Data-Dependent Ensemble of Magnitude Spectrum Predictions for Single Channel Speech Enhancement

Pasi Pertilä
Faculty of Information Technology and Communication Sciences
Tampere University, Tampere, Finland
pasi.pertila@tuni.fi

***Abstract*—Applying a predicted time-frequency mask to the noisy speech spectrogram and directly predicting the clean speech magnitude spectrogram are two common deep learning-based speech enhancement approaches. Ensemble techniques such as averaging and neural network-based fusion of magnitude spectra obtained with these two approaches have been shown to improve the objective perceptual quality of speech using synthetic mixtures of data. This work generalizes the averaging ensemble approach by proposing neural network layers to predict time-frequency varying weights for the combination of the two magnitude spectra obtained by time-frequency masking and by direct prediction. In order to combine the best individual magnitude spectrum estimates, the proposed weight prediction layers are trained after the time-frequency mask and magnitude spectrum networks layers have been separately trained for their corresponding objectives and their weights have been fixed. Using the publicly available CHiME3-challenge data, which consists of both simulated and real speech recordings in everyday environments with noise and interference, the proposed approach leads to significantly higher noise suppression in terms of segmental source-to-distortion ratio over the alternative approaches. In addition, the approach achieves similar improvements in the average objective instrumentally measured intelligibility scores with respect to the best achieved scores.**

I. Introduction

Speech enhancement has been approached through different strategies in the past such as spectral subtraction and Wiener filtering [1], [2]. In the time-frequency masking (TFM) approach the noisy input signal is multiplied with a predicted mask to attenuate the time-frequency regions affected by noise. A neural network can be trained to learn a time-frequency dependent mask in order to remove noise and interference from the target speech. Many mask variants have been developed, ranging from ideal binary masking (IBM) [3] to continuous-valued masking such as ideal-ratio-masking (IRM) [4], phase-sensitive [5], and complex-valued masks [6].

Another strategy to enhance the signal is direct magnitude spectrum (MS) prediction of the clean MS from the noisy MS by a trained neural network [7]. The signal approximation (SA) approach predicts the TFM but defines the loss function using the masked noisy input and the clean MS [8]. A comparison [9] with instantaneous synthetic mixtures suggests that the TFM can outperform the MS prediction approach in low SNR conditions with various types of added interference signals in terms of objective perceptual quality and noise suppression.

In addition to investigating training targets, different input features have been studied for improving performance for TFM prediction in [10], [11]. Most of all, large performance benefits have been observed by changing the network structure from a traditional neural network to a deep neural network (DNN) with several stacked layers (often with frame stacking) or to a recurrent neural network (RNN) (see e.g., [8]) with a specific type of memory cell such as long short-term memory (LSTM) [12] or gated recurrent unit (GRU) [13]. These memory cells can alleviate issues related to training RNNs [14]. In addition to traditional techniques to prevent over-fitting such as early stopping and weight-decay, dropout regularization [15] removes a neuron's output and connections at probability p during training, while retaining all neurons during testing with appropriate scaling.

In [16], strategies for TFM and MS prediction and related tasks are presented and compared using DNNs and RNNs. The authors of [16] propose a neural network structure that uses a single joint loss function, which accounts for both MS and TFM errors. The approach was contrasted to the ensemble of the two spectral magnitude predictions. The ensemble was the best approach to improve the short-time objective intelligibility (STOI) of the enhanced speech using synthetic mixtures. In addition, the RNN approach using LSTMs was found to increase STOI over the DNN approach.

Multi-context networks are proposed in [17] for speech separation. The approach consists of fusing DNNs, which are trained in separate contexts. The authors of [17] also proposed multi-context stacking, by concatenating the predictions of the DNNs with input features as modules on top of each other. Along a similar path, in [18] a DNN-based fusion method is proposed that consists of three separate target predictions for IRM, IBM, and MS. A non-recurrent fusion layer is added that takes the reconstructed magnitude spectra as its input and outputs a final MS prediction. Using synthetic mixtures, the combination network is shown to outperform individual types of enhancement methods and the basic ensemble of the magnitude spectrograms in terms of objective perceptual metrics (STOI [19], and perceptual evaluation of speech quality (PESQ) [20]).

The ensemble approach generally produces better results than the individual estimators and is known to work best when

978-1-7281-1817-819$31.00 ©2019 IEEE

© 2019 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

DOI: 10.1109/MMSP.2019.8901800

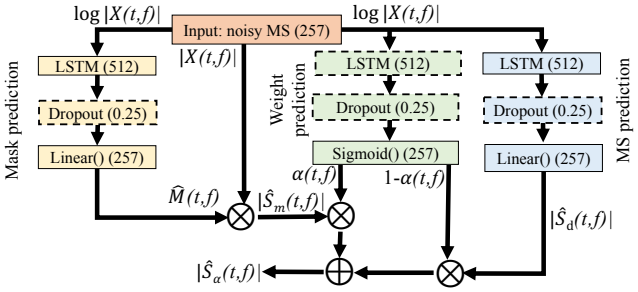

Fig. 1. Structure of the proposed weighted MS prediction. The left path depicts the TFM prediction $\hat{M}(t,f)$, while the right path depicts the MS prediction. The weight prediction value $\alpha(t,f)$ is applied to the masked input $|\hat{S}_m(t,f)|$ and the inverse-weight $(1-\alpha(t,f))$ is applied to the MS prediction $|\hat{S}_d(t,f)|$. These weighted MS estimates are then summed to result in the MS estimate $|\hat{S}_\alpha(t,f)|$. Optionally applied dropout procedures and the optional LSTM layers are drawn with dashed lines.

the combined estimators are both accurate and make their prediction errors in different parts of the input space [21]. Since TFM and MS-based estimators can behave differently with respect to input SNR [17], the ensemble approach is chosen. Since the basic ensemble can be improved by data-dependent weighting [22], this paper proposes a data-dependent ensemble of the predicted magnitude spectra. Here, LSTMs are applied in TFM and MS prediction instead of DNNs, since they have been shown to perform better in the speech enhancement task, see e.g. [8], [16]. Real and synthetic data from the CHiME3-challenge [23] are used to contrast the proposed approach to other state-of-the-art multi-target approaches. The benefits of using recurrent layers for the combination of targets and the effects of applying dropout regularization [15] in the TFM and MS sub-networks and in their combination was investigated. The results show that the proposed approach achieves significant improvement in noise reduction, while simultaneously retaining a high objective measure of intelligibility of the enhanced speech.

The rest of the paper is organized as follows. Section II describes the signal model and the proposed method, followed by a summary of the contrasted approaches. Section III describes the used database and the used objective evaluation metrics. Section IV describes the obtained results and is followed by a discussion. Section V concludes the discussion.

II. Combining The Magnitude Spectra

The relationship between the captured signal $X(t,f)$ and the reference signal $S(t,f)$ is modeled with additive noise $N(t,f)$ in the short-time Fourier transform (STFT) domain

$$X(t,f) = S(t,f) + N(t,f), \tag{1}$$

where t and f denote time frame and frequency indices, respectively. The neural network's input features are the noisy signal's MS values $|X(t,f)|$. A Discrete Fourier Transform of length 512 (32 ms at 16 kHz sampling rate) was used of which first 257 values were kept to represent the real part of the spectrum. The clean signal's MS $|S(t,f)|$ is used as the target for the MS prediction, and to define a ratio-mask as the target for the TFM prediction $M(t,f) = |S(t,f)|/|X(t,f)|$, where the maximum ratio is bounded to avoid numerical instability.

The training of the network consists of two stages. In the first stage, the mask prediction $\hat{M}(t,f)$ and MS prediction $|\hat{S}_\mathrm{d}(t,f)|$ are trained independently of each other with their corresponding targets $M(t,f)$ and $|S(t,f)|$. Both sub-networks follow a similar structure, where the input is fed into an LSTM layer (with 512 nodes) that optionally uses dropout ($p = 0.25$), followed by an output layer with linear activation functions (257 nodes). The predicted values are in the linear domain.

The proposed weight prediction sub-network uses the noisy input MS to predict time-frequency dependent weight values $\alpha(t,f) \in [0,1]$. It consists of either an LSTM (with 512 nodes) followed by a non-recurrent sigmoid-activation layer (257 nodes), or just the non-recurrent sigmoid-activation layer (257 nodes) without the LSTM layer. In the second stage of training, the parameters of the weight prediction sub-network are learned while keeping the TFM and MS prediction network parameters fixed. The motivation is that the TFM and MS networks would keep producing the best estimates of these variables and not be skewed to produce sub-optimal estimates in order to compensate for errors introduced in later stages. The weight prediction layer therefore learns to combine the MS estimate obtained via masking $|\hat{S}_\mathrm{m}(t,f)| = \hat{M}(t,f) \cdot |X(t,f)|$ with the directly predicted clean MS estimate $|S_\mathrm{d}(t,f)|$ based on the noisy input MS. The reconstructed signal is the weighted combination of the two MS estimates with the original signal's phase

$$\hat{S}_\alpha(t,f) = \Big(|S_\mathrm{d}(t,f)|\cdot(1-\alpha(t,f)) + |S_\mathrm{m}(t,f)|\cdot\alpha(t,f)\Big)\cdot e^{i\angle X(t,f)}, \tag{2}$$

where $\angle(\cdot)$ denotes phase angle. Refer to Fig. 1. In contrast to the basic ensemble of using a fixed value $\alpha(t,f) = 0.5$, the prediction of $\alpha(t,f)$ values is learned during training and is inferred during testing. Due to the multiplicative nature of masking, the TFM signal cannot contain high magnitude values that are not originally present in the input, while the MS prediction can contain any values. Therefore, the approach can be viewed as a regularization of the MS prediction towards the TFM signal. Feeding additionally the pre-trained predicted mask and MS into the weight prediction sub-network did not improve performance and is therefore not considered further.

The loss function for the weight prediction structure is defined between $|\hat{S}_\alpha(t,f)|$ and the clean MS. A logarithm is sometimes applied to the magnitude since it is related to the human sensitivity to sound amplitude. Therefore, the mean squared logarithmic error (MSLE) function $\mathrm{MSLE}(a,b) = (\log(a+\epsilon) - \log(b+\epsilon))^2$ is used as the loss function, where the default ϵ value of 1.0 of [24] was used. Note that the approach predicts the MS values in the linear domain, and therefore reduces the overall complexity by not taking the log and inverse log operations during testing. The same loss function is used for the mask training, since it was observed to result in a more monotonic decay of the loss function in contrast to using mean squared error (MSE).

Fig. 2 illustrates the predicted $\alpha(t,f)$ values during test-time, the ratio-mask, MS prediction, and the resulting weighted

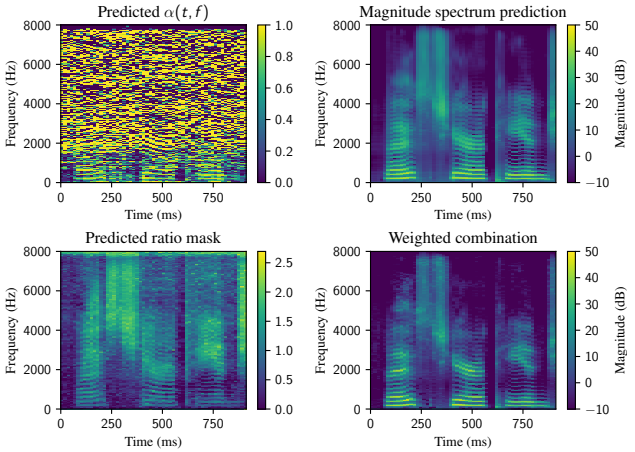

Fig. 2. Example predictions of the ratio mask $\hat{M}(t,f)$, MS $|\hat{S}_\mathrm{d}|$, weight values $\alpha(t,f)$, and resulting weighted combination $|\hat{S}_\alpha(t,f)|$.

combination. Note that $\alpha(t,f)$ varies in time and frequency.

A. Description of the Contrast Methods

The first contrasted approach is the average of resulting MS estimates, first proposed in [16], where the predicted TFM is multiplied with the input signal to produce a MS estimate to be averaged with the directly predicted MS estimate, see Fig. 3. This basic ensemble signal can be written also using Eq. (2) by fixing $\alpha(t,f)=0.5$, i.e., $\hat{S}_\mathrm{avg}(t,f)=\hat{S}_{\alpha=0.5}(t,f),\ \forall(t,f)$.

The second contrasted approach adds a fusion layer on top of trained TFM and MS sub-networks to produce the final MS estimate, inspired[1] by [18]. The first added layer is either a recurrent (here LSTM) or a regular fully connected layer with the hyperbolic tangent (tanh) activation function. This layer contains 512 nodes and is followed by a layer with 257 linear nodes for output reconstruction. The approach can be described by learning a fusion function $g(\cdot)$

$$\hat{S}_\mathrm{f}(t,f) = g(|S_\mathrm{d}(t,f)|, |S_\mathrm{m}(t,f)|) \cdot e^{i\angle X(t,f)}. \qquad (3)$$

Similarly to other methods, the fusion layer is trained by using the MSLE loss while keeping the weights of the already trained MS and the TFM sub-networks fixed. See Fig. 3 for illustration. Feeding additionally the noisy input features to the fusion layer was implemented and the results are reported using the best input feature combination.

B. Implementation of the Methods

The neural networks were trained using the Keras deep learning library [24] with the Python programming language v. 2.7. The LSTM sequence length was set to 64 frames. Both the analysis and the synthesis used square-root Hann windowing with 50% overlap between sequential frames. The Adam optimizer [25] was used with default parameter values, and the mini-batch size was set to 750 sequences. Training was stopped if the validation error did not decrease in 50 consecutive epochs or 1500 epochs were reached. The

[1]We use LSTMs for the TFM and MS prediction in contrast to DNNs with frame stacking and omit the IBM. In addition, we explore the use of LSTMs for the fusion layer.

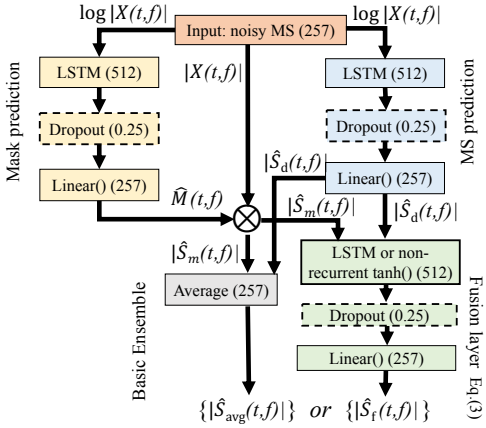

Fig. 3. The structure of the two contrast methods, averaging of MS estimates [16], and the fusion of estimates inspired by [18] are illustrated. Optional application points of dropout are drawn with dashed lines. Note that unlike the LSTM of the weighting layer in Fig. 1, the fusion layer is either LSTM or non-recurrent (tanh activation).

maximum ratio-mask value was clipped to 2.0 during training for numerical stability.

A noise floor was observed for the MS prediction output, which could be caused by leaked noise in the clean signal of the real recordings. Since noise is not deterministic the network compensates for the noise by adding a bias to the MS prediction. Hence, a minimum-statistics-based noise floor subtraction [1] was applied for the predicted MS values, where the noise floor was estimated as the 5th percentile for each frequency bin of the MS over the processed sentence.

III. Data and Evaluation

The CHiME3 database was used in the experiments [23]. A tablet device with external microphones embedded into its edges was used to record spoken WSJ0 corpus sentences in four different environments: bus (BUS), street (STR), cafeteria (CAF), and pedestrian area (PED). We arbitrarily selected the 1st device microphone channel as the input signal and omitted the rest of the channels. The reference signal was captured with a head-worn close-talk microphone (CTM). Here, all signals were RMS normalized as a pre-processing step.

The database has been divided into training (TR), development (DT), and evaluation (ET) sets in the CHiME3-challenge. The TR set consists of approximately 15 hours of simulated data (7138 utterances) from 83 unique speakers and 3 hours of real recorded data (1600 utterances) from four speakers (two male, two female). The DT and ET data sets consist of 3280 and 2640 utterances, respectively, both sets each contain four unique different speakers, and 50% of the utterances were simulated and the rest were captured in real environments. The amount of data is balanced between the environments. Features were normalized with the TR set mean and standard deviation.

For training purposes, the TR set was further divided into three subsets (TR-train, TR-validation, and TR-test) to train the networks using k-fold cross-validation. A single fold of the TR-train subset contained 27 unique speakers for the simulated data, and three unique speakers for the real data. The TR-

TABLE I
AVERAGE SEGMENTAL SDR IMPROVEMENT (DB) OVER A SINGLE CHANNEL (CH. 1) OF THE CHIME3 CHALLENGE DATA [23].

Sub-net type	TFM $\hat{S}_{\text{m}}$		MS $\hat{S}_{\text{d}}$		$\hat{S}_{\text{avg}}$		Fusion $\hat{S}_{\text{f}}(t,f)$, Eq.(3)						Proposed $\hat{S}_{\alpha}$, Eq.(2)				
LSTM-layer	yes	yes	yes	yes	-	-	yes	no	no	no	yes	yes	yes	no	no	yes	yes
D.O.(TFM)	0	.25	-	-	0	.25	0	0	.25	0	0	.25	0	0	.25	0	.25
D.O.(MS)	-	-	0	.25	0	.25	0	0	.25	0	0	.25	0	0	.25	0	.25
D.O.(Comb)	-	-	-	-	-	-	0	0	0	.25	.25	0	0	-	-	.25	0
TR (Real)	4.0	3.9	1.9	2.7	3.5	3.8	3.0	1.7	1.9	1.7	3.6	3.1	4.6	4.6	**4.8**	4.6	4.7
DT (Real)	4.4	4.3	3.0	3.6	4.2	4.3	3.8	2.7	2.9	2.7	4.4	4.0	5.0	5.1	**5.2**	5.1	5.1
ET (Real)	3.8	3.9	2.1	3.0	3.4	3.8	3.2	2.0	2.3	2.0	3.9	3.3	4.3	4.3	**4.5**	4.4	**4.5**
AVG (Real)	4.1	4.0	2.3	3.1	3.7	3.9	3.3	2.1	2.4	2.1	4.0	3.4	4.6	4.6	**4.8**	4.7	**4.8**
TR (Simu)	6.3	6.3	1.9	2.9	5.3	5.5	3.4	1.5	1.6	1.5	4.3	3.3	6.3	6.3	**6.4**	**6.4**	**6.4**
DT (Simu)	4.8	4.8	1.5	2.4	4.2	4.4	2.6	1.0	1.1	1.0	3.4	2.6	5.0	4.9	**5.1**	4.9	5.0
ET (Simu)	5.8	5.6	2.5	3.6	5.0	5.2	3.9	2.3	2.6	2.3	4.9	3.9	6.2	6.2	**6.3**	6.2	6.2
AVG (Simu)	5.6	5.6	1.9	3.0	4.8	5.0	3.3	1.6	1.8	1.6	4.2	3.2	5.8	5.8	**5.9**	5.8	**5.9**

TABLE II
RELATIVE ESTOI IMPROVEMENT PERCENT OVER A SINGLE CHANNEL (CH. 1) OF THE CHIME3 CHALLENGE DATA [23].

Sub-net type	TFM $\hat{S}_{\text{m}}$		MS $\hat{S}_{\text{d}}$		$\hat{S}_{\text{avg}}$		Fusion $\hat{S}_{\text{f}}(t,f)$, Eq.(3)						Proposed $\hat{S}_{\alpha}$, Eq.(2)				
LSTM-layer	yes	yes	yes	yes	-	-	yes	no	no	no	yes	yes	yes	no	no	yes	yes
D.O.(TFM)	0	.25	-	-	0	.25	0	0	.25	0	0	.25	0	0	.25	0	.25
D.O.(MS)	-	-	0	.25	0	.25	0	0	.25	0	0	.25	0	0	.25	0	.25
D.O.(Comb)	-	-	-	-	-	-	0	0	0	.25	.25	0	0	-	-	.25	0
TR (Real)	9	9	12	14	13	14	17	12	12	12	**18**	16	14	15	16	16	16
DT (Real)	9	8	12	13	13	12	14	10	11	10	**16**	15	13	14	14	13	13
ET (Real)	1	3	4	6	6	7	6	3	4	3	**9**	7	7	6	8	7	7
AVG (Real)	6	6	10	11	11	11	12	8	9	8	**14**	13	11	12	13	12	12
TR (Simu)	23	22	20	22	24	24	26	20	20	20	**27**	25	24	26	25	26	25
DT (Simu)	14	15	12	15	16	17	16	12	13	12	17	17	16	16	**18**	16	17
ET (Simu)	18	20	19	24	22	24	24	20	23	20	26	**27**	23	24	26	24	25
AVG (Simu)	18	19	17	20	21	22	22	17	19	17	**23**	**23**	21	22	**23**	22	**23**

validation and TR-test shared the fourth real speaker, but split the environments, while the remaining simulated environments both contained 28 unique speakers. Four different folds were trained by rotating the speaker identities to cover all TR set data in the training. For each approach, the test-time output used an ensemble of the predictions of the separately trained folds, which is expected to increase the results slightly [21]. Note that the ensemble, fusion, and weighted ensemble approaches combine the TFM and MS estimates of the same fold and are subject to averaging across folds only during test time.

The segmental source-to-distortion ratio (SDR) [26] measures the amount of achieved noise reduction with respect to the amount of introduced artifacts. The amount of target signal energy is estimated using a time-varying convolution filter between the estimate and the reference in 1 s frames with 500 ms overlap with the mir_eval-toolbox [27]. The extended short-time objective intelligibility (ESTOI) metric is a measure of intelligibility of non-linearly distorted and noise contaminated signal with respect to the reference signal [28].

IV. RESULTS AND DISCUSSION

Table I lists the average SDR improvement obtained over the unprocessed microphone channel 1 for the CHiME3 challenge data. The results are averaged over the four different background environments and listed separately for the different data sets. Note that the TR data set is used to train the network parameters, while DT and ET data sets are unseen. The "LSTM-layer" row indicates whether the sub-network contained an LSTM layer. Note that the ensemble average ($\hat{S}_{\text{avg}}$) cannot contain dropout or LSTM. The fusion layer approach contained either an LSTM or a regular layer with tanh() activation, which was followed by a linear layer. The notation "D.O.(layer)" rows indicates dropout probability p for the specific layer. Recall that the proposed weighted ensemble consisted of either an LSTM layer followed by a sigmoid layer, or just a sigmoid layer, in which case dropout was not used.

In Table I, the second and third columns represents TFM results, which achieve 4.1 dB and 5.6 dB average SDR improvement for the real and simulated data, respectively. Using dropout slightly decreases the SDR. In contrast, dropout improves the SDR of MS by increasing the average SDR of the real data from 2.3 dB to 3.1 dB, and the simulated data from 1.9 dB to 3.0 dB. By including dropout in TFM and MS sub-networks, the ensemble layer's SDR increases from 3.7 dB to 3.9 dB and 4.8 dB to 5.0 dB for real and simulated data, respectively. However, the TFM sub-network still outperforms the basic ensemble approach in terms of SDR improvement.

The best fusion layer results are obtained by not using dropout in the MS and TFM sub-networks and by including dropout in the LSTM-type fusion layer. Using the LSTM layer instead of the non-recurrent layer significantly improves the results. However, the fusion layer does not improve SDR over the TFM approach and performs somewhat similarly or worse to the ensemble layer, resulting in 4.0 dB and 4.2 dB improvements for real and simulated data, respectively.

The proposed weighted ensemble method achieves the best SDR improvement by including dropout in the TFM and MS sub-networks while applying a single non-recurrent layer to merge the TFM-based MS and the direct MS estimates. An average improvement of 4.8 dB and 5.9 dB for real and simulated data over the single channel was reached, respectively.

Table II lists the scores for the intelligibility metric ESTOI. Note that the results are shown as relative improvement percentage over the 1st channel. TFM improvement in ESTOI is lowest among the methods, and is slightly improved by using dropout, while the ESTOI of MS prediction is higher. The ensemble average approach slightly improves ESTOI over the MS prediction by reaching 11% improvement in the real data, and 22% improvement in the simulated data. The LSTM -type fusion layer (with dropout only in the fusion layer) was able to improve the ESTOI the most, marginally outperforming the proposed approach in the real data case by reaching 14% improvement for real data (proposed method obtains 13%), and 23% improvement for the simulated data, which is similar to the proposed method. Only the best performing fusion variant benefited by including the noisy input features as well.

The ensemble average method $\hat{S}_{\text{avg}}$ was as successful or better in terms of SDR improvement than the best fusion layer method $\hat{S}_{\text{f}}$, but slightly worse in its ESTOI improvement. Since the ensemble averaging is data-independent model averaging with $\alpha(t,f) = 0.5$ the improvements brought by weighted ensemble $\hat{S}_{\alpha}$ can be attributed to the proposed data-dependent modeling of the weights $\alpha(t,f)$.

Using the MS as the fusion layer's output can explain its low SDR scores, since the SDR of MS prediction $\hat{S}_d$ was lower than that of TFM $\hat{S}_m$.

V. Conclusions

This work proposed a data-dependent weighted ensemble of two common deep learning approaches for speech enhancement – the time-frequency masking and magnitude spectrum estimation. The approach was implemented as a single deep neural network structure and was trained in stages. For each time-frequency point, the weighting sub-network linearly combines the two separately estimated magnitude spectra with a predicted weight value. The approach was contrasted with two alternative state-of-the-art multi-target approaches, including averaging of predicted magnitude spectra (basic ensemble), and adding a fusion layer for the estimated magnitude spectra. The proposed approach had clearly the best noise reduction performance (measured with segmental SDR) while it achieved similar objective intelligibility improvements with respect to the fusion approach (measured with ESTOI).

A single sigmoid layer was the best sub-network structure to predict the weights for the ensemble combination of the magnitude spectra obtained from TFM and MS prediction paths, where both of the paths used an LSTM layer and the ensemble benefited from including dropout in both paths.

The results show that the method generalizes well for new speakers using a similar capture setting, but generalization to different scenarios requires further validation and possibly more training data matching such new scenarios.

References

[1] R. Martin, "Spectral Subtraction Based on Minimum Statistics," in *Proc. 7th Eur. Signal Processing Conf. EUSIPCO*, 1994, pp. 1182–1185.

[2] E. Diethorn, "Subband Noise Reduction Methods for Speech Enhancement," in *Audio Signal Processing for Next-Generation Multimedia Communication Systems*, Y.(A.) Huang and J. Benesty, Eds., chapter 4, pp. 91–115. Kluwer Academic Publishers, 2004.

[3] 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, July 2013.

[4] A. Narayanan and D. Wang, "Ideal Ratio Mask Estimation Using Deep Neural Networks for Robust Speech Recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vancouver, Canada, 2013, pp. 7092–7096.

[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 *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 708–712.

[6] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for monaural speech separation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 24, no. 3, pp. 483–492, March 2016.

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

[8] F. Weninger, J. R. Hershey, J. Le Roux, and B. Schuller, "Discriminatively trained recurrent neural networks for single-channel speech separation," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, 2014, pp. 577–581.

[9] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 22, no. 12, pp. 1849–1858, Dec. 2014.

[10] M. Delfarah and D. Wang, "A feature study for masking-based reverberant speech separation.," in *Annual Conference of the International Speech Communication Association (Interspeech)*, 2016, pp. 555–559.

[11] M. Delfarah and D. Wang, "Features for masking-based monaural speech separation in reverberant conditions," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 25, no. 5, pp. 1085–1094, 2017.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *In Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1724—-1734.

[14] R. Pascanu, T. Mikolov, and Y. Bengio, "Understanding the exploding gradient problem," *Computing Research Repository (CoRR)*, vol. abs/1211.5063, 2013.

[15] N. Srivastava, G. E Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting.," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[16] L. Sun, J. Du, L.R. Dai, and C.H. Lee, "Multiple-target deep learning for LSTM-RNN based speech enhancement," in *Hands-free Speech Communications and Microphone Arrays (HSCMA), 2017*. IEEE, 2017, pp. 136–140.

[17] X. Zhang and D. Wang, "A deep ensemble learning method for monaural speech separation," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 24, no. 5, pp. 967–977, 2016.

[18] H. Zhang, X. Zhang, and G. Gao, "Multi-target ensemble learning for monaural speech separation," *Annual Conference of the International Speech Communication Association (Interspeech)*, pp. 1958–1962, 2017.

[19] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 19, no. 7, pp. 2125–2136, 2011, (Software: http://www.ceestaal.nl/stoi.zip).

[20] "Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs," in *ITU-T, P.862, SERIES P: TELEPHONE TRANSMISSION QUALITY, TELEPHONE INSTALLATIONS, LOCAL LINE NETWORKS. Methods for objective and subjective assessment of quality.* 02/2001.

[21] D. W. Opitz and R. Maclin, "Popular Ensemble Methods: An Empirical Study," *J. Artif. Intell. Res.(JAIR)*, vol. 11, pp. 169–198, 1999.

[22] S. Hashem, "Optimal linear combinations of neural networks," *Neural Networks*, vol. 10, no. 4, pp. 599 – 614, 1997.

[23] J. Barker, R. Marxer, E.l Vincent, and S. Watanabe, "The third 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines," in *IEEE 2015 Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2015.

[24] F. Chollet et al., "Keras," https://github.com/fchollet/keras, 2015.

[25] D.P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *the International Conference on Learning Representations (ICLR)*, 2015.

[26] E. Vincent, R. Gribonval, and C. Fevotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 14, no. 4, pp. 1462–1469, July 2006.

[27] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, "mir_eval: A Transparent Implementation of Common MIR Metrics," in *Proceedings of the 15th International Conference on Music Information Retrieval*, 2014.

[28] J. Jensen and C. H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers," *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, vol. 24, no. 11, pp. 2009–2022, Nov 2016.