SUT
SEATUC2023 | The 17th South East Asian Technical University Consortium
"Empowering Communities through Innovative Engineering, Science and Technology"

CNN-based Speech Enhancement by using Amplitude Modulation Spectrogram

Backgrounds

- Voice chat and Video Meeting is commonly used in our lives.
- Strong noises from outside can interfere with speech intelligibility.
- Conventional methods could not suppress strong noises efficiently.

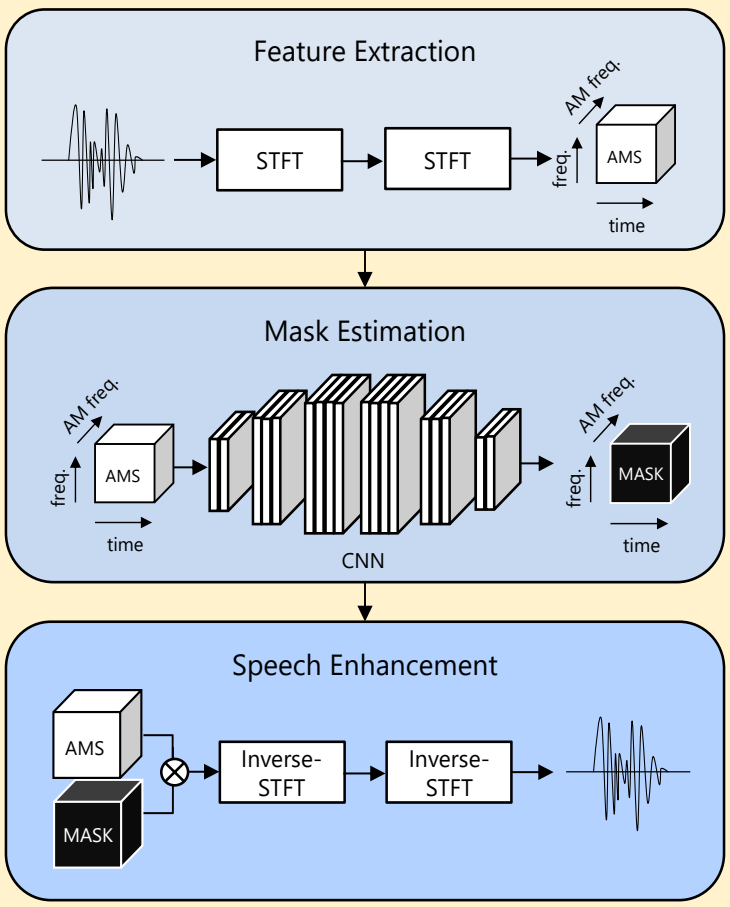
Related Works

- Time-Frequency spectrogram estimation with Convolutional Neural Network (CNN) [1] is one of the modern speech enhancement methods.
- Amplitude Modulation (AM) features of speech signals could be effective hints to suppress noisy signals [2].
- Human Auditory System uses Amplitude Modulation features to percept various sounds from a mixture source [3].

Objective

- Improve speech quality in noisy sources by using Deep Learning-based speech enhancement with AM features.

Proposed Method



- We utilize Amplitude Modulation Spectrogram (AMS).
- A couple of Short-Time Fourier Transformation (STFT) processes generate AMS from noisy speech waveform.
- CNN estimates mask to extract speech components from noisy AMS.

Discussion

- AM feature is effective for speech enhancement in noisy conditions where input SNR is under 0 dB.
- CNN could estimate masks on the AMS feature map, however, there is room for improvement.

Future Work

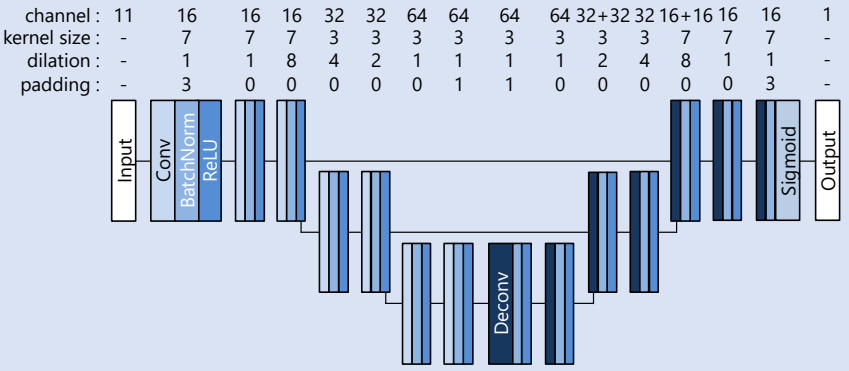
- Apply our method to many kinds of noise and inspect how AM features affect mask estimation with CNN.

Experiment Setup

- Using VCTK version.0.92 dataset [4].
  - 5000 utterances for training
  - 200 utterances for validation
  - 200 utterances for testing
- For the training set, add a random amount of white noise whose Signal-to-Noise Ratio (SNR) is between -10dB and 10dB .
- Parameters to generate AMS

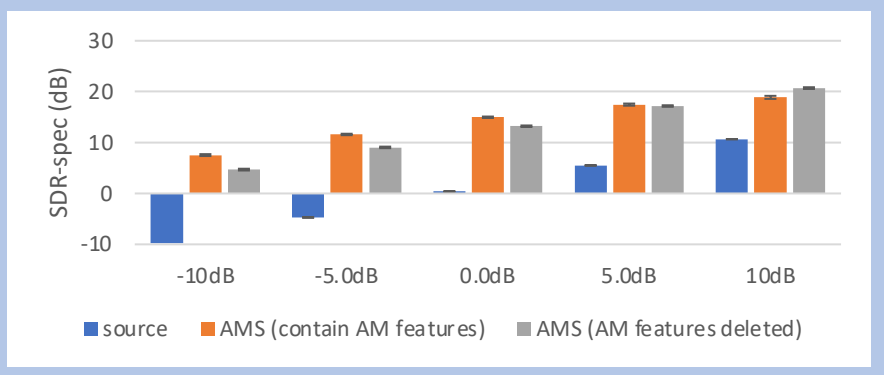
	frame length	hop length	zero-padding
1st STFT	256 (16ms)	32 (2ms)	0
2nd STFT	32 (64ms)	4 (8ms)	224

- CNN Architecture for the Mask Estimation

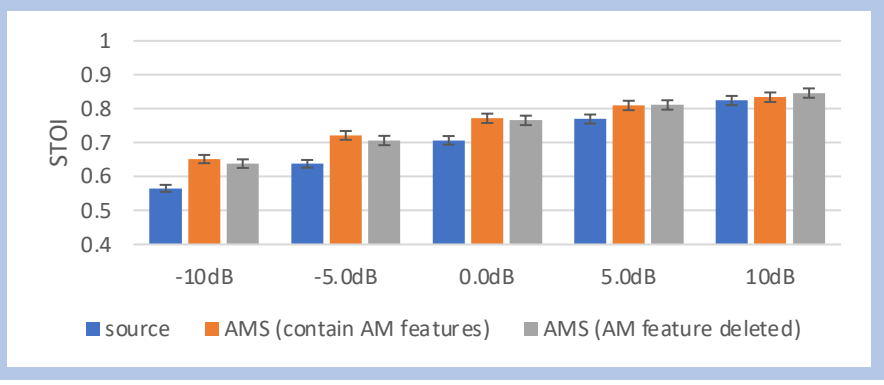


Results

- Evaluation results of enhanced speech by Source-to-Distortion Ratio (SDR)



- Evaluation results of enhanced speech by Short-Time Objective Intelligibility measure (STOI)



References

[1] Y. Xu, et. al., "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, 2015.
[2] K. Fujioka, et. al., "A Noise Reduction Method of Speech Signals Using Running Spectrum Filtering," The IEICE Transactions D, vol. J88-D2, no. 4, pp. 695-703, 2005.
[3] B. Moore, "An Introduction to the Psychology of Hearing - Sixth Edition," Brill Academic Pub, 2013.
[4] J. Yamagishi, et. al., "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92)," University of Edinburgh. The Centre for Speech Technology Research (CSTR), 2019.

Authors Information

* Masaki Wakabayashi Email : mf21140@shibaura-it.ac.jp
Graduate School of Engineering and Science
Shibaura Institute of Technology, Japan

Kazunori Mano
Graduate School of Engineering and Science
Shibaura Institute of Technology, Japan