Intelligibility-Oriented (I-O) Audio-visual Speech Enhancement

TASSADAQ HUSSAIN, MANDAR GOGATE, KIA DASHTIPUR, AMIR HUSSAIN

School of Computing

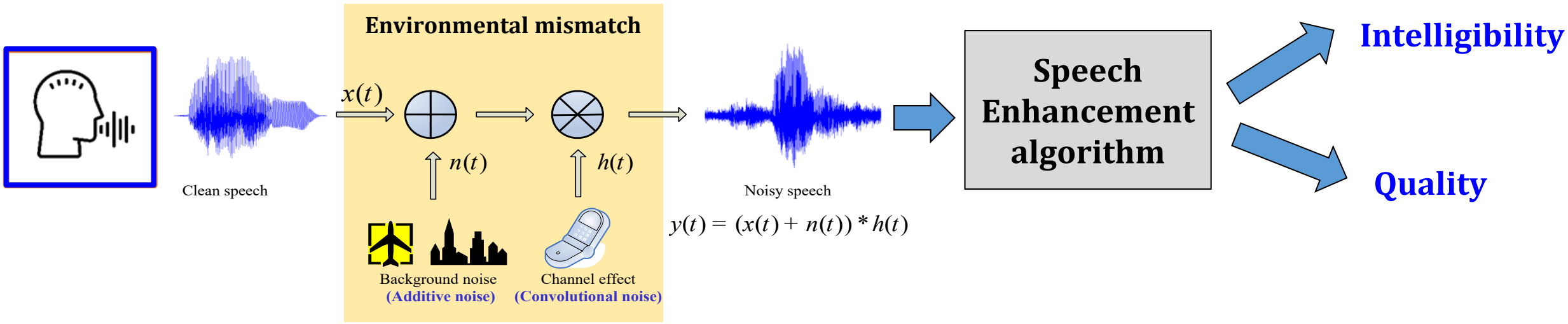
Edinburgh Napier University

t.hussain@napier.ac.uk

Outline

- Overview
- Speech Assessment Measures
- Deep Learning Based Speech Enhancement
- Objective Functions
- Proposed Intelligibility-oriented (I-O) Speech Enhancement Framework
- Experimental Setup and Preliminary Results
- Challenges and Ongoing Work

Overview

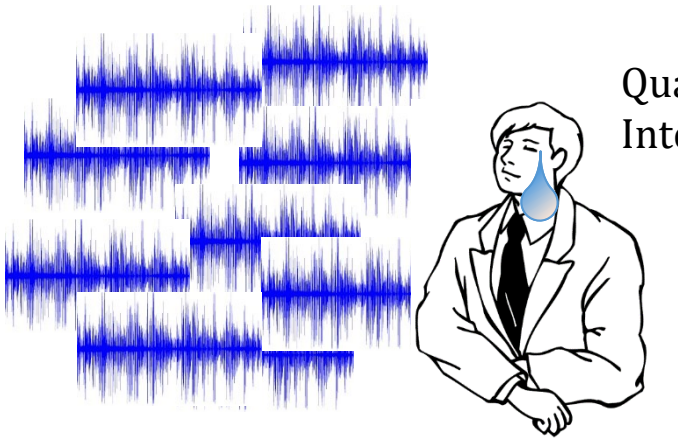


- Over the last few decades, a great amount of research has been done on various aspects and properties of speech signals.

- However, improving the intelligibility for both human listening and machine recognition in real acoustic conditions remains a highly challenging task.

Speech Assessment Measures

Speech perception
- Quality
- Intelligibility



Quality: 3
Intelligibility: 0.8

Quality:3.13
Intelligibility: 0.75

Quality:2.8
Intelligibility: 0.72

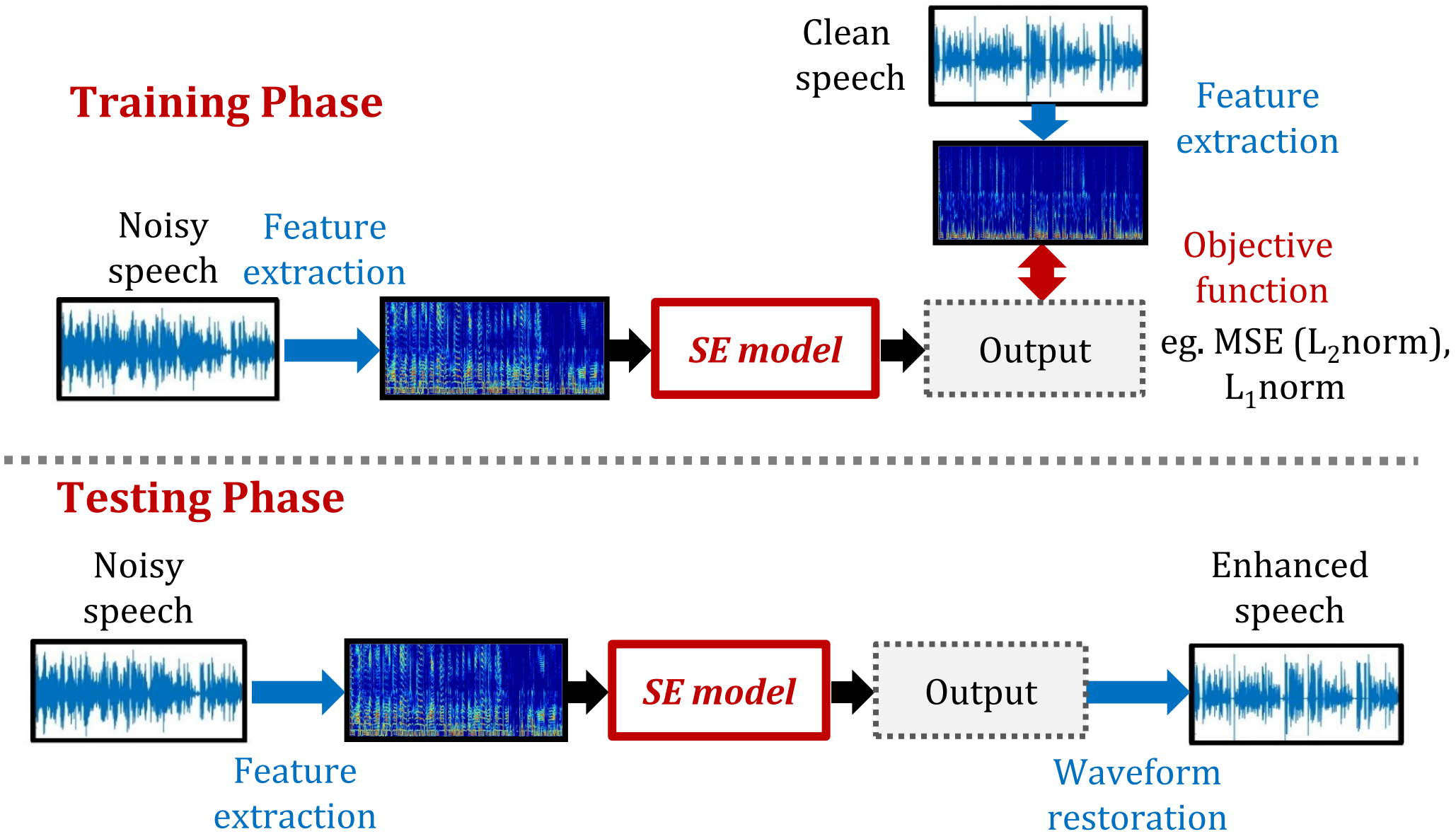
Quality:3.1
Intelligibility: 0.72

Quality:2.5

Quality:1.6
Intelligibility: 0.72

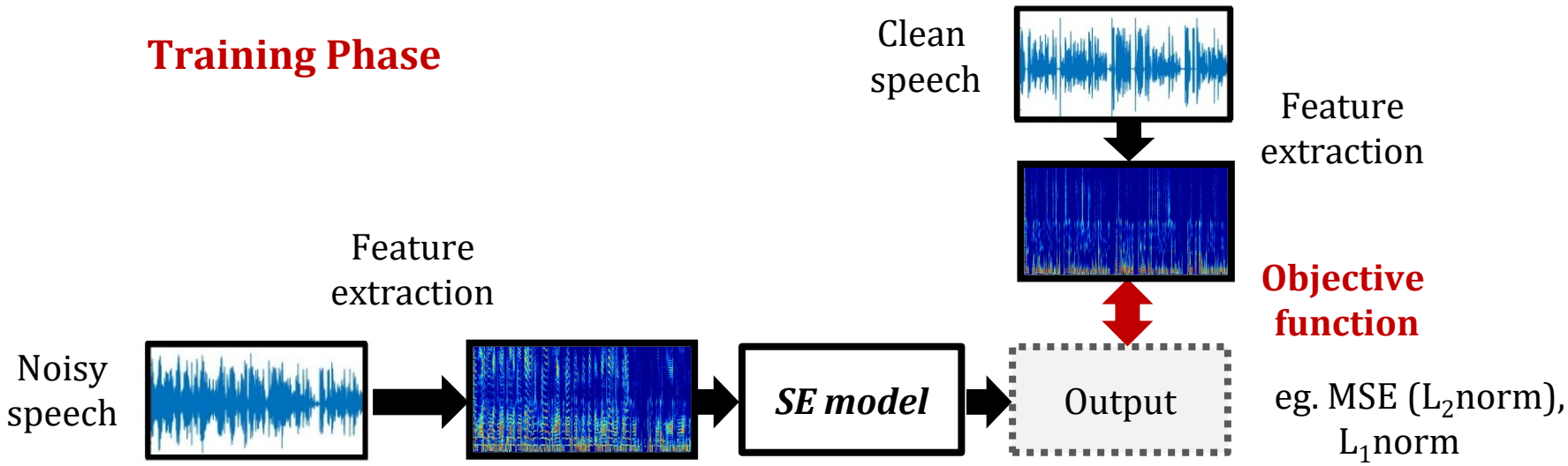
Intelligibility: 0.89

Deep Learning based Speech Enhancement (SE)



Objective Functions

- Conventional deep learning based speech enhancement models are trained using the L1 norm or L2 norm (MSE)
- Mean squared error (MSE) and L1 losses aim to minimize the difference between enhanced and target speech signals and do not directly consider human perception and ASR performance.
- In this work, we aim to exploit well-known short-time objective intelligibility (STOI) metric as an objective function to train audio-visual models to increase speech intelligibility.

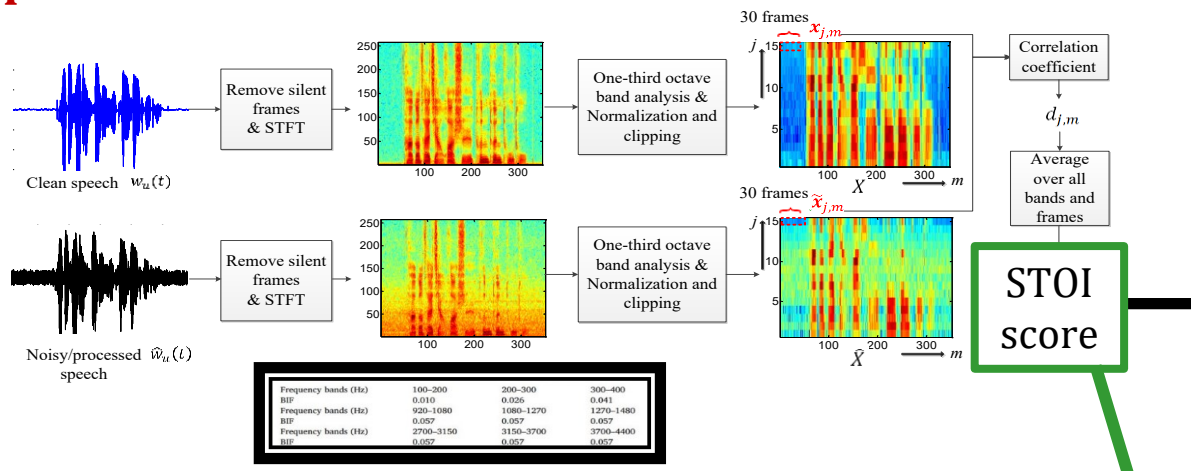


Mean squared error (MSE) and L1 losses aim to minimize the differences of enhanced and target and do not directly consider human perception and ASR performance.

Objective Function

- **STOI-based Objective Function [1][2]**

STOI Computation



The intermediate intelligibility measure is defined as the correlation coefficient between the temporal envelopes of the clean and degraded speech

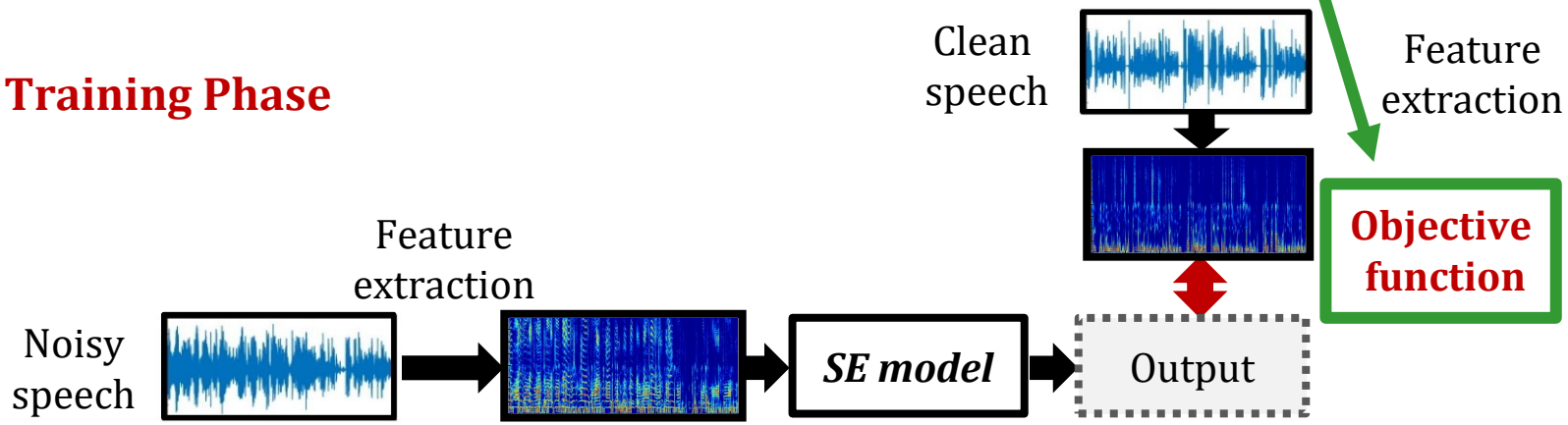
$$d_{j,m} = \frac{\left(\boldsymbol{x}_{j,m} - \mu_{\boldsymbol{x}_{j,m}}\right)^T \left(\tilde{\boldsymbol{x}}_{j,m} - \mu_{\tilde{\boldsymbol{x}}_{j,m}}\right)}{\left\|\boldsymbol{x}_{j,m} - \mu_{\boldsymbol{x}_{j,m}}\right\|_2 \left\|\tilde{\boldsymbol{x}}_{j,m} - \mu_{\tilde{\boldsymbol{x}}_{j,m}}\right\|_2}$$

The objective function can be represented as

$$\mathrm{O} = -\frac{1}{U}\sum_{u} stoi\left(w_u\left(t\right), \hat{w}_u\left(t\right)\right)$$

where $w_u(t)$ and $\hat{w}_u(t)$ are the clean and estimated utterance with index u, respectively, and U is the total number of training utterances. $stoi(.)$ is the function that includes the five steps

Training Phase



[1] 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.

[2] 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.

Proposed I-O Audio-visual (AV) Speech Enhancement Framework

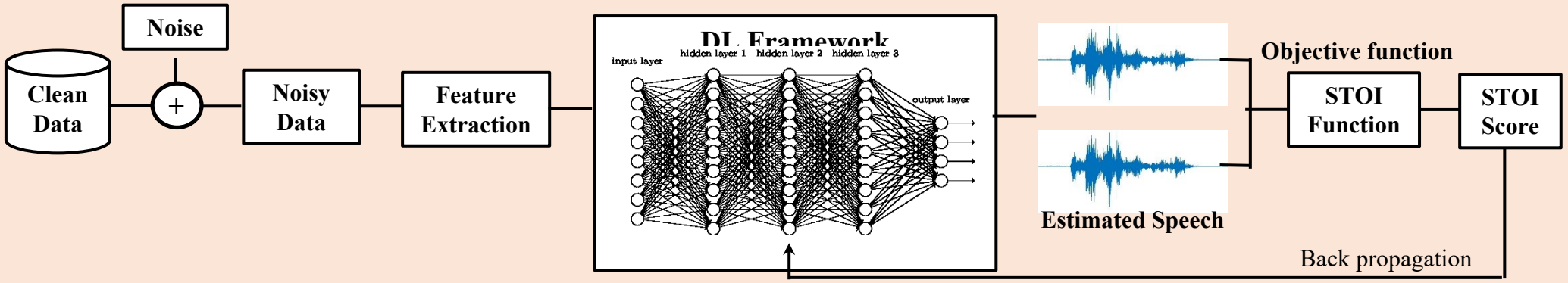


Fig. 1. Intelligibility-oriented Audio-only Speech Enhancement Framework

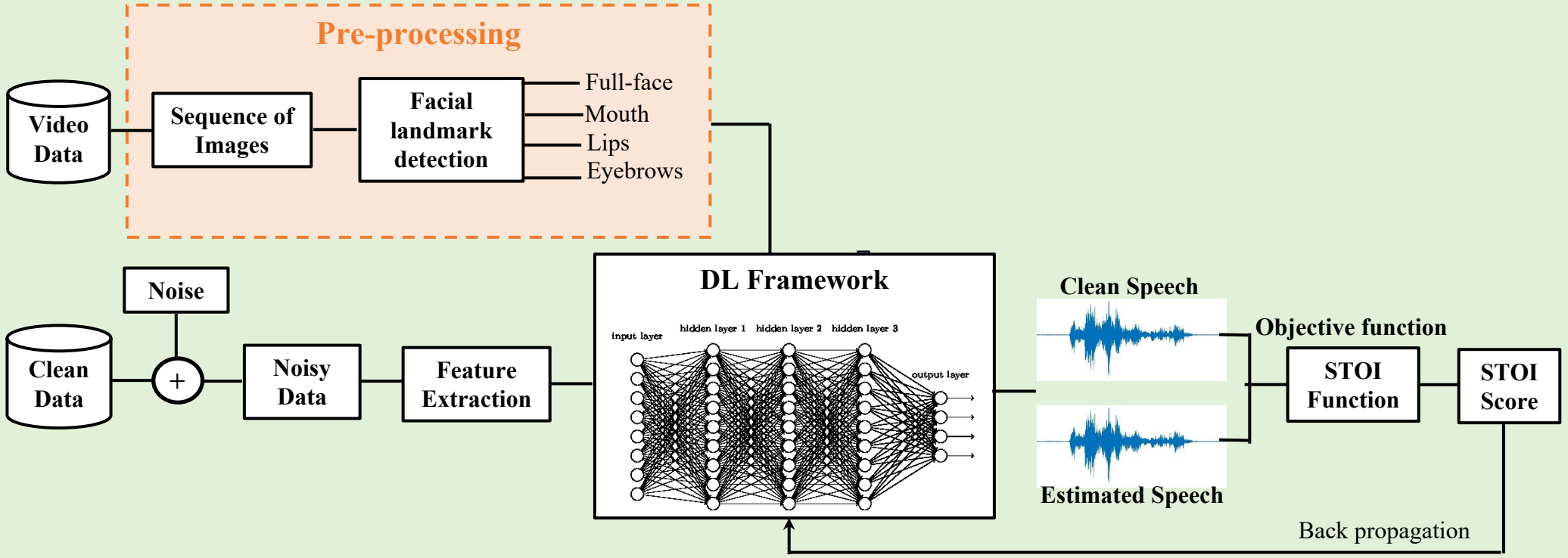
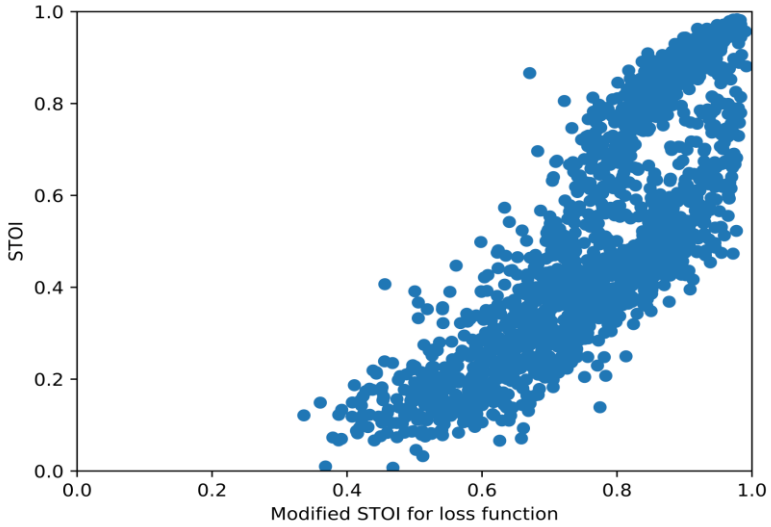
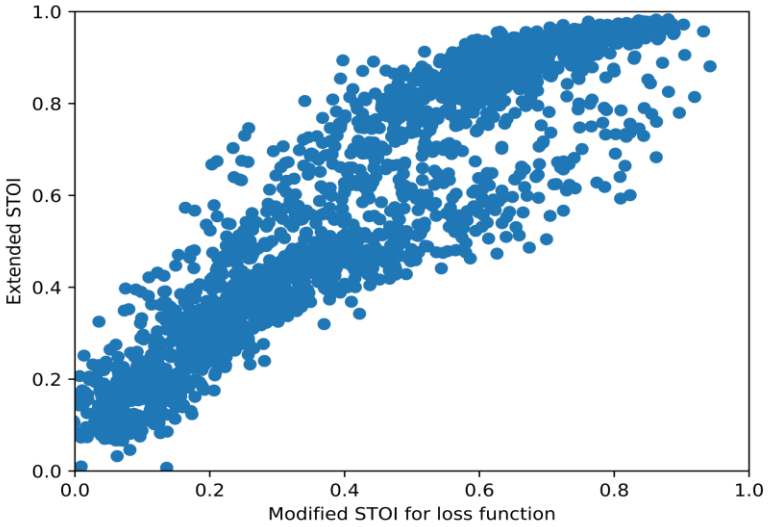


Fig. 2. Proposed Intelligibility-oriented Audio-visual (AV) Speech Enhancement Framework

Classical STOI, Extended STOI vs Modified STOI

- Unlike the original (classical and extended) STOI measures, which down sample signals to 10kHz, carry out silent frame removal, and then apply short-time Fourier transform (STFT), we modified STOI to account for 16kHz signals in the frequency domain and ignored silent frame removal.
- Scatter plots below show our **modified STOI** correlates well with **extended STOI** and can be used for training AV DL models





- C.H.Taal, R.C.Hendriks, R.Heusdens, J.Jensen. "A Short-Time Objective Intelligibility Measure for Time-Frequency Weighted Noisy Speech." in Proc. ICASSP 2010, Texas, Dallas.
- J. Jensen and C. H. Taal, "An Algorithm for Predicting the Intelligibility of Speech Masked by Modulated Noise Maskers", IEEE Transactions on Audio, Speech and Language Processing, 2016.

Experimental Setup

I-O AV speech enhancement setup

Training:

GRID/AV Speech dataset:

Total Number of Speakers: 34/33
Number of Utterances/speaker = 1000

Noise Types = BUS, STR, PED, CAF (CHiME-3 background noises)
SNR = [-12, 9] step of 3 dB

Training set: 1000 (utt.) x 20 (speakers) = 20,000 noisy utterances.
Validation set: 20,000 noisy utterances x 0.1 = 2000 noisy utterances.

Testing:

ASPIRE database

Experimental Setup

I-O AV DL Architecture

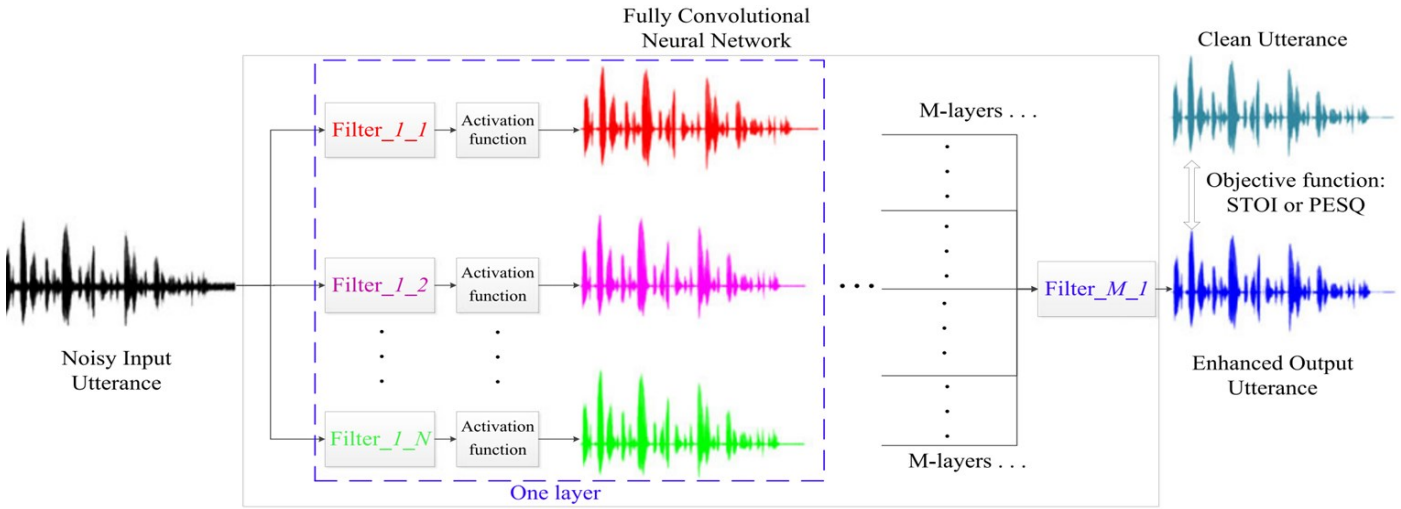


Fig. 3. Utterance-based raw waveform enhancement by fully convolutional neural network (FCN) [1].

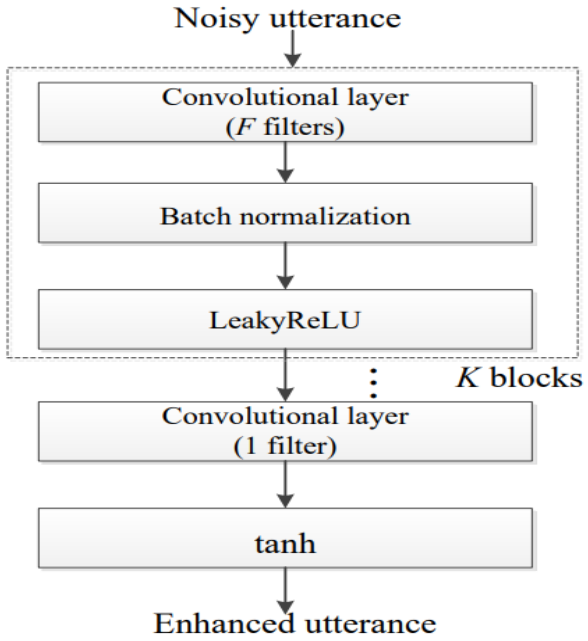


Fig. 4. The FCN structure used in the paper. For the GRID dataset, we used the same setup as reported in [1], i.e., K=7 and F=30.

For I-O AV SE, we are currently extending our MSE based benchmark AV CochleaNet model ([2] [3])

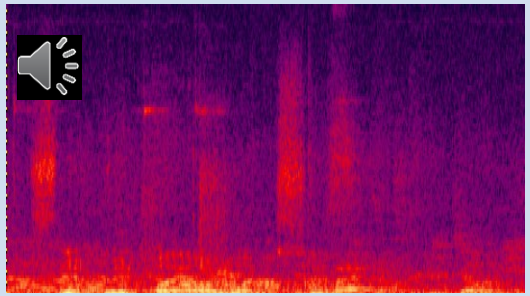
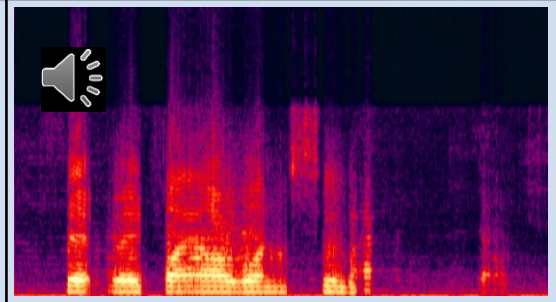
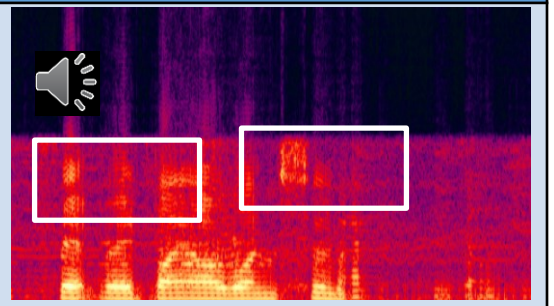
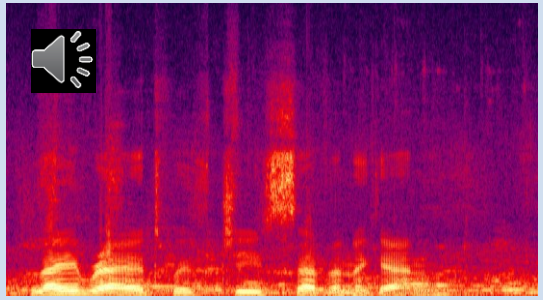
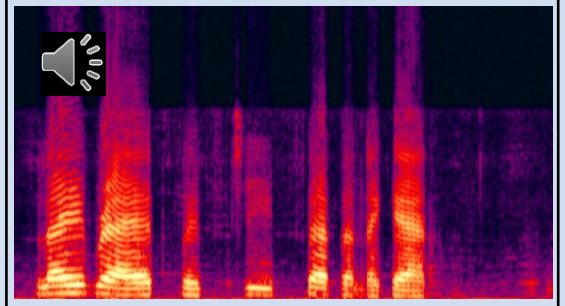
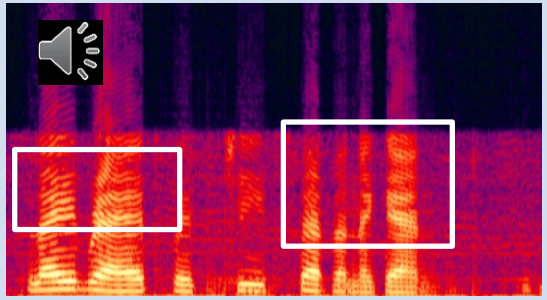
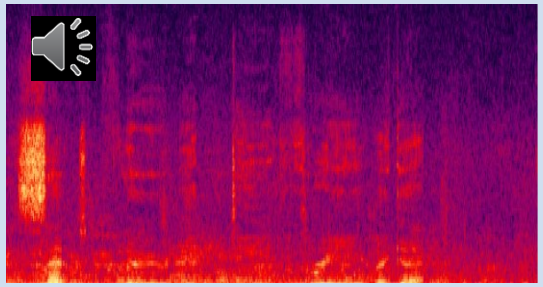
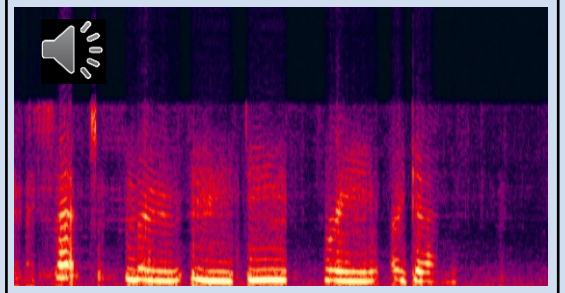
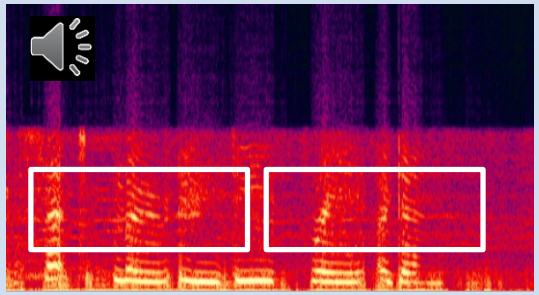
[1] S. –W. Fu, T. –W. Wang, Y. Tsao, X. Lu, and H. Kawai. "End-to-end waveform utterance enhancement for direct evaluation metrics optimization by fully convolutional neural networks." *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* vol. 26, no. 9, pp. 1570-1584, 2018.

[2] A. Adeel, M. Gogate, A. Hussain and W. M. Whitmer, (2021) "Lip-Reading Driven Deep Learning Approach for Speech Enhancement," in*IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 3, pp. 481-490, June 2021

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

Preliminary Results

A-only MSE vs STOI

	Noisy	FCN (obj = MSE)	FCN (obj = STOI)
Speaker 1			
	Noisy	**MSE**	**STOI**
Speaker 2			
	Noisy	**MSE**	**STOI**
Speaker 3			

Challenges/Ongoing work

- Extended A-only to AV – preliminary results

- Example I-O AV enhanced utterance:

Noisy  Enhanced 

- Full comparative results of I-O vs MSE based AV SE will be reported in the full paper (including an on-line evaluation demo)

ACKNOWLEDGEMENT:

- This work is funded by the UK Engineering and Physical Sciences Research Council (EPSRC) programme grant: COG-MHEAR (http://cogmhear.org)