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

Multi-target Ensemble Learning for Monaural Speech Separation

Hui Zhang, Xueliang Zhang, Guanglai Gao

Department of Computer Science, Inner Mongolia University, China
alzhu.san@163.com, {cszxl, csggl}@imu.edu.cn

Abstract

Speech separation can be formulated as a supervised learning problem where a machine is trained to cast the acoustic features of the noisy speech to a time-frequency mask, or the spectrum of the clean speech. These two categories of speech separation methods can be generally referred as the masking-based and the mapping-based methods, but none of them can perfectly estimate the clean speech, since any target can only describe a part of the characteristics of the speech. However, the estimated masks and speech spectrum can, sometimes, be complementary as the speech is described from different perspectives. In this paper, by adopting an ensemble framework, a multi-target deep neural network (DNN) based method is proposed, which combines the masking-based and the mapping-based strategies, and the DNN is trained to jointly estimate the time-frequency masks and the clean spectrum. We show that as expected the mask and speech spectrum based targets yield partly complementary estimates, and the separation performance can be improved by merging these estimates. Furthermore, a merging model trained jointly with the multi-target DNN is developed. Experimental results indicate that the proposed multi-target DNN based method outperforms the DNN based algorithm which optimizes a single target.

Index Terms: speech separation, multi-target, ensemble learning

1. Introduction

Speech separation is helpful to improve speech intelligibility [1] or the accuracy of the automatic speech recognition (ASR) [2]. When only one speaker is of interest, it aims to extract the speech signal from the interfering background noises. Recently, great progresses have been achieved by solving the problem in a supervised learning framework. Given the noisy speech and based on the selected training targets, a machine is trained to map from the input features to the output targets. Depending on the training targets, supervised speech separation methods can be categorized as (a) masking-based and (b) mapping-based approaches.

Masking-based approaches employ an ideal time-frequency (T-F) mask as the computational target, and the commonly used ideal masks include the ideal binary mask (IBM) [3] and the ideal ratio mask (IRM) [4]. The IBM value of a T-F unit is assigned to 1 if the local signal-to-noise ratio (SNR) is above a local criterion, and 0 otherwise, hence non-zero IBM units indicate the target speech dominance. Early learning-based IBM estimation methods include the Gaussian mixture models (GMM) based method [1], support vector machines (SVM) based method [5], and multilayer perceptron (MLP) based method [6] which adopts one hidden layer in the neural network. Recently the deep neural network (DNN) consisting of two or more hidden layers have been adopted [7], and the separation performance is significantly improved because of the superior model representation capability of DNN. The IRM of one T-F unit is defined as the ratio between the powers of the target signal and mixture, and taking the IRM rather than the IBM as target leads to a better speech quality [7]. In [7], a DNN based IRM estimation method has been proposed.

On the other hand, mapping-based methods directly predict the clean spectrum from the mixture. A typical DNN and mapping-based method in this category is developed in [8], and it is extended by [9] in which the global variance equalization, dropout, and noise-aware training are exploited to improve the generalization ability to new speakers. In [10], in addition to estimating the target speech, the DNN is trained to jointly estimate the speech and noise.

The targets can only model the speech characteristics from a certain perspective. Thus the masking-based and mapping-based methods perform differently on speech separation. According to [11], mapping-based methods are more robust to the snr variation, while the masking methods can use the training data more efficiently, because they explore the mutual information between target speech and interfering noise. besides the t-f units in which the target can be well estimated by either both or none of these two categories, there always will be some "gray areas", in which the target can only be accurately estimated by one of the two categories. In other words, the masking-based and mapping-based methods provide complementary estimation results for the t-f units in the gray areas.

In this paper, we integrate the masking- and mapping-based targets into a single DNN, and propose an ensemble learning framework to combine complementary estimation results for better speech separation. A multi-target DNN is first trained, which jointly predicts both the clean speech spectrogram and the T-F masks including the IBM and IRM, and an output averaging strategy is utilized to improve the separation performance. In order to merge the DNN outputs which correspond to different targets more effectively, an additional one-hidden-layer MLP is further developed. In the final step, the multi-target DNN and the merging MLP are jointly trained to achieve overall optimization. By conducting experiments in different noisy conditions, we demonstrate the effectiveness of the proposed method over other DNN based speech separation methods.

2. System description

2.1. System overview

The core of the proposed system is a multi-target DNN, as illustrated in Fig. 1. With the multi-target DNN, the input features are mapped to three different output targets: the clean speech spectrogram, the IBM and the IRM.

The multi-target DNN learns to estimate three different targets at the same time. Therefore it is able to model the commonalities and differences across different targets. This

http://dx.doi.org/10.21437/Interspeech.2017-240

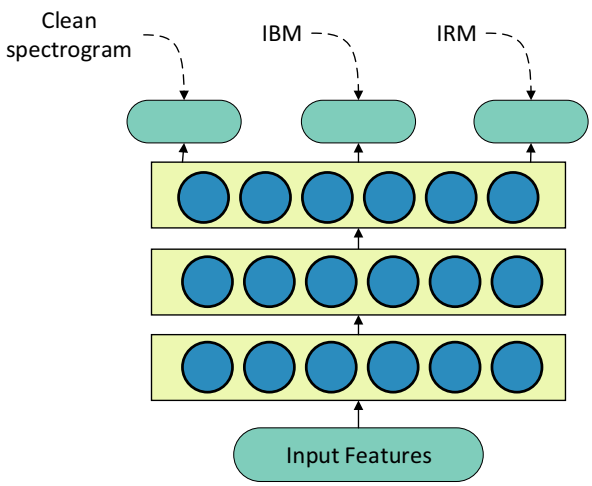

Figure 1: *Schematic diagram of the multi-target DNN, where dashed lines indicate the training targets.*

method can improve the learning efficiency and prediction accuracy for the task-specific models, when compared to training the models separately [12].

2.2. Inputs and outputs

We compute the input features based on the short time Fourier transform (STFT) of the mixture signal. Under the sampling frequency of 16 kHz, first the STFT is obtained using the 320-point (20 ms) hamming window with 50% overlap. As the STFT is conjugate symmetric, in each frame, a preliminary feature vector is formed using the amplitude of only the first 161 STFT coefficients. Then the vector is cubic-rooted and normalized to be zero-mean and unit-variance. The STFT based feature vectors for the DNN are finally generated by concatenating the feature vector of current frame with those of the previous 2 and subsequent 2 frames, thus a $161 \times 5 = 805$ dimensional feature vector is obtained for each frame.

Now let us consider the training targets. For each frame, as we want to recover the clean speech, the clean speech spectrogram is simply the 161-dimensional STFT vector of the clean speech signal without cubic-rooting and normalization. The IBM and IRM based target vectors are also 161-dimensional, whose elements are defined respectively as in (1) and (2):

$$IBM(t,f) = \begin{cases} 1 & \text{if } |S(t,f)|^2 > |N(t,f)|^2 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

$$IRM(t,f) = \sqrt{\frac{|S(t,f)|^2}{|S(t,f)|^2 + |N(t,f)|^2}}, \quad (2)$$

Where $S(t,f)$ and $N(t,f)$ denote the STFT of speech and noise in the T-F unit (t,f), respectively.

2.3. Model configurations

The model is trained to minimize a mean square error (MSE) based loss function formulated by equally weighting these three different training targets. In the output layer, the output transformation function is chosen as the rectified linear units (ReLU) for the clean speech spectrogram based target, since the amplitude spectrogram are all greater than 0, and as sigmoid for the IBM/IRM based targets since both $IBM(t,f)$ and $IRM(t,f)$ are in $[0,1]$. The DNN has three hidden layers containing 1024 ReLU nodes.

2.4. Outputs merging

With the three estimated targets of the multi-target DNN, the mapping based, IBM based and IRM based estimations of the separated speech amplitude spectrogram, can be individually obtained. We denote them as $\hat{S}_M(t,f)$ (mapping based), $\hat{S}_B(t,f)$ (IBM based) and $\hat{S}_R(t,f)$ (IRM based), respectively. $\hat{S}_M(t,f)$ is simply the estimated clean amplitude spectrogram in the output layer, and $\hat{S}_B(t,f)$ and $\hat{S}_R(t,f)$ are obtained by applying the estimated IBM and IRM on the mixture signal, which is done by conducting the element-wise multiplication between the noisy spectrogram and the mask. To get the final estimation, the three estimations are merged by taking the average:

$$\hat{S}_{AVG}(t,f) = \frac{\hat{S}_M(t,f) + \hat{S}_B(t,f) + \hat{S}_R(t,f)}{3}, \quad (3)$$

where the subscript "AVG" means exploiting the averaging strategy.

We should note the merging strategy in (3) is not guaranteed to be optimal since equal importance is assigned to each of the three estimations, but the correctness of each estimation is unknown in advance. Here, a better merging method based on MLP learning is proposed, by which the final estimation is expressed as a function of the above three preliminary estimations:

$$\hat{S}_{MLP}(t,f) = g\left(\hat{S}_M(t,f), \hat{S}_B(t,f), \hat{S}_R(t,f)\right), \quad (4)$$

where $g(\cdot)$ is the transformation function represented by the MLP. The MLP has only one hidden layer consisting of 1600 ReLU nodes, and takes the MSE between the estimation and target as the loss function. By using the preliminary estimates as inputs and the real clean amplitude spectrogram as the target, the MLP is trained to yield a final amplitude spectrogram estimation from the multi-target DNN outputs and the noisy amplitude spectrograms. Normalization is not applied to the inputs which make it easy to connect it to the multi-target DNN.

The diagram of the system is illustrated in Fig. 2. In fact, using MLP can be seen as adding additional layers to the multi-target DNN, and the whole system can be regarded as a new DNN, but the intermediate connections between nodes are more meticulously designed and trained. We can also notice that the noisy amplitude spectrogram is needed to generate the inputs of the MLP, this is equivalent to adding a connection between the input layer to the hidden layer containing the multi-target outputs. In this way, by concatenating the multi-target DNN and MLP and forming a new DNN system, a better performance can be achieved.

3. Experiments

3.1. Dataset and Evaluation

We construct a dataset with the IEEE corpus [13] as speech and the NOISE-92 [14] database as noise. The IEEE corpus consists of 720 utterances from a single male. The NOISE-92 has 15 noise types, with each recording approximately 4 minutes long.

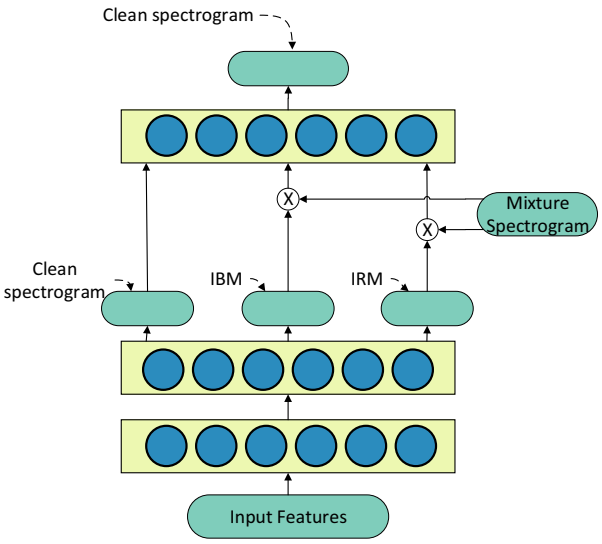

Figure 2: *Schematic diagram of the proposed joint model, where dashed lines indicate the training targets.* $\bigotimes$ *denotes the element-wise multiplication.*

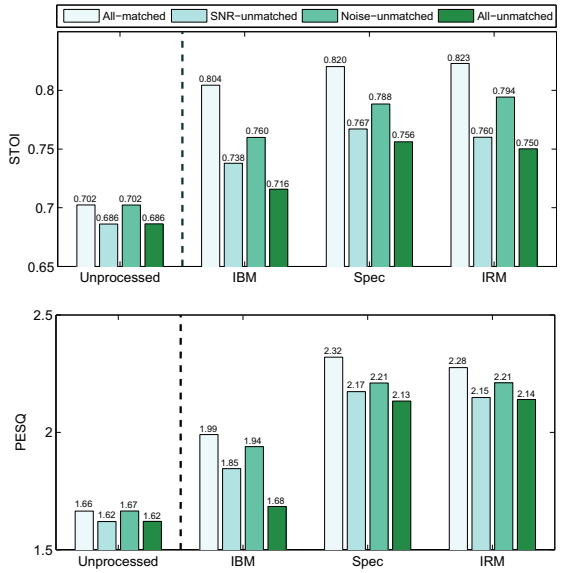

Figure 3: *Comparison on different single target systems.*

All signals in the generated dataset are re-sampled to 16 kHz sampling rate.

A training set is collected by firstly randomly selecting 500 speech utterances and then mixing them with 10 randomly selected types of noise at SNRs = $\{-3, 0, 3\}$ dB. Then totally 500 (speech) $\times$ 10 (noise) $\times$ 3 (SNR) = 15000 mixtures are generated in the training set. The development set utilizes another 60 speech utterances and adopts the same noise data as in the training set. In the test set, 60 utterances are randomly selected from the remaining subset of the speech corpus, and are contaminated with all 15 noise types at SNRs = $\{-6, -3, 0, 3, 6\}$ dB, resulting in $60 \times 15 \times 5 = 4500$ mixtures. In the experiment, to test the generalization ability, for each selected noise recording, the first half the signal is used in the training and development set, and the second half is used in the test set.

We report the performance under the all-matched, SNR-unmatched, noise-unmatched and all-unmatched conditions. We can note that 5 noise types and two SNRs = $\{-6, 6\}$ dB are not included in the training set. The all-matched case means that both the tested SNRs and noise types are present in the training set, while the all-unmatched case means the totally opposite setup. The other two possible combinations in terms of noise type and SNR are denoted as SNR-unmatched and noise-unmatched respectively.

We evaluate the speech separation performance in terms of short-time objective intelligibility (STOI) [15] and perceptual evaluation of speech quality (PESQ) [16]. Both metrics are widely used in the speech separation research. STOI is a standard objective metric for speech intelligibility by computing the correlation of short-time temporal envelopes between the clean and separated speech. The STOI varies in [0, 1], and a higher value indicates the better speech intelligibility. PESQ measures speech quality by computing disturbance between the clean speech and the separated speech using cognitive modeling, which ranges in [-0.5, 4.5], and is large if the speech quality is high.

3.2. Results

Since the proposed method is based on multi-target DNN, three baseline systems based on single-target DNN are developed. These baseline systems use the clean amplitude spectrogram, IBM and IRM as the target, and are denoted as "spec", "IBM" and "IRM" respectively for simplicity. Each baseline DNN model contains three 1024-node ReLU hidden layers, which is the same as the multi-target DNN in the proposed method. In the output layer, the "spec" DNN uses ReLU as the transformation function, and the "IBM" and "IRM" DNNs use the sigmoid function instead. All models are trained 200 epochs with the Adam optimizer [17] and use the development set for early stopping control.

The separation performances of these baselines are summarized in Fig. 3. We can see that all baseline methods can improve the speech intelligibility and quality. The IBM based baseline has significantly higher STOI, but only improves slightly on PESQ. The IRM-based and the mapping-based methods show comparable results. In detail, the mapping-based method can achieve better PESQ, and shows better generalization ability when the SNR changes. The IRM-based method also shows a slightly better generalization ability on the noise type variation.

The separation performance of each estimated target of the proposed multi-target DNN is shown in Fig. 4, and compared with the best results of the single target DNN under each condition. It can be observed that the outputs from the proposed multi-target DNN outperform the corresponding single target model.

Finally, we compare the merging performance and the results are shown in Fig. 5. The vertical axis is zoomed in for clarity. It is clear that generally merging improves the separation performance, and this shows the effectiveness of multi-target learning. Compared with the single-target based methods, the simple averaging merging strategy yields better results, and merging the outputs with fixed multi-target DNN using the MLP further increases the STOI and PESQ. By using a joint model consisting of a multi-target DNN and MLP, the

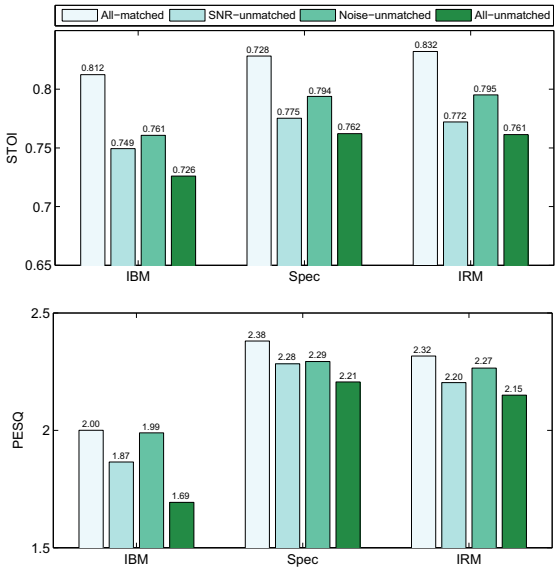

Figure 4: *Comparison on different outputs of the proposed multi-target DNN.*

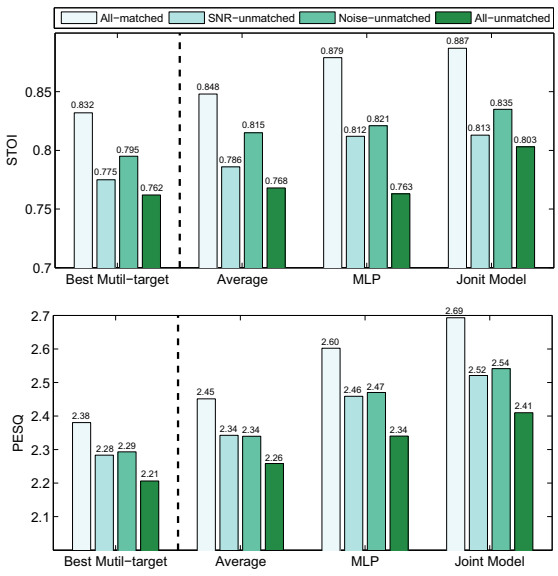

Figure 5: *Comparison on merging methods.*

system achieves the best performance amongst all evaluated systems. The parameter size of the comparison models shown in Fig. 5 are comparable, because we remove one hidden layer from the multi-target DNN when it is working together with a MLP.

4. Conclusions

In this work, we combine the masking- and mapping-based speech separation methods into a multi-target DNN, and estimate both T-F masks and the clean amplitude spectrogram. The multi-target DNN outperforms any of the model trained with a single target. The separation performance can be improved by further merging the multi-target DNN's outputs, because the estimated time-frequency masks and the clean spectrogram are partly complementary. A MLP based merging method is proposed, and a new DNN and MLP based model is developed to jointly train the DNN and MLP. Experiments show that the multi-target DNN outperforms the single-target based model, and the proposed joint model obtains the best separation performance.

5. Acknowledgments

This research was supported in part by the China national nature science foundation (No. 61365006, No. 61263037).

6. References

[1] G. Kim and P. Loizou, "Improving speech intelligibility in noise using environment-optimized algorithms," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 18, no. 8, pp. 2080–2090, 2010.

[2] Y. Shao, S. Srinivasan, Z. Jin, and D. Wang, "A computational auditory scene analysis system for speech segregation and robust speech recognition," *Computer Speech & Language*, vol. 24, no. 1, pp. 77–93, 2010.

[3] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

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

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

[6] Z. Jin and D. Wang, "A supervised learning approach to monaural segregation of reverberant speech," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 625–638, 2009.

[7] 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.

[8] Y. Xu, J. Du, L. Dai, and C. Lee, "An experimental study on speech enhancement based on deep neural networks," *IEEE Signal Processing Letters*, vol. 21, no. 1, pp. 65–68, 2014.

[9] J. Du, Y. Tu, Y. Xu, L. Dai, and C.-H. Lee, "Speech separation of a target speaker based on deep neural networks," in *2014 12th International Conference on Signal Processing (ICSP)*, Oct 2014, pp. 473–477.

[10] Y. Tu, J. Du, Y. Xu, and L. Dai, "Speech separation based on improved deep neural networks with dual outputs of speech features for both target and interfering speakers," in *International Symposium on Chinese Spoken Language Processing*, 2014, pp. 250 – 254.

[11] X. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 5, pp. 967–977, May 2016.

[12] R. Caruana, "Multitask learning," in *Learning to learn*. Springer, 1998, pp. 95–133.

[13] "IEEE recommended practice for speech quality measurements," in *IEEE Trans. Audio and Electroacoustics*, 1969, pp. 225–246.

[14] A. Varga and H. J. Steeneken, "Assessment for automatic speech recognition II: NOISEX-92: a database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication*, vol. 12, no. 3, pp. 247–251, 1993.

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

[16] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *icassp*, 2001, pp. 749–752.

[17] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.