## Summary

- Proposing adversarial domain adaptive VAD (ADA-VAD), which is a deep neural network (DNN)-based VAD method highly robust to audio samples with various noise types and low SNRs.
- Trains DNN models for a VAD task in a supervised manner.
- Simultaneously, the adversarial domain adaptation method adopted to match the domain discrepancy between noisy and clean audio stream in an unsupervised manner.
- ADA-VAD achieves an average of 3.65%p and 7.45%p higher AUC than models trained with manually extracted features on the AVA-speech dataset and a speech database synthesized with an unseen noise database.

## Method

### Model components

- **Feature Extractor**: Extracting acoustic features with learnable filters.
- **VAD classifier**: Predicting VAD labels for each frame.
- **Discriminator**: Classifying whether its input noisy or clean speech signal.

### Baseline methods

- VAD methods based on deep neural networks
  - DNN [7], bDNN [8] and LSTM [9]
  - Trained with manually extracted acoustic features such as the mel-spectrogram
  - Trained on the Train_D1 dataset
- End-to-end domain-adversarial voice activity detection (DA-VAD) [1]
  - Trained on the Train_D2 dataset

### Results

- Impact of the adversarial domain adaptation
  - Achieving 1.85%p higher AUC compared to the LSTM-FL (Same model architecture without adversarial domain adaptation method)
  - Comparison to a DNN-based model that learned Mel-spectrograms
    - Achieving 9.06% higher AUC compared to bDNN
    - Achieving 13.77% higher AUC than LSTM in extremely low SNR level such as -10
  - The lower the SNR levels, the higher AUCs score gap between ADA-VAD and other VAD methods

### Updating the weights of the feature extractor and the VAD classifier

- Extract latent feature \( z^{clean} \) by feed-forwarding noisy speech \( x^{noisy} \) through the feature extractor.
- Predict VAD label \( y^{noisy} \).
- Fool the discriminator.

### Updating the weight of the discriminator

- Minimize auto-encoding loss from feature from clean speech signal \( z^{clean} \).
- Maximize auto-encoding loss from feature from corrupted speech signal \( z^{noisy} \).
- Mean squared error loss (MSE) used as the objective loss function.
- Adopting Boundary-Equilibrium GAN approach (BEGIN) [3]

\[
\min_{\Phi} \mathcal{V}_{BG}(D) = \mathcal{L}_{1}^{\Phi}(D_{\Phi}^{clean}) + \mathcal{L}_{1}^{\Phi}(D_{\Phi}^{noisy}) - \mathcal{L}_{2}^{\Phi}(D_{\Phi}^{clean}) - \mathcal{L}_{2}^{\Phi}(D_{\Phi}^{noisy}),
\]

where \( k_1 < 0 \), \( k_2 = 0 \).

### Experimental Setup

- **Dataset preparation**
  - TIMIT corpus [4] as the speech database for training and test dataset.
  - Train dataset
    1. **Train_D1**: TIMIT corrupted with Sound effect library as noise dataset.
    2. **Train_D2**: TIMIT corrupted with randomly selected noise data of the Sound effect library into 18 classes.
  - Both train datasets are synthesized in randomly selected SNR level from -12 to 10.
  - Test dataset
      - Synthesize in 4 different SNR levels: -10, -5, 0, 5.
    2. **Test_D2**: AVA-speech dataset [6]
      - Human annotated VAD label.

- **Baseline methods**
  - VAD methods based on deep neural networks
    - DNN [7], bDNN [8] and LSTM [9]
    - Trained with manually extracted acoustic features such as the mel-spectrogram
    - Trained on the Train_D1 dataset.
  - End-to-end domain-adversarial voice activity detection (DA-VAD) [1]
    - Trained on the Train_D2 dataset.

### References

[1] End-to-end domain-adversarial voice activity detection, Laukoch et al., 2020, INTERSPEECH