Interspeech 2020
Lite Audio-Visual Speech Enhancement

Shang-Yi Chuang[1], *Yu Tsao*[1], *Chen-Chou Lo*[2], *Hsin-Min Wang*[3]

[1]Research Center for Information Technology Innovation, Academia Sinica
[2]EAVISE, Dept. of Electrical Engineering, KU Leuven, Belgium
[3]Institute of Information Science, Academia Sinica

Outline

- Introduction
- Related Works
- Proposed Lite Audio-Visual Speech Enhancement (LAVSE) System
- Experiments
- Conclusion

LAVSE on GitHub⭐

Introduction

Speech Enhancement (SE)

- Improve speech quality and intelligibility
- Front-end processing of speech-related applications
  - Automatic speech recognition
  - Assistive hearing technologies
  - Speaker recognition
- Deep-learning models in SE
  - Outstanding nonlinear mapping properties
  - Easy to fuse multimodal data

Introduction

Audio-Visual SE (AVSE)

- Visual information has been adopted as auxiliary information to facilitate better SE performance
- Two AVSE problems
  - Additional processing costs for visual input
  - Privacy problems of face or lip images

Introduction

Proposed Lite AVSE (LAVSE)

- Two visual data compression techniques
  - Autoencoder (AE)-based compression network
    - Reduces the size of the visual input
    - Extracts highly informative visual information
  - A quantization data compression scheme
    - Reduces the bits of the extracted representation
- LAVSE yields better performance than an audio-only SE baseline
- The user identity can be removed from the compressed visual data

Related Works

Multimodal Deep Convolutional Neural Networks (AVDCNN) [1]

- AVDCNN is adopted as the basic AVSE system in this study
- Receives noisy audio and lip images as the input
- Generates enhanced audio and lip images as the output

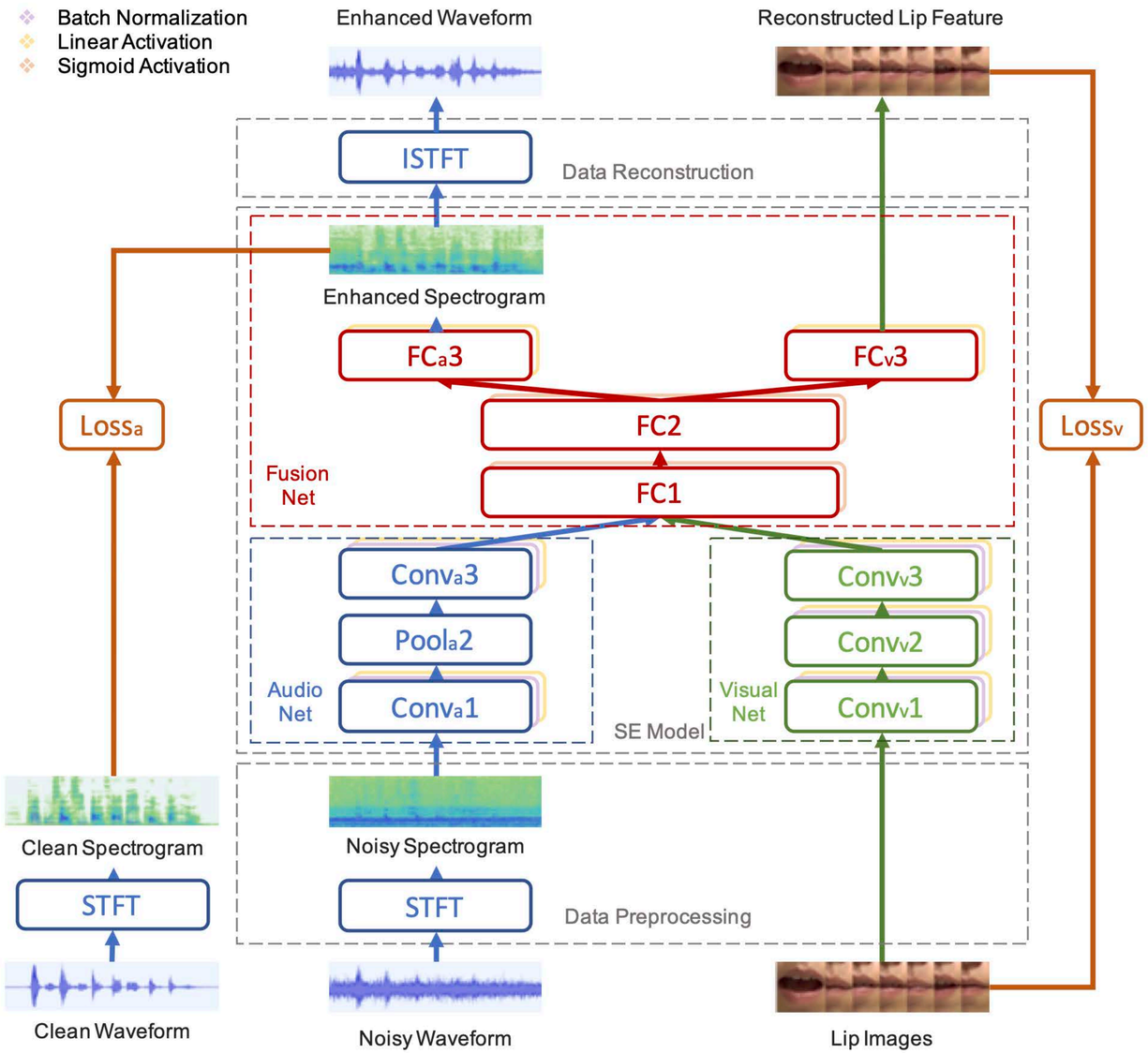


Figure 1: The AVDCNN architecture.

Related Works

Bit-wise Data Compression

- 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
    - Significant figures
- Exponent-only floating-point (EOFP) format [2]
  - No mantissa bit
  - Does not change the represented value itself
  - Only reduces the precision

The Proposed LAVSE

- LAVSE architecture
  - Data preprocessing
  - SE model
  - Data reconstruction
- Two visual data compression techniques
  - Encoder_{AE}
  - Qua_{latent} (EOFP)
- Features
  - Audio: log1p magnitude spectrum
  - Visual: AE+EOFP

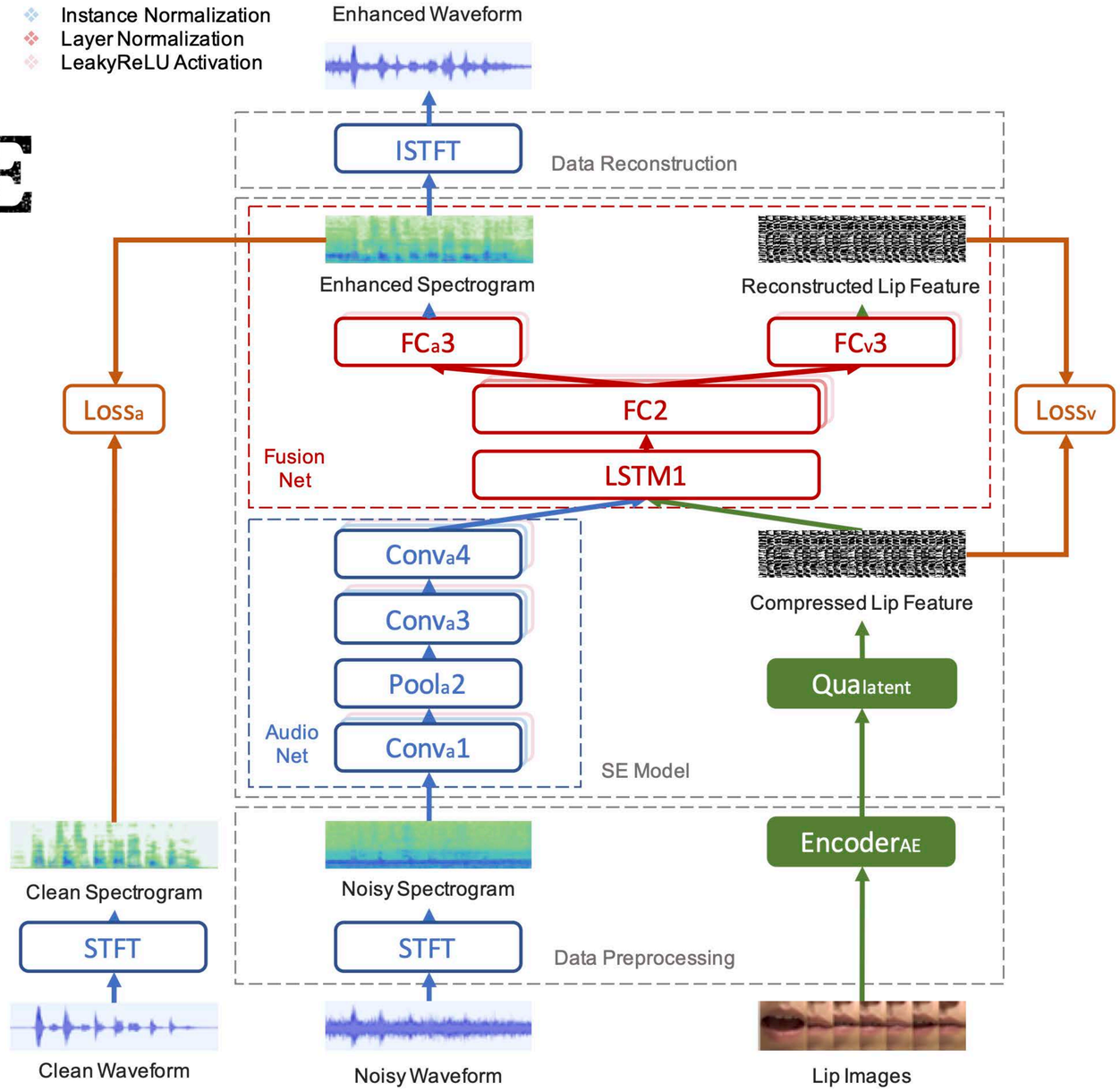
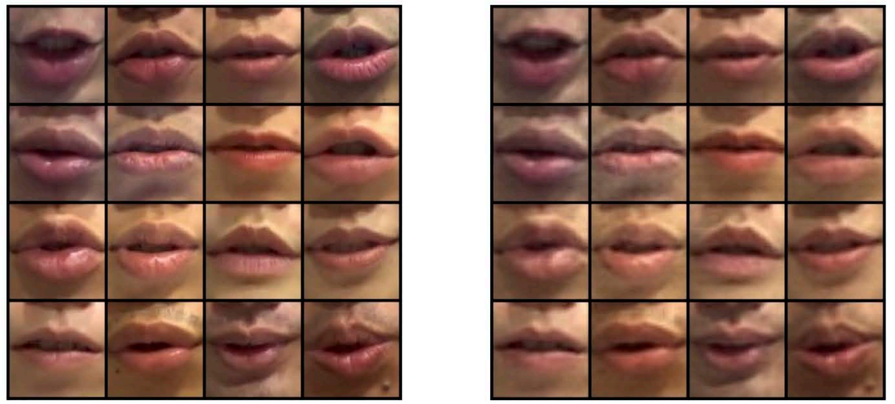


Figure 2: The LAVSE architecture with two visual data compression units (Encoder_{AE} and Qua_{latent}).

The Proposed LAVSE

Encoder$_{AE}$

- Dimension
  - Original image: 3×64×64
  - AE feature: 2048 (= 32×8×8)
- Only 16.67% of the size after AE compression



(a) Original lip images. (b) AE reconstructed images.

Figure 4: Original and AE reconstructed lip images.

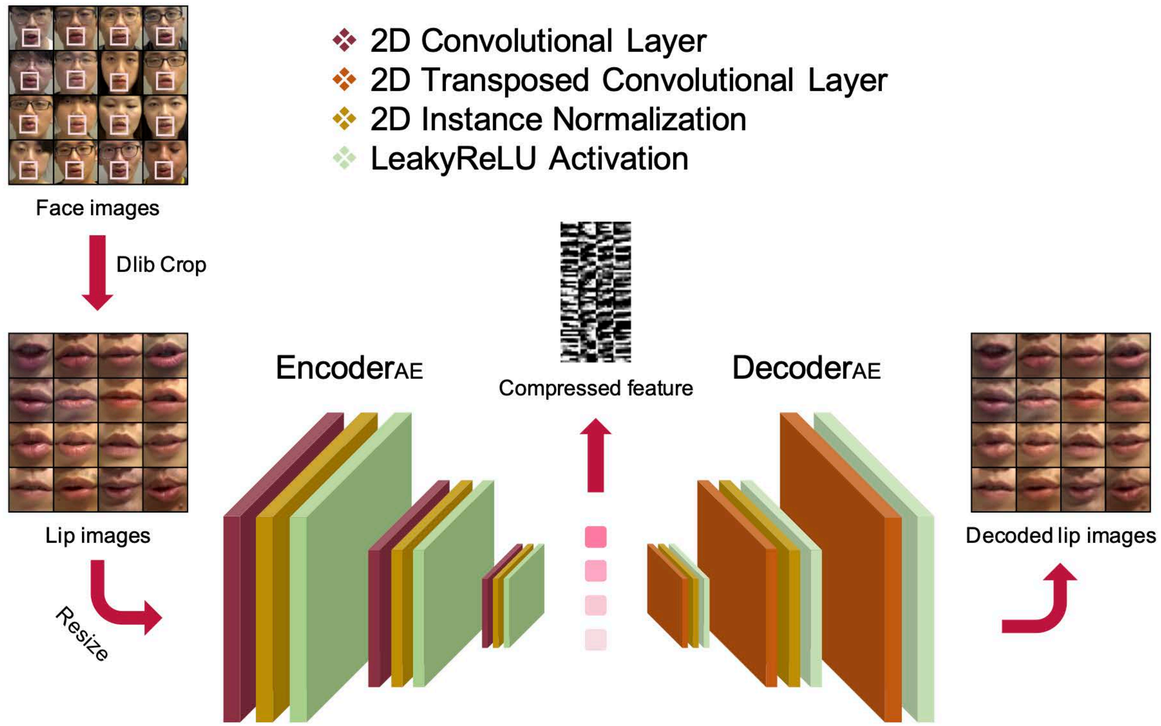


Figure 3: The AE model for visual data compression.

The Proposed LAVSE

Qua$_{latent}$

- Bits
  - 32-bit floating-point
    - 1 sign bit
    - 8 exponential bits
    - 23 mantissa bits
  - 4-bit EOFP
    - 1 sign bit
    - 3 exponential bits
    - 0 mantissa bit
- Only 12.5% of the size after applying EOFP
- User identity has been removed

(a) AE feature.

(b) AE+EOFP feature.

Figure 6: Visual latent features of lips.

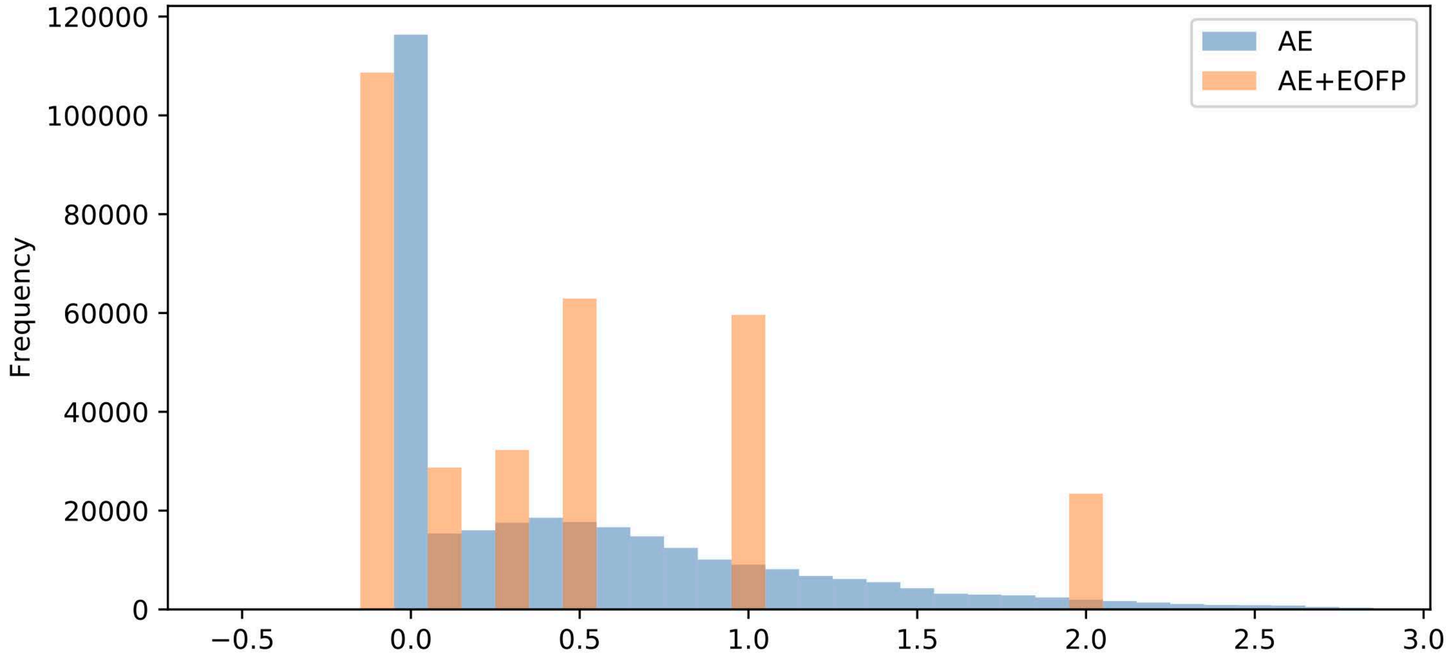


Figure 5: The distributions of visual features before and after applying Qua$_{latent}$.

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 [3]
    - 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 [4]
- Evaluation metrics
  - Perceptual evaluation of speech quality (PESQ) [5]
  - Short-time objective intelligibility measure (STOI) [6]

Experiments

Experimental Results

- Investigate the effects of visual information
- LAVSE(AE): Proposed LAVSE with Encoder$_{AE}$
- Baselines
  - Audio-only SE system
  - AVSE with different visual features
    - AVSE(VGGface): face features processed by VGGface [7]
    - AVSE(face): raw face images
    - AVSE(lip): raw lip images

	PESQ	STOI
Noisy	1.001	0.587
Audio-only	1.283	0.610
AVSE(VGGface)	0.797	0.492
AVSE(face)	1.270	0.616
AVSE(lip)	1.337	0.641
LAVSE(AE)	**1.374**	**0.646**

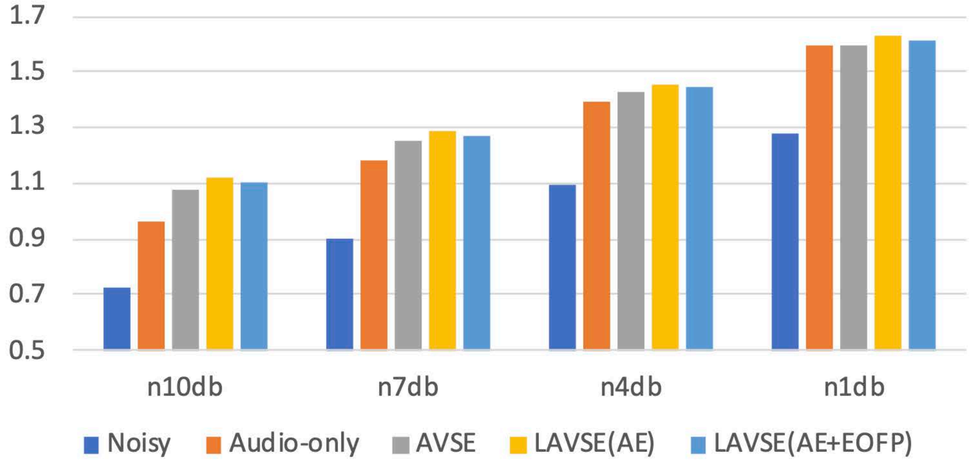
Table 1: PESQ and STOI scores of the LAVSE(AE) system and the baselines.

Experiments

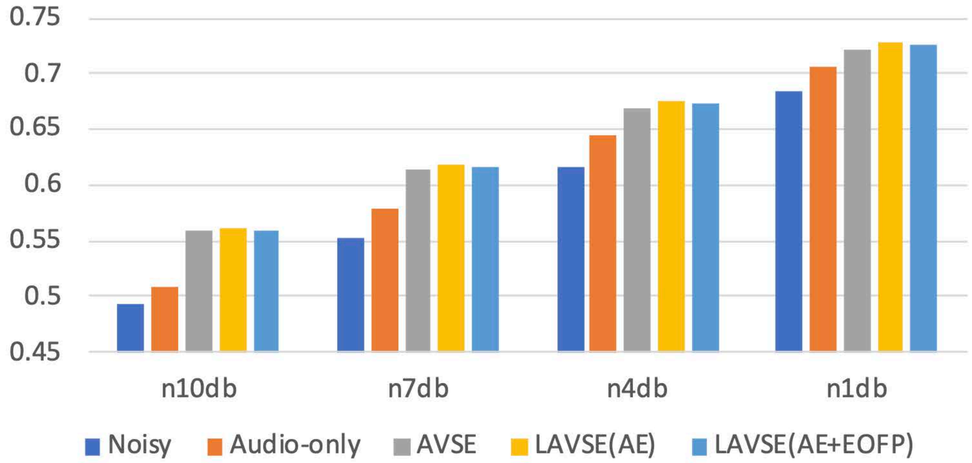
Experimental Results

- Investigate the effects of Encoder$_{AE}$ and Qua$_{latent}$
- Compression ratio
  - Encoder$_{AE}$: $R_{AE} = \frac{3\times64\times64}{2048} = 6$
  - Qua$_{latent}$: $R_{Qua} = \frac{1+8+23}{1+3+0} = 8$
  - Overall: $R_{Comp} = R_{AE} \times R_{Qua} = 48$
- LAVSE(AE+EOFP): LAVSE with Encoder$_{AE}$ and Qua$_{latent}$
  - PESQ: 1.358, STOI: 0.643
  - PESQ and STOI maintain
  - Robust over different SNRs

Refer to Table 1	PESQ	STOI
Noisy	1.001	0.587
Audio-only	1.283	0.610
AVSE(lip)	1.337	0.641
LAVSE(AE)	**1.374**	**0.646**



(a) PESQ.



(b) STOI.

Figure 7: PESQ and STOI scores at specific SNR levels.

Conclusion

- The contributions of this study are threefold
- Verified the effectiveness of incorporating visual information into SE system
- The compressed visual data can still provide significant complementary information for the SE task
- The proposed compression modules can moderately address the privacy problems

Thank you!

References

[1] 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.

[2] 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.

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

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

[5] 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.

[6] 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.

[7] O. M. Parkhi, A. Vedaldi, and A. Zisserman, Deep face recognition. British Machine Vision Association, 2015.