ADA-VAD: UNPAIRED ADVERSARIAL DOMAIN ADAPTATION FOR NOISE-ROBUST VOICE ACTIVITY DETECTION

Taesoo Kim⋆‡, Jiho Chang† and Jong Hwan Ko⋆

ji5u1031@g.skku.edu⋆‡ jiho.chang@kriss.re.kr † jhko@skku.edu ⋆

Department of Electrical and Computer Engineering, Sungkyunkwan University, Republic of Korea⋆

KT Corporation, Republic of Korea ‡

Korea Research Institute of Standards and Science, Republic of Korea†

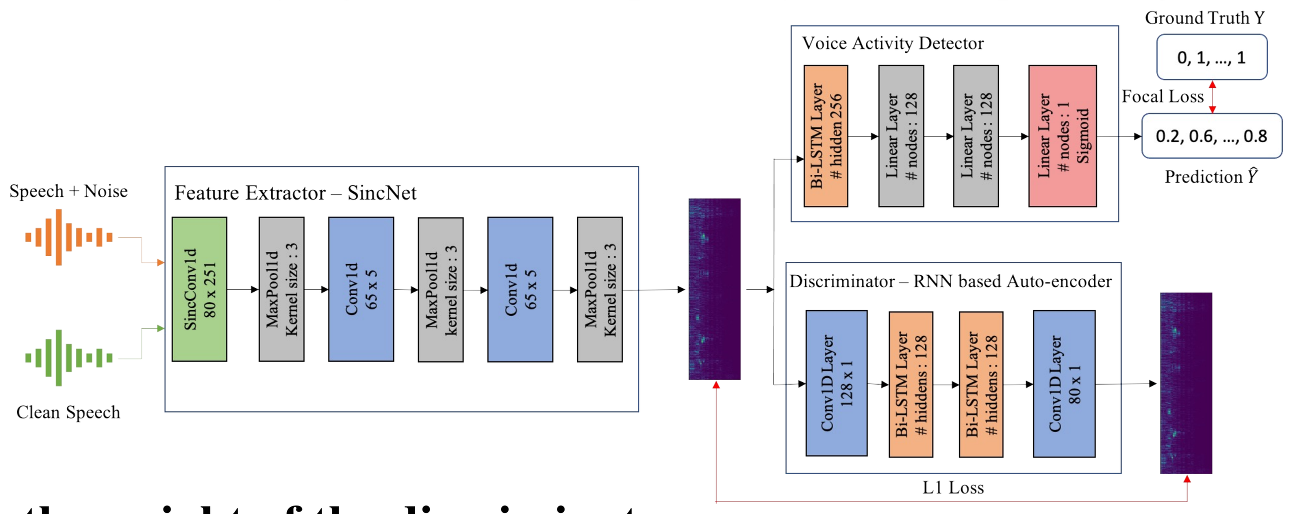
Summary

- 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 3.6%p and 7%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 frames
- Discriminator : Classifying whether its input noisy or clean speech signal



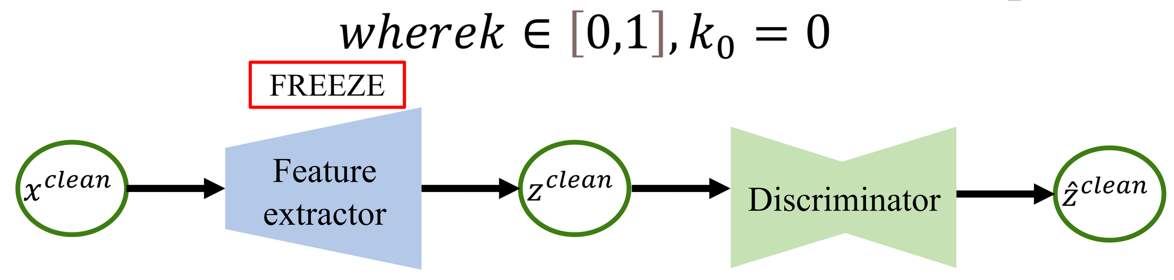
Updating the weight of the discriminator

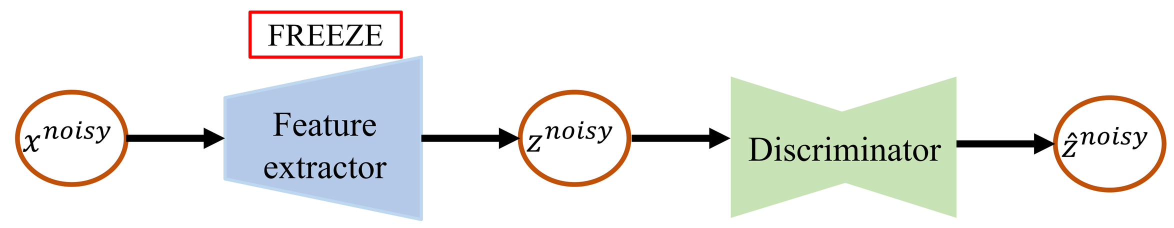
- Minimize auto-encoding loss from feature from clean speech signal z^{clean}
- Maximize auto-encoding loss from feature from corrupted speech signal z^{noisy}
- Mean squared error loss (MSE) used as the objective loss function
  - Adopting Boundary-Equilibrium GAN approach (BEGAN) [3]

$$min_D V_{BG}(D) = E_{z_T^{clean} \sim p_{z^{clean}}}\left[l_D\left(z_T^{clean}\right)\right] - k_t * E_{z_T^{noisy} \sim p_{z^{noisy}}}\left[l_D\left(z_T^{noisy}\right)\right]$$

$$k_{t+1} = k_t + \lambda_k \left(\gamma E_{z_T^{clean} \sim p_{z^{clean}}}\left[l_D\left(z_T^{clean}\right)\right]\right) - E_{z_T^{noisy} \sim p_{z^{noisy}}}\left[l_D\left(z_T^{noisy}\right)\right],$$

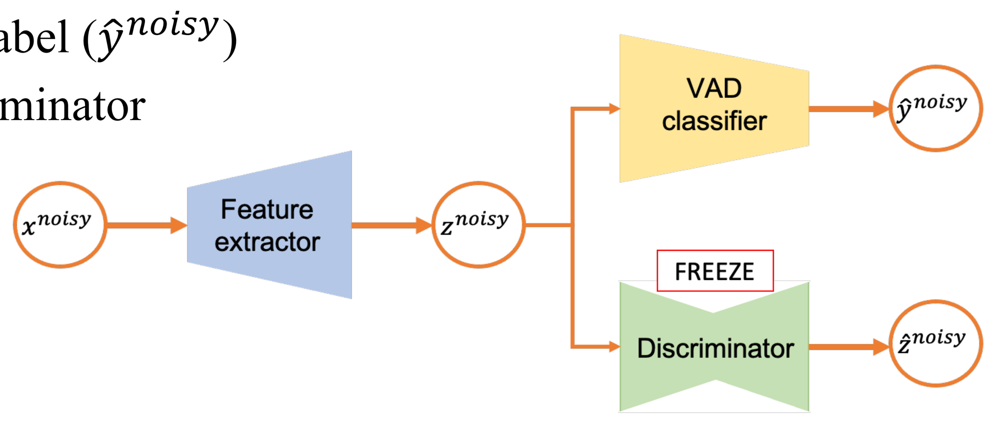
$$where k \in [0,1], k_0 = 0$$





Updating the weights of the feature extractor and the VAD classifier

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



$$L_{total} = L_{VAD} + \tau L_{BEGAN}$$

Experimental Setups

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
  1. **Test_D1** : NOISEX-92 database [5] as noise dataset for 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

Results

- Impact of the adversarial domain adaptation
  - Achieving 1.8 %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

SNRs	*DNN*	*bDNN*	*LSTM*	*LSTM-F*	*ADA-VAD*
-10	67.46	70.44	73.12	83.64	**86.89**
-5	67.46	80.46	81.44	91.85	**94.36**
0	85.4	88.6	87.73	95.46	**97.01**
5	91.73	93.62	91.44	97.04	**98.01**
10	95.69	96.32	93.11	97.8	**98.48**
AVG	83.4	85.89	85.37	93.16	**94.95**

Train_D1 as the training set. AUC(%) on the Test_D1

Model	*DNN*	*bDNN*	*LSTM*	*LSTM-F*	*ADA-VAD*
AUC(%)	67.46	70.44	73.12	83.64	**86.89**

Train_D1 as the training set. AUC(%) on the Test_D2

- Comparison to the DA-VAD method [1]
  - Achieving 1.1 % higher AUC compared to DA-VAD
  - Achieving 7.52 % higher AUC on Test_D2 (AVA-Speech dataset)

Test Set	*SNRs*	*DA-VAD*	*ADA-VAD*
Test_D1	*-10*	85.02	**87.42**
	-5	93.53	**94.95**
	0	96.8	**97.74**
	5	98.17	**98.73**
	10	98.8	**99.6**
	AVG	94.47	**95.6**
Test_D2	-	71.58	**79.1**

Train_D2 as the training set. AUC(%) of DA-VAD and ADA-VAD for each SNR levels on the Test_D1 and the Test_D2

Conclusion

- Proposing a VAD model trained with the adversarial domain adaptation technique
- Matching distribution discrepancy between clean speech signal and speech signal corrupted by background noises
- Able to extract acoustic feature that is more suitable for VAD task
- Audio recordings with multiple background noise types available as training dataset

References

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

[2] Ravanelli, Mirco, and Yoshua Bengio. "Speaker recognition from raw waveform with sincnet." 2018 IEEE Spoken Language Technology Workshop (SLT). IEEE, 2018.

[3] Berthelot, D., Schumm, T., Metz, L. (2017). Began: Boundary equilibrium generative adversarial networks. arXiv preprint arXiv:1703.10717.

[4] Zue, V., Seneff, S., Glass, J. (1990). Speech database development at MIT: TIMIT and beyond. Speech communication, 9(4), 351-356.

[5] Varga, A., Steeneken, H. J. (1993). Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems. Speech communication, 12(3), 247-251.

[6] Sourish Chaudhuri et al., “Ava-speech: A densely labeled dataset of speech activity in movies,” arXiv preprint arXiv:1808.00606, 2018.

[7] Xiao-Lei Zhang, “Deep belief networks based voice activity detection,” IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 4, pp. 697–710, 2012.

[8] Xiao-Lei Zhang, “Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection,” in Fifteenth annual conference of the international speech communication association, 2014.

[9] Florian Eyben, “Real-life voice activity detection with lstm recurrent neural networks and an application to hollywood movies”, ICASSP 2013