

Attention-based multi-task learning for speech-enhancement and speaker-identification in multi-speaker dialogue scenario

Chiang-Jen Peng*, Yun-Ju Chan*, Cheng Yu†, Syu-Siang Wang†‡, Yu Tsao† and Tai-Shih Chi*

*Department of Electrical and Computer Engineering, National Chiao Tung University, Hsinchu, Taiwan
Email: tschi@mail.nctu.edu.tw

†Research Center for Information Technology Innovation, Academia Sinica, Taipei, Taiwan
Email: yu.tsao@citi.sinica.edu.tw, chengyu_citi@citi.sinica.edu.tw

‡Department of Electrical Engineering, Yuan Ze University, Taoyuan, Taiwan
Email: syddbhee@saturn.yzu.edu.tw

Abstract—Multi-task learning (MTL) and attention mechanism have been proven to effectively extract robust acoustic features for various speech-related tasks in noisy environments. In this study, we propose an attention-based MTL (ATM) approach that integrates MTL and the attention-weighting mechanism to simultaneously realize a multi-model learning structure that performs speech enhancement (SE) and speaker identification (SI). The proposed ATM system consists of three parts: SE, SI, and attention-Net (AttNet). The SE part is composed of a long-short-term memory (LSTM) model, and a deep neural network (DNN) model is used to develop the SI and AttNet parts. The overall ATM system first extracts the representative features and then enhances the speech signals in LSTM-SE and specifies speaker identity in DNN-SI. The AttNet computes weights based on DNN-SI to prepare better representative features for LSTM-SE. We tested the proposed ATM system on Taiwan Mandarin hearing in noise test sentences. The evaluation results confirmed that the proposed system can effectively enhance speech quality and intelligibility of a given noisy input. Moreover, the accuracy of the SI can also be notably improved by using the proposed ATM system.

Index Terms—Speech enhancement, speaker identification, multi-task learning, attention weighting, neural network.

I. INTRODUCTION

Speech signals propagating in an acoustic environment are inevitably distorted by noise. Such distortions may considerably degrade the performance of the target speech-related tasks, such as assistive hearing systems [1], [2] automatic speech recognition [3], [4], and speaker recognition [5], [6], [7]. To address this issue, speech enhancement (SE), which aims to extract clean acoustic signals from noisy inputs, has been widely used. Conventional SE techniques, including signal subspace [8], power spectral subtraction [9], Wiener filtering [10], and minimum mean square error based estimations [11], [12], perform well in stationary noise environments, where the statistical assumptions on environmental noises and human speech hold adequately [13], [14], [15]. For environments involving non-stationary noises, conventional SE techniques may not provide satisfactory performance. Recently, deep learning (DL)-based SE methods have been widely studied, and notable performance improvements have been observed over conventional techniques. In general, the DL-based SE methods aim to transform the noisy source to a clean target using nonlinear mapping, in which no statistical property assumption of noise and speech acoustic signals is required [16], [17], [18]. In [19], [20], the authors proposed a deep denoising autoencoder (DDAE) SE system that encodes input noisy signals into a series of frame-level speech codes and then performs a decoding process to retrieve the enhanced signals from the system output. Another study in [3] applied a long short-term memory (LSTM) model to integrate the context information to carry out SE for improving speech quality and intelligibility and achieving a low word error rate in an ASR system. In [21], the transformer model that utilizes an attention

mechanism [22] to compute attention weights is used to emphasize and fuse related context symbols to obtain clean components.

The SE system can be used as a front-end processor for specific applications by placing it in front of a main speech-signal-processing system. By jointly minimizing the losses from the SE and the main system, the overall system is considered to be optimized in a multi-task learning (MTL) manner [23], [24], [25]. In such systems, the MTL aims to purify the representations along with the goal to boost the performance of the main task [26], [27], [28]. In [29], [30], visual information is treated as the second task to promote the SE capability. Experimental results show that audio and visual cues can be jointly considered to derive more representative acoustic features in a DL-based SE system.

MTL has also been used in speaker recognition, namely speaker identification (SI) and speaker verification (SV), systems [31], [32], [33]. The recognition accuracy of an SI task is highly dependent on the quality of speaker feature extraction. Therefore, most existing systems aim to compute a decent speaker representation from speech signals. A well-known speaker recognition system is a combination of an i -vector with a probabilistic linear discriminant analysis [34]. This system has been widely used and yields satisfactory performance in numerous speaker recognition tasks. More recently, d -vector [35] and x -vector [36] features extracted by DL models have been proven to provide more abundant speaker information and, therefore, show superior recognition performances to the i -vector.

Inspired by the transformer model structure, this study proposes a novel system, namely the attention-based MTL (ATM), to extract the shared information between SE and SI to attain improved performance for individual tasks. The outputs of the ATM system are enhanced speech and identification results, and the input is noisy speech signals. In addition, an attention-based network (AttNet) is used to integrate both speech and speaker cues between SE and SI models to extract robust features. The ATM consists of three DL-based models: the first LSTM enhances the noisy input and the other two DNNs are used to identify the speaker identity and extract the attention weights. We tested the proposed system on the Taiwan Mandarin hearing in noise test (TMHINT) sentences [37]. The experimental results show that the proposed ATM can not only enhance the quality and intelligibility of noisy speech but also the improve SI accuracy.

The remainder of this paper is organized as follows. Section II reviews the related work, including LSTM-based SE and DNN-based SI. Section III introduces the proposed ATM architecture. Experimental results and analyses are provided in Section IV. Finally, Section V presents the conclusions and directions for future research.

II. RELATED WORKS

This section briefly reviews the related works of the LSTM-SE and DNN-SI systems. Considering noisy speech signals are obtained from contaminating clean speech signals with additive noise signals. With short-time Fourier transform (STFT) and several feature processing steps, we can obtain noisy and clean logarithmic power spectra (LPS), \mathbf{Y} and \mathbf{S} , respectively, from the noisy and clean speech signals. We assumed that there are N frames in the paired (\mathbf{Y} - \mathbf{S}). The context feature of noisy LPS by concatenating the adjacent $2M$ feature frames of the target feature vector $\mathbf{Y}[n]$, namely, $\mathbb{Y}[n] = [\mathbf{Y}'[n-M], \dots, \mathbf{Y}'[n], \dots, \mathbf{Y}'[n+M]]'$ are used.

A. Speech enhancement

In this study, the baseline SE system is composed of an L -hidden-layer LSTM and a feed-forward layer and is denoted as LSTM-SE. The input-output relationship $(\mathbf{z}_{\ell+1}[n], \mathbf{z}_{\ell}[n])$ at the n -th frame and the ℓ -th hidden layer is formulated as

$$\mathbf{z}_{\ell+1}[n] = LSTM_{\ell}\{\mathbf{z}_{\ell}[n]\}, \quad \ell = 1, 2, \dots, L. \quad (1)$$

The input of the first LSTM layer is \mathbf{Y} , i.e. $\mathbf{z}_1[n] = \mathbf{Y}[n]$, and the output $\mathbf{z}_{L+1}[n]$ is

$$\hat{\mathbf{S}}[n] = \mathbf{W}\mathbf{z}_{L+1}[n] + \mathbf{b}, \quad (2)$$

where \mathbf{W} and \mathbf{b} are the weight matrix and bias vector, respectively. In the training stage, the parameters of the LSTM-SE system are updated to minimize the difference between $\hat{\mathbf{S}}[n]$ and $\mathbf{S}[n]$ in terms of the mean square error (MSE). In the testing stage, the output $\hat{\mathbf{S}}$ of the LSTM-SE is combined with the phase from the noisy speech signals to produce the enhanced signals $\hat{\mathbf{s}}$ in the time domain.

B. Speaker identification

The objective of the DNN-SI is to classify an input speech signal $\mathbb{Y}[n]$ at the n -th frame to a specific speaker identity. We categorized the non-speech segments as a single virtual speaker. Therefore, the dimension of DNN-SI output is the number of speakers plus one, namely $K + 1$. The reference target for the DNN training is a one-hot $(K + 1)$ -dimensional vector $\mathbf{I}[n]$, where a single non-zero element corresponds to the target speaker identity.

The DNN-SI contains D layers, and the input-output relationship $(\mathbf{z}_d[n], \mathbf{z}_{d+1}[n])$ at the d -th layer and the n -th frame can be formulated as

$$\mathbf{z}_{d+1}[n] = \sigma_d(F_d(\mathbf{z}_d[n])), \quad d = 1, \dots, D, \quad (3)$$

where $\sigma_d(\cdot)$ and $F_d(\cdot)$ are the activation and linear transformation functions, respectively. In this study, the softmax function is used as the activation function for the output layer, and the rectified linear units (ReLU) function is used for all hidden layers. Meanwhile, the input and output of DNN is $\mathbf{z}_1[n] = \mathbb{Y}[n]$ and $\mathbf{z}_{D+1}[n] = \hat{\mathbf{I}}$, respectively. The categorical cross-entropy loss is used to compute the DNN parameters in Eq. (3).

III. THE PROPOSED APPROACH

Figure 1 shows the block diagram of the proposed ATM system. From the figure, the input to the ATM system is a noisy LPS feature, \mathbf{Y} , while the outputs have enhanced LPS feature in SE and speaker identity vector in SI. Between SE and SI tasks, an AttNet is employed to reshape the feature size from SI and extract more compact speaker cues for SE. Based on the ways of incorporating attention mechanism, two ATM architectures have been proposed, namely ATM_{bef} and ATM_{ide} , which are detailed in the following two sub-sections.

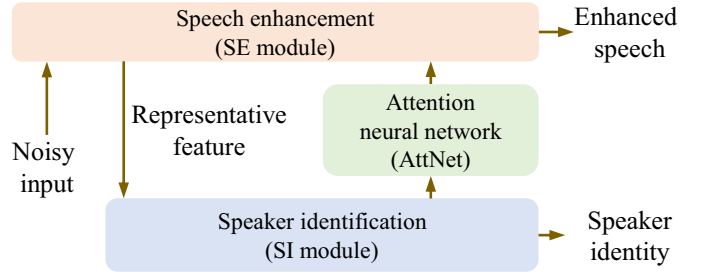


Fig. 1. The block diagram of the proposed ATM system, in which the input is noisy speech while the outputs are enhanced speech and the recognized speaker identity.

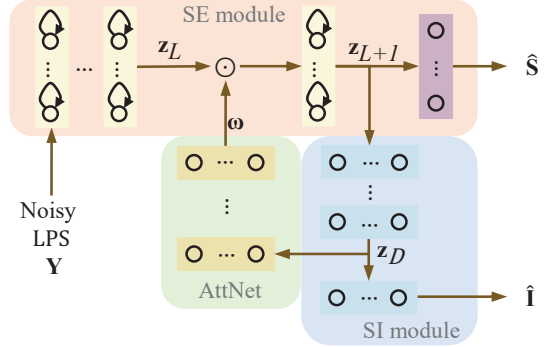


Fig. 2. The architecture of ATM_{bef} . The output of the $(L + 1)$ -th layer of LSTM-SE is used to compute ω , which is then used to weight the representative features at the L -th layer.

A. The ATM_{bef} system

Figure 2 illustrates the block diagram of the ATM_{bef} system. As shown in the figure, the SE model is used to provide the embedded speech code vector, $\mathbf{z}_{L+1}[n]$, from the output of the L -th LSTM hidden layer. We then created the context information of speech by concatenating the adjacent vectors of $\mathbf{z}_{L+1}[n]$ to obtain $[\mathbf{z}'_{L+1}[n-M], \dots, \mathbf{z}'_{L+1}[n], \dots, \mathbf{z}'_{L+1}[n+M]]'$ to form the input of SI to compute the speaker feature (from the output of the last hidden layer). Then, AttNet, which is a J -layer DNN model, takes the speaker feature as the input to compute the weighting vector, ω , to weight the LSTM-SE by performing $\omega[n] \odot \mathbf{z}_L[n]$, where \odot is an element-wise multiplication operator. Finally, enhanced speech, $\hat{\mathbf{S}}$, and recognized speaker identity, $\hat{\mathbf{I}}$, are obtained. This system is referred to ATM_{bef} because the attention operation is performed before extracting the acoustic feature representation.

To train ATM_{bef} , we prepare noisy LPS features as the input and the corresponding speaker-identity vectors and clean LPS features as the two outputs. Then, an iterative training procedure is applied to train SI and SE-AttNet models using the following steps: (1) The categorical cross-entropy loss is used to train the SI model, where the model input and output are the contextual embedding features and the speaker-identity vectors, respectively. (2) The speaker features, \mathbf{z}_D , using the SI model was then extracted. (3) The training proceeds with \mathbf{Y} and \mathbf{z}_D on the input side of SE and AttNet, respectively, to produce an enhanced output that approximates \mathbf{S} . Notably, the SE and AttNet models are jointly trained based on the MSE loss.

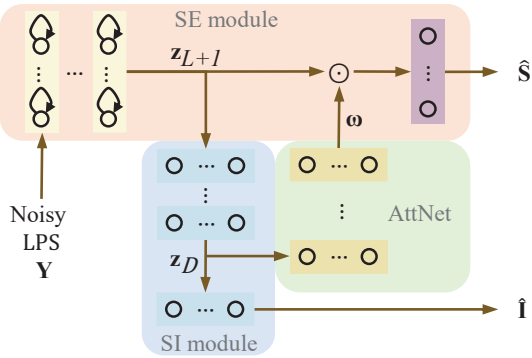


Fig. 3. The architecture of ATM_{ide} . The shared representation of SE and SI and the weighting operation are performed in the same layer of LSTM-SE.

B. The ATM_{ide} system

Figure 3 shows the block diagram of ATM_{bef} . Different from ATM_{bef} , ATM_{ide} performs feature extraction and weighting operation in the same layer of LSTM-SE. More specifically, ATM_{ide} performs four steps to obtain \hat{S} and \hat{I} . First, the acoustic code $\mathbf{z}_{L+1}[n]$ is first computed by passing the noisy LPS $\mathbf{Y}[n]$ to LSTM-SE. Next, the SI model generates $\hat{I}[n]$ in the output and the speaker code $\mathbf{z}_D[n]$, which is served as the input of the AttNet, to obtain the weighting vector, $\omega[n]$. Then, the weighting vector, $\omega[n]$, is applied to weight the acoustic code (that is, $\omega \odot \mathbf{z}_{L+1}[n]$). Finally, a transformation is applied to the weighted acoustic code to generate the enhanced output $\hat{S}[n]$. For ATM_{ide} , the weighting vector ω is extracted to introduce the speaker-dependent characteristics in the acoustic feature and guide the SE to generate the output corresponding to the target speaker. The proposed ATM systems (ATM_{ide} and ATM_{bef}) can be considered as multi-model systems because the speaker characteristics are used to guide SE to achieve better performance.

The dynamic weighted loss function has been proposed to address the scaling issue of classification and regression tasks [38], [39]. The loss is formulated in Eq. (4) with two additional trainable parameters, α and β .

$$L(\Theta, \alpha, \beta) = \frac{1}{2\alpha^2} L_1(\Theta) + \frac{1}{\beta^2} L_2(\Theta) + \log\alpha + \log\beta, \quad (4)$$

where L_1 and L_2 are the MSE and the categorical cross-entropy loss, respectively; Θ represents all the parameters in ATM_{ide} .

IV. EXPERIMENTS AND ANALYSES

In the following subsections, we first introduce the experimental setup and then provide the experimental results along with a discussion about our findings.

A. Experimental setup

We evaluated the proposed ATM system on TMHINT sentences. The training and testing utterances were recorded by eight speakers at a 16 kHz sampling rate in a noise-free meeting room. A total of 1,560 clean utterances were pronounced by three males and three females ($K = 6$ in Section II-B), each providing 260 utterances, for the training set. From these clean utterances, we randomly concatenated three utterances to simulate the dialogue scenario and subsequently generated 520 clean training utterances, where each utterance contained exactly three different speakers. Noisy utterances were generated by artificially adding 100 different types of noises [40] at six signal-to-noise ratio (SNR) levels (15, 10, 5, 0, -5, and -10 dB) to the prepared 520 clean training utterances, and thus generating

312,000 ($= 520 \times 100 \times 6$) noisy-clean training pairs. Among them, we randomly selected 500 noisy-clean pairs to form the validation set. Meanwhile, two different testing configurations were prepared for the SE and SI tasks. For SE, the testing set contained additional male and female speech. We randomly concatenated one utterance of the male speaker and one utterance of the female speaker and generated 60 clean utterances for testing. Noisy testing utterances were then prepared by deteriorating these clean utterances with four additive noises (“engine”, “pink”, “street”, and “white”) at three SNRs (5, 0, and -5 dBs). Accordingly, we prepared 720 ($= 60 \times 4 \times 3$) noisy testing utterances. In contrast to the SE testing set, the utterances used to evaluate the SI part were from the same speakers in the training set. We prepared 120 clean dialogue utterances for testing, with each utterance containing segments from three different speakers. Then, we added four additive noises (“engine”, “pink”, “street”, and “white”) at three SNRs (5, 0, and -5 dBs) to these clean testing utterances to form the noisy utterances. Accordingly, we prepared 1440 noisy utterances for testing the SI performance. Please note that we did not consider overlapped speech in this study, and thus there were no overlapped segments in the utterances for all the training and testing sets.

To apply the STFT, we used a window with frame size and shift of 32 ms and 16 ms, respectively. Then, a 257-dimensional LPS vector was obtained. The context feature was created by $M = 5$ thus with the dimension of $2,827 = 257 \times (2 \times 5 + 1)$. Accordingly, the input- and output-layer sizes of SE were both 257, and those of SI were 2,827 and 7 (i.e., $K + 1 = 6 + 1$), respectively. For the overall ATM system, the input size were 257 and the output sizes was 257 for SE and 7 for SI. The detailed network configuration is as follows:

- The SE model consisted of two LSTM layers ($L = 2$) with 300 cells in each layer, followed by a 257-node feed-forward layer.
- The SI model comprised four hidden layers ($D = 4$) in the order of 1024, 1024, 256, and 7 nodes.
- The AttNet comprised two hidden layers ($J = 2$) with each layer having 300 nodes.

In this study, we applied three metrics to evaluate the proposed system: perceptual evaluation of speech quality (PESQ) [41], short-time objective intelligibility (STOI) [42], and segmental SNR index (SSNRI) [43]. The score ranges of PESQ and STOI are $[-0.5, 4.5]$ and $[0, 1]$, respectively. Higher PESQ and STOI scores indicate better speech quality and intelligibility. Meanwhile, a higher SSNRI score indicates better signal-level SE performance.

B. Experimental results

In this subsection, we split the evaluation results into two parts. We first report the SE evaluation results and then the SI performance.

1) *SE results*: Table I lists the averaged PESQ, STOI, and SSNRI results with respect to all testing utterances of the noisy baseline (denoted as “Noisy”) and the enhanced speech obtained by conventional LSTM-SE, ATM_{bef} , and ATM_{ide} . In addition, the results of MTL,

TABLE I
AVERAGED PESQ, STOI AND SSNRI SCORES OF NOISY, LSTM-SE, MTL, ATM_{bef} , AND ATM_{ide} .

	Noisy	LSTM-SE	MTL	ATM_{bef}	ATM_{ide}
PESQ	1.25	1.86	1.86	1.94	1.98
STOI	0.72	0.73	0.74	0.74	0.75
SSNRI	–	7.39	7.61	7.57	8.05

which is composed of only SE and SI models (without AttNet) in Fig. 1, are also listed for comparison. From the table, most evaluation metrics using the MTL criterion, that is, MTL, ATM_{bef} , and ATM_{ide} , show better results than those provided by LSTM-SE, except the PESQ score of MTL. The results confirm the effectiveness of MTL-based models in improving the speech quality, intelligibility, and background noise reductions. In addition, both ATM_{bef} and ATM_{ide} provide better results than MTL for all evaluation metrics. The results confirm that the MTL-based SE system can be further improved by applying the attention-weighting mechanism. In addition, ATM_{ide} yields scores superior to ATM_{bef} implying that a suitable attention mechanism further promotes the system capability.

To further analyze the benefits of the proposed systems, we report the detailed PESQ and STOI scores of Table I in Tables II and III, respectively. We compared the performance of Noisy, LSTM, MTL, ATM_{bef} , and ATM_{ide} with respect to four testing noise environments over all SNR levels. From both tables, we observe that all DL-based SE approaches provide better PESQ and STOI scores on all evaluated conditions than the noisy baseline, while ATM_{ide} performs the best. The results verify the capability of the proposed ATM approach to extract robust features for SE, thus further improving speech quality and intelligibility.

2) *SI results*: Figure 4 illustrates the frame-wise SI accuracy of the DNN-SI baseline, MTL, ATM_{bef} , and ATM_{ide} . The evaluations were conducted on testing utterances involving “engine”, “pink”, “street”, and “white” noise backgrounds, among which “street” is considered to be the most complex noise type. From the figure, it can be observed that MTL-based approaches (MTL, ATM_{bef} , and ATM_{ide}) provide higher SI accuracies than those achieved by conventional DNN-SI. In addition, ATM_{ide} shows the highest recognition accuracy in the street background, and competes with MTL in other noise environments. The results demonstrate that the MTL architecture can effectively enhance the SI performance and can be further improved by incorporating the attention-weighting mechanism.

Next, we analyze the speaker features using DNN-SI and ATM_{ide} based on t-SNE analyses [44]. The t-SNE analysis is a widely

TABLE II

THE AVERAGED SCORES OF PESQ WITH RESPECT TO FOUR DIFFERENT NOISE ENVIRONMENTS OVER ALL SNR LEVELS, ACHIEVED BY NOISY, LSTM-SE, MTL, ATM_{bef} , AND ATM_{ide} SYSTEMS.

	Noisy	LSTM-SE	MTL	ATM_{bef}	ATM_{ide}
WHITE	1.25	2.01	2.00	2.08	2.13
PINK	1.28	1.88	1.88	1.96	2.02
STREET	1.32	1.84	1.83	1.89	1.92
ENGINE	1.16	1.72	1.71	1.81	1.84

TABLE III

THE AVERAGED SCORES OF STOI WITH RESPECT TO DIFFERENT NOISE ENVIRONMENTS OVER ALL SNR LEVELS, ACHIEVED BY NOISY, LSTM-SE, MTL, ATM_{bef} , AND ATM_{ide} .

	Noisy	LSTM-SE	MTL	ATM_{bef}	ATM_{ide}
WHITE	0.75	0.75	0.75	0.76	0.77
PINK	0.72	0.72	0.73	0.73	0.74
STREET	0.72	0.74	0.75	0.75	0.76
ENGINE	0.69	0.70	0.71	0.71	0.73

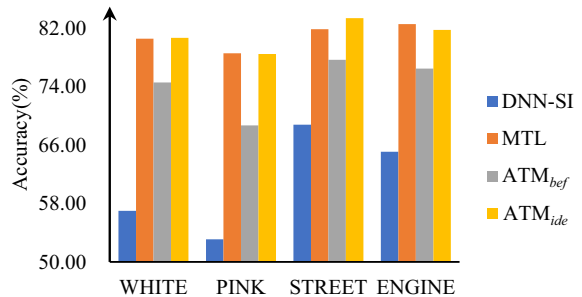


Fig. 4. The frame-wise SI accuracy of DNN-SI, MTL, ATM_{bef} , and ATM_{ide} in four testing noise environments.

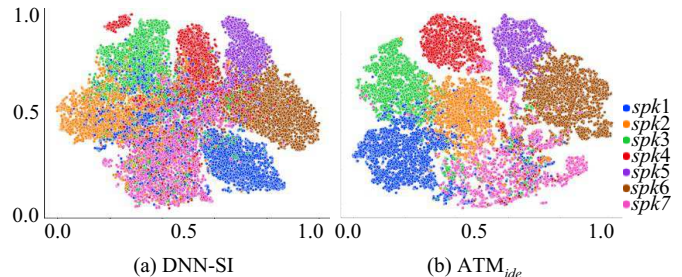


Fig. 5. The distributions of (a) DNN-SI and (b) ATM_{ide} extracted speaker features with t-SNE analyses.

used technique that provides visualized feature clusters from high-dimensional spaces. In this study, seven speakers were involved in the training set (including one non-speech virtual speaker). The analysis was carried out by first placing all SI-testing noisy utterances on the input of DNN-SI or ATM_{ide} to derive the associated speaker features. Then, these high-dimensional DNN-SI- and ATM_{ide} -extracted speaker features were processed by t-SNE to yield two-dimensional representations. Fig. 5 illustrates the distributions of these dimension-reduced (a) DNN-SI and (b) ATM_{ide} features with associated speaker identities. In the figure, it can be observed that the ATM_{ide} system provides a larger inter-class distance and a clearer class boundary than those of the DNN-SI baseline. The results show that a combination of MTL and AttNet techniques can extract more representative features for the SI task.

V. CONCLUSION

In this study, we proposed a novel ATM approach that integrates the MTL and the attention-weighting mechanism to carry out SE and SI tasks simultaneously. The overall ATM system is composed of SE, SI, and AttNet modules, and is able to extract representative and robust acoustic features in a noisy environment. Experimental results on the simulated dialog conditions confirm that the proposed ATM can significantly reduce the noise components from the noisy speech, thereby improving speech quality and intelligibility for the SE task. Meanwhile, a suitable attention mechanism performed in ATM could further improve the enhancement performance. On the other hand, the recognition accuracy of the SI system can be further improved through the proposed ATM approach. In the future, we plan to test the ATM system with other languages. We will also explore ATM by using other types of SE and SI models. Finally, we will test the proposed ATM architecture on speaker-diarization and speech-source separation tasks.

REFERENCES

- [1] DeLiang Wang, "Deep learning reinvents the hearing aid," *IEEE spectrum*, vol. 54, no. 3, pp. 32–37, 2017.
- [2] Ying-Hui Lai, Fei Chen, Syu-Siang Wang, Xugang Lu, Yu Tsao, and Chin-Hui Lee, "A deep denoising autoencoder approach to improving the intelligibility of vocoded speech in cochlear implant simulation," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 7, pp. 1568–1578, 2016.
- [3] Felix Weninger, Hakan Erdogan, Shinji Watanabe, Emmanuel Vincent, Jonathan Le Roux, John R Hershey, and Björn Schuller, "Speech enhancement with lstm recurrent neural networks and its application to noise-robust asr," in *Proc. LVA/ICA*, 2015, pp. 91–99.
- [4] Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.
- [5] Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen, "Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification," in *Proc. SLT*, 2016, pp. 305–311.
- [6] Suwon Shon, Hao Tang, and James Glass, "Voiceid loss: Speech enhancement for speaker verification," *arXiv preprint arXiv:1904.03601*, 2019.
- [7] K A Al-Karawi, A H Al-Noori, Francis F Li, Tim Ritchings, et al., "Automatic speaker recognition system in adverse conditions—implication of noise and reverberation on system performance," *International Journal of Information and Electronics Engineering*, vol. 5, no. 6, pp. 423–427, 2015.
- [8] Philipos C Loizou, *Speech enhancement: theory and practice*, CRC press, 2013.
- [9] Steven Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 27, no. 2, pp. 113–120, 1979.
- [10] Jae Soo Lim and Alan V Oppenheim, "Enhancement and bandwidth compression of noisy speech," *Proceedings of the IEEE*, vol. 67, no. 12, pp. 1586–1604, 1979.
- [11] Yariv Ephraim and David Malah, "Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator," *IEEE Transactions on acoustics, speech, and signal processing*, vol. 32, no. 6, pp. 1109–1121, 1984.
- [12] Yariv Ephraim and David Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *IEEE transactions on acoustics, speech, and signal processing*, vol. 33, no. 2, pp. 443–445, 1985.
- [13] P. C. Loizou, "Speech enhancement based on perceptually motivated bayesian estimators of the magnitude spectrum," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 5, pp. 857–869, 2005.
- [14] Kuldip Paliwal, Kamil Wójcicki, and Belinda Schwerin, "Single-channel speech enhancement using spectral subtraction in the short-time modulation domain," *Speech communication*, vol. 52, no. 5, pp. 450–475, 2010.
- [15] Thomas Lotter and Peter Vary, "Speech enhancement by map spectral amplitude estimation using a super-gaussian speech model," *EURASIP journal on applied signal processing*, vol. 2005, pp. 1110–1126, 2005.
- [16] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.
- [17] Yan Zhao, Zhong-Qiu Wang, and DeLiang Wang, "Two-stage deep learning for noisy-reverberant speech enhancement," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 1, pp. 53–62, 2018.
- [18] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee, "Dynamic noise aware training for speech enhancement based on deep neural networks," in *Proc. INTERSPEECH*, 2014, pp. 2670–2674.
- [19] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori, "Speech enhancement based on deep denoising autoencoder," in *Proc. Interspeech*, 2013, vol. 2013, pp. 436–440.
- [20] Cheng Yu, Ryandhimas E Zezario, Syu-Siang Wang, Jonathan Sherman, Yi-Yen Hsieh, Xugang Lu, Hsin-Min Wang, and Yu Tsao, "Speech enhancement based on denoising autoencoder with multi-branched encoders," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2756–2769, 2020.
- [21] Jaeyoung Kim, Mostafa El-Khamy, and Jungwon Lee, "T-gsa: Transformer with gaussian-weighted self-attention for speech enhancement," in *Proc. ICASSP*, 2020, pp. 6649–6653.
- [22] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *arXiv preprint arXiv:1706.03762*, 2017.
- [23] Zhuo Chen, Shinji Watanabe, Hakan Erdogan, and John R Hershey, "Speech enhancement and recognition using multi-task learning of long short-term memory recurrent neural networks," in *Proc. INTERSPEECH*, 2015, pp. 3274–3278.
- [24] Geon Woo Lee and Hong Kook Kim, "Multi-task learning u-net for single-channel speech enhancement and mask-based voice activity detection," *Applied Sciences*, vol. 10, no. 9, pp. 3230, 2020.
- [25] Jing Shi, Jiaming Xu, and Bo Xu, "Which ones are speaking? speaker-inferred model for multi-talker speech separation," in *Proc. INTERSPEECH*, 2019, pp. 4609–4613.
- [26] Sebastian Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.
- [27] Yu Zhang and Qiang Yang, "A survey on multi-task learning," *arXiv preprint arXiv:1707.08114*, 2017.
- [28] Michael Crawshaw, "Multi-task learning with deep neural networks: A survey," *arXiv preprint arXiv:2009.09796*, 2020.
- [29] Giovanni Morrone, Daniel Michelsanti, Zheng-Hua Tan, and Jesper Jensen, "Audio-visual speech inpainting with deep learning," *arXiv preprint arXiv:2010.04556*, 2020.
- [30] Jen-Cheng Hou, Syu-Siang Wang, Ying-Hui Lai, Yu Tsao, Hsiu-Wen Chang, and Hsin-Min Wang, "Audio-visual speech enhancement using multimodal deep convolutional neural networks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 2, pp. 117–128, 2018.
- [31] Nanxin Chen, Yanmin Qian, and Kai Yu, "Multi-task learning for text-dependent speaker verification," in *Proc. INTERSPEECH*, 2015, pp. 185–189.
- [32] Zhiyuan Tang, Lantian Li, and Dong Wang, "Multi-task recurrent model for speech and speaker recognition," in *Proc. APSIPA*, 2016, pp. 1–4.
- [33] Gueorgui Pironkov, Stéphane Dupont, and Thierry Dutoit, "Speaker-aware multi-task learning for automatic speech recognition," in *Proc. ICPR*, 2016, pp. 2900–2905.
- [34] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.
- [35] Ehsan Variani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Proc. ICASSP*, 2014.
- [36] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *Proc. ICASSP*, 2018, pp. 5329–5333.
- [37] MW Huang, "Development of taiwan mandarin hearing in noise test," *Master thesis, Department of speech language pathology and audiology, National Taipei University of Nursing and Health science*, 2005.
- [38] Alex Kendall, Yarin Gal, and Roberto Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in *Proc. CVPR*, 2018.
- [39] Ozan Sener and Vladlen Koltun, "Multi-task learning as multi-objective optimization," *arXiv preprint arXiv:1810.04650*, 2018.
- [40] Guoning Hu and DeLiang Wang, "A tandem algorithm for pitch estimation and voiced speech segregation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 8, pp. 2067–2079, 2010.
- [41] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra, "Perceptual evaluation of speech quality (pesq)—a new method for speech quality assessment of telephone networks and codecs," in *Proc. ICASSP*. IEEE, 2001.
- [42] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper 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.
- [43] Jingdong Chen, Jacob Benesty, Yiteng Arden Huang, and Eric J Dithorn, "Fundamentals of noise reduction," in *Springer Handbook of Speech Processing*, pp. 843–872. Springer, 2008.
- [44] Laurens van der Maaten and Geoffrey Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. 11, pp. 2579–2605, 2008.