# Conditional Generative Adversarial Networks for Speech Enhancement and Noise-Robust Speaker Verification

Daniel Michelsanti and Zheng-Hua Tan

Department of Electronic Systems, Aalborg University, Denmark
dmiche15@student.aau.dk, zt@es.aau.dk

INTERSPEECH 2017, Stockholm, Sweden

Conference of the International Speech Communication Association

## Introduction

- Improving speech system performance in noisy environments remains a challenging task.
- Speech enhancement (SE) has been performed using statistical methods like short-time spectral amplitude minimum mean square error (STSA-MMSE) or deep learning techniques, such as deep neural networks (DNNs).
- This work proposes the use of generative adversarial networks (GANs) for SE.
- GANs [1] consist of two players:
  - A generative model, or generator (G), that represents a mapping function from a random noise vector $\mathbf{z}$ to an output sample $G(\mathbf{z})$, ideally indistinguishable from the real data $\mathbf{x}$.
  - A discriminative model, or discriminator (D), that tries to distinguish the samples presented to it between real and fake.
- Our purpose is to use a general-purpose conditional GAN (cGAN) framework, Pix2Pix [2], to perform spectral SE.

## Pix2Pix Framework for Speech Enhancement

- In cGANs, both G and D are conditioned on some extra information $\mathbf{y}$, and trained following a min-max game with the objective:

$$L(D, G) = \mathbb{E}_{\mathbf{x},\mathbf{y}\sim\, p_{data}(\mathbf{x},\mathbf{y})}[\log(D(\mathbf{x}, \mathbf{y}))] + \mathbb{E}_{\mathbf{z}\sim\, p_{\mathbf{z}}(\mathbf{z}),\mathbf{y}\sim\, p_{data}(\mathbf{y})}[\log(1 - D(G(\mathbf{z}, \mathbf{y}), \mathbf{y}))].$$

- Pix2Pix does not use $\mathbf{z}$ (Fig. 1), and the L1 distance between $G(\mathbf{y})$ and the ground truth is used in addition to $L(D, G)$.

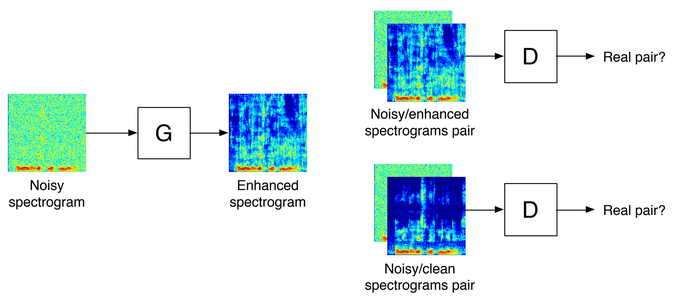


**Fig. 1:** Pix2Pix framework for speech enhancement.

- The spectrogram has been obtained by computing the magnitude of a 512-point STFT of the speech signal.
- The noisy phase is used to reconstruct the enhanced signal.

## Experiments

- Evaluation metrics:
  - PESQ (measure for speech quality).
  - STOI (measure for speech intelligibility).
  - EER of a GMM-UBM speaker verification system.
- Baseline methods:
  - STSA-MMSE.
  - IRM-based DNN speech enhancement algorithm (DNN-SE) [3].
- Datasets:
  - TIMIT (to train the UBM).
  - RSR2015.
  - 5 noise types (airplane, babble, cantine, market, white Gaussian).
- Setup:
  - 6 Pix2Pix front-ends: 5 noise specific (NS-Pix2Pix) and 1 noise general (NG-Pix2Pix).
  - 6 DNN-SE front-ends: 5 noise specific (NS-DNN) and 1 noise general (NG-DNN).
  - Training at 10 and 20 dB SNR. Testing at 0, 5, 10, 15, and 20 dB SNR.
  - 3 tests:
    1. Computation of PESQ and STOI.
    2. Computation of EER on clean speaker model.
    3. Computation of EER on multi-condition speaker model.

## Results

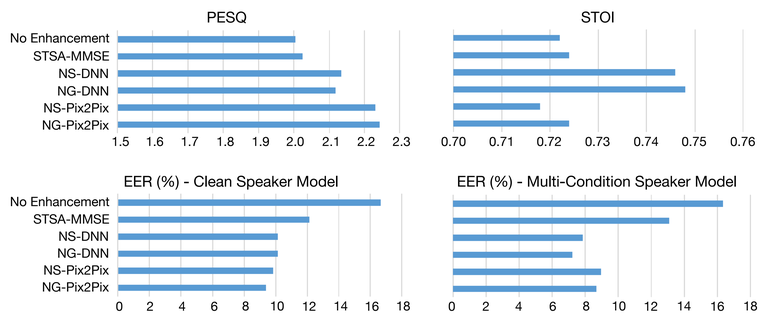


**Fig. 2:** Results.

Fig. 2 shows the average results of the front-ends for the conducted experiments. In general, Pix2Pix can be considered competitive with DNN-SE and overall superior to STSA-MMSE.

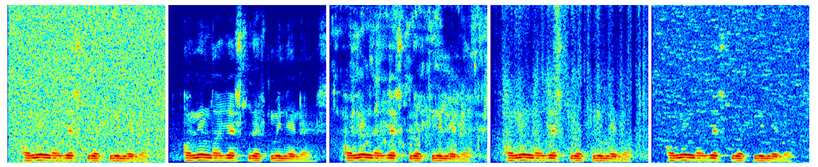

**Fig. 3:** From left to right: noisy spectrogram (white noise at 0 dB SNR); clean spectrogram; spectrogram of the signal enhanced with NG-Pix2Pix; spectrogram of the signal enhanced with NG-DNN; spectrogram of the signal enhanced with STSA-MMSE.

Fig. 3 shows the spectrograms of a noisy utterance, together with its clean and enhanced versions with NG-Pix2Pix, NG-DNN, and STSA-MMSE. It is observed that the spectrogram enhanced by the cGAN approach preserves the structure of the original signal better than the other SE techniques.

## Conclusions

- cGANs are a promising technique for speech denoising, being globally superior to the classical STSA-MMSE algorithm, and comparable to a DNN-SE algorithm.
- Future work includes:
  - Evaluation of the framework in more critical SNR situations.
  - Some modifications:
    - A model with G generating a small size output window from a fixed number of successive frames.
    - A specific perceptual loss to be added to the cGAN loss.

## References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014.

[2] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint arXiv:1611.07004*, 2016.

[3] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech intelligibility potential of general and specialized deep neural network based speech enhancement systems" *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, 2017.