IEEE/ACM TASLP
Improved Lite Audio-Visual Speech Enhancement

Shang-Yi Chuang[1],*Hsin-Min Wang*[2], *Yu Tsao*[1]

[1]Research Center for Information Technology Innovation, Academia Sinica
[2]Institute of Information Science, Academia Sinica

Outline

- Introduction
- Related Works
- Improved Lite Audio-Visual Speech Enhancement (iLAVSE) System
- Experiments
- Conclusion

iLAVSE Demo🚗

Introduction

Speech Enhancement (SE)

- Improve speech quality and intelligibility
- Front-end processing of speech-related applications
  - Automatic speech recognition
  - Speaker recognition
  - Speech coding
  - Hearing aids
  - Cochlear implants
- Deep-learning models in SE
  - Better performance than traditional statistical and machine-learning methods
  - Flexibility to fuse data from different domains

Introduction

Lite Audio-Visual SE (LAVSE) [1]

- Ability to handle immense visual data and potential privacy issues
  - Autoencoder (AE)-based compression network
  - Latent feature quantization unit
- Three practical issues when implementing AVSE systems in real-world scenarios
  - The additional cost of processing visual data
    - Usually much higher than the cost of processing audio data
    - Computing power or memory, and visual sensors
  - Audio-visual asynchronization
  - Low-quality visual data

Introduction

Improved Lite Audio-Visual SE (iLAVSE)

- Three-unit data compression module CRQ
- Data augmentation scheme on asynchronization
- Zero-out training scheme

Related Works

- AVSE: Multimodal Deep Convolutional Neural Networks (AVDCNN) [2]
- Lite Audio-Visual SE (LAVSE) [1]
- Bit-wise Data Compression: Exponent-only floating-point (EOFP) format [3]

Related Works

AVDCNN [2]

- AVSE systems like AVDCNN require additional visual input, which causes additional hardware and computational costs
- Facial or lip images may cause privacy issues

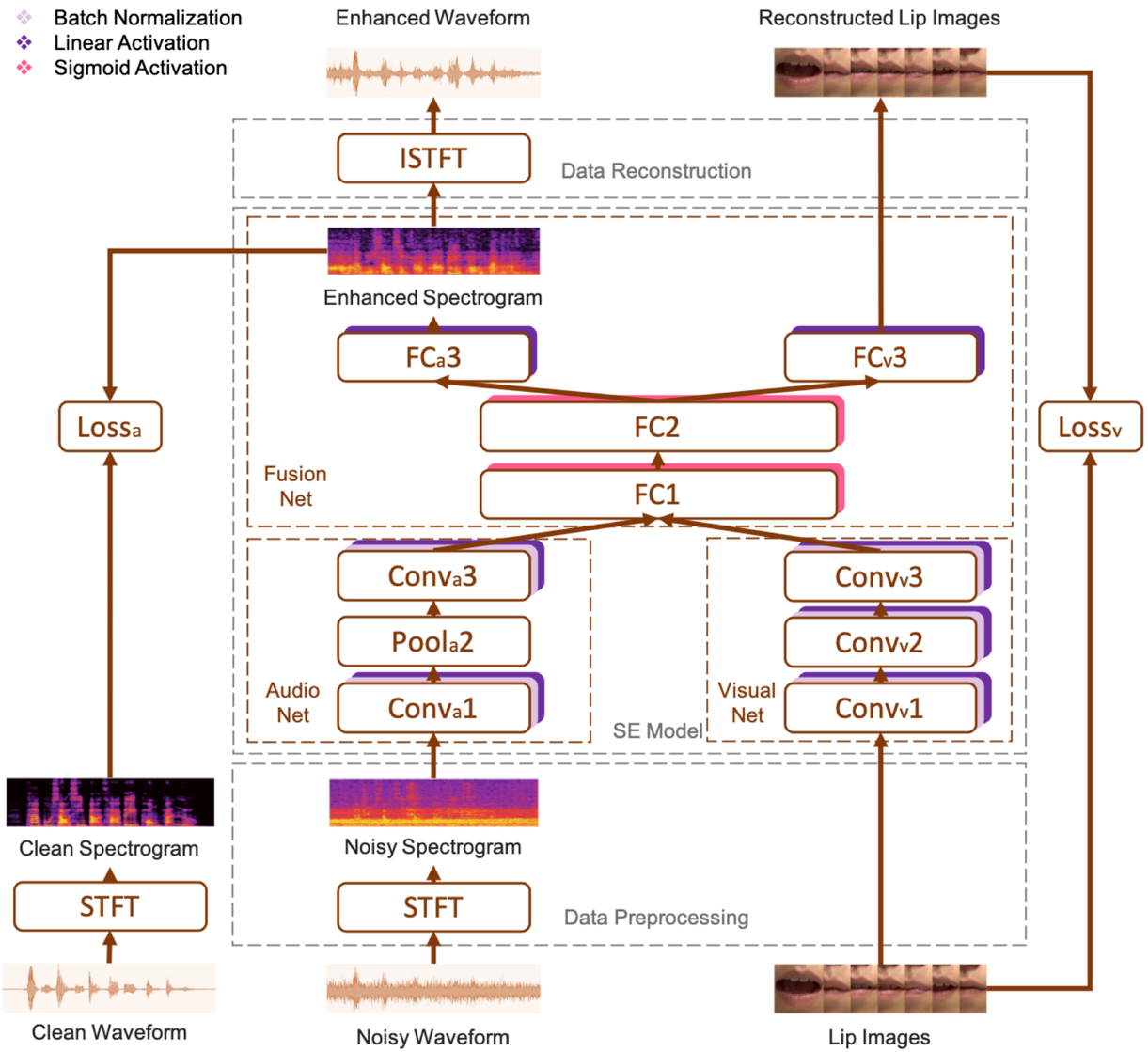


Figure 1: The AVDCNN system.

Related Works

LAVSE [1]

- Used a pre-trained AE to extract meaningful and compact representations of visual data
  - To reduce computational costs
  - Appropriately solve the privacy problem in facial information
- The AE is pre-trained in an unsupervised learning manner
  - Can be trained on a richer unimodal dataset

Related Works

EOFP [3]

- Single-precision floating-point format
  - 1 sign bit
    - The value is positive or negative
  - 8 exponential bits
    - Representation range of the value
  - 23 mantissa bits
    - Precision
- Exponent-only floating-point (EOFP) format [3]
  - No mantissa bit
  - Does not change the represented value itself
  - Only reduces the precision

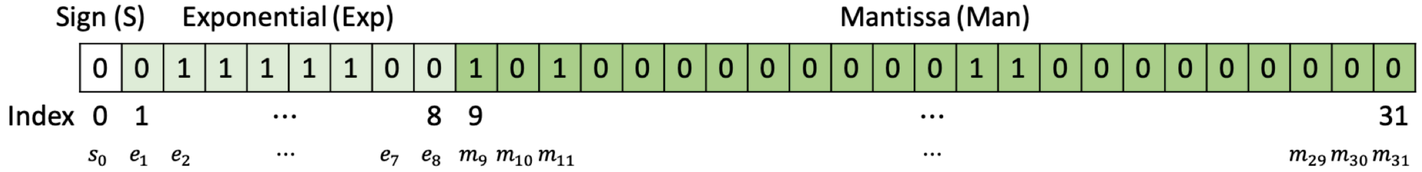


Figure 2: Single-Precision Floating-Point Format.

$$value_{10} = (-1)^{S} \times 2^{(Exp_{10} - bias)} \times Man_{10}$$

$$S = s_0$$

$$Exp_2 = e_1 e_2 e_3 e_4 e_5 e_6 e_7 e_8$$

$$Exp_{10} = \sum_{i=1}^{8} e_i \times 2^{(8-i)}$$

$$Man_2 = m_9 m_{10} \ldots m_{31}$$

$$Man_{10} = \sum_{i=9}^{31} m_i \times 2^{(8-i)}$$

The Proposed iLAVSE

- The iLAVSE system
  - LAVSE (Encoder$_{AE}$ and Qua$_{latent}$)
  - CRQ visual data compression
  - Compensation on audio-visual asynchronization
  - Zero-out training
- Features
  - Audio: log1p magnitude spectrum
  - Visual: CRQ+AE+EOFP

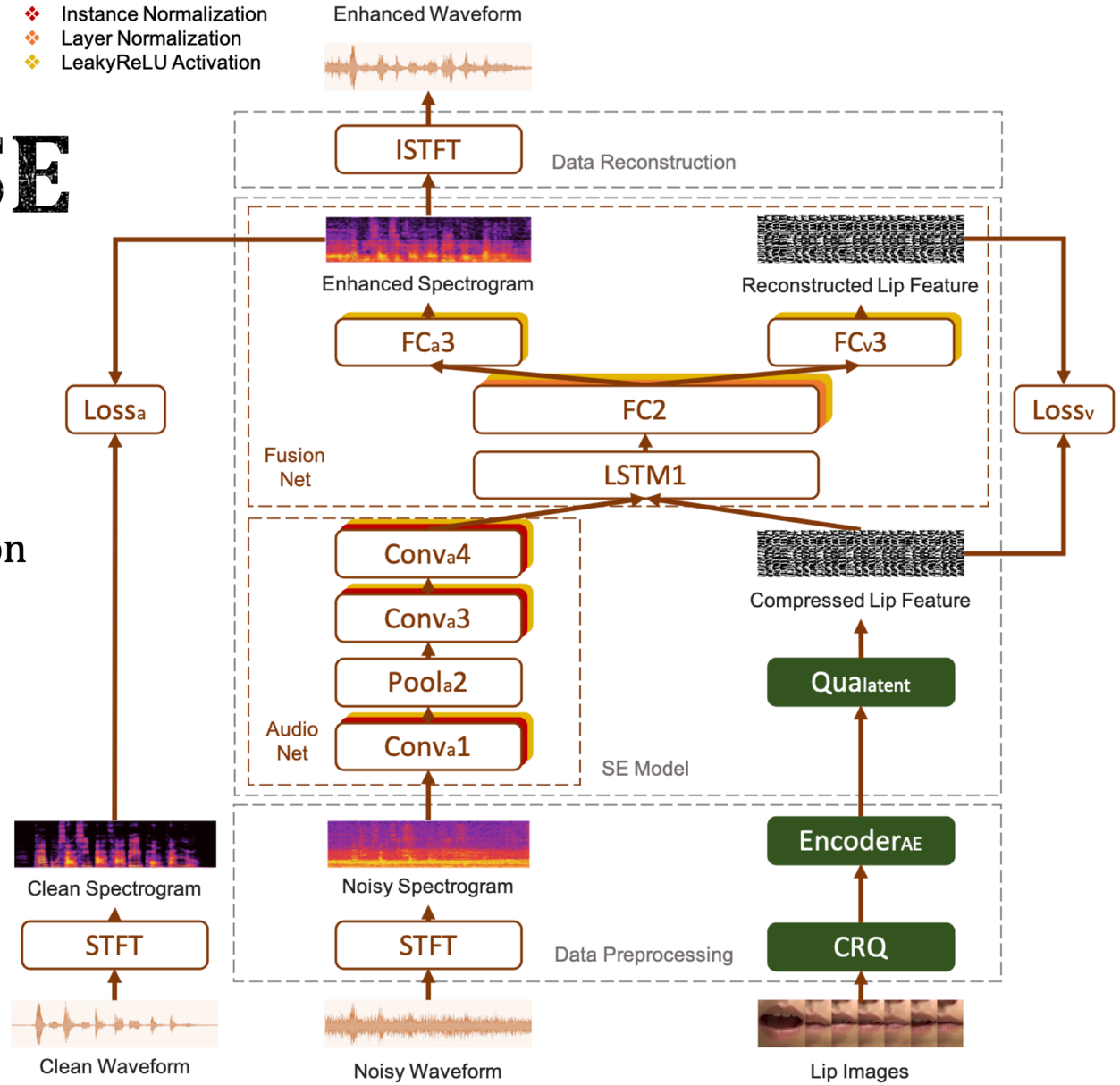


Figure 3: The proposed iLAVSE system.

The Proposed iLAVSE

CRQ

- A three-unit data compression module for channel reduction, resolution reduction, and bit quantization

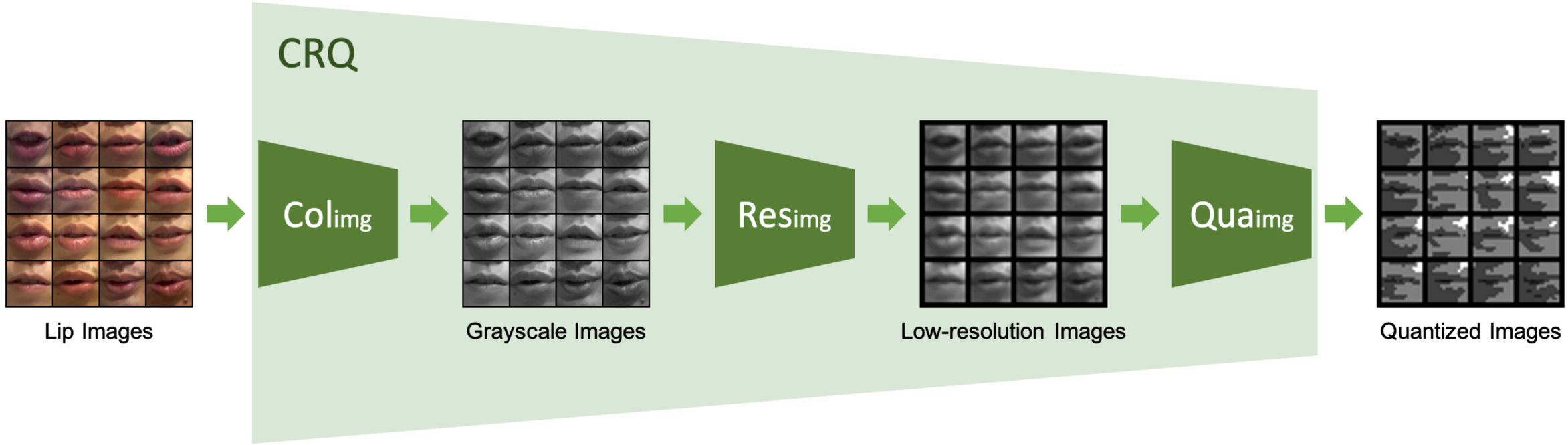


Figure 4: The proposed CRQ module.

The Proposed iLAVSE

CRQ integrated with Encoder$_{AE}$

- The AE trained in frame-wise manner
  - Input: CRQ processed lip images
  - Output: grayscale low-resolution images

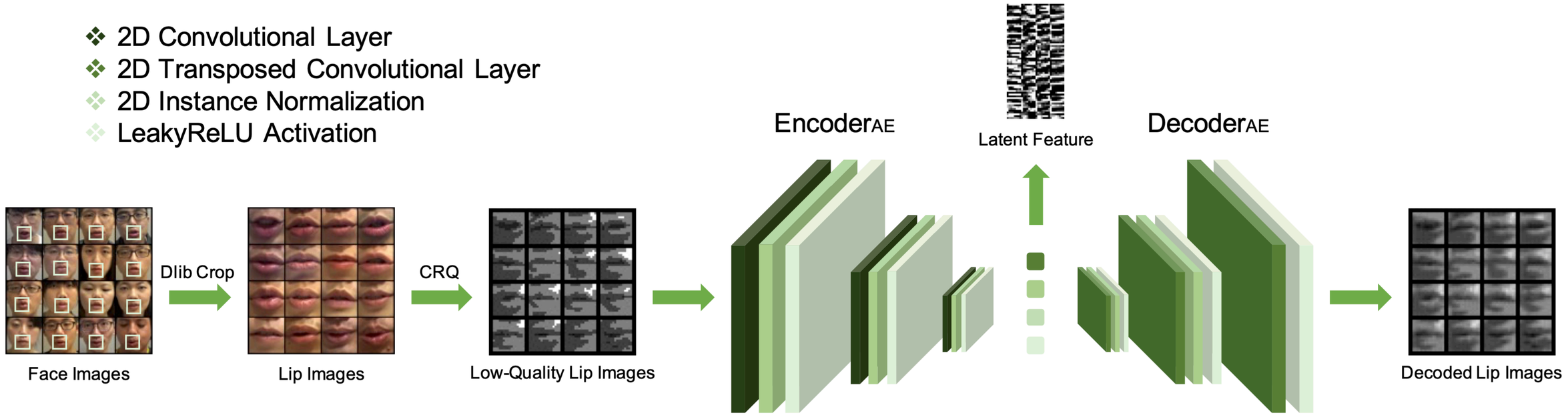


Figure 5: The AE model for visual input data compression.

The Proposed iLAVSE

CRQ integrated with Encoder$_{AE}$ and Qua$_{latent}$

- Bits
  - 32-bit floating-point
    - 1 sign bit
    - 8 exponential bits
    - 23 mantissa bits
  - 3-bit EOFP
    - 1 sign bit
    - 2 exponential bits
    - 0 mantissa bit

(a) 32-bit AE features. (b) EOFP 3-bit AE features.

Figure 6: Original and quantized visual latent features.

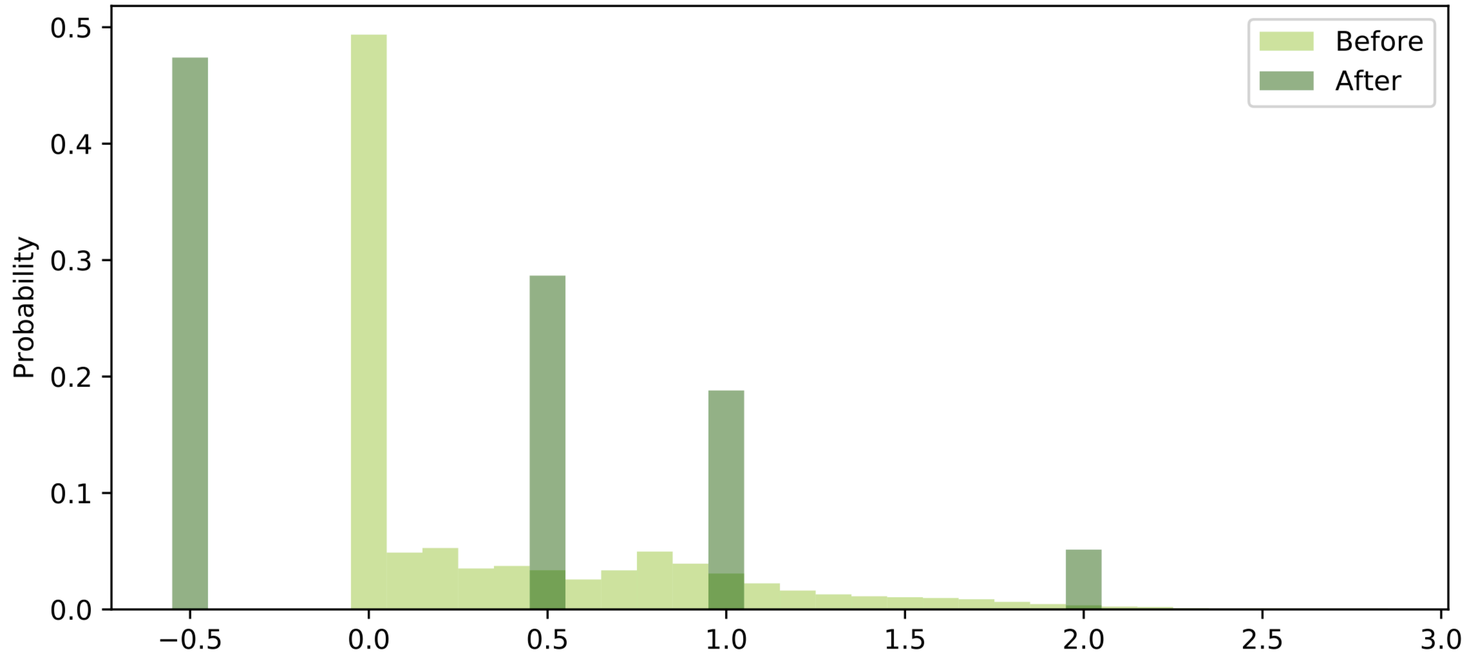
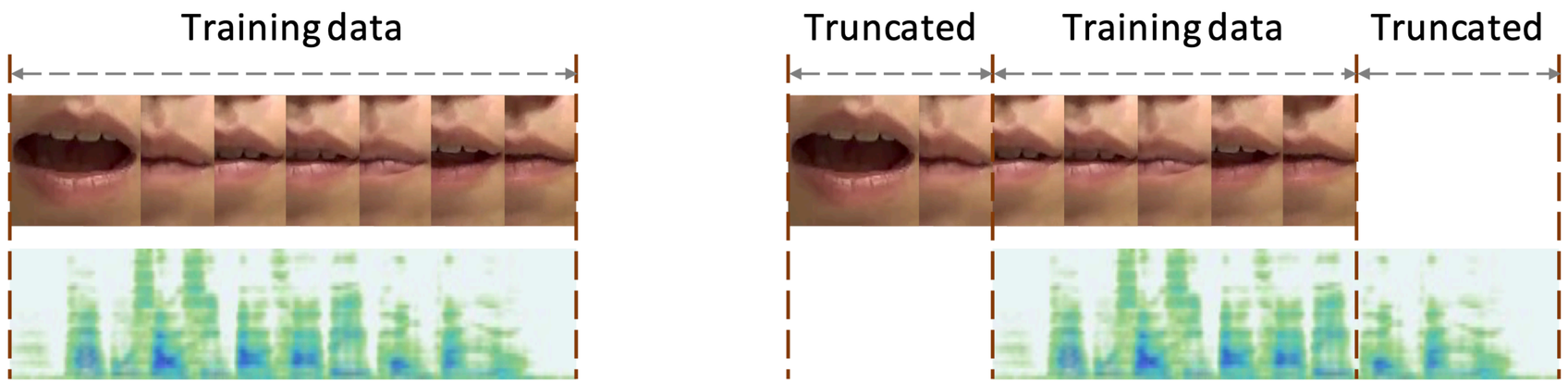


Figure 7: The distributions of visual features before and after applying Qua$_{latent}$.

The Proposed iLAVSE

Compensation of Audio-Visual Asynchronization

- Artificially simulate various asynchronous audio-visual data



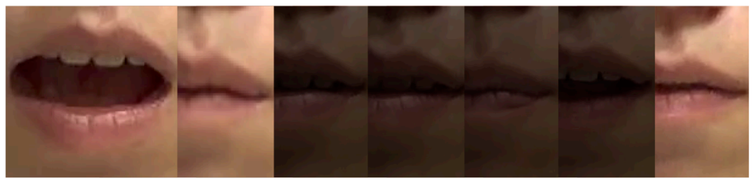
(a) Synchronous. (b) Asynchronous.

Figure 8: Synchronous and asynchronous audio and visual data.

The Proposed iLAVSE

Zero-out Training

- The quality of video frames may be poor in poor lighting conditions, such as in a tunnel or at a night market
- Let iLAVSE dynamically decide whether video data should be used



(a) Low-quality lip images.

(b) Low-quality latent features.

Figure 9: Low-quality visual data.

Experiments

Experimental Setup

- The dataset of Taiwan Mandarin speech with video (TMSV)
- Mismatched speakers, noise types, and SNR levels in training and testing sets
  - Training set
    - 4 males, 4 females
    - The 1st to the 200th utterance
    - 100 types of noise [4]
    - SNRs: from -12 dB to 12 dB with a step of 6 dB
  - Testing set (car driving scenario)
    - 1 male, 1 female
    - The 201st to the 320th utterance
    - Noise types
      - Cries of a baby
      - Engine noise
      - Background talkers
      - Music
      - Pink noise
      - Street noise
    - SNRs: -1, -4, -7, -10 dB

Experiments

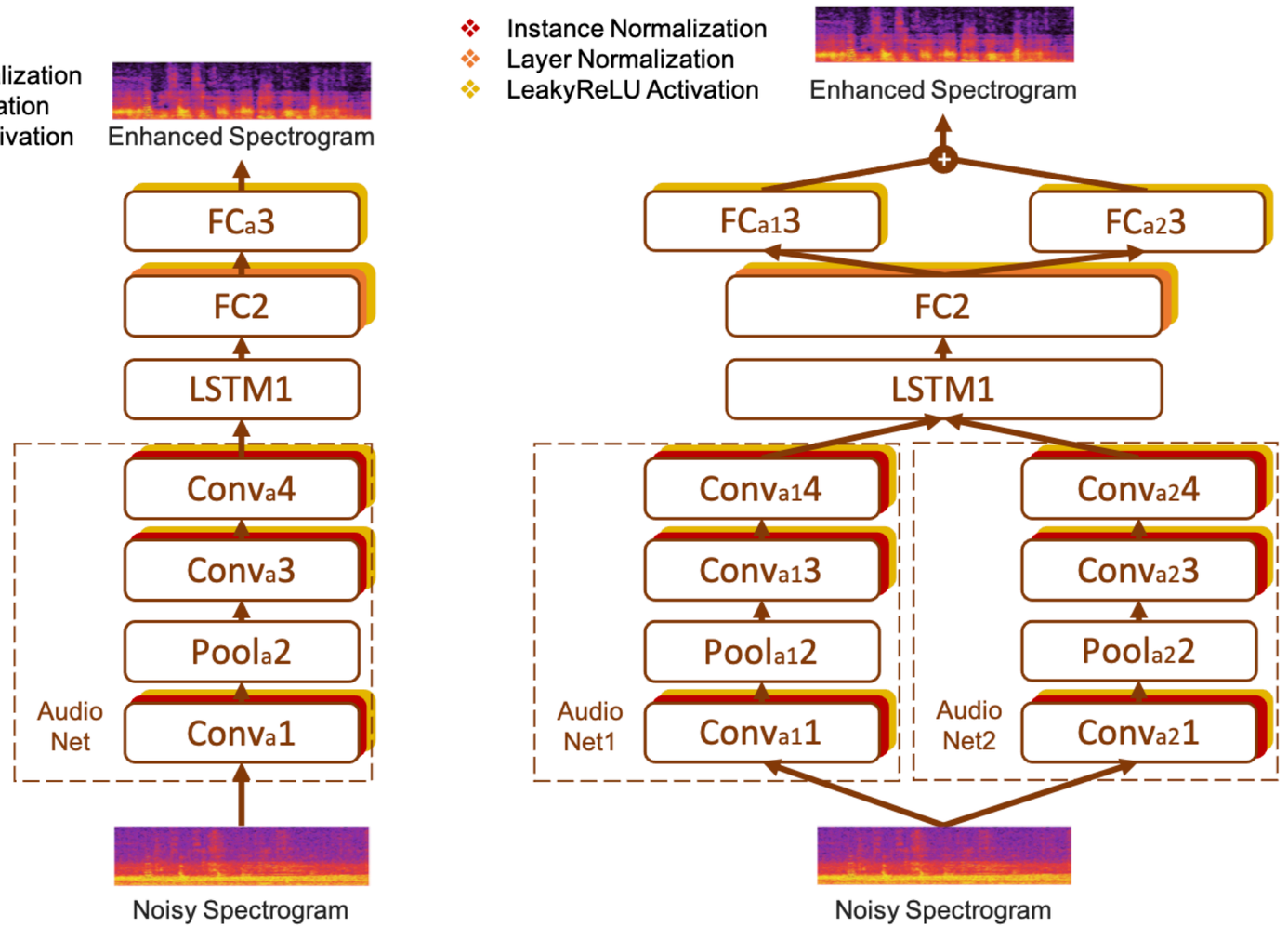
Experimental Setup

- The lip or face image contours were positioned using Dlib [5]
- Evaluation metrics
  - Perceptual evaluation of speech quality (PESQ) [6]
  - Short-time objective intelligibility measure (STOI) [7]

Experiments

Experimental Results

- Audio-only baselines
- AOSE(DP) has a similar number of model parameters to LAVSE



(a) Audio-only SE. (b) Dual-path-audio-only SE.

Figure 10: Architectures of two audio-only SE systems.

Experiments

Experimental Results

- Baselines
  - Audio-only SE systems
  - AVSE systems
- Compared to AOSE and AOSE(DP), all the AVSE systems yield higher PESQ and STOI
- The proposed iLAVSE can maintain SE performance comparable to LAVSE(AE)

	PESQ	STOI
Noisy	1.001	0.587
AOSE	1.282	0.616
AOSE(DP)	1.283	0.610
AVDCNN [2]	1.337	0.641
LAVSE(AE) [1]	1.374	**0.646**
LAVSE(AE+EOFP4bit) [1]	1.358	0.643
iLAVSE(CRQ)	1.387	0.639
iLAVSE(CRQ+AE)	1.398	0.641
iLAVSE(CRQ+AE+EOFP3bit)	**1.410**	0.641

Table I: Average PESQ and STOI scores of the two audio-only SE systems and the AVSE systems over SNRs of -1, -4, -7 and -10 dB.

Experiments

Experimental Results

- Data Compression of CRQ and Qua_{latent}
- {Col_{img}, Res_{img}, Qua_{img}, Qua_{latent}} = {**A**, **B**, C, D}
  - **A: *RGB* or *GRAY* (for grayscale)**
  - **B: image resolution**
  - C: image data quantization
  - D: latent feature quantization

	PESQ		STOI	
	R	**G**	**R**	**G**
AOSE(DP)	1.283		0.610	
iLAVSE *64*	<u>1.374</u>	**1.378**	<u>0.646</u>	0.646
iLAVSE *32*	1.371	1.375	0.644	0.645
iLAVSE *16*	1.374	1.358	0.646	**0.649**

Table II: The performance of iLAVSE using lip images with reduced channel numbers and resolutions, **R**: {*RGB*} and **G**: {*GRAY*}. The underlined scores are the same as those of LAVSE in Table I because the iLAVSE with the {*RGB*, *64*} setup is equivalent to LAVSE.

Experiments

Experimental Results

- Data Compression of CRQ and Qua_{latent}
- $\{Col_{img}, Res_{img}, Qua_{img}, Qua_{latent}\} = \{A, B, \mathbf{C}, D\}$
  - A: *RGB* or *GRAY* (for grayscale)
  - B: image resolution
  - **C: image data quantization**
  - D: latent feature quantization

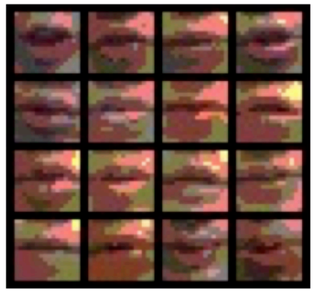
	PESQ		STOI	
Total bits	**R**	**G**	**R**	**G**
1	1.333	1.296	0.619	0.615
3	1.250	1.295	0.628	0.613
5	1.361	**1.398**	0.644	0.641
7	1.374	1.379	0.640	0.644
9	1.386	1.387	0.642	0.642
32	1.374	1.358	0.646	**0.649**

Table III: The performance of iLAVSE with or without image quantization (the original image is with 32 bits), **R**: {*RGB*, *64*} and **G**: {*GRAY*, *16*}. The underlined scores are the same as those of LAVSE in Table I.

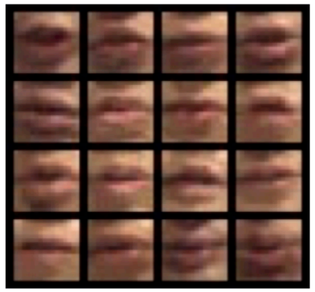
Experiments

Experimental Results

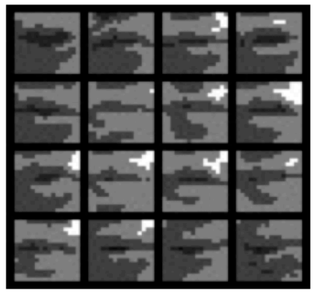
- Data Compression of CRQ and Qua_{latent}
- $\{Col_{img}, Res_{img}, Qua_{img}, Qua_{latent}\} = \{A, B, C, D\}$
  - A: *RGB* or *GRAY* (for grayscale)
  - B: image resolution
  - **C: image data quantization**
  - D: latent feature quantization
- Compression ratio R_{comp} of CRQ
  - {*RGB*, *64*, 32bits(i)} to {*GRAY*, *16*, 5bits(i)}
  - $\frac{3}{1} \times \frac{64^2}{16^2} \times \frac{32}{5} = 307.2$



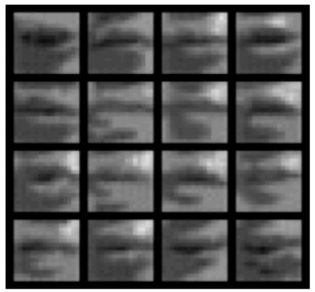
(a) {RGB, 16, 5bits(i)} input.



(b) {RGB, 16, 5bits(i)} output.



(c) {GRAY, 16, 5bits(i)} input.



(d) {GRAY, 16, 5bits(i)} output.

Figure 11: AE lip images in 5 bits (1 sign bit and 4 exponential bits).

Experiments

Experimental Results

- Data Compression of CRQ and Qua_{latent}
- $\{Col_{img}, Res_{img}, Qua_{img}, Qua_{latent}\} = \{A, B, C, \mathbf{D}\}$
  - A: *RGB* or *GRAY* (for grayscale)
  - B: image resolution
  - C: image data quantization
  - **D: latent feature quantization**
- Choose **{*GRAY*, *16*, 5bits(i), 3bits(l)}**
- Baselines
  - AOSE(DP) (PESQ = 1.283 and STOI = 0.610)
  - LAVSE (PESQ = 1.374 and STOI = 0.646)

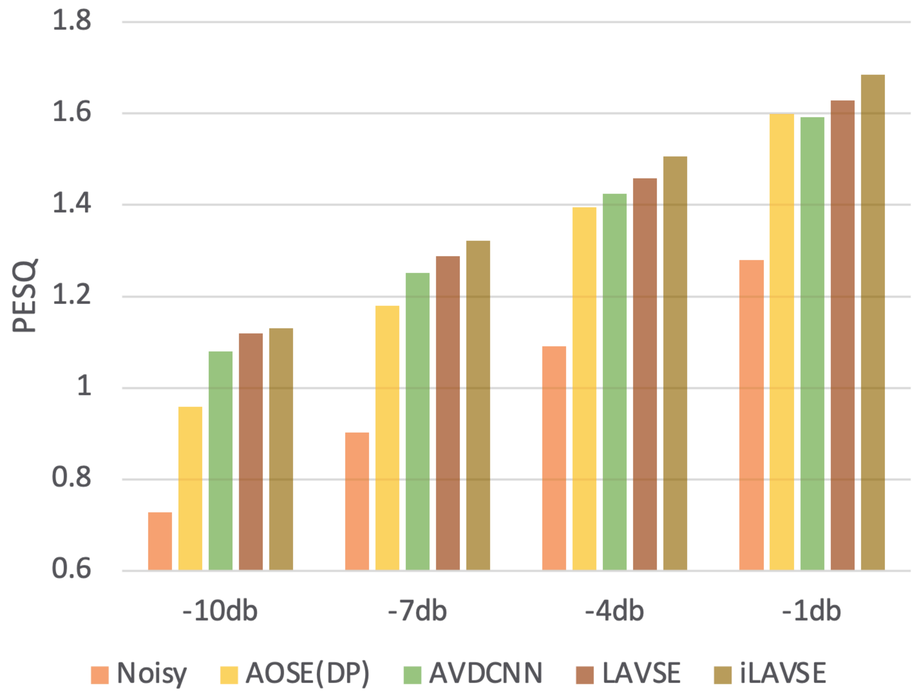
	PESQ		STOI	
Total bits	**R**	**G**	**R**	**G**
1	1.365	1.374	0.642	0.642
3	1.337	1.410	0.642	0.641
5	1.343	**1.413**	0.643	0.641
7	1.357	1.391	0.643	0.641
9	1.362	1.373	0.643	0.643
32	1.374	1.398	**0.646**	0.641

Table IV: The performance of iLAVSE with or without latent quantization, **R**: {*RGB*, *64*, 32bits(i)} and **G**: {*GRAY*, *16*, 5bits(i)} (1 sign bit + 4 exponential bits).

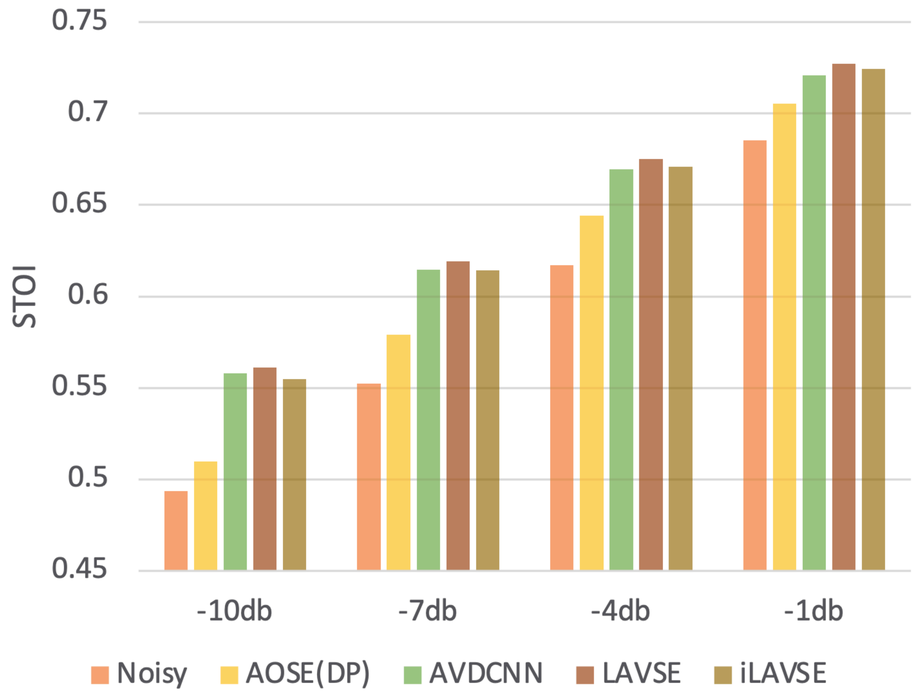
Experiments

Experimental Results

- Further Analysis
  - AVDCNN: original high-quality images
  - LAVSE: {*RGB*, *64*, 32bits(i), 32bits(l)}
  - iLAVSE: {*GRAY*, *16*, 5bits(i), 3bits(l)}



(a) PESQ.



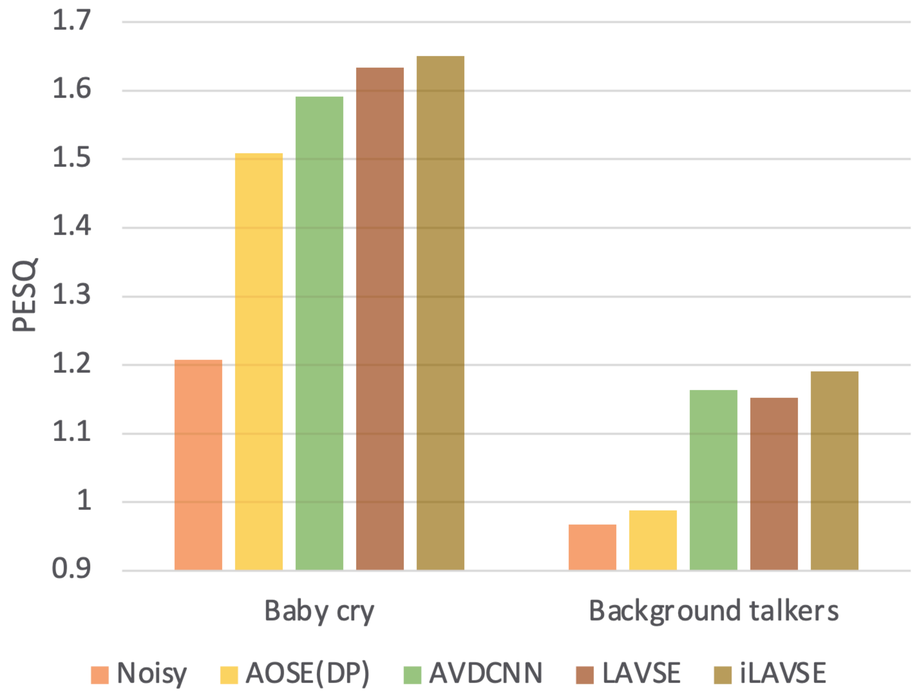
(b) STOI.

Figure 12: The performance of different SE systems at different SNR levels. LAVSE: {*RGB*, *64*, 32bits(i), 32bits(l)}, iLAVSE: {*GRAY*, *16*, 5bits(i), 3bits(l)}.

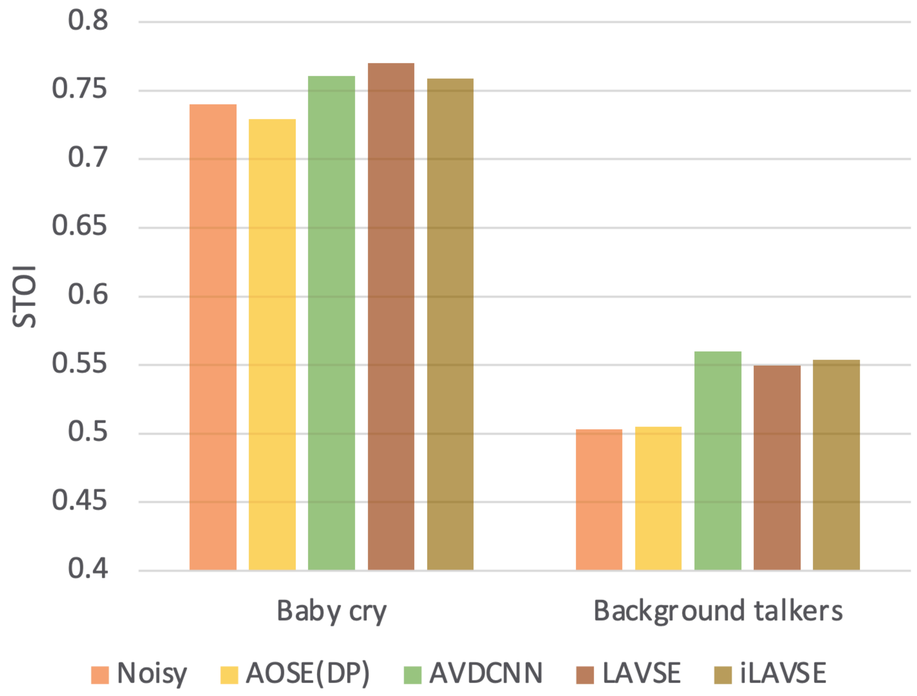
Experiments

Experimental Results

- Further Analysis
  - AVDCNN: original high-quality images
  - LAVSE: {*RGB*, *64*, 32bits(i), 32bits(l)}
  - iLAVSE: {*GRAY*, *16*, 5bits(i), 3bits(l)}



(a) PESQ.



(b) STOI.

Figure 13: The performance of different SE systems on different human-voiced noises. LAVSE: {RGB, 64, 32bits(i), 32bits(l)}, iLAVSE: {GRAY, 16, 5bits(i), 3bits(l)}.

Experiments

Experimental Results

	PESQ		STOI	
SNRs	AOSE(DP)	iLAVSE	AOSE(DP)	iLAVSE
Poor	1.387	1.544	0.699	0.734
Low	1.629	1.757	0.760	0.783
Mild	1.886	1.966	0.812	0.823

(a) Baby cry.

	PESQ		STOI	
SNRs	AOSE(DP)	iLAVSE	AOSE(DP)	iLAVSE
Poor	0.793	1.009	0.435	0.487
Low	1.183	1.372	0.575	0.621
Mild	1.575	1.733	0.702	0.733

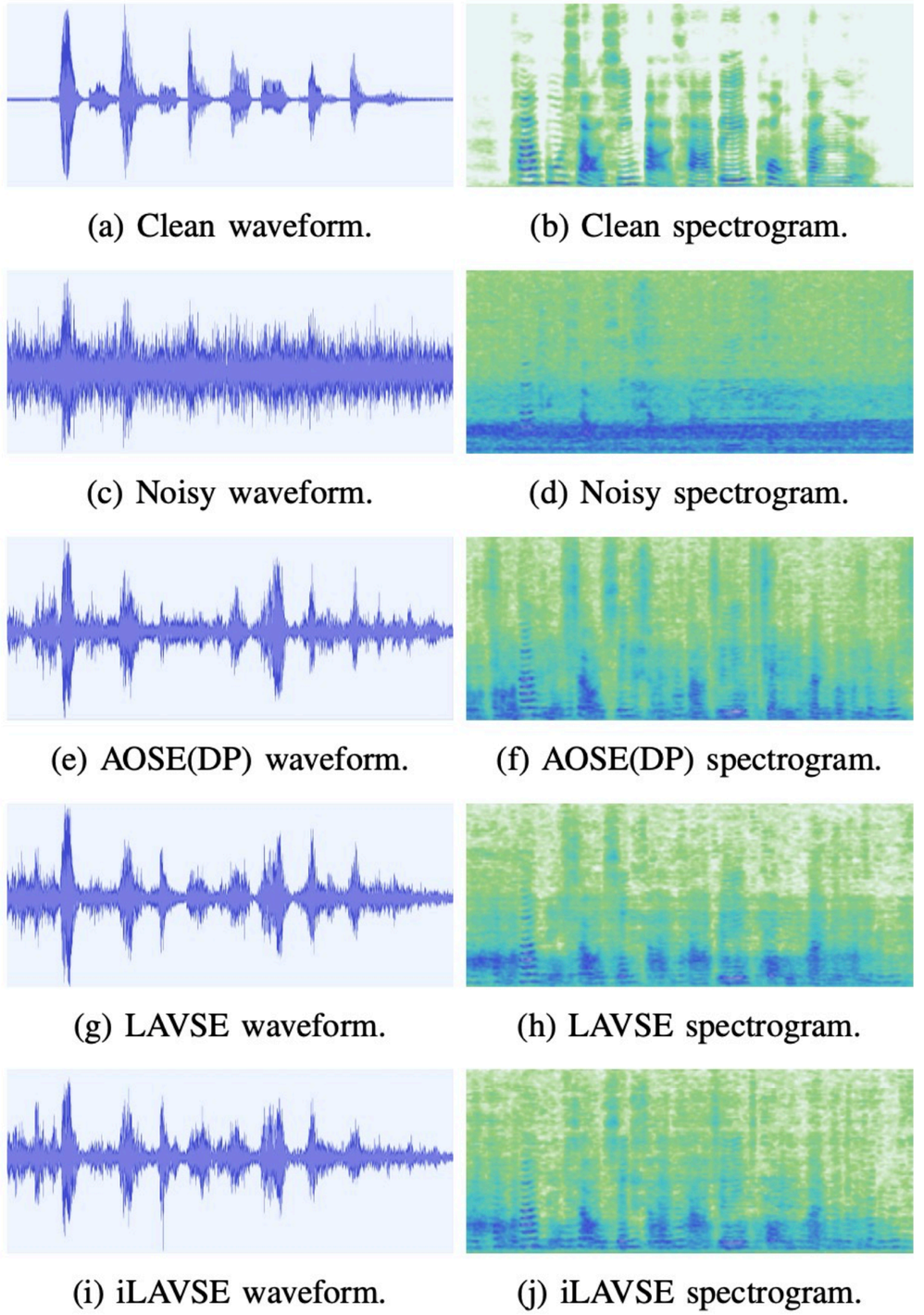
(b) Background talkers.

Table V: The performance of AOSE(DP) and iLAVSE on different human-voiced noises at different SNR levels. Poor: -10db and -7db, Low: -4 and -1db, Mild: 2db and 5db. iLAVSE: {GRAY, 16, 5bits(i), 3bits(l)}.

Experiments

Experimental Results

- Spectrograms and waveforms
  - Example of -7 dB street noise
  - iLAVSE can suppress the noise more effectively than AOSE(DP)
  - iLAVSE and LAVSE are very similar



(a) Clean waveform. (b) Clean spectrogram.

(c) Noisy waveform. (d) Noisy spectrogram.

(e) AOSE(DP) waveform. (f) AOSE(DP) spectrogram.

(g) LAVSE waveform. (h) LAVSE spectrogram.

(i) iLAVSE waveform. (j) iLAVSE spectrogram.

Figure 14: The waveforms and spectrograms of an example speech utterance under the condition of street noise at -7 dB. The vertical axis of the waveform figure represents the normalized amplitude (-0.1~0.1), and the vertical axis of the spectrogram figure represents the frequency (0k~8k Hz). The horizontal axis is time. The example utterance is 3 seconds long.

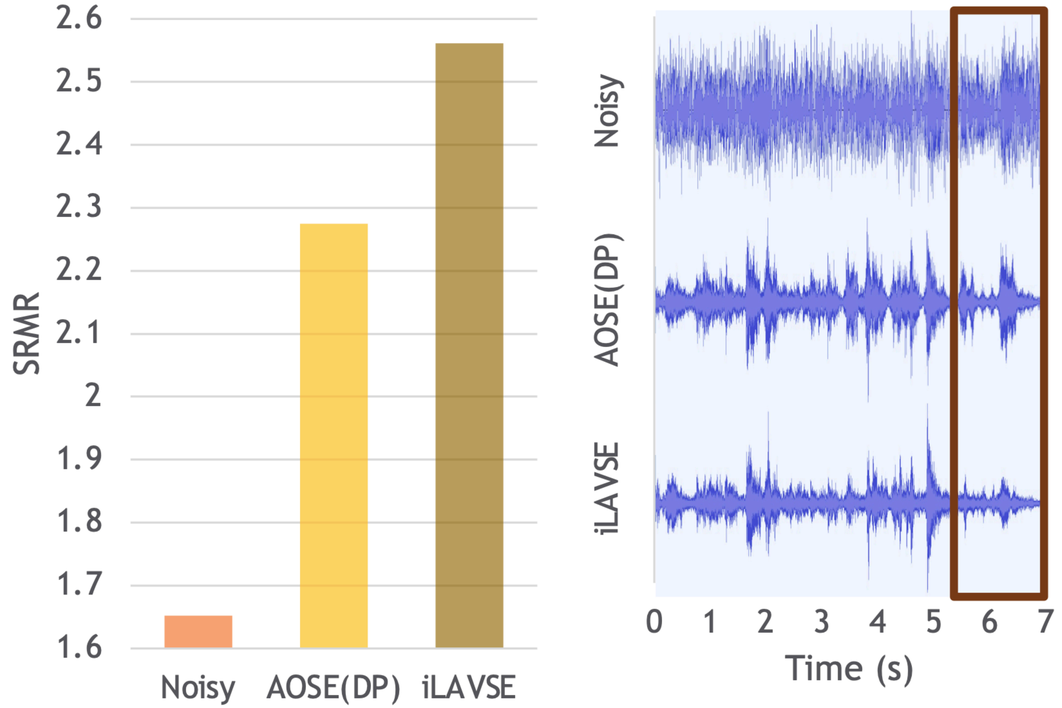
Experiments

Experimental Results

- Real-world data
  - 10 video clips in a real car-driving scenario
  - Background music and car-driving noise
  - Speech-to-reverberation modulation energy ratio (SRMR) [8]
  - No speech in brown box, only music
  - The closed lips can help iLAVSE remove the background music better than AOSE(DP)

Figure 15: The real-world car-driving scenario.

iLAVSE Demo🚗



(a) SRMR. (b) Waveforms.

Figure 16: The average SRMR scores and sample processed waveforms obtained by AOSE(DP) and iLAVSE for the real- world videos. iLAVSE: {*GRAY*, *16*, 5bits(i), 3bits(l)}.

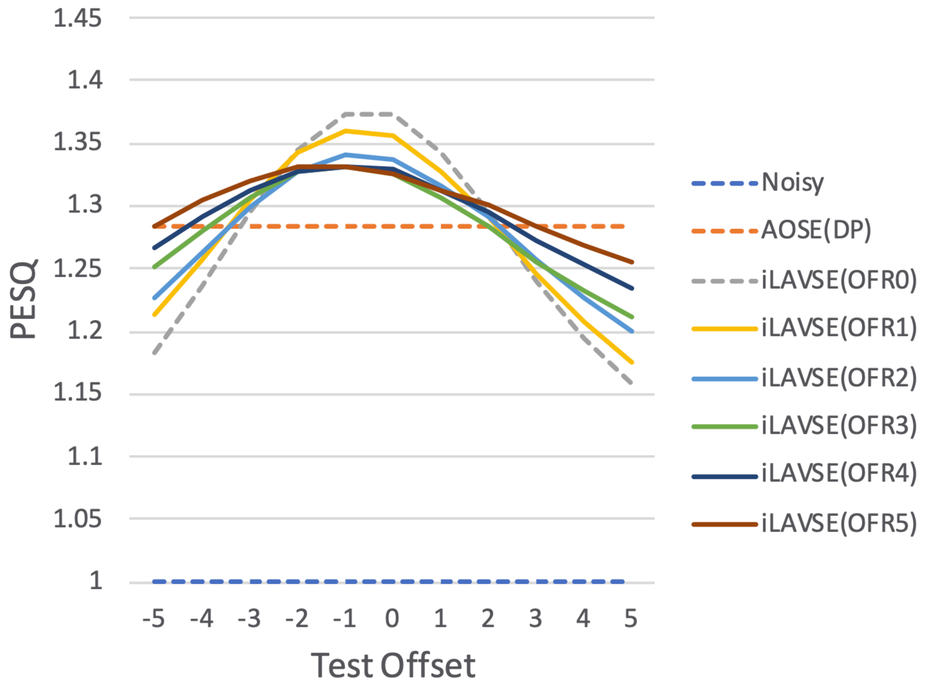
Experiments

Experimental Results

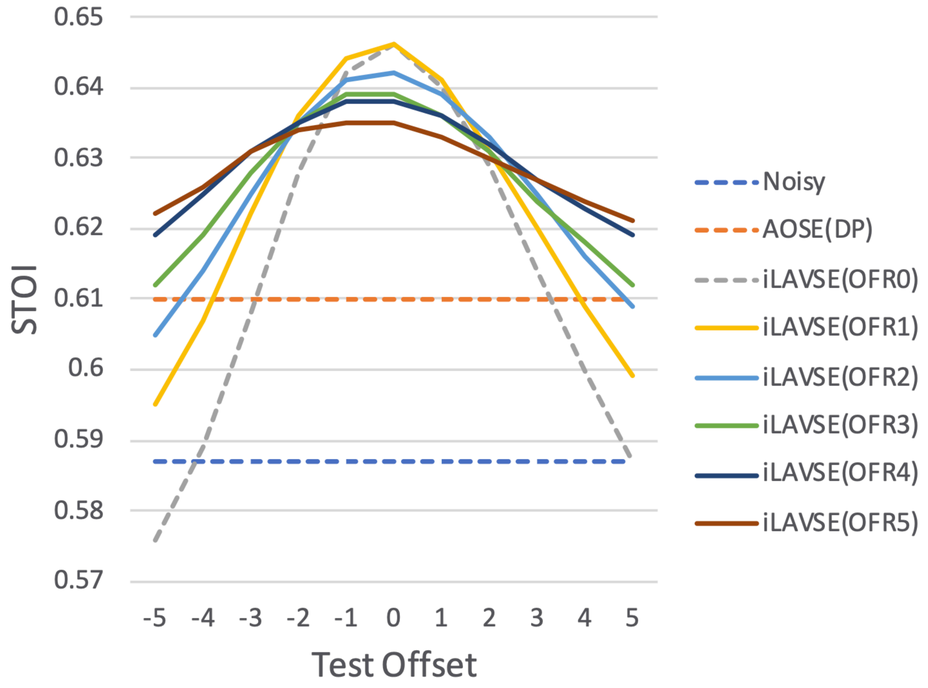
- Asynchronization Compensation
  - 5 specific offset ranges (OFR): [-1, 1], [-2, 2], [-3, 3], [-4, 4], and [-5, 5]
  - Take iLAVSE(OFR1) for example
    - OFR = [-1, 1]
    - An offset of -1, 0, or 1 frame (each frame = 20ms) was randomly selected (with equal probability)
  - For testing, fixed offsets in [-5, 5] is used, contained 11 different degrees of asynchronization
  - Used original visual data
  - “Test Offset = -5” and “Test Offset = 5” are the most severe conditions
    - Audio and visual signals are misaligned for 5 frames (100 ms) in both cases

Experiments

Experimental Results



(a) PESQ.



(b) STOI.

Figure 17: The PESQ and STOI scores of iLAVSE trained and tested with different audio-visual asynchronous data.

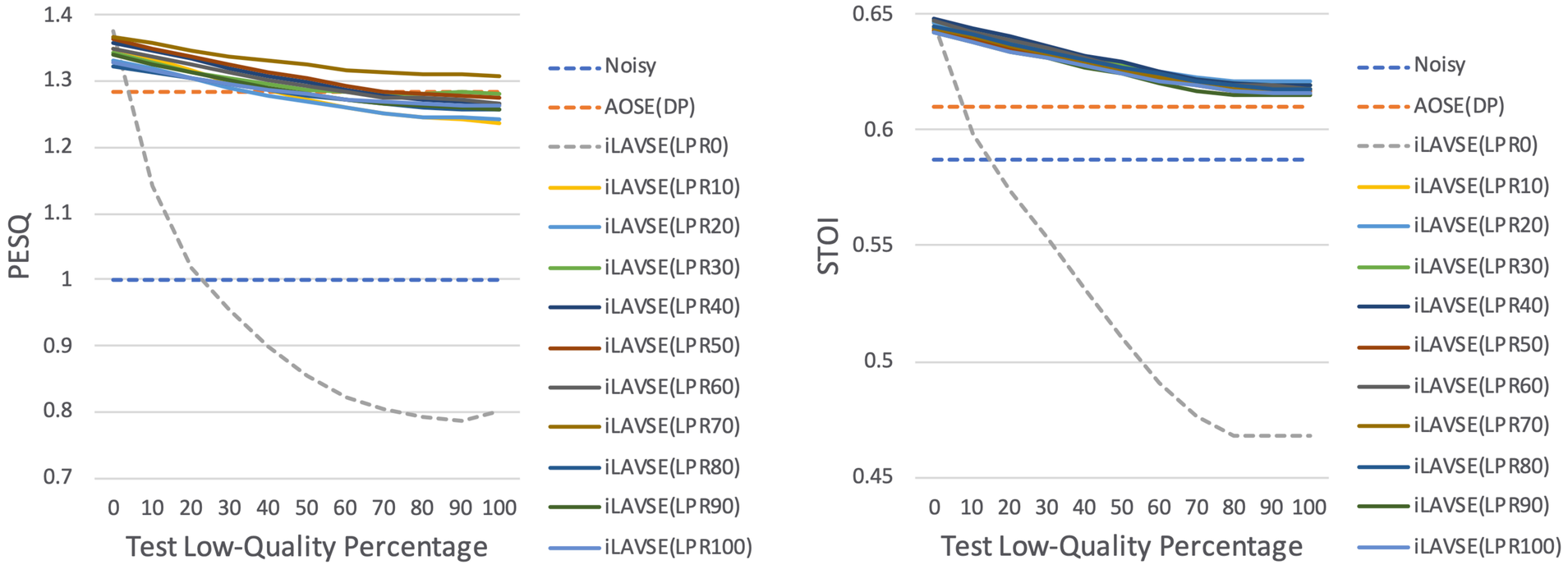
Experiments

Experimental Results

- Zero-Out Training
  - Low-quality percentage (LP): the percentage of missing frames in the visual data
  - Low-quality percentage range (LPR): the range of randomly assigned LPs for each batch
  - Take iLAVSE(LPR10) for example with a batch of 150 frames
    - LP will be randomly selected from 0% to 10%
    - If LP is set to 4%, a sequence of 6 (150×4%) frames of the visual data will be replaced with zeros
  - For training, LPRs ∈ {0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100}
  - For testing, LPs ∈ {0, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100}
  - The starting point of the missing visual part was randomly assigned for each batch

Experiments

Experimental Results



(a) PESQ.

(b) STOI.

Figure 18: The PESQ and STOI scores of iLAVSE trained with different LPRs and tested on specific LP conditions.

Conclusion

- For practical AVSE systems
  - Decreased the cost of visual data by using preprocessing modules (CRQ and AE)
  - Solved audio-visual asynchronization by a data augmentation scheme
  - Addressed low-quality visual data issues with a zero-out training approach
- The proposed iLAVSE system is robust under adverse conditions and can be appropriately implemented in real-world applications

Thank you!

References

[1] S.-Y. Chuang, Y. Tsao, C.-C. Lo, and H.-M. Wang, “Lite audio-visual speech enhancement,” in Proc. INTERSPEECH 2020.

[2] J.-C. Hou, S.-S. Wang, Y.-H. Lai, Y. Tsao, H.-W. Chang, and H.-M. 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.

[3] Y.-T. Hsu, Y.-C. Lin, S.-W. Fu, Y. Tsao, and T.-W. Kuo, “A study on speech enhancement using exponent-only floating point quantized neural network (eofp-qnn),” in Proc. SLT 2018.

[4] G. Hu, “100 nonspeech environmental sounds,” 2004, available: http://web.cse.ohio-state.edu/pnl/corpus/HuNonspeech/HuCorpus.html.

[5] D. E. King, “Dlib-ml: A machine learning toolkit,” Journal of Machine Learning Research, vol. 10, pp. 1755–1758, 2009.

[6] 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 2001.

[7] 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.

[8] T. H. Falk, C. Z., and W.-Y. Chan, “A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 18, no. 7, pp. 1766–1774, 2010.