



# Adversarial Multi-Task Deep Learning for Noise-Robust Voice Activity Detection with Low Algorithmic Delay

Claus M. Larsen, Peter Koch, Zheng-Hua Tan

Department of Electronic Systems, Aalborg University, Denmark

zt@es.aau.dk

## Abstract

Voice Activity Detection (VAD) is an important pre-processing step in a wide variety of speech processing systems. VAD should in a practical application be able to detect speech in both noisy and noise-free environments, while not introducing significant latency. In this work we propose using an adversarial multi-task learning method when training a supervised VAD. The method has been applied to the state-of-the-art VAD *Waveform-based Voice Activity Detection*. Additionally the performance of the VAD is investigated under different algorithmic delays, which is an important factor in latency. Introducing adversarial multi-task learning to the model is observed to increase performance in terms of Area Under Curve (AUC), particularly in noisy environments, while the performance is not degraded at higher SNR levels. The adversarial multi-task learning is only applied in the training phase and thus introduces no additional cost in testing. Furthermore the correlation between performance and algorithmic delays is investigated, and it is observed that the VAD performance degradation is only moderate when lowering the algorithmic delay from 398 ms to 23 ms.

**Index Terms:** Voice Activity Detection, adversarial multi-task learning, algorithmic delay, deep learning, noise robustness

## 1. Introduction

Voice Activity Detection (VAD) aims to detect which segments of an audio stream contains speech and the segments are typically 10 ms each [1], [2], [3]. It is widely used as a pre-processing step in more complex audio signal processing tasks such as speech recognition [4], speaker verification [5] or speech enhancement [6], but can also be applied on its own to reduce the computational cost of downstream processing. While detecting speech in a noise-free environment is a trivial task, the difficulty in classification arises in noisy environments [7].

VAD algorithms can generally be categorized into classes of supervised and unsupervised methods. The unsupervised methods can be energy based [8], however, this approach is very sensitive to noisy conditions. More complex unsupervised methods are generally based on assumptions of speech and noise characteristics, e.g. using Mel Frequency Cepstral Coefficients (MFCCs) [9], [10], perceptual spectral flux [11] or pitch detection and spectral flatness [1].

In recent years supervised methods for VAD has gained increased popularity within the field of research [12], [13]. The supervised methods require large amounts of labelled speech data and their performance are highly dependent on the quality of the labelled data used for training and testing. Some supervised methods contains a pre-processing step which aims to extract useful features from the audio such as MFCCs [14], while other methods resolves to the raw waveform as their input [3], [15]. A benefit of using the raw waveform as input to the su-

pervised VAD is that the method will potentially find the most optimal features to be used for classification on its own, and is therefore able to utilise both the magnitude and the phase of the audio [3]. Using the raw waveform as features to the VAD is an active area of research and has shown appealing performance in terms of noise-robustness [3], [15].

Two important factors in a VAD algorithm are how noise-robust it is and how much latency it introduces. Adversarial multi-task learning has proven to be effective to make applications invariant to noise and thereby more noise robust, e.g. for speech recognition in [16] and speech enhancement in [17]. Noise robustness of speaker verification is largely boosted by adversarial training in [18]. When applying VAD in a real-world application, often the latency is of great concern. Even though VAD is an active area of research, the algorithmic delay it introduces is rarely explored.

In this work we aim to investigate if the noise-robustness can be increased even further by introducing an additional sub-network which aims to classify the noise types to a supervised VAD, and train the VAD adversarially to these. The supervised VAD method presented in [3], which is based on fully convolutional neural networks (CNN) and shows state-of-the-art performance on the AURORA2 dataset [19], will be used as the framework in this work. An additional discriminative sub-network for adversarial multi-task training, inspired by the work of [20] and further refined for the speech recognition task in [16] will be introduced. The additional network used for adversarial multi-task learning is only introduced in the training phase and thus introduces no additional cost in testing. Furthermore, we investigate the impact of different algorithmic delays on VAD performance and realise this by varying CNN kernel sizes. The source code for this work is publicly available on GitHub<sup>1</sup>.

## 2. Proposed method

In this work we propose to introduce adversarial multi-task learning to enhance the robustness of deep model based VAD. Specifically, an additional sub-network for adversarial training is introduced to a state-of-the-art waveform-based VAD using a CNN model. The entire framework is illustrated in Figure 1, in which the adversarial-training sub-network is shown by the green box, the algorithmic delay is investigated by modifying the blue block, and the waveform-based VAD [3] consists of the blue and red blocks.

### 2.1. Framework

In realising the adversarial multi-task learning for VAD, we consider the model presented in [3] with our own implementation in Pytorch [21]. The method resorts to a fully convolu-

<sup>1</sup><https://github.com/aau-es-ml/VAD-with-adversarial-multi-task-learning>

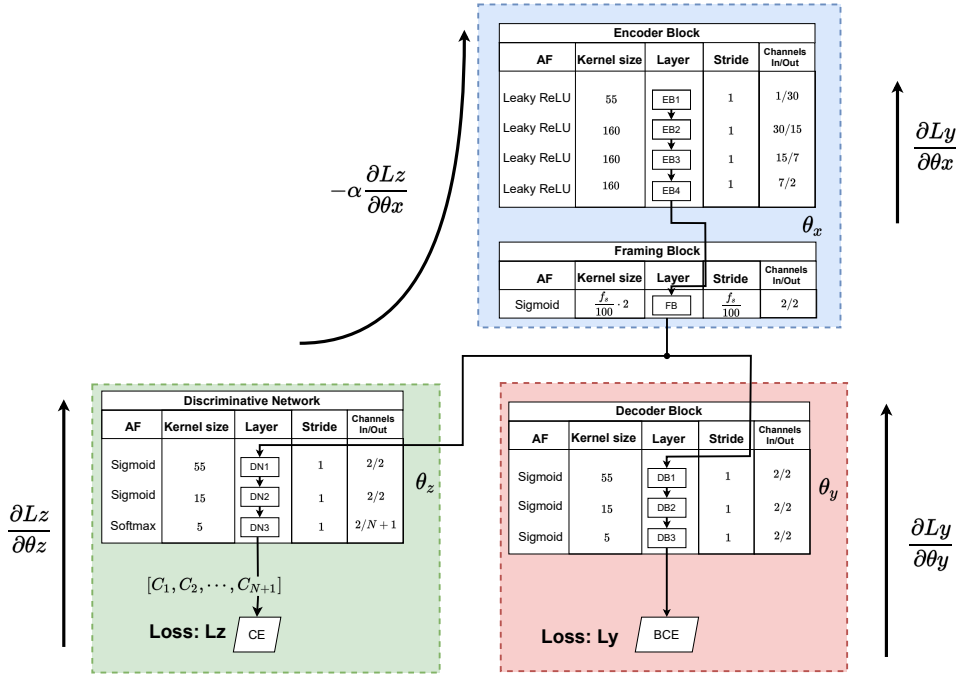


Figure 1: Overview of the proposed discriminative network (the green block) applied to the waveform-based VAD (the blue and red blocks) proposed in [3].

tional neural network. It consists of an Encoder Block (EB), a Framing Block (FB) and a Decoder Block (DB). The EB is taking as input the raw waveform of audio and the FB generates features on a 10 ms basis. The DB further refines these features into the VAD output, where the larger value of the two channels, *speech* and *non-speech*, determines the VAD label as speech or non-speech on a 10 ms basis.

In this work we introduce the additional Discriminative Network (DN). It takes as input the output from the FB and aims to learn to correctly classify the different noise types as well as to adversarially train the EB and FB part to make it noise robust.

## 2.2. Adversarial multi-task learning

The first part of this work is to introduce adversarial learning. The proposed DN is implemented in a similar way to the DB, where each layer resolves time-frequency representations in the channels of feature maps and the shrinking kernel sizes (55, 15, 5) reflect the decreasing modulation frequency (1.83 Hz, 6.66 Hz, 20 Hz) with the segment rate being 100 Hz [3]. The discriminative network generates softmax probabilities for each of the  $N + 1$  channels (i.e. outputs), where  $N$  is the number of different noise types in the training set, while the remaining channel is for clean speech. The labels used for the DN is the noise types on a 10 ms basis, similarly to the VAD labels.

Following the DN, the cross-entropy loss is calculated between the noise types predicted by the DN and the true noise types. Following the DB is the binary-cross-entropy loss calculated between the VAD labels and the truth labels. The losses are expressed as:

$$L_z = - \sum_i t_i \log(p_i) \quad (1)$$

$$L_y = - [t \log(p) + (1 - t) \log(1 - p)] \quad (2)$$

where  $t$  is the true labels and  $p$  is the scores output by the networks on a 10 ms basis.

When backpropagating the error through the model, the gradients of the DN are updated based only on the loss  $L_z$ , the gradients of the DB are updated based only on the loss  $L_y$  while the gradients of the EB and the FB are updated based on both losses. However, the sign of the gradients calculated from  $L_z$  is flipped such that the EB and FB are trained adversarially to the DN and friendly to the DB. Additionally the magnitude of this gradient is multiplied by a scalar  $\alpha$  that determines the contribution from this sub-network. The key idea behind the method is that the FB will then output features that are invariant to the noise type which in turn will lead to a more noise-robust VAD and hence better VAD performance. The DN is illustrated in Figure 1. The gradients are noted as the partial derivatives of the loss function  $L$  with respect to the parameters  $\theta$ .

The convolutional operations of the DN can be expressed as:

$$y_{[c]}^{[l]}(\tau) = \text{AF} \left( \left( \mathbf{F}_{[c]}^{[l]} * y_{[c]}^{[l-1]}(\tau) \right) + \mathbf{b}_{[c]}^{[l]} \right) \quad (3)$$

where  $c$  denotes the channel,  $l$  denotes the layer,  $\mathbf{F}_{[c]}^{[l]} \in \mathbb{R}^{C \times k}$  is the convolutional kernel,  $\mathbf{b}_{[c]}^{[l]} \in \mathbb{R}^{C \times 1}$  is the bias,  $y_{[c]}^{[l]} \in \mathbb{R}^{C \times \max(\tau)}$  is the feature map and AF is the activation function.

## 2.3. Algorithmic delay

Second part of this work focuses on reducing the latency of the VAD. Only the algorithmic delay is considered, i.e. it is investigated how the VAD performance is affected based on how many future samples are used in the predictions, from here on referred to as *future context*. The amount of future temporal context used for a given classification is calculated based on the fact that the feature map through a 1-dimensional convolutional layer will shrink as stated by Eq. 4.

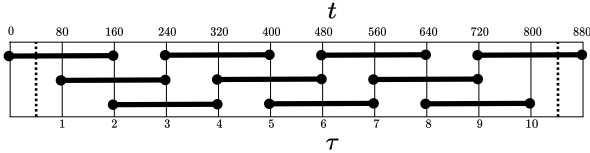


Figure 2: Illustration of the context needed to generate an output from the FB. In this example 10 outputs are generated from an 8 kHz signal, but because of the 50% overlap between frames, an additional  $\frac{f_s}{100} - 1$  samples of context is needed to generate an output.

$$n^{[l]} - k + 1 = n^{[l+1]} \quad (4)$$

where  $n^{[l]}$  is the size of the feature map generated by the  $l^{th}$  layer and  $k$  is the kernel size. The number of samples by which the feature map is shrinking will have to be considered as context where half of it is past and the other half is future. The algorithmic delay introduced by the network is found by calculating how much context is needed in each layer and finally summing them together. The context introduced in the EB layers is simply found as in Eq. (4), while the context introduced in the framing block is more complex and best illustrated by Figure 2. The context introduced by the DB is dependent on the stride of the FB and once again calculated using Eq. (4). The total algorithmic delay in seconds is found by dividing the context with two times the sampling rate and is calculated as Eq. (5).

$$AD = \frac{\sum_{n=1}^4 (k_{EBn} - 1) + \frac{f_s}{100} - 1 + \sum_{i=1}^3 (k_{DBi} - 1) \cdot \frac{f_s}{100}}{2f_s} \quad (5)$$

### 3. Speech corpora

In this work two speech corpora are used. The AURORA2 [19] database and the TIMIT [22] database.

#### 3.1. AURORA2

First is the AURORA2 [19] database which is used for multi-condition training at a sampling frequency of 8 kHz. The training set consists of 8440 utterances with four noise types artificially added at SNR levels of 5 dB, 10 dB, 15 dB, 20 dB and *clean*. The four noise types used are *subway*, *babble*, *car* and *exhibition hall*. For each combination of noise type and SNR level 422 utterances are used.

AURORA2 contains three test sets, of which two are used in this work. Test set A uses the same noise types as the training set, and test set B uses four noise types unknown to the training set. These are *restaurant*, *street*, *airport* and *train station*. In the test sets the following SNR levels are used: -5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB and *clean*. Each test set consists of 4004 utterances which are evenly distributed on the four noise types and repeated for each SNR level. The true VAD labels are from the open-source rVAD repository [1].

#### 3.2. TIMIT

Secondly, the TIMIT [22] speech corpus is used. The training set and test set, respectively, consists of 4620 and 1680 spoken sentences. In this work, 6 noise types are artificially added by

the authors at SNR similar to those of the AURORA2 test sets. Different instances of the same noise type are used for training and testing sets such that no instance of noise is ever repeated. Furthermore, in this work the test set is split into validation and test sets using a 1/3, 2/3 split. The purpose of the validation split is to find the optimal hyperparameter  $\alpha$  shown in Figure 1, while the test split is used to obtain the rest of the results in this work. The noise types used for training and testing are the same, meaning the noise types will be known under testing. The noise types *babble*, *bus*, *caf* and *pedestrian* are from the CHiME3 dataset [23] while *babble* and *speech shaped noise* are generated by the authors of [24]. The noise is artificially added as described in section 2.A of [25]. The VAD labels for the TIMIT database is generated using the wrd-formatted files in the database which states, at which time stamps speech is present. These labels are shared along with the source code of this work which is publicly available on GitHub.

## 4. Experimental setup and results

In order to evaluate the performance of the proposed method for adversarial training, the model is trained and tested on both the TIMIT database and the AURORA database. When evaluating Eq. (5) with the kernel sizes shown in Figure 1, the context to generate a VAD label spans 398 ms to both sides. This means that to generate the first/last VAD label of each file, 398 ms of context is missing. In combination with the short duration of the files of AURORA2 (typically 0.8-2 seconds) and TIMIT (typically 2-4 seconds) leads to that a large part of the VAD outputs of each file will be generated without sufficient context. Additionally, the files in these data sets are all following the same structure. That is a short duration of silence in the beginning and the end, while the middle part contains speech. This leads to that the VAD algorithm learns to recognize this structure based on the missing context in the beginning and the end. To deal with this problem, during training and testing 10 files of the same noise type and SNR level are randomly chosen and concatenated leading to longer inputs to the VAD algorithm and therefore a smaller part of the VAD outputs will be generated based on insufficient context. The reason for using 10 files is that the computer on which the model is trained runs into memory problems when using more files. This forward CNN calculation step is performed three times before each backward step leading to a mini-batch size of 30 audio files. For each three forward steps a single backward step is performed using the RMSprop optimiser. The model is trained over 30 epochs, the learning rate is initialised as 0.01 and the learning rate is multiplied by 0.7 after each epoch.

Table 1: AUC values on the validation sets using different values of  $\alpha$

$\alpha$	0	0.01	0.1	1	10	100
AURORA2 A	93.90	94.15	<b>95.18</b>	94.74	94.30	94.24
AURORA2 B	91.15	91.16	92.49	<b>92.69</b>	91.13	90.38
TIMIT	88.78	89.98	<b>90.32</b>	89.3	88.91	87.94

#### 4.1. Adversarial multi-task learning

First the optimum value of the scalar  $\alpha$  is found experimentally on both data sets by using validation data. Given that the AURORA2 is labelled as 73% speech and TIMIT is labelled as 85% speech, the results will be given by calculating the Area Under Curve (AUC) of their respective Receiver Operating Characteristics (ROC) curves while the accuracy will be disregarded as it

can be misleading. For finding the optimum value of  $\alpha$  the average AUC over the noise types of the validation split at each SNR level is calculated. In each experiment the model is initialised using the same set of parameters to remove the randomness that can potentially be introduced by different initialisations. The Leaky ReLU layers are initialised using He [26] while the parameters of the sigmoid layers are initialised using Xavier [27]. The Leaky ReLU layers use a slope coefficient of 0.01. The average AUC at  $\alpha = 10^n, n \in [-2, -1, 0, 1, 2]$  on the validation sets are presented in Table 1. It is seen that the performance of the VAD is increased by a wide range of values of  $\alpha$  and in general the addition of a discriminative network for adversarial multi-task learning outperforms the baseline model in terms of AUC. In the case of all 3 test sets, and thereby also to the model both known and unknown noise types, it is found that a wide range of values of  $\alpha$  results in an increase in performance. In two of three cases a value of 0.1 is found to be optimal, thus this will be the value used for further experiments in this work.

The performance of the model using an  $\alpha$  value of 0.1, which was found optimal on the validation splits, is further investigated using the test splits of each data set. The results are seen in Table 2. Once again the models are trained from the same initial values using the same simulation settings as described earlier and it is clearly seen that while the performance at high SNR levels is approximately the same with and without adversarial multi-task learning, at lower SNR levels the models trained using adversarial multi-task learning performs better and proves to be more noise-robust. This is the case both when it comes to noise types known to the model (TIMIT and AURORA2 test set A) and noise types unknown to the model (AURORA2 test set B).

Table 2: AUC values on the test sets of AURORA2 and TIMIT with (W) and without (W/O) adversarial multi-task learning. When adversarial multi task learning is used, an  $\alpha$  value of 0.1 is used

	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	-5 dB	Mean
AURORA2 A - W/O	97.96	97.95	97.85	97.62	96.59	92.12	78.79	94.08
AURORA2 A - W	97.91	97.90	97.88	97.66	96.94	93.44	84.49	95.18
AURORA2 B - W/O	97.96	97.96	97.10	95.05	89.56	79.01	63.79	88.64
AURORA2 B - W	97.91	97.90	97.54	96.54	94.20	89.11	74.25	92.49
TIMIT - W/O	95.44	95.41	94.63	92.64	88.60	82.67	71.96	88.76
TIMIT - W	95.62	95.49	94.92	93.91	90.75	84.94	74.26	89.99

#### 4.2. Algorithmic delay

The second part of this work is to evaluate the performance of the proposed method at lower algorithmic delays. The model is trained with the optimum value of  $\alpha = 0.1$  while the kernel sizes of the decoder block is reduced. The algorithmic delay is calculated as a function of the sampling frequency and the kernel sizes as in Eq. (5). The majority of the algorithmic delay is introduced by the DB, thus only these kernel sizes will be varied. The kernel sizes and their corresponding algorithmic delays as used in the experiments are presented in Table 3. The AUC is calculated as the average over the 7 SNR levels and the noise types of each test set.

It is seen that the performance of the VAD decreases as the algorithmic delay is lowered, however this performance decrease is not drastic. It is in particular interesting to note that even when completely disregarding the decoder block with an algorithmic delay of 23 ms the VAD still performs well. When decreasing the algorithmic delay from 398 ms to 23 ms, only a performance decrease of 7% AUC is seen. The performance of

different algorithmic delays at each SNR level for AURORA2 test set B is presented in Figure 3. In particular the performance at clean speech seems to be unaffected by the low algorithmic delay.

Table 3: VAD performance in terms of AUC at different kernel sizes and algorithmic delays tested on AURORA2 test sets A and B

DB1	DB2	DB3	AD [ms]	AURORA2 B	AURORA2 A
55	15	5	398	92.11	94.44
45	15	5	348	91.91	94.38
35	15	5	298	90.81	93.44
25	15	5	248	91.04	93.32
15	10	5	173	90.95	93.01
10	7	5	133	88.09	90.84
7	5	5	98	89.01	91.85
5	3	3	78	87.14	89.78
3	3	2	63	85.77	90.00
2	2	2	53	85.26	87.56
2	2	0	43	85.03	87.40
2	0	0	33	85.66	87.90
0	0	0	23	85.07	87.27

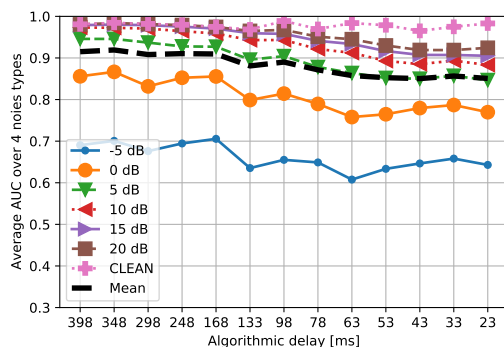


Figure 3: Average AUC at decreasing algorithmic delays on the (to the model) unknown noise types of AURORA2 test set B

## 5. Conclusions

In this paper we proposed a novel approach of training a supervised VAD using adversarial multi-task learning, where the model is trained friendly to the VAD labels and adversarially to the noise types aiming to make the VAD more invariant to noise. This is done by introducing an additional sub-network which aims to classify the noise types to the model. The VAD is then trained adversarially to these. It is found that the adversarial multi-task training is capable of increasing the VAD performance especially in more noisy environments, and it is shown to increase performance both when presented to unknown and already known noise types. At lower SNR levels the performance is boosted for both AURORA2 sets A and B. On the TIMIT set an increase is also observed, however less significant. This multi-task learning is only used when training the model and disregarded under testing, meaning the proposed method is cost-less once training has finished.

Furthermore it was investigated if this method can be useful in a low-latency application. This was done by reducing the kernel sizes of the DB resulting in lower algorithmic delays. It was found that even at an algorithmic delay as low as 23 ms, at which point the DB is completely disregarded, the performance of the method was still good. When decreasing the algorithmic delay from 398 ms to 23 ms the performance is only reduced by 7% AUC.

## 6. References

- [1] Z.-H. Tan, A. Sarkar, and N. Dehak, "rvad: An unsupervised segment-based robust voice activity detection method," *Computer Speech and Language*, vol. 59, pp. 1–21, 2020. [Online]. Available: <https://github.com/zhenghuatan/rVAD>
- [2] J. Ramírez, J. Gorriz, and J. Segura, *Voice Activity Detection. Fundamentals and Speech Recognition System Robustness*, 06 2007, vol. 6(9).
- [3] C. Yu, K.-H. Hung, I.-F. Lin, S.-W. Fu, Y. Tsao, and J.-w. Hung, "Waveform-based voice activity detection exploiting fully convolutional networks with multi-branched encoders," *arXiv preprint arXiv:2006.11139*, 2020.
- [4] A. Baeovski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," *Advances in Neural Information Processing Systems*, vol. 34, 2021.
- [5] B. Chettri, E. Benetos, and B. L. Sturm, "Dataset artefacts in anti-spoofing systems: a case study on the asvspoof 2017 benchmark," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 3018–3028, 2020.
- [6] P. Hoang, Z.-H. Tan, J. M. De Haan, and J. Jensen, "The minimum overlap-gap algorithm for speech enhancement," *IEEE Access*, vol. 10, pp. 14 698–14 716, 2022.
- [7] H. Dinkel, S. Wang, X. Xu, M. Wu, and K. Yu, "Voice activity detection in the wild: A data-driven approach using teacher-student training," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 1542–1555, 2021.
- [8] C.-C. Hsu, C. Kah Meng, T.-S. CHI, and Y. Tsao, "Robust voice activity detection algorithm based on feature of frequency modulation of harmonics and its dsp implementation," *IEICE Transactions on Information and Systems*, vol. E98.D, pp. 1808–1817, 10 2015.
- [9] T. Kinnunen and P. Rajan, "A practical, self-adaptive voice activity detector for speaker verification with noisy telephone and microphone data," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7229–7233.
- [10] P. Mahalakshmi, "A review on voice activity detection and mel-frequency cepstral coefficients for speaker recognition (trend analysis)," *Asian Journal of Pharmaceutical and Clinical Research*, vol. 9, p. 360, 12 2016.
- [11] S. O. Sadjadi and J. H. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE signal processing letters*, vol. 20, no. 3, p. 197, 2013.
- [12] D. Rho, J. Park, and J. H. Ko, "Nas-vad: Neural architecture search for voice activity detection," *arXiv preprint arXiv:2201.09032*, 2022.
- [13] Y. Lee, J. Min, D. K. Han, and H. Ko, "Spectro-temporal attention-based voice activity detection," *IEEE Signal Processing Letters*, vol. 27, pp. 131–135, 2019.
- [14] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*, vol. 21, no. 4, p. 697, 2013.
- [15] R. Zazo-Candil, T. N. Sainath, G. Simko, and C. Parada, "Feature learning with raw-waveform cldnns for voice activity detection," in *INTERSPEECH*, 2016.
- [16] Y. Shinohara, "Adversarial Multi-Task Learning of Deep Neural Networks for Robust Speech Recognition," in *Proc. Interspeech 2016*, 2016, pp. 2369–2372.
- [17] Z. Meng, J. Li, Y. Gong, and B.-H. F. Juang, "Adversarial feature-mapping for speech enhancement."
- [18] H. Yu, Z.-H. Tan, Z. Ma, and J. Guo, "Adversarial network bottleneck features for noise robust speaker verification," *Proc. Interspeech 2017*, pp. 1492–1496, 2017.
- [19] D. Pearce and H.-G. Hirsch, "The aurora experimental framework for the performance evaluations of speech recognition systems under noisy condition," vol. 4, 01 2000, pp. 29–32.
- [20] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," 2015. [Online]. Available: <https://arxiv.org/abs/1505.07818>
- [21] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems* 32. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: <http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf>
- [22] J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, N. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," *Linguistic Data Consortium*, 11 1992.
- [23] J. Barker, R. Marxer, E. Vincent, and S. Watanabe, "The third 'chime' speech separation and recognition challenge: Dataset, task and baselines," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 504–511.
- [24] M. Kolbæk, Z.-H. Tan, and J. Jensen, "Speech enhancement using long short-term memory based recurrent neural networks for noise robust speaker verification," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 305–311.
- [25] 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, pp. 153–167, 2017.
- [26] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *CoRR*, vol. abs/1502.01852, 2015. [Online]. Available: <http://arxiv.org/abs/1502.01852>
- [27] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *AISTATS*, 2010.