INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Polishing the Classical Likelihood Ratio Test by Supervised Learning for Voice Activity Detection

Tianjiao Xu, Hui Zhang, Xueliang Zhang

Department of Computer Science, Inner Mongolia University, Hohhot, China, 010021

xtj@mail.imu.edu.cn {cszh, cszxl}@imu.edu.cn

Abstract

Voice activity detection (VAD) is essential for speech signal processing system, which desires low computational cost and high real-time processing. Likelihood ratio test (LRT) based VAD is a widely used and effective approach in many applications. However, it is still a challenge in low signal-to-noise ratio (SNR) and non-stationary noisy scenario. To cope with this challenge, we propose a supervised masking-based parameter estimation module with an adaptive threshold to improve the performance of a state-of-the-art LRT based VAD. Moreover, considering real-time processing, we compared the proposed with corresponding end-to-end supervised learning approaches in various model sizes. Experimental results show that the proposed method leads to consistently better performance than both of the existing LRT based method and end-to-end supervised learning based approaches.

Index Terms: voice activity detection, likelihood ratio test, adaptive threshold

1. Introduction

Voice activity detection (VAD) is widely used in many applications, such as speaker recognition, voice wake-up and automatic speech recognition. The task of VAD is to make a determination of speech or non-speech intervals, which poses many challenges due to the non-stationary nature of speech and/or background noises. To cope with these challenges, three types of methods were proposed in literature, including: acoustic features based methods, statistical model based methods and supervised learning based methods.

Conventionally, acoustic features based methods [1,2] were widely used in early researches which identified the difference between speech and noise in feature space, which relied on a small amount of labeled data. To improve the performance of VAD, statistical model based methods have become the mainstream [3–5]. One of the most widely used methods is proposed by *Sohn et al.* [6], i.e. SohnVAD. SohnVAD assumed that the spectrum of the input signal at each frequency bin obeys Gaussian distribution and made VAD decision by Bayes rule called likelihood ratio test (LRT) with a given threshold. One important thing in SohnVAD is to estimate the spectrum of the noise which is done by minimum mean square error (MMSE) [7–9]. To prevent the clipping on weak speech tails, it used an effective hang-over scheme based on hidden Markov model (HMM) which considered both historical and current decisions. However, SohnVAD is particularly weakness in non-stationary noise scenarios, e.g. a sudden rise of noise power may be misinterpreted as a speech onset, changes of noise power during speech presence may bring a certain delay of speech detected [9].

From the perspective of supervised learning, VAD is usually treated as a binary classification problem with pre-marked labels on each frame. The state-of-the-art algorithms are usually based on deep learning [10–12].

In this study, we try to improve a statistical model based method SohnVAD with supervised learning methods. Actually, there are two core problems of SohnVAD, the first is how to estimate the parameters of the likelihood ratio, and the second is how to set the decision threshold. For the first problem, the MMSE estimator does badly at the beginning of the speech signal or a sudden change of noise. For the second problem, it is known that the decision threshold of VAD is usually determined on-the-fly rather than a fixed value [13, 14]. So, an adaptive decision threshold is necessary in VAD algorithms.

With these observations, we propose a method which improves SohnVAD by applying supervised learning to parameter estimation and corresponding adaptive threshold. On the one hand, SohnVAD simplifies the problem of discriminating between speech and noise and it is less data-dependent than end-to-end supervised learning methods. On the other hand, with the help of data-driven supervised learning methods, the poor performance of SohnVAD on non-stationary noisy conditions can be solved, simultaneously. Specifically, we apply a convolutional recurrent network (CRN) to estimate the T-F mask of noisy speech, which has demonstrated effective in speech enhancement tasks [15]. Then the estimated T-F mask is used to calculate the parameters and the decision threshold, which solves the two core problems of SohnVAD, respectively. Furthermore, considering real-time processing, we compared the proposed with corresponding end-to-end supervised learning approaches in different scale of model. Our experiments suggest that the proposed leads to consistently better than existing LRT based and end-to-end supervised learning based approaches.

2. Method

2.1. Overview of SohnVAD

In SohnVAD, by assuming that the clean speech S is corrupted by some additive noise N, the noisy speech Y is defined as:

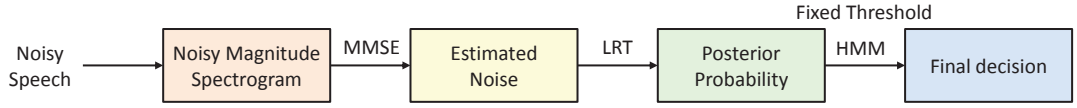
$$Y_{t,f} = S_{t,f} + N_{t,f} \tag{1}$$

where subscript t and f denotes the frame and frequency band index of their short-time Fourier transform (STFT) spectrum, respectively.

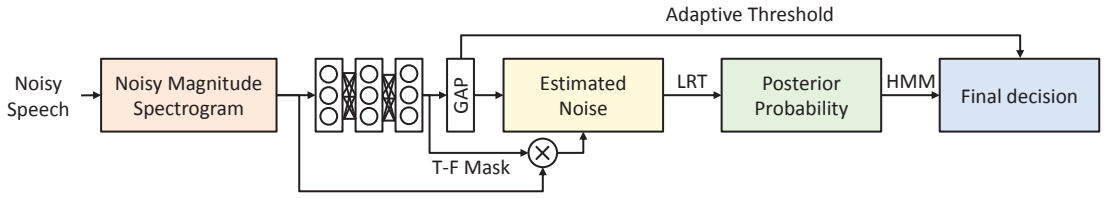
SohnVAD formalizes VAD as a likelihood ratio test (LRT) problem with two hypotheses: H_N for speech absent and H_S for speech present. For a certain frame t, the speech is either present or absent. If the speech is absent, only the noise is

This research was supported in part by the China National Nature Science Foundation (No. 61876214, No. 61866030).

http://dx.doi.org/10.21437/Interspeech.2020-1177



(a) Overview of the SohnVAD



(b) Overview of the proposed method

Figure 1: *Overview of the SohnVAD and the proposed method.*

observed, i.e. $Y_t = N_t$. This situation is denoted as hypothesis H_N. If the speech is present, both of the speech and noise are observed, i.e. $Y_t = S_t + N_t$, which denoted as H_S.

By modeling each frequency band as a single Gaussian distribution, the probability density function of each hypothesis is given:

$$p(Y_{t,f}|H_N) = \frac{1}{\pi\lambda_N(f)} \exp\left\{-\frac{|Y_{t,f}|^2}{\lambda_N(f)}\right\} \tag{2}$$

$$\begin{aligned} p(Y_{t,f}|H_S) =& \frac{1}{\pi[\lambda_N(f) + \lambda_S(f)]} \\ & \cdot \exp\left\{-\frac{|Y_{t,f}|^2}{\lambda_N(f) + \lambda_S(f)}\right\} \end{aligned} \tag{3}$$

where $|\cdot|$ denotes the amplitude. $\lambda_N(f)$ and $\lambda_S(f)$ denote the variances of the $N_{\cdot,f}$ and $S_{\cdot,f}$, respectively. Thus, the likelihood ratio is given by:

$$\Lambda_{t,f} = \frac{p(Y_{t,f}|H_S)}{p(Y_{t,f}|H_N)} = \frac{1}{1+\xi_{t,f}} \exp\left\{\frac{\gamma_{t,f}\xi_{t,f}}{1+\xi_{t,f}}\right\} \tag{4}$$

where $\xi_{t,f}$ and $\gamma_{t,f}$ are the prior and posterior SNRs, which are:

$$\xi_{t,f} \triangleq \lambda_S(f)/\lambda_N(f) \tag{5}$$

$$\gamma_{t,f} \triangleq |Y_{t,f}|^2/\lambda_N(f) \tag{6}$$

Assuming $\lambda_N(f)$ can be estimated, then $\gamma_{t,f}$ is known, since $Y_{t,f}$ is observed. A maximum likelihood (ML) estimator of $\xi_{t,f}$ is given by:

$$\hat{\xi}_{t,f} = \gamma_{t,f} - 1 \tag{7}$$

So estimating the likelihood ratio $\Lambda_{t,f}$ is converted to estimate the variances of noise spectrum $\lambda_N(f)$, which is the first core problem of SohnVAD.

After obtaining the likelihood ratio in frequency bands, the frame-level likehood ratio is written as

$$\mathcal{L}(t) = \log \hat{\Lambda}_t = \frac{1}{F}\sum_{f=1}^{F} \log \hat{\Lambda}_{t,f} \tag{8}$$

where F is the number of frequency bands. Here, we use the log likelihood ratio to prevent underflow caused by multiplication.

Considering the continuity of the speech, $\mathcal{L}(t)$ can be smoothed along the time axis with some filters or time sequence model. SohnVAD uses an HMM-based hang-over scheme to handle it. Then, a sigmoid transformation is performed to limit the range of logarithmic likelihood ratio into the range of $[0, 1]$ and obtain the final decision statistics, and the decision rule is established as

$$\mathcal{F}(\mathcal{L}(t)) = \frac{1}{1+e^{-(\mathcal{L}(t))}} \underset{H_N}{\overset{H_S}{\gtrless}} \eta. \tag{9}$$

where $\mathcal{F}$ denotes sigmoid function, η is the decision threshold. $\mathcal{F}(\mathcal{L}(t)) \underset{H_N}{\overset{H_S}{\gtrless}} \eta$ means we take the H_S hypothesis if $\mathcal{F}(\mathcal{L}(t)) > \eta$ else we take the H_N hypothesis.

There are two problems left over. The first one is how to estimate $\lambda_N(f)$, and the second one is how to set the decision threshold η. As illustrated in Fig. 1 (a), SohnVAD employed MMSE to estimate the noise spectrum [6]. It is well known that MMSE works poorly for non-stationary noise [9]. After obtained the noise spectrum estimation, the log-likelihood ratio is calculated as (8) and smoothed with the HMM-based hang-over scheme. The final decision is made by a fixed threshold η. Using a fixed decision threshold is also not the optimal solution. In this study, we solve both of these problems with a supervised masking-based noise estimation method.

2.2. Proposed Method

To improve the accuracy of the SohnVAD, we replace the MMSE-based noise spectrum estimator by a supervised method. The noise spectrum estimation is obtained from the T-F mask estimated with some supervised model. Specifically, a convolutional recurrent network (CRN) is used to estimate the T-F mask of noisy speech, where we use the ratio of speech and noise as the T-F mask target, i.e. ideal ratio mask (IRM) [15]. IRM is ranged in $[0, 1]$, and can be regarded as the probability of the presence of speech. The used T-F mask is described as:

$$M_{(t,f)} = \sqrt{\frac{|S|^2}{|S|^2 + |N|^2}} \in [0, 1] \tag{10}$$

As shown in Fig. 1 (b), we first apply STFT on noisy speech to get the magnitude spectrogram, and subsequently feed it into a CRN to predict a T-F mask $M_{(t,f)}$, which can be viewed as the probability of speech presence of each T-F unit. The probability of speech absence is $1 - \hat{M}_{(t,f)}$. Finally, it is multiplied to the noisy speech to get the estimated noise.

Table 1: *Detailed configuration of different models.*

Model name	Methods	Model configuration details	Total number of parameters
model64	Pro.	conv(8,8,8,16,16), lstm(64,64), deconv(16,8,8,8,1)	70825
	End-to-end	conv(8,8,8,16,16), lstm(64,64), fc(64,1)	71889
model32	Pro.	conv(4,4,4,8,8), lstm(32,32), deconv(8,4,4,4,1)	17844
	End-to-end	conv(4,4,4,8,8), lstm(32,32), fc(32,1)	18153
model16	Pro.	conv(2,2,2,4,4), lstm(16,16), deconv(4,2,2,2,1)	4617
	End-to-end	conv(2,2,2,4,4), lstm(16,16), fc(16,1)	4917

The final decision statistics of SohnVAD can be interpreted as the SNR of the noisy speech, which are in the range of $[0, 1]$ at (8), similarly. While IRM is SNR of each T-F unit precisely. Therefore, using the estimated IRM $\hat{M}_{(t,f)}$, the performance of proposed can be improved further.

In a certain speech block, the global SNR, i.e. IRM can measure the probability of speech frame. So the global SNR is obtained by the global average pooling (GAP) of the estimated mask, which is given by

$$M = \frac{1}{TF} \sum_{t=1}^{T} \sum_{f=1}^{F} M_{(t,f)} \tag{11}$$

where T is the number of frames, F is the number of frequency bands.

So we use the estimated IRM to determine the decision threshold. As shown in Fig. 1 (b), once the T-F mask is estimated, a global average pooling layer is applied to get the probability of speech, which is subsequently transformed to the threshold η as follows:

$$\eta = \mathcal{F}\left(\log \frac{M}{1-M}\right) \in [0, 1] \tag{12}$$

where $\mathcal{F}$ denotes sigmoid transformation, which is same to equation (9).

3. Experiments

3.1. Experimental Setting

All experiments are conducted on the TIMIT database [16]. TIMIT corpus has a time-aligned word transcription file associated with each utterance, in which the word boundaries were aligned with the phonetic segments in the time axis. We convert it to the ground-truth labels which can correspond with features for each frame.

We randomly selected 2000 clean utterances from the training set, and use the TIMIT core test set as our test utterances. The TIMIT core test set contains 192 utterances, 8 from each of 24 speakers. We concatenate the selected train utterances with some silence segments of random length, which makes the ratio of speech frames account for around 60%. Then these utterances are mixed with a *speech shape noise (SSN)* and 4 other types of noise from the NOISEX-92 dataset [17]: *babble*, *factory*, *destroy engine*, and *destroyer operations room* at SNRs of -5, 0, 5 dB for training. Each noise is divided into two non-overlapping segments for training and testing respectively. To make the sample general and diverse, we intercept noise segments from long noise randomly. Besides these noise, another four types of noise are used for unseen noise test, which includes an unseen factory, buccaneer from NOISEX-92 and bus, street from CHiME-4 dataset [18]. All signal is resampled to 16 kHz before mixing.

To evaluate the performance of the class imbalance problem like VAD, we use the area under the curve (AUC) as the evaluation metric, which is the area under the receiver operating characteristic (ROC) curve [19]. AUC is considered as an overall metric of the performance of VAD rather than the detection accuracy. Higher value means better performance.

3.2. Model Configuration

For the proposed noise estimation model, we use noisy speech magnitude spectrum as input feature, which utilizes the STFT after dividing speech signal into frames using 20 ms Hamming window with 10 ms overlap. So the input of each frame is 161-dimensional. A $\log$ operation is applied to compress the dynamic range and facilitate training.

The proposed approach constructs the noise estimation model based on CRN. The input feature is encoded by 5 layers of 2-D convolutional, which increases the number of channels while reducing the size of the feature map. Then the sequence of feature vectors are modeled by two LSTM layers. Subsequently, the output sequence of the LSTM layers is decoded back to the output feature by 5 layers of 2-D deconvolutional. Moreover, the skip layer connection is applied to connect each encoder layer to the corresponding decoder layer, which can improve the flow of gradients when the network is deep. For all the convolutions and the deconvolutions, the kernel size is 1×3 and the zero-padding is 1×2, which means operating in the time direction but not in the frequency direction. We apply exponential linear units (ELUs) [20] to all convolutional and deconvolutional layers except the output layer.

It is optimized using Adam optimizer [21] with a mini-batch size of 64. We use a constant dropout rate of 0.4 at LSTM layers.

3.3. Comparison Methods

We compared the proposed method with the original SohnVAD and a supervised end-to-end model. Since VAD is a pre-processing task for speech signal processing system. The storage and computational ability of the device limits the amount of trainable parameters of the model. In order to investigate the influence of model size on performance, we set proposed model and corresponding end-to-end model under three model sizes.

The detailed configuration is provided in Tab.1. Limited by space, the network structure will not be displayed, but you can see it in [15]. The model configuration details are specified in $LayerType(UnitNumers)$ format. It can be seen from the Tab.1 that the first seven layers of the proposed model and the end-to-end model are exactly the same. Such setting can reduce the performance changes caused by the model structure, and pay attention to the method itself. In order to investigate

Table 2: *AUC (%) comparison among the proposed and comparison methods on seen and unseen noisy conditions. Bold font indicates the best performance.*

	seen noise				unseen noise			
SNR	-5dB	0dB	5dB	Avg.	-5dB	0dB	5dB	Avg.
End-to-end	89.73	91.71	92.47	91.30	86.95	90.55	92.01	89.84
SohnVAD	56.13	62.83	68.51	62.49	54.73	61.52	68.18	61.48
SohnVAD+Mask	90.74	92.98	92.66	92.13	88.92	92.67	93.57	91.72
SohnVAD+Mask+Adapted	**91.39**	**93.56**	**93.63**	**92.86**	**91.81**	**93.02**	**94.20**	**93.01**

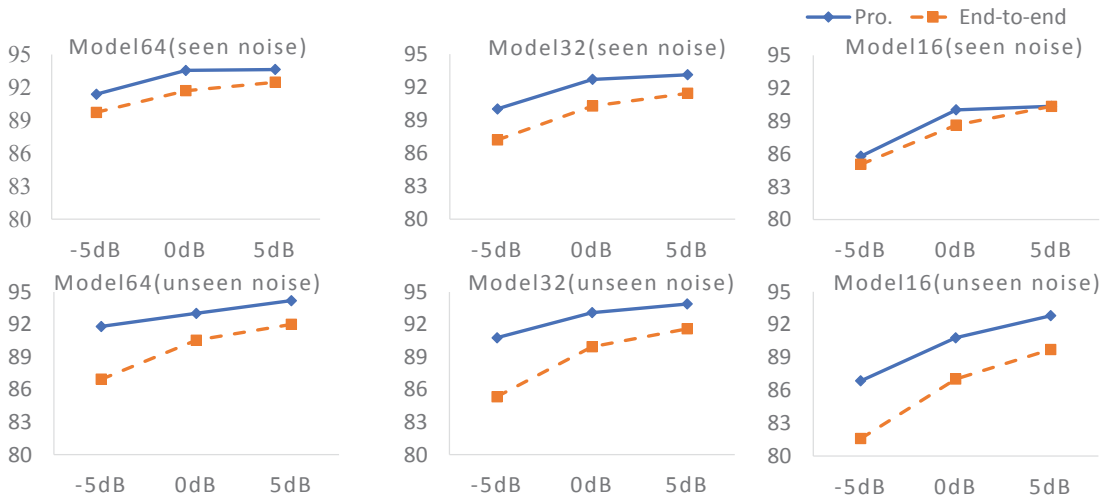


Figure 2: *Performance comparison of models with different sizes on seen and unseen noisy conditions.*

the influence of parameter quantity on model performance, we set up three IRM estimation models and corresponding end-to-end models, which are denoted as model64, model32 and model16, respectively. In these three configurations, the parameter amount of the end-to-end model is always slightly larger than the IRM estimation model.

3.4. Experimental Results

We evaluate the proposed method in two steps. First, we replace the original noise spectrum estimation of SohnVAD with the masking-based method, which is denoted as SohnVAD+Mask. Second, we apply the decision threshold adaptation method, which is denoted as SohnVAD+Mask+Adapted. The results are listed in Tab. 3, where the mask is the ground-truth calculated by equation (10). From Tab. 3, we can see that the proposed two steps can improve the performance compared with the original SohnVAD from 64.34% to 94.79%.

In practice, the mask should be estimated. In Tab. 2, we shows the results of the proposed method using estimated mask which is estimated by model 64 (detailed configuration is shown in Tab. 1) and compare it with the SohnVAD and End-to-end approach under noisy seen and unseen conditions with different SNRs. In general, the performance of supervised approaches is better than the SohnVAD. Specifically, with better noise estimation, SohnVAD+Mask provides over 47.43% and 49.18% relative improvement than SohnVAD under seen and unseen noisy conditions on average. SohnVAD+Mask+Adapted provides over 1.70% and 1.40% relative improvement compared with SohnVAD+Mask under seen and unseen noisy conditions on average, respectively.

Compared with End-to-end approach, we observe that the supervised parameter estimation from the T-F mask is not only beneficial for SohnVAD but also outperforms end-to-end. The relative improvement are 1.56% and 3.52% under seen and unseen noisy conditions on average.

Table 3: *AUC (%) of the baseline and the proposed improving methods with the ideal mask*

SNR	-5dB	0dB	5dB	Avg.
SohnVAD	58.07	64.51	70.44	64.34
SohnVAD+Mask	92.88	92.67	92.37	92.64
SohnVAD+Mask+Adapted	94.37	94.90	95.12	94.79

Fig. 2 shows the performance of the proposed method and the End-to-End approach with different sizes. It can be seen that the performance of both methods decrease when the number of parameters being small. However, the proposed method leads consistently better performance than End-to-end approach under all conditions, particularly for unseen noisy conditions, which shows better generalization ability of the proposed method.

4. Conclusions

In this work, we employ deep learning to improve the performance of the classical LRT-based VAD, SohnVAD, under low-SNR conditions. The proposed approach converts the parameter estimation and threshold adaptation to T-F mask estimation process, in a uniform T-F mask estimation process, simply and effectively. Experimental results show that the proposed algorithm can benefit from both of the conventional signal processing and the deep learning.

5. References

[1] L. R. Rabiner and M. R. Sambur, "An algorithm for determining the endpoints of isolated utterances," *The Bell System Technical Journal*, Feb 1975, vol. 54, no. 2, pp. 297–315.

[2] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the LPC residual domain," *Transactions on Speech and Audio Processing*, March 2001, vol. 9, no. 3, pp. 217–231.

[3] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *Signal Processing Letters*, Jan 1999, vol. 6, no. 1, pp. 1–3.

[4] Joon-Hyuk Chang, Nam Soo Kim, and S. K. Mitra, "Voice activity detection based on multiple statistical models," *Transactions on Signal Processing*, June 2006, vol. 54, no. 6, pp. 1965–1976.

[5] J. Ramirez, J. C. Segura, C. Benitez, L. Garcia, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *Signal Processing Letters*, Oct 2005, vol. 12, no. 10, pp. 689–692.

[6] J. Sohn and W. Sung, "Voice activity detector employing soft decision based noise spectrum adaptation," in *International Conference on Acoustics*, 1998.

[7] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error log-spectral amplitude estimator," *Transactions on Acoustics, Speech, and Signal Processing*, April 1985, vol. 33, no. 2, pp. 443–445.

[8] R. C. Hendriks, R. Heusdens, and J. Jensen, "MMSE based noise psd tracking with low complexity," in *International Conference on Acoustics, Speech and Signal Processing*, March 2010, pp. 4266–4269.

[9] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *Transactions on Audio, Speech, and Language Processing*, May 2012, vol. 20, no. 4, pp. 1383–1393.

[10] X. L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *Transactions on Audio, Speech, and Language Processing*, April 2013, vol. 21, no. 4, pp. 697–710.

[11] X. L. Zhang and D. L. Wang, "Boosted deep neural networks and multi-resolution cochleagram features for voice activity detection," *Interspeech*, 2014, pp. 1534–1538.

[12] L. Mateju, P. Cerva, and J. Zdansky, "Study on the use of deep neural networks for speech activity detection in broadcast recordings," *International Conference on Signal Processing and Multimedia Applications*, 2016, pp. 45–51.

[13] R. M. Debayan Ghosh and S. Gurugopinath, "Robust voice activity detection using frequency domain long-term differential entropy," *Interspeech*, 2018, pp. 1220–1224.

[14] Z. Fan, Z. Bai, X. Zhang, S. Rahardja, and J. Chen, "AUC optimization for deep learning based voice activity detection," in *International Conference on Acoustics, Speech and Signal Processing*, May 2019, pp. 6760–6764.

[15] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement," in *Interspeech*, 2018, pp. 3229–3233.

[16] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren, and V. Zue, "TIMIT acoustic-phonetic continuous speech corpus," in *Linguistic Data Consortium*, 1993.

[17] A. Varga and H. J. M. Steeneken, "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*, 1993, vol. 12, no. 3, pp. 247–251.

[18] E. Vincent, S. Watanabe, A. A. Nugraha, J. Barker, and R. Marxer, "An analysis of environment, microphone and data simulation mismatches in robust speech recognition," *Computer Speech & Language*, 2017, vol. 46, pp. 535–557.

[19] J. A. Hanley and B. J. Mcneil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve." *Radiology*, 1982., vol. 143(1), pp. 29–36

[20] D. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2016.

[21] D. P. Kingma and J.Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.