Towards Intelligibility-Oriented Audio-Visual Speech Enhancement

Tassadaq Hussain[1], *Mandar Gogate*[1], *Kia Dashtipour*[1], *Amir Hussain*[1]

[1]School of Computing, Edinburgh Napier University

{t.hussain, m.gogate, k.dashtipour, a.hussain}@napier.ac.uk

Extended Abstract

Speech is the most effective and natural medium of communication in human-to-–human interaction. The goal of speech enhancement (SE) is to improve quality and intelligibility of original speech in real world environments where speech is often distorted by additive or convolutive noises. In the literature, extensive research has been carried out to propose a number of SE methods, such as, speech coding [1-3], assistive hearing devices [4] [5], automatic speech recognition (ASR) [6] [7]. In recent years, machine-learning-based SE approaches have caught great attention. These approaches generally use a mapping function, which aims to transform the noisy to clean speech based on a machine-learning-based model, to reconstruct clean speech from the noisy input. Notable machine-learning-based SE approaches include sparse coding [8], robust PCA (RPCA) [9], and non-negative matrix factorization (NMF) [10] [11].

Recently, deep learning (DL) based models have been applied to the SE field and yielded outstanding performance. For example, a deep denoising autoencoder (DDAE) framework has been developed by stacking multiple denoising autoencoders and demonstrated promising SE performance [12]. Subsequently, a deep neural network (DNN) is adopted to handle a wide range of additive noises for the SE task [13]. In addition to standard feed-forward neural networks, different structures of convolutional neural networks (CNNs) have also been employed to scrutinize the performance of the system for SE tasks. In [14], a CNN is trained in an encoder-decoder style having an additional temporal convolutional module to provide real-time ASE. In [15], a fully convolutional neural network (FCN) is exploited to effectively recover enhanced speech waveforms for ASE in an end-to-end manner. Different from traditional DL-based approaches, the authors in [16] adopted a novel strategy and trained an FCN using an objective evaluation-based cost function for better speech perception. Research has shown that visual modality carries important information, such as lip motions and mouth articulations that can help discriminate similar speech sounds in noisy conditions. Recent examples in the use of multimodal approaches for addressing speech related issues by leveraging auditory and visual information to improve the overall performance include DL models for building an AV SE system and have improved the noise reduction performance [17-20].

Despite the excellent performance achieved by DL approaches, the parameters of DL-based approaches are optimized using mean squared error (MSE) which may not be an optimal performance metric for speech-related applications because the MSE estimates the distance of two signals and does not consider human perception. We believe that optimizing the human perception-based evaluation metrics directly may lead to optimal results corresponding to the target task. To assess the performance of SE methods, we usually use some evaluation metrics that are derived based on human auditory perception. However, MSE is commonly used as the objective function for SE model optimization. There are two widely used popular metrics, perceptual evaluation of speech quality (PESQ) [21] and short-time objective intelligibility (STOI) [22], which are used to approximate the subjective speech quality and intelligibility, respectively. Apart from conventional MSE-based DL approaches, a number of STOI-metric based DL approaches have been proposed and shown to be effective in improving the intelligibility. For example, in [16], the authors utilized the STOI measure as an objective function to optimize an audio (A)-only fully convolutional network (FCNN) model for SE. Fig. 1. shows the basic I-O A-only SE system with DL framework. In addition, the authors in [23] proposed a DL-based speech intelligibility assessment model formed by the combination of a CNN and bidirectional long short-term memory (BLSTM) architecture with a multiplicative attention mechanism. More recently, the authors in [24] studied the influence of six different loss functions (including STOI-based cost function) and evaluated them in a structured manner with end-to-end time-domain DL-based SE systems.

In this study, we plan to combine visual and acoustic inputs to further optimize the speech intelligibility in listening environments in which traditional A-only STOI-based systems prove ineffective (see Fig. 2). The study aims to address a full range of challenges associated with the co-creation of I-O AV SE in accordance with the listening test. The plan is to derive advanced intelligibility-based algorithms by investigating the STOI as an objective function for AV input using DL architectures. The overall goal is to address key limitations of the most popular I-O SE frameworks, through the development of alternate novel multimodal DL algorithms for intelligibility enhancement. The newly proposed I-O AV SE frameworks will be trained with a synthetic mixture of Grid and CHiME3 noises and evaluated with the ASPIRE corpus, which was recorded in real noisy situations [17]. In our full paper submission at the workshop, we will compare the performance of our proposed framework against traditional MSE-based A-only and AV SE frameworks, as well as recently proposed STOI-based A-only methods, in terms of standardized objective and subjective evaluations for synthetic and real noise conditions under matched and mismatched testing conditions.

Index Terms: speech enhancement, audio-visual speech enhancement, objective intelligibility, deep learning

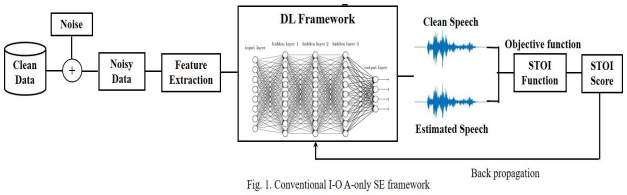

Fig. 1. Conventional I-O A-only SE framework

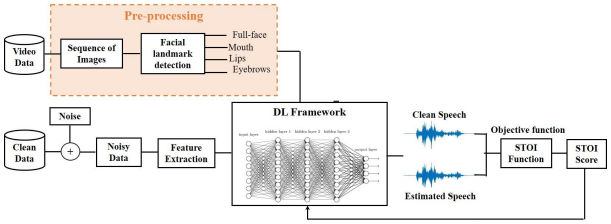

Fig. 2. Proposed I-O AV SE framework

References

[1] J. Li, S. Shuichi, S. Hongo, M. Akagi, and Y. Suzuki. "Two-stage binaural speech enhancement with Wiener filter for high-quality speech communication." *Speech Communication* 53, no. 5 (2011): 677-689.

[2] J. Li, L. Yang, J. Zhang, Y. Yan, Y. Hu, Masato Akagi, and Philipos C. Loizou. "Comparative intelligibility investigation of single-channel noise-reduction algorithms for Chinese, Japanese, and English." *The Journal of the Acoustical Society of America* 129, no. 5 (2011): 3291-3301.

[3] J. S. Lim, and A. V. Oppenheim. "Enhancement and bandwidth compression of noisy speech." *Proceedings of the IEEE* 67, no. 12 (1979): 1586-1604.

[4] A. Chern, Y.-H. Lai, Y.-P. Chang, Y. Tsao, R. Y. Chang, and H.-W. Chang. "A smartphone-based multi-functional hearing assistive system to facilitate speech recognition in the classroom." *IEEE Access* 5 (2017): 10339-10351.

[5] D. Wang, "Deep learning reinvents the hearing aid." *IEEE spectrum* 54, no. 3 (2017): 32-37.

[6] R. Li, X. Wang, S. H. Mallidi, S. Watanabe, T. Hori, and H. Hermansky, "Multi-stream end-to-end speech recognition, " IEEE/ACM Transactions on Audio, Speech, and Language Process., vol. 28, pp. 646–655, 2020.

[7] S. Wang, W. Li, S. M. Siniscalchi, and C. Lee, "A cross-task transfer learning approach to adapting deep speech enhancement models to unseen background noise using paired senone classifiers, " in Proc. ICASSP, 2020.

[8] Y. He, G. Sun, and J. Han. "Spectrum enhancement with sparse coding for robust speech recognition." *Digital Signal Processing* 43 (2015): 59-70.

[9] C. Sun, Q. Zhang, J. Wang, and J. Xie. "Noise reduction based on robust principal component analysis." *Journal of Computational Information Systems* 10, no. 10 (2014): 4403-4410.

[10] H.-T. Fan, J.-W. Hung, X. Lu, S.-S. Wang, and Y. Tsao. "Speech enhancement using segmental nonnegative matrix factorization." In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4483-4487. IEEE, 2014.

[11] P. Smaragdis, C. Fevotte, G. J. Mysore, N. Mohammadiha, and M. Hoffman. "Static and dynamic source separation using nonnegative factorizations: A unified view." *IEEE Signal Processing Magazine* 31, no. 3 (2014): 66-75.

[12] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder," in Proc. INTERSPEECH, pp. 436-440, 2013.

[13] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," IEEE/ACM Transactions on Audio, Speech, Language Processing, vol. 23, no. 1, pp. 7–19, 2015.

[14] A. Pandey and D. Wang, "TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain," in Proc. ICASSP, pp. 6875–6879, 2019.

[15] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in Proc. APSIPA, pp. 6–12, 2017.

[16] S.-W. Fu, T.-W. Wang, Y. Tsao, X. Lu, and H. Kawai, "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks," IEEE/ACM Transactions on Audio, Speech and Language Processing, vol. 26, no. 9, pp. 1570-1584, 2018.

[17] M. Gogate, K. Dashtipour, A. Adeel, and A. Hussain, "CochleaNet: A robust language-independent audio-visual model for real-time speech enhancement," Information Fusion, vol. 63, pp.273-285, 2020.

[18] M. Gogate, K. Dashtipour, P. Bell, and A. Hussain, "Deep neural network driven binaural audio visual speech separation, " in Proc. IJCNN, pp. 1-7, 2020.

[19] A. Adeel, M. Gogate, and A. Hussain, "Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments," Information Fusion, vol. 59, pp.163-170, 2020.

[20] M. Gogate, K. Dashtipour, and A. Hussain, "Visual Speech In real Noisy environments (VISION): A novel benchmark dataset and deep learning-based baseline system," Proc. INTERSPEECH, pp.4521-4525, 2020.

[21] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in Proc. ICASSP, vol. 2, pp. 749–752, 2001

[22] 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, Language Process., vol. 19, no. 7, pp. 2125–2136, 2011.

[23] R. E. Zezario, S.-W. Fu, C. -S. Fuh, Y. Tsao, and H. -M. Wang. "STOI-Net: A Deep Learning based Non-Intrusive Speech Intelligibility Assessment Model." In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 482-486. IEEE, 2020.

[24] M. Kolbæk, Z. -H. Tan, S. H. Jensen, and J. Jensen. "On loss functions for supervised monaural time-domain speech enhancement." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 28 (2020): 825-838.