Towards Intelligibility-Oriented Audio-Visual Speech Enhancement

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

Edinburgh Napier University

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

Abstract

Existing deep learning (DL) based approaches are generally optimised to minimise the distance between clean and enhanced speech features. These often result in improved speech quality however they suffer from a lack of generalisation and may not deliver the required speech intelligibility in real noisy situations. In an attempt to address these challenges, researchers have explored intelligibility-oriented (I-O) loss functions and integration of audio-visual (AV) information for more robust speech enhancement (SE). In this paper, we introduce DL based I-O SE algorithms exploiting AV information, which is a novel and previously unexplored research direction. Specifically, we present a fully convolutional AV SE model that uses a modified short-time objective intelligibility (STOI) metric as a training cost function. To the best of our knowledge, this is the first work that exploits the integration of AV modalities with an I-O based loss function for SE. Comparative experimental results demonstrate that our proposed I-O AV SE framework outperforms audio-only (AO) and AV models trained with conventional distance-based loss functions, in terms of standard objective evaluation measures when dealing with unseen speakers and noises. [1]

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

1. Introduction

The main goal of a speech enhancement (SE) system is to improve the quality and intelligibility of speech in real world environments where speech is often distorted by multiple-competing additive or convolutive noises. In the literature, extensive research has been carried out to develop SE methods for speech coding [1, 2, 3], assistive hearing devices [4] [5] and automatic speech recognition (ASR) [6] [7]. In recent years, machine-learning-based SE approaches have received great attention due to their ability to outperform state-of-the-art SE models. These approaches generally use machine-learning based mapping functions to reconstruct the clean speech from 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 exploited in the SE field and yielded enhanced performance. For example, a deep denoising autoencoder (DDAE) framework has demonstrated promising SE performance compared to traditional methods [12]. Subsequently, a deep neural network (DNN) was 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 been employed in an attempt to improve the generalisation performance for SE. In [14], an audio-only (AO) CNN was trained in an encoder-decoder style with an additional temporal convolutional module to provide real-time SE. In [15], a fully convolutional neural network (FCN) was exploited to effectively recover the enhanced speech waveform for AO SE in an end-to-end manner. Different from traditional DL-based approaches, authors in [16] adopted a novel strategy and trained FCN using an objective evaluation-based cost function for enhanced speech perception. Research has shown that the visual modality carries important information (such as lip motions and mouth articulations) that can help discriminate similar speech sounds in noisy conditions. Recent examples on the use of multimodal approaches to address speech related issues by leveraging AV information to improve performance, include DL based AV SE systems, which have shown significant improvement in noise reduction. [17, 18, 19, 20, 21, 22, 17, 23].

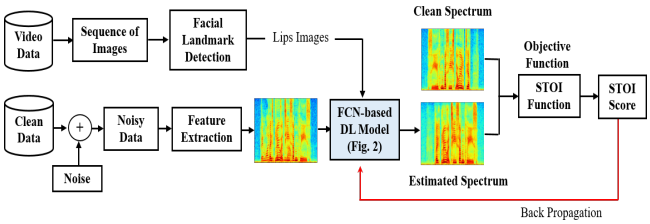


Figure 1: *Block diagram of our proposed STOI-based audio-visual SE model*

Despite the excellent performance achieved by DL based SE models, the parameters of such approaches are often optimized using distance-based loss functions including mean squared error (MSE) and mean absolute error (MAE). However, these may not be optimal performance evaluation metrics for speech-related applications as they are not based on human auditory perception. In addition, we believe that the optimizing human perception-based evaluation metrics directly may lead to more optimal results corresponding to the target task. In the context of SE, researchers usually employ number of performance evaluation metrics that are inspired by human auditory perception. There are two widely used metrics, specifically, perceptual evaluation of speech quality (PESQ) [24] and short-time objective intelligibility (STOI) [25], which are used to approximate subjective speech quality and intelligibility, respectively. Apart from conventional MSE/MAE-based DL approaches, a number of intelligibility-oriented (I-O) STOI-metric based DL approaches have also been proposed and shown to be effective for SE. For example, in [16], authors utilized the STOI measure as an objective function to optimize an AO fully-convolutional network (FCNN) model for SE. The results demonstrated that the STOI-based SE framework can perform significantly better than a conventional MSE-based SE system due to increased consistency between the training and evaluation target. In addition, authors in [26] proposed a DL-based

[1] Equal contribution.

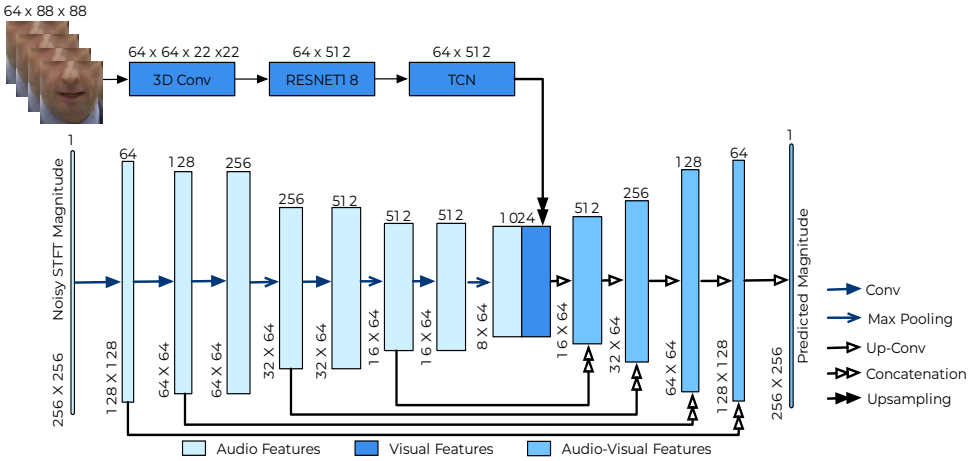


Figure 2: *The FCN-based U-Net framework used to optimize the STOI-based audio-visual SE model*

speech intelligibility assessment model by combining a CNN and bidirectional long short-term memory (BLSTM) architecture with a multiplicative attention mechanism. More recently, authors in [27] studied the influence of six different loss functions (including the STOI-based cost function) and evaluated them in a structured manner with end-to-end time-domain DL-based SE systems.

Motivated by the promising performance achieved by STOI-based SE systems, we further develop and extend conventional STOI-based AO SE approaches by incorporating visual information to jointly optimize the AV SE system in listening environments in which traditional methods can prove ineffective. The aim of this study is to investigate the effectiveness of STOI as an objective function to train DL-based AV SE architectures to overcome limitations of current frameworks. Unlike previously proposed STOI-based systems which process speech in an end-to-end manner to construct an utterance-based (time-domain) SE system, we process the signal in a frame-wise manner in the frequency domain by focusing on magnitude spectra of noisy and clean speech utterances. We next use a FCN to learn the spectral mapping for AV input data and perform SE using a STOI-based cost function. This entails modifying conventional (classical and extended) STOI measures to account for signals in the frequency domain. In addition to formulating the modified STOI-based cost function, we comparatively evaluate the effectiveness of two conventional distance-based cost functions, namely MSE and MAE, to optimise AV SE system performance. All AV SE frameworks are trained and evaluated using a two-speaker synthetic mixture of the benchmark GRID corpus [28], at random Signal-to-Noise Ratios (SNR). Note that the task of suppressing speech interference is more challenging than suppressing non-speech noises. Experimental results show that our proposed AV framework optimized with a modified STOI-based cost function can achieve significant SE performance improvement over both MSE and MAE-based AO and AV frameworks, as well as recently proposed STOI-based AO methods, under mismatched testing conditions, using a range of standardized objective measures: namely, the perceptual evaluation of speech quality (PESQ), STOI, scale-invariant signal-to-distortion ratio (SI-SDR) [29], and Virtual Speech Quality Objective Listener (VISQOL) [30].

The remainder of the paper is organised as follows. Section 2 describes the methodology and our proposed STOI-based AV SE system. Section 3 presents the experimental setup including dataset description, AV feature extraction and comparative evaluation results. Finally, concluding remarks are presented in Section 4.

2. Methodology

Conventional DL-based SE models are trained using MSE and MAE loss functions. Thus, we first consider the most commonly used MSE loss function to train a FCN based AV SE model, which is implemented using the benchmark U-Net framework [31]. The MSE can be computed as follows:

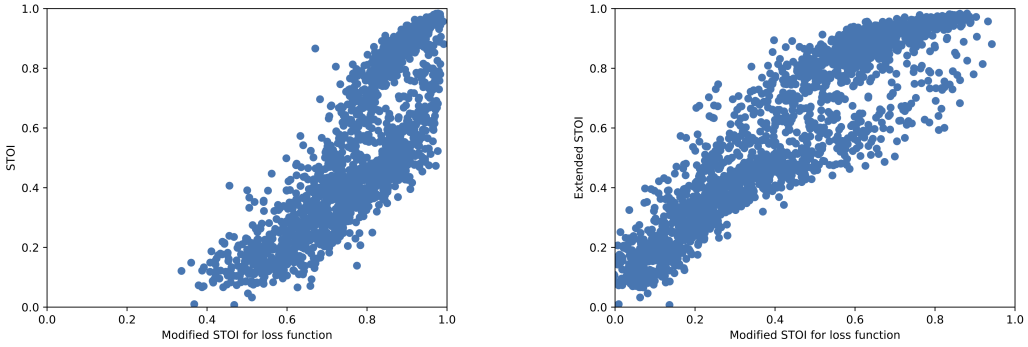
$$\mathcal{L}_{MSE} = min(\frac{1}{M}\sum_{m=1}^{M} \|\hat{Y}_m - Y_m\|_2) \quad (1)$$

where M is the total number of speech frames, $\hat{Y}_M$ is the estimated magnitude spectrum, Y_M is the reference magnitude spectrum of the utterance, and $\|\cdot\|_2$ denotes L2-normalization.

Recent studies have shown that the STOI metric is highly correlated to human perception and performs more optimally compared to MSE, suggesting it could be employed as an alternative loss function in speech-related applications [16] [27]. In this paper, we evaluate the effectiveness of a STOI based loss function to optimise the performance of a DL-based AV SE system. Specifically, we adopt a deep FCN-based U-Net architecture that takes the noisy magnitude spectrum as input and exploits a modified STOI loss function to optimally learn the spectral mapping and estimate the enhanced magnitude spectrum. Figure 1 shows the block diagram of our proposed I-O AV SE framework, and Figure 2 illustrates the FCN-based U-Net framework used to optimize our STOI-based AV SE model. Our key focus is to explore how incorporating visual information into an AO SE model and optimisation using a modified STOI loss function, impacts the quality and intelligibility of the enhanced speech signal as evaluated with a range of standardized objective measures

2.1. Short-time Objective Intelligibility (STOI)

Here, we develop and employ a modified version of a well-known STOI intelligibility measure as an objective function to train AV SE models. The STOI is an intrusive



(a) *STOI vs Modified STOI* (b) *Extended STOI vs Modified Extended STOI*

Figure 3: *Scatter Plot between (a) STOI vs Modified STOI and (b) Extended STOI vs Modified Extended STOI*

measure that requires both estimated speech and reference (clean) speech signals and ranges from 0 to 1, with 1 denoting the highest intelligibility of the speech signal. The STOI function takes the clean and estimated speech signals as input and computes the score in five steps: i) Removal of silent regions from clean and estimated speech signals, ii) Application of the short-time Fourier transform (STFT); iii) Estimation of the short-time envelope of clean and noisy speech using one-third octave-band analysis of the STFT frames; iv) Normalization and clipping to compensate for global level differences and stabilisation of the STOI evaluation; and v) Intelligibility measure computation: the correlation coefficient between the two spectral envelopes is estimated using the equations below.

$$d_{i,j} = \frac{(y_{i,j} - \mu_{y_{i,j}})^{\intercal}(\hat{y}_{i,j} - \mu_{y_{i,j}})^{\intercal}}{\|y_{i,j} - \mu_{y_{i,j}}\|_2 \|\hat{y}_{i,j} - \mu_{\hat{y}_{i,j}}\|_2} \tag{2}$$

where y and $\hat{y}$ are the short-time spectral envelope of the reference clean and estimated speech signals, $\mu_{y_{i,j}}$ and $\mu_{\hat{y}_{i,j}}$ are the corresponding sample mean vectors, and $\|\cdot\|_2$ represents the L2-normalization. The final STOI is the average of the intelligibility measure over all bands and frames.

$$d_{STOI} = \frac{1}{I(M-N+1)} \sum_{i=1}^{I} \sum_{j=1}^{J} d_{i,j} \tag{3}$$

where I = 15 is the number of one-third octave band and $M - N + 1$ is the total number of short-time temporal envelope vector. For a more detailed setting of each step, please refer to [32]. The computation of STOI is differentiable, thus, it can be used as the objective function directly to optimize the AV SE model.

$$\mathcal{L}_{STOI} = -\frac{1}{M} \sum_{m=1}^{M} d_{STOI}(\hat{Y}_m, Y_m) \tag{4}$$

where $d_{STOI}(\hat{Y}_m, Y_m)$ measures the STOI score between the estimated and clean magnitude spectra of audio utterances. Unlike the MSE, where the goal is to reduce the distance, we want to maximize the STOI score to enhance speech intelligibility.

2.2. Proposed Audio-Visual SE Framework

This section presents the DL models used for our I-O AV SE framework as depicted in Fig. 2. Specifically the network architecture for required AV feature extraction, fusion and speech resynthesis pipeline is outlined below.

2.2.1. Audio feature extraction

The audio feature extraction stage utilises a U-net [31] style network consisting of an encoder and decoder block modified for AV SE. The input to the network is the magnitude of noisy speech Short-Time Fourier Transform (STFT) of dimension $F \times T$ where F and T are frequency and time dimension of the spectrogram. The input is fed to two convolutional layers with filter size of 4 and stride of 2 to downsample the time-frequency dimension until the time dimension is equal to 64. The downsampled features are passed through three convolutional blocks each consisting of two convolutional layers with filter size of 3 and stride of 1, followed by a frequency pooling layer that reduces the frequency dimension by 2. Note that the spatial dimension is preserved during the processing of convolutional blocks.

2.2.2. Visual feature extraction

The visual feature extraction stage of the pipeline comprises a 3D convolutional layer with filter size of $5 \times 7 \times 7$ and stride of $1 \times 2 \times 2$, followed by RESNET-18 [33]. The residual network features are then fed to a temporal convolutional network (TCN) as described in [34]. The input to the network is a time-series of lip cropped images of size $N \times 88 \times 88$, where N is the number of frames. The visual feature network outputs a 512-D vector for each lip image. The visual features are upsampled to match the audio feature sampling rate.

2.2.3. Multimodal fusion

The upsampled visual features and audio features are concatenated and fed to a U-net decoder as shown in Fig. 2. The decoder consists of 3 up convolutional blocks each consisting of two upsampling layers that upsample the time dimension by 2, followed by convolutional layers with a filter size of 3 and stride of 1. The AV features are then fed to two transposed convolutional layers with filter size of 4 and stride of 2 to upsample the time-frequency dimension, until the

time-frequency dimension is equal to the input. Next we use a sigmoid layer to map the output in the range of 0 to 1. The predicted mask is then multiplied with the input spectrogram to generate the masked spectrogram as output.

2.2.4. Speech Resynthesis

The proposed model estimates the clean spectrogram when the noisy spectrogram and cropped lip images are fed as input. The estimated magnitude is combined with the noisy phase to generate enhanced speech using an inverse STFT.

3. Experiments and Results

3.1. Experimental Setup

For initial testing, we trained our proposed I-O AV SE model using a small vocabulary AV corpus to assess how the STOI loss function affects the overall SE performance. Specifically, the performance of the framework was evaluated using the benchmark GRID corpus [28]. The dataset contained AV recordings of clean utterances from 34 male and female speakers, each with 1000 utterances lasting around three seconds. The AV utterances were recorded in a quiet room with sufficient background lighting, with the speaker filmed facing the camera. The visual data was captured at a frame rate of 25 frames per second (fps) while the audio data was recorded at 48 kHz which was then resampled to 16 kHz. We randomly selected 23 speakers for the training set and 4 speakers each for the, validation and test sets. The split ensured speaker independence criteria i.e. there was no overlap of speakers between the training, validation and test sets. The test and validation set comprised 2 male and 2 female speakers. The clean utterances were mixed with randomly selected clean speech utterances from the respective sets, at randomly selected SNRs ranging from [0 to 20] dB with 1 dB increments. In total, training, validation and testing had 46000, 4000 and 4000 utterances respectively. To improve generalisation during training a single clean utterance was mixed with two different randomly selected interferences. In order to objectively measure the quality of denoised speech, a number of state-of-the-art evaluation metrics were used, including PESQ, STOI, SI-SDR and VISQOL. Furthermore, as proposed in [35], three additional measures were used to: compute the signal distortion of the speech signal (termed CSIG), predict the background noise in the estimated signal (termed CBAK), and predict the overall quality of the estimated speech (termed COVL)

3.2. Audio and Visual Features

We used the STFT with a frame length of 25ms and a frameshift of 10ms to process the audio speech signals. For the visual features, we converted each video into a sequence of images at a frame rate of 25 fps. The lip region of size 88 x 88 was extracted using the Dlib library and extracted lip regions were converted to greyscale.

3.3. STOI vs Modified STOI

Unlike the original (classical and extended) STOI measures, which initially down sample speech signals to 10kHz, carry out silent frame removal, and then apply STFT, we formulated a modified version of STOI (termed modified STOI) to account for 16kHz signals in the frequency domain while ignoring downsampling and silent frame removal steps. To examine the behaviour of the modified STOI, we plotted correlations between the modified STOI and original STOI (classical and extended) scores as shown in Fig. 3. Specifically, Fig. 3 (a) and (b) present scatter plots for STOI (modified STOI vs original STOI) and (modified extended STOI vs extended STOI) scores, respectively. From the figures, we can note that the modified STOI scores are strongly correlated with the original and extended STOI scores, demonstrating that our modified extended STOI correlated well with the extended STOI and can be directly used as a loss function to train and optimize the DL models for AV SE.

3.4. Objective Evaluation

We investigated the impact of our modified STOI loss function on the performance of the frequency-domain AV SE system in terms of PESQ, STOI, SI-SDR, CSIG, CBAK, COVL, and VISQOL. We used the same setup to train three AO and AV SE frameworks utilizing three different loss functions, namely $\mathcal{L}_{MSE}$, $\mathcal{L}_{MAE}$, and $\mathcal{L}_{STOI}$, and evaluated their performance using the GRID dataset (see Sec. 3.1).

Table I shows the performance comparison of AO and AV SE frameworks trained using three different loss functions. It can be seen from Table I that, both AO and AV SE frameworks, when trained with different loss functions, enhanced the original noisy speech utterances with a reasonable margin in terms of all performance measures. In short, both frameworks optimized using three loss functions proved to be effective for SE. We note that in contrast to AO SE frameworks trained with $\mathcal{L}_{MSE}$ and $\mathcal{L}_{MAE}$ functions, the AO framework trained with the $\mathcal{L}_{STOI}$ loss function demonstrated better performance in terms of PESQ, STOI, SI-SDR, CSIG, CBAK, COVL, and VISQOL scores, respectively.

Further, we note that despite the excellent performance of AO SE systems optimised using three loss functions, it can be seen from Table I that incorporating visual information into the AO systems and training with $\mathcal{L}_{MSE}$, $\mathcal{L}_{MAE}$, and $\mathcal{L}_{STOI}$ functions, further improves not only primary objective evaluation metrics like the PESQ, STOI, SI-SDR, and VISQOL, but also other metrics like the CSIG, CBAK, and COVL. In particular, we observe that our proposed AV SE framework, when trained with $\mathcal{L}_{STOI}$, improves the performance significantly in terms of PESQ, STOI, SI-SDR, and VISQOL. However, the framework under-performs when compared with $\mathcal{L}_{MAE}$ loss for the CSIG, CBAK, and COVL measures. It is to be noted that the performance of the AO SE framework optimised using the STOI loss function is similar to the AV SE framework trained using the MSE loss function. This shows that the performance improvement achieved using a distance-based metric and AV information is similar to one achieved using an I-O loss function.

Finally, we plot the spectrograms of randomly selected clean and noisy speech signals and compare them with enhanced speech signals estimated by AO and AV SE frameworks using MAE, MSE and STOI loss functions. Figures 4(a) and (b) display the spectrogram of a noisy test utterance contaminated by a female speaker's speech at 3 dB SNR and corresponding clean speech signal. Figures 4(c) and (d) show the spectrograms of the enhanced speech signal for AO and AV SE frameworks optimised using $\mathcal{L}_{MSE}$. Similarly, Fig. 4(e), (f), (g), and (h) present the spectrogram of the enhanced speech signal for the two frameworks optimised using $\mathcal{L}_{MAE}$ and $\mathcal{L}_{STOI}$. It can be seen that despite the excellent performance achieved by MSE and MAE-based AO and AV SE frameworks

Table 1: PERFORMANCE COMPARISON OF AUDIO-ONLY AND AUDIO-VISUAL DNN MODELS USING STANDARDISED OBJECTIVE EVALUATION METRICS UNDER SPEAKER-INDEPENDENT CONDITIONS.

Framework	**Loss Function**	**Objective Evaluation Metrics**							**Avg.**
		PESQ	STOI	SI-SDR	CSIG	CBAK	COVL	VISQOL	
Noisy	–	2.414	0.828	8.067	3.159	2.441	2.373	3.213	3.214
Audio-only	MSE	2.712	0.851	10.441	3.515	2.815	2.783	3.268	3.769
	MAE	2.789	0.852	10.794	3.705	3.004	3.017	3.301	3.923
	STOI	3.005	0.884	11.246	3.852	2.781	3.132	3.376	4.039
Audio-Visual	MSE	2.724	0.857	10.640	3.644	2.847	2.861	3.284	3.836
	MAE	3.008	0.887	11.753	**3.991**	**3.143**	**3.274**	3.403	4.208
	STOI	**3.206**	**0.914**	**12.403**	3.863	2.844	3.184	**3.478**	**4.270**
IRM	–	3.432	0.872	7.383	4.683	2.389	3.902	3.501	3.737

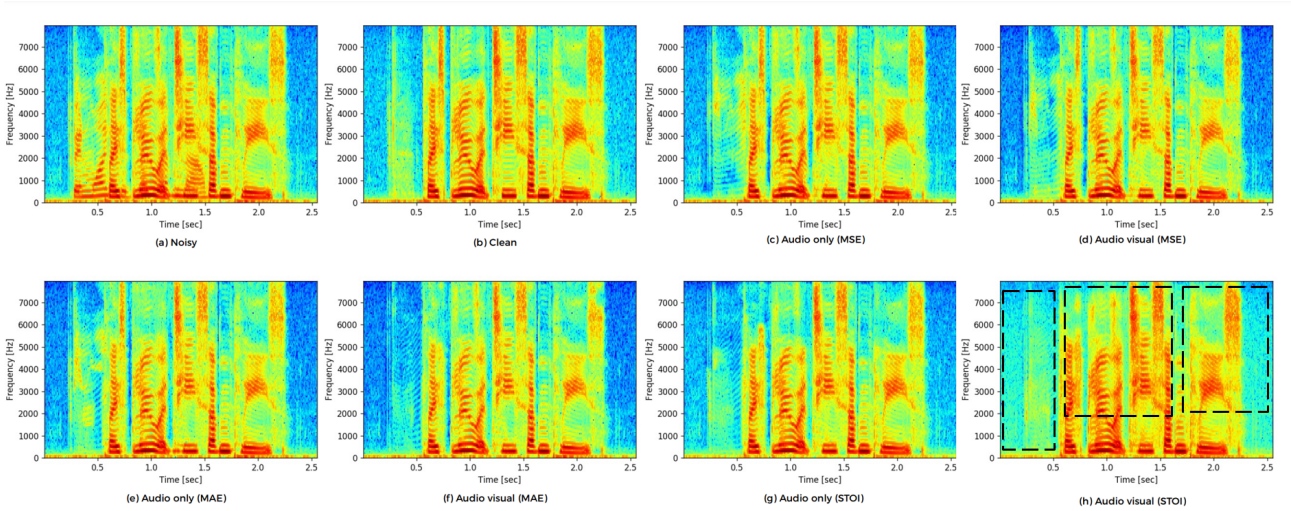


Figure 4: *Spectrogram of a randomly selected utterance from the test set. It can be seen that our modified STOI based AV SE model recovered more speech regions than conventional AO and AV SE models*

in terms of objective evaluation measures and noise suppression capabilities (evident from Fig. 4(c)–4(f)), these frameworks were unable to capture some middle and high-frequency regions when compared with STOI-based frameworks (Fig. 4(g)-(h)). Specifically, our proposed STOI-based AV SE framework (Fig. 4(h)) restored more (low-mid-high frequency region) speech components compared to MSE- and MAE-based AO and AV SE frameworks (as illustrated with dashed boxes in Fig. 4).

4. Conclusion

In this paper, we proposed a novel I-O AV SE paradigm to enhance the performance of conventional AO SE systems by exploiting an intelligibility-based evaluation metric as an alternative cost function. Specifically, we developed and utilised a modified version of the conventional STOI loss function to train AV SE models, that can effectively account for signals in the frequency domain as opposed to conventional I-O frameworks that require down sampling signals in the time-domain. Comparative experimental findings show that incorporating visual information as part of a modified STOI-based AV DL framework can estimate the output signal with enhanced speech quality and intelligibility. In summary, we found that an I-O based loss function achieves good general performance and produces better results for a variety of SE evaluation metrics, implying that the modified STOI is a promising choice to optimize frequency-domain AV SE applications. Ongoing work is aimed at evaluating our I-O AV SE system with more challenging real-world AV corpora and subjective listening tests for speech and hearing-aid applications.

5. Acknowledgements

This work is supported by the UK Engineering and Physical Sciences Research Council (EPSRC) programme grant: COG-MHEAR (Grant reference EP/T021063/1).

6. References

[1] J. Li, S. Sakamoto, S. Hongo, M. Akagi, and Y. Suzuki, "Two-stage binaural speech enhancement with wiener filter for high-quality speech communication," *Speech Communication*, vol. 53, no. 5, pp. 677–689, 2011.

[2] J. Li, L. Yang, J. Zhang, Y. Yan, Y. Hu, M. Akagi, and P. C. Loizou, "Comparative intelligibility investigation of single-channel noise-reduction algorithms for chinese, japanese, and english," *The Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. 3291–3301, 2011.

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

[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*, vol. 5, pp. 10 339–10 351, 2017.

[5] D. Wang, "Deep learning reinvents the hearing aid," *IEEE spectrum*, vol. 54, no. 3, pp. 32–37, 2017.

[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 Processing*, vol. 28, pp. 646–655, 2019.

[7] S. Wang, W. Li, S. M. Siniscalchi, and C.-H. Lee, "A cross-task transfer learning approach to adapting deep speech enhancement models to unseen background noise using paired senone classifiers," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6219–6223.

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

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

[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)*. IEEE, 2014, pp. 4483–4487.

[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*, vol. 31, no. 3, pp. 66–75, 2014.

[12] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Interspeech*, vol. 2013, 2013, pp. 436–440.

[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, and Language Processing*, vol. 23, no. 1, pp. 7–19, 2014.

[14] A. Pandey and D. Wang, "Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6875–6879.

[15] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2017, pp. 006–012.

[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] D. Michelsanti, Z.-H. Tan, S.-X. Zhang, Y. Xu, M. Yu, D. Yu, and J. Jensen, "An overview of deep-learning-based audio-visual speech enhancement and separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[18] 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.

[19] W. Wang, C. Xing, D. Wang, X. Chen, and F. Sun, "A robust audio-visual speech enhancement model," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7529–7533.

[20] M. Gogate, K. Dashtipour, P. Bell, and A. Hussain, "Deep neural network driven binaural audio visual speech separation," in *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2020, pp. 1–7.

[21] 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.

[22] M. Gogate, K. Dashtipour, and A. Hussain, "Visual speech in real noisy environments (VISION): A novel benchmark dataset and deep learning-based baseline system." in *Interspeech*, 2020, pp. 4521–4525.

[23] M. Gogate, A. Adeel, R. Marxer, J. Barker, and A. Hussain, "Dnn driven speaker independent audio-visual mask estimation for speech separation," in *Interspeech 2018*. ISCA, 2018, pp. 2723–2727.

[24] 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 *2001 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 01CH37221)*, vol. 2. IEEE, 2001, pp. 749–752.

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

[26] 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)*. IEEE, 2020, pp. 482–486.

[27] 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*, vol. 28, pp. 825–838, 2020.

[28] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *The Journal of the Acoustical Society of America*, vol. 120, no. 5, pp. 2421–2424, 2006.

[29] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr–half-baked or well done?" in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[30] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "Visqol: an objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 1–18, 2015.

[31] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

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

[33] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[34] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *ICASSP*, 2020.

[35] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.