeting room; (b) lecture room.*

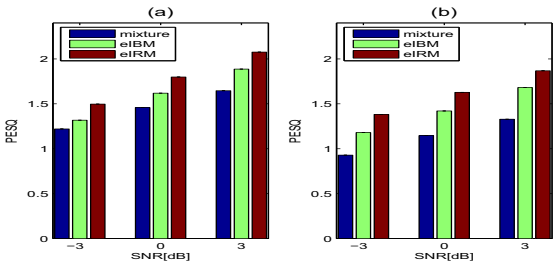

Figure 5: *Mean PESQ of the masked speech in the babble noise with two reverberant conditions: (a) meeting room; (b) lecture room.*

of -3, 0 and 3 dB, generating the noisy reverberant mixtures. Finally, there were 2400×3(SNRs)×2(noises)=14400 utterances for training, and 200×3(SNRs)×2(noises)=1200 utterances for validation.

To generate the test set, 50 utterances randomly chosen from the test set of the TIMIT corpus were used as the test utterances. Three RIRs (without overlapping RIRs used in the training/validation set) in the three rooms (meeting, office and lecture) were used to obtain the reverberant signal. Three noises, which consisted of babble, SSN and factory, were employed to generate the test set. Both the speech and noise signal were first convolved with each RIR and then mixed at the SNRs of -3, 0 and 3 dB to generate the noisy reverberant mixtures. Finally, there were 50×3(RIRs)×3(SNRs)×3(noises)=1350 test sentences. In order to test the generality of the trained DNN, a new RIR (office) and a new noise type (factory) unseen in the training set were used to generate the test set.

Short-time objective intelligibility (STOI) [22] and perceptual evaluation of speech quality (PESQ) [23] were employed to evaluate speech intelligibility and quality, respectively. For evaluation metrics, higher values mean the better segregation quality.

3.2. Results Analysis

In the experiments, we also present results obtained using an IBM estimation algorithm. In the estimation algorithm, the IBM defined in reverberant conditions [11] is used as the target during training. It uses DNNs trained similarly to the proposed IRM estimation system except that the IBM is used as the target.

First we examine the results in matched conditions. Figs 2-3 show the STOI and PESQ results for unprocessed mixtures (mixture), the processed speech signals by the estimated IBM (eIBM) and the processed speech by the estimated IRM (eIRM) in the two reverberant conditions (meeting,lecture) with the SSN noise. The mean scores are computed for the 50 test utterances in each SNR conditions. It is obvious that both the speech intelligibility and quality increase when the eIBM and eIRM are used to segregate the mixture signals. In comparison of the STOI and PESQ in the two reverberant conditions, it is shown that the speech intelligibility and quality of the mixtures degrade as the amount of reverberation increases. In addition, the STOI obtained by the eIBM and eIRM are almost the same, while the eIRM obtains higher PESQ scores than the eIBM. The results are in agreement with the results obtained in [13]. Figs 4-5 show the STOI and PESQ results in the two reverberant conditions (meeting, lecture) with the babble noise. Similar to the results in the Figs 2-3, the eIBM and eIRM increase the STOI and PESQ compared with the mixtures, and also the STOI and PESQ degrade as the amount of reverberation increases.

Comparing the Fig.2 (a) with the Fig.4 (a), which are the mean STOI in the meeting condition with the two noises (SSN, babble), the results show that the STOI of the mixture signals in the SSN condition is slightly worse than that in the babble noise condition. After the masked processing, both eIBM and eIRM provide much more gains in STOI for the SSN condition than the babble noise condition. The reason is partly due to that the SSN is stationary and the trained DNN could model it well. The STOI in the Fig.1 (b) and Fig.3 (b) show similar results.

Next we examine the segregation results in unmatched conditions. Figs 6-7 show the STOI and PESQ results in the two reverberant conditions (meeting, lecture) with the new factory noise. Figs 8-9 show the STOI and PESQ results in the new reverberant condition (office) with the two seen noise (SSN, babble). It is shown that the eIBM and eIRM improve the speech intelligibility and quality compared with the mixtures in the new office reverberant condition while do not improve the speech in-

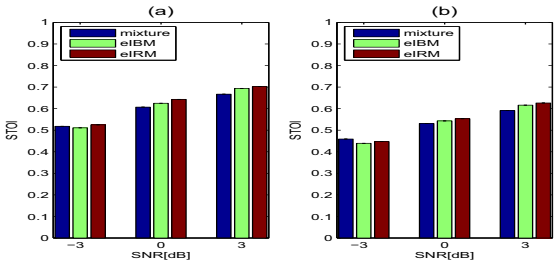

Figure 6: *Mean STOI of the masked speech in the factory noise with two reverberant conditions: (a) meeting room; (b) lecture room.*

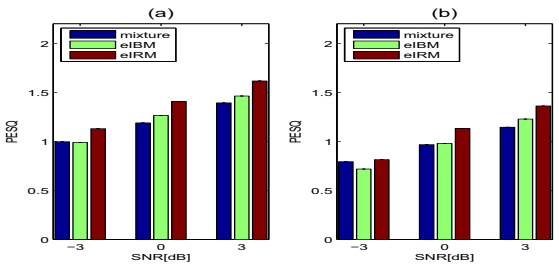

Figure 7: *Mean PESQ of the masked speech in the factory noise with two reverberant conditions: (a) meeting room; (b) lecture room.*

telligibility and quality much in the new noise condition. The degradation of the results is mainly attributed to the dominant impact of the new noise than the new reverberation, since the reverberation time of the office RIR is not large. The spectral characteristic of the factory noise is much different from the spectral characteristic of the SSN and babble noise in the training set. Thus the trained DNN could not model the factory noise well.

4. Conclusions

In this paper, we extend the IRM to the reverberant conditions where the direct sound and early reflections of target speech are regarded as the desired signal. DNNs is employed to estimate the extended IRM in the noisy reverberant conditions for speech segregation. Experiments show that the estimated IRM improves the performance in the noisy reverberant conditions over the mixture signals. In addition, the estimated IRM shows more performance improvement than the estimated IBM for the speech quality. For the future work, we expect to handle the generalization problem of the proposed algorithm in the unseen noisy reverberant conditions.

5. Acknowledgements

This work is partially supported by the National 973 Program (2013CB329302), the National Natural Science Foundation of China (Nos. 10925419, 90920302, 61072124, 11074275, 11161140319, 91120001), the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant Nos. XDA06030100, XDA06030500) and the National 863 Program (No. 2012AA012503).

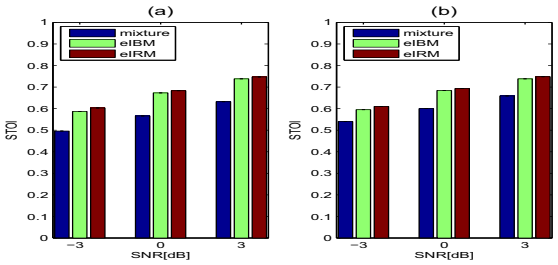

Figure 8: *Mean STOI of the masked speech in the office reverberant condition with two noise conditions: (a) SSN; (b) babble noise.*

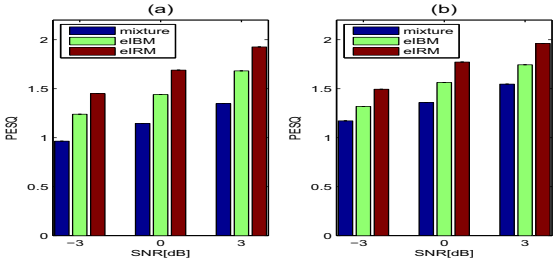

Figure 9: *Mean PESQ of the masked speech in the office reverberant condition with two noise conditions: (a) SSN; (b) babble noise.*

6. References

[1] K. Han and D. Wang, "A classification based approach to speech segregation," *The Journal of the Acoustical Society of America*, vol. 132, no. 5, pp. 3475–3483, 2012.

[2] G. Kim, Y. Lu, Y. Hu, and P. C. Loizou, "An algorithm that improves speech intelligibility in noise for normal-hearing listeners," *The Journal of the Acoustical Society of America*, vol. 126, no. 3, pp. 1486–1494, 2009.

[3] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech separation by humans and machines*. Springer, 2005, pp. 181–197.

[4] D. Wang and G. J. Brown, *Computational auditory scene analysis: Principles, algorithms, and applications*. Wiley-IEEE Press, 2006.

[5] Y. Wang and D. Wang, "Towards scaling up classification-based speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 7, pp. 1381–1390, 2013.

[6] E. W. Healy, S. E. Yoho, Y. Wang, and D. Wang, "An algorithm to improve speech recognition in noise for hearing-impaired listeners," *The Journal of the Acoustical Society of America*, vol. 134, no. 4, pp. 3029–3038, 2013.

[7] P. Assmann and Q. Summerfield, "The perception of speech under adverse conditions," in *Speech processing in the auditory system*. Springer, 2004, pp. 231–308.

[8] J. Meyer, L. Dentel, and F. Meunier, "Speech recognition in natural background noise," *PloS one*, vol. 8, no. 11, p. e79279, 2013.

[9] H. Kuttruff, *Room acoustics*. Crc Press, 2016.

[10] J. Bradley, H. Sato, and M. Picard, "On the importance of early reflections for speech in rooms," *The Journal of the Acoustical Society of America*, vol. 113, no. 6, pp. 3233–3244, 2003.

[11] N. Roman and J. Woodruff, "Intelligibility of reverberant noisy speech with ideal binary masking," *The Journal of the Acoustical Society of America*, vol. 130, no. 4, pp. 2153–2161, 2011.

[12] J. Li, R. Xia, Q. Fang, A. Li, J. Pan, and Y. Yan, "Effect of the division between early and late reflections on intelligibility of ideal binary-masked speech," *The Journal of the Acoustical Society of America*, vol. 137, no. 5, pp. 2801–2810, 2015.

[13] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1849–1858, 2014.

[14] S. Srinivasan, N. Roman, and D. Wang, "Binary and ratio time-frequency masks for robust speech recognition," *Speech Communication*, vol. 48, no. 11, pp. 1486–1501, 2006.

[15] A. Narayanan and D. Wang, "Ideal ratio mask estimation using deep neural networks for robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7092–7096.

[16] C. Hummersone, T. Stokes, and T. Brookes, "On the ideal ratio mask as the goal of computational auditory scene analysis," in *Blind Source Separation*. Springer, 2014, pp. 349–368.

[17] J. Chen, Y. Wang, and D. Wang, "A feature study for classification-based speech separation at low signal-to-noise ratios," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1993–2002, 2014.

[18] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[19] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. Jul, pp. 2121–2159, 2011.

[20] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, 1993.

[21] M. Jeub, M. Schafer, and P. Vary, "A binaural room impulse response database for the evaluation of dereverberation algorithms," in *Digital Signal Processing, 2009 16th International Conference on*. IEEE, 2009, pp. 1–5.

[22] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 7, pp. 2125–2136, 2011.

[23] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (pesq), an objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs," *ITU-T Recommendation*, vol. 862, 2001.