On training targets for noise-robust voice activity detection

1st Sebastian Braun
Microsoft Research
Redmond, WA, USA
sebastian.braun@microsoft.com

2nd Ivan Tashev
Microsoft Research
Redmond, WA, USA
ivantash@microsoft.com

***Abstract*—The task of voice activity detection (VAD) is an often required module in various speech processing, analysis and classification tasks. While state-of-the-art neural network based VADs can achieve great results, they often exceed computational budgets and real-time operating requirements. In this work, we propose a computationally efficient real-time VAD network that achieves state-of-the-art results on several public real recording datasets. We investigate different training targets for the VAD and show that using the segmental voice-to-noise ratio (VNR) is a better and more noise-robust training target than the clean speech level based VAD. We also show that multi-target training improves the performance further.**

***Index Terms*—voice activity detection, convolutional recurrent neural network, real-time inference**

I. Introduction

Classifiers detecting speech presence in audio signals, widely known as a voice activity detector (VAD), are a mature but still popular research topic. There are countless applications that require or benefit from a VAD, like all kinds of audio processing modules [1], [2] like speech enhancers, localizers, echo controllers, automatic gain controls, speech level or signal-to-noise ratio (SNR) measurement, speech-rate or silence-rate estimation, pre-processing, gating or segmentation for speech recognizers, speech-related classifiers for emotion [3], gender, age, identity, anomaly, toxicity, or speech encoding and transmission [4], and so forth.

Traditional statistical models, widely used in speech enhancement due to their simplicity, often rely only on the non-stationarity of speech [5]–[7]. Better accuracy can be achieved by integrating models of features like pitch, harmonicity [8], [9], modulation [10], spectral shape [11], [12], etc. [13]. Dealing with increasing numbers of features becomes more difficult with human-crafted models, often resulting in diminishing performance gains. Therefore, data-driven approaches, especially neural networks [14]–[16], are an attractive choice and have shown substantial performance boosts [17].

A major design point of VAD methods is the temporal resolution of the classification. While for some applications like speech segmentation, emotion classifiers etc. a larger granularity is sufficient, speech processing applications usually require a VAD in frame-level resolution of typical lengths between 5 - 20 ms. While some deep learning (DL) based VADs designed for the first class of applications use coarser temporal resolutions of 0.2 - 10 s, which additionally can improve robustness and performance, this limits the use as general purpose VAD for the second class of applications. Furthermore, most applications have real-time requirements, and are deployed on computationally and power-constrained edge devices, which poses challenges on the latency and computational efficiency of the algorithms.

In this work, we propose a neural network architecture for real-time VAD on a typical short audio frame basis. The network provides predictions per frame without look-ahead, and is reasonably small and efficient to be executed on standard CPUs of typical battery-powered devices. The main contribution of this work is, however, the investigation of noise robust VAD training targets. We show that training on the clean speech level-based VAD is less robust in low SNR conditions than using the frequency-weighted segmental voice-to-noise ratio (VNR). While the ground truth depending only on the clean speech level is ill-conditioned in very low SNR, where the speech might not even be audible, the VNR label accounts for the audibility of speech in noise. Additionally, we propose a multi-target training, using speech level-based VAD and VNR as joint targets, which improves the performance further. While in [18] a method for frame-level SNR estimation has been proposed, to the best of our knowledge this is the first work to use SNR as training target for VAD.

II. Signal model and training targets

We assume that noisy training data is generated by mixing possibly reverberant speech with noise, resulting in the training mixture signal

$$y(t) = h(t) \star s(t) + v(t), \qquad (1)$$

where $h(t), s(t), v(t)$ are the acoustic impulse response (AIR), non-reverberant speech signal and noise signal, t is the time index, and $\star$ denotes the convolution operator. As we are interested in detecting speech, be it reverberant or not, but do not consider reverberant tails as desired speech information, we define the target speech signal

$$x(t) = h_{\text{win}}(t) \star s(t), \qquad (2)$$

where $h_{\text{win}}(t)$ is a windowed version of the full AIR $h(t)$ removing the late reverberation tail. We chose an exponentially decaying window corresponding to a decay rate of 60 dB$/0.3\ s$ starting from the direct path of the AIR.

ISBN: 978-9-0827-9706-0

The typical VAD training target is defined by the clean speech level threshold T_{level} as

$$VAD(n) = \begin{cases} 1 & \text{if } \|\mathbf{W}_{\text{VAD}}\,\mathbf{x}(n)\|^2 > T_{\text{level}} \\ 0 & \text{if } \|\mathbf{W}_{\text{VAD}}\,\mathbf{x}(n)\|^2 \leq T_{\text{level}} \end{cases}, \tag{3}$$

where the vector $\mathbf{x}(n)$ contains the frequency-domain representation of the target speech signal $x(t)$ at frame n and the diagonal matrix $\mathbf{W}_{\text{VAD}}$ is an optional frequency weighting such as bandpass filters or loudness weighting.

We propose an alternative training target, i. e. the Mel-weighted segmental VNR

$$VNR(n) = 10 \log_{10} \frac{\|\mathbf{W}_{\text{VNR}}\,\mathbf{x}(n)\|^2}{\|\mathbf{W}_{\text{VNR}}\,\mathbf{v}(n)\|^2}, \tag{4}$$

where $\mathbf{v}(n)$ is the frequency-domain representation of the noise signal $v(t)$ and the matrix $\mathbf{W}_{\text{VNR}}$ is an auditory or loudness weighting, e. g. a Mel-filterbank. Speech presence can then be computed similarly as in (3) by comparing the continuous VNR values to a threshold. Note that in contrast to the speech level only dependent VAD, the VNR also takes noise into account, providing additional information to voice presence about the background noise and *the audibility of the speech in noise*. This avoids practically wrong VAD labels (3) for highly negative SNR, i. e. when the speech may not even be audible. Furthermore, the VNR attributes a physical and interpretable meaning to the speech detection threshold. By adjusting the detection threshold for a VNR based VAD, we can detect either any audible speech in noise using a very low threshold, or only target well-audible foreground speech by choosing a positive VNR threshold. The term VNR is used in analogy to VAD, to highlight measurement of voice only, opposed to the general SNR.

III. Proposed single- and multi-target losses

Having defined the two training targets, the binary $VAD(n)$, and the continuous $VNR(n)$ in the previous section, we define the considered loss functions in the following. A natural choice for a binary classification like $VAD(n)$ is the binary cross-entropy (BCE) [19, ch. 2.8.1]

$$\mathcal{L}_{\text{BCE}}(z) = -\frac{1}{N} \sum_n z_n \log(\hat{z}_n) + (1 - \hat{z}_n) \log(z_n), \tag{5}$$

where z_n and $\hat{z}_n$ are labels and predictions, respectively, and N is the number of frames. On the other hand, continuous distributions like $VNR(n)$ are often fitted using the mean-squared error (MSE) or mean absolute error (MAE) [18]. We indeed determined in preliminary experiments that the BCE performed best when training $VAD(n)$, while the MAE loss performed best when training to predict $VNR(n)$.

As multi-target training can be conveniently realized in deep learning and often helps generalization and robustness, we propose to combine $VAD(n)$ and $VNR(n)$ training targets. We explored applying BCE and MAE loss to the training targets, resulting in the two multi-target training loss options

$$\mathcal{L}_{\text{VADVNR,1}} = (1-\alpha)\mathcal{L}_{\text{BCE}}(VAD) + \alpha\mathcal{L}_{\text{MAE}}(VNR) \tag{6}$$

$$\mathcal{L}_{\text{VADVNR,2}} = \mathcal{L}_{\text{BCE}}(VAD) + \mathcal{L}_{\text{BCE}}(VNR), \tag{7}$$

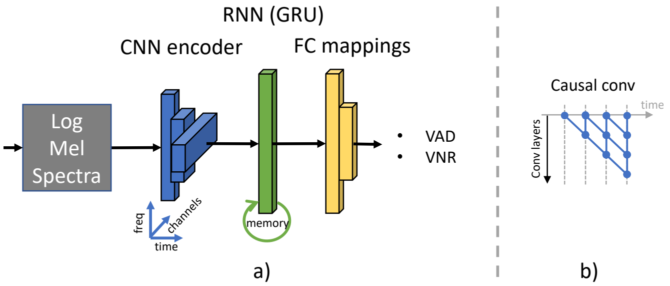


Fig. 1. a) convolutional recurrent network architecture, b) efficient causal convolutions with temporal kernel size 2.

where the weighting factor $\alpha = 0.2$ to balance BCE and MAE was optimized on the validation set. We finally consider four different loss functions:

1) Predict VAD with BCE loss: $\mathcal{L}_{\text{BCE}}(VAD)$
2) Predict VNR with MAE loss: $\mathcal{L}_{\text{MAE}}(VNR)$
3) Predict both VAD, VNR with BCE and MAE: $\mathcal{L}_{\text{VADVNR,1}}$
4) Predict both VAD and VNR with BCE: $\mathcal{L}_{\text{VADVNR,2}}$

IV. Proposed network architecture

Due to the success of convolutional recurrent network (CRN) structures for efficient noise suppression [20], [21], we adapt a similar structure for the VAD classification task. Figure 1a) shows the proposed network architecture. To reduce the input dimensionality, the input features are log-Mel energies. The features are encoded by a few 2D convolutional layers extracting spectro-temporal information. The frequency axis is reduced by a stride of 2 every layer, while the channel dimension is doubled. The convolution over time is causal, meaning no future information is used to infer the current frame as illustrated in Fig. 1b). The kernel size along time axis is only 2, which is very efficient, but still extracts temporal information across $N_{\text{CNN}} + 1$ frames, where N_{CNN} is the number of convolution layers. The output of the convolutional encoder is reshaped to a single vector and fed to a recurrent layer, specifically a gated recurrent unit (GRU) [22]. Finally, two fully connected (FC) layers are used to obtain the desired output of one (single-target case) or two values (multi-target case) per time frame. All convolutional and FC hidden layers use parametric rectified linear unit (PReLU) activations, while the output layer uses a sigmoid to obtain constrained values.

V. Experimental validation

A. Training data

We a use large-scale augmented synthetic training set to ensure generalization to real-world signals. The training set uses 544 h of high mean opinion score (MOS) rated speech recordings from the LibriVox corpus, 247 h noise recordings from Audioset, Freesound, internal noise recordings and 1 h of colored stationary noise. Except for the 65 h internal noise recordings, the data is available publicly as part of the 2nd DNS challenge[1] [23]. Non-reverberant speech files were

[1] https://github.com/microsoft/DNS-Challenge

TABLE I
NEURAL NETWORK LAYER DIMENSIONS.

layer type	hyperparameters	activation
conv2D	1 → 16, (2,3), (1,2), (1,0,1,1)	PReLU
conv2D	16 → 32, (2,3), (1,2), (1,0,1,1)	PReLU
conv2D	32 → 64, (2,3), (1,2), (1,0,1,1)	PReLU
conv2D	64 → 128, (2,3), (1,2), (1,0,1,1)	PReLU
reshape	128×4 → 512	–
GRU	512 → 512	Sigmoid, tanh
FC	512 → 256	PReLU
FC	256 → 1/2	Sigmoid

augmented with acoustic impulse responses randomly drawn from a set of 7000 measured and simulated responses from several public and internal databases. 20% non-reverberant speech is not reverb augmented to represent conditions such as close-talk microphones or professionally recorded speech.

Time shifted speech and noise sequences were mixed with a Gaussian SNR distribution with $\mathcal{N}(5, 10)$ dB and augmented to different microphone signal levels with $\mathcal{N}(-28, 10)$ dBFS. The total training set comprised roughly 1000 h of 10 s mixture-target signal pairs. A more detailed description of the dataset generation can be found in [21]. The training targets $VAD(n)$ and $VNR(n)$ are obtained as described in Section II.

For training monitoring and hyper-parameter tuning, we generated a synthetic validation set in the same way as above, using speech from the DAPS dataset [24], and room impulse responses (RIRs) and noise from the QUT[2] database.

B. Experimental setup

Training targets and features were computed from 32 ms frames with 16 ms frame shift. The VAD weighting $\mathbf{W}_{\text{VAD}}$ was a bandpass filter between $[150, 5000]$ Hz, and we used a signal-dependent threshold $T_{\text{level}} = 0.01 \cdot \max\left[|\mathbf{W}_{\text{VAD}}\,\mathbf{x}(n)|^2\right]$. $\mathbf{W}_{\text{VNR}}$ was a 32 band Mel weighting, and $VNR(n)$ was limited to the range $[-15, 40]$ dB and mapped to the range $[0, 1]$ for training purposes. Both targets, $VAD(n)$ and $VNR(n)$ were temporally smoothed to remove measurement errors and obtain smoother predictions, using a centered moving average window smoothing with length 0.2 s.

The network input features were 64 log-Mel energy bins in the range 0-8 kHz. Table I shows the network layer dimensions. Convolutional layers have parameters $\{$*inChannels* $\rightarrow$ *outChannels, kernelSize, stride, padding*$\}$, where *kernelSize* and *stride* are defined as *(time,frequency)*, and *padding* is defined as *(timeLeft,timeRight,freqLow,freqHigh)*. Reshaping, uni-directional GRU, and FC layers are defined by $\{$*inShape* $\rightarrow$ *outShape*$\}$. The proposed networks are trained using the AdamW optimizer [25] with a learning rate of $5 \cdot 10^{-5}$, weight decay 0.01, batch size of 50 sequences of 10 s length, and adaptive gradient norm clipping [26]. Training is stopped when there is no improvement anymore on the validation set.

C. Metrics, test sets, and post-processing for evaluation

We used the area under curve (AUC) of false positive rate vs. true positive rate as metric for evaluation, training monitoring and tuning. In contrast to the often used precision and recall or F-score, AUC is threshold-independent, and we believe it is therefore more holistic.

As DL methods can only be convincing when generalization on real data can be proven, we use three different public datasets consisting only of real recordings in the wild. We chose the datasets to be from public domains, having been used in prior studies to benchmark VAD methods for comparability, and to be representative for common real-world scenarios and challenging SNR distributions. The three test sets are described in the following. The *KAIST* dataset[3] consists of four 30-min field-recordings in noisy environments (bus stop, construction site, park, room) performed by the authors of [27]. Speech on- and off-sets are human-labelled by the authors. The *HAVIC* database [28] is a collection of 72 h videos of everyday life situations like children playing, kitchens, living rooms, traffic, office, sports events, etc., with human annotations of acoustic events. We categorized all speech-related events as target, while the labels {noise, background noise, music, unintelligible, baby, TV, singing} were considered as non-speech. Although baby and singing are voice, it was excluded as it is not represented in our speech training data. The *AVA speech v1.0* dataset[4] is a 30 h collection of crowd-source labeled YouTube clips (we were only able to obtain 120 of the 160 15-min clips from YouTube). Most of the AVA speech clips seem to originate from movies, so this dataset should be considered as a biased subsample of possible scenarios.

As all test sets are human-labeled, this might involve annotation errors, and exhibit coarse temporal granularity (pauses between words and sentences are usually still labeled as speech), and temporal label uncertainty. This might interfere with our fine-grained frame-wise VAD predictions, that will detect short pauses between words. To mitigate these test label inaccuracies, we apply a post-processing to the frame-wise predictions $VAD(n)$, $VNR(n)$ by smoothing the predictions with a fixed window of 0.4 s length. The window involves no look-ahead, computing the 90-th percentile within the window, in order to detect speech, even when only a fraction of the frame within the 0.4 s window contains speech activity. This post-processing achieved up to 6% relative AUC improvement on the given datasets, while still being real-time compatible.

D. Results

We compare our models to a state-of-the-art baseline ACAM [27] employing a recursive temporal attention module and features with look-ahead of 190 ms. Table II shows the AUC results for baseline and proposed models. We can see that our best model achieves comparable performance on KAIST and better performance on the diverse and noisy HAVIC dataset, while being real-time without look-ahead, and using a

[2] https://research.qut.edu.au/saivt/databases/qut-noise-databases-and-protocols/

[3] https://github.com/jtkim-kaist/VAD

[4] http://research.google.com/ava/index.html

TABLE II
AUC RESULTS (%) ON TEST SETS.

target	loss	output	KAIST	HAVIC	AVA
ACAM [27]			**99.2**	73.4	n. a.
VAD (3)	BCE (5)	VAD	95.9	86.8	92.4
VNR (4)	MAE	VNR	98.8	88.1	92.3
VAD,VNR	BCE,MAE (6)	VNR	98.8	**88.3**	**92.6**
VAD,VNR	BCE,MAE (6)	VAD	95.9	86.5	92.0
VAD,VNR	BCE (7)	VNR	**99.0**	**88.9**	**92.4**
VAD,VNR	BCE (7)	VAD	93.9	86.1	91.4

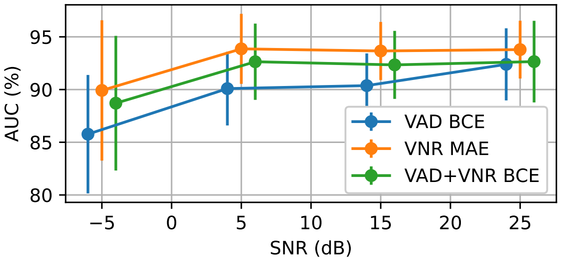
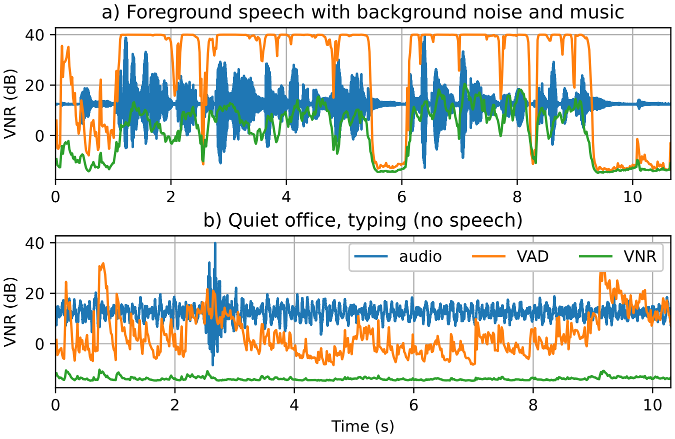


Fig. 2. AUC (mean, std. dev.) on validation set for training targets over SNR.

less complex model during inference time. The ONNX model of our network processes 1 s audio in 7 ms on a laptop CPU at 2.5 GHz, while for ACAM [27] reported 10 ms without, and over 200 ms with feature extraction on a GPU workstation.

The first and second colums in Table II indicates the training targets and loss functions of the proposed networks. As described in Section III, we trained four different networks, two single-target and two multi-target predictors. Since the multi-target networks provide two outputs per frame, they are evaluated in two ways, either using the VAD or the VNR output. While the multi-target training improves performance on the test sets, we did not find useful improvements by combining the VAD and VNR outputs of the model, compared to only using VNR output. The two top results per dataset are printed boldface. We can observe that predicting VNR yields significantly better results than VAD on KAIST and HAVIC, while the difference on AVA is minor. The VNR output of the multi-target predictors show consistently equal or better results than the single-target predictors, while the VAD output of the multi-target networks performs still worse than the VNR single target. This illustrates both the advantage predicting VNR over VAD, and also the slight gain by multi-target training. Note that the label errors of the HAVIC dataset prevent achieving high results.

Further analysis of AUC on the validation set grouped by the mixing SNR shown in Fig. 2 reveals that indeed all training targets perform similarly at high SNR, while VNR and VAD+VNR targets provide better performance at medium to low SNR. This proves the robustness to noise of using VNR as training target compared to the common clean speech level based activity. Interestingly, on the validation set the VNR only (orange) loss shows slightly better results than the multi-target loss (green line) in Fig. 2, while all three test sets in Table II show a slight advantage of the multi-target training. We still can conclude that the multi-target loss helps generalization due to better results on three real recording test sets in contrast to the synthetically mixed validation set, which the networks also were optimized on, and is therefore indirectly seen data.


a) Foreground speech with background noise and music

b) Quiet office, typing (no speech)

Fig. 3. Outputs of the proposed model trained to output VAD and VNR with BCE loss for a) an easy scenario with prominent foreground speech, b) a recording without active speech.

The two outputs of the best proposed model, multi-target prediction with BCE loss, is shown in Fig. 3 for two test recordings. The waveform is shown in blue, the VAD output in orange, and the VNR in green. Waveform and VAD (range [0,1]) are scaled to the VNR dB y-axis for illustration purposes. For the easy scenario in Fig. 3a), both outputs provide good indications of active and non-active speech frames. Both predictors show low values even in short pauses between words. However, a recording in a quiet office, containing only slight ambient noise and keyboard typing in Fig. 3b) reveals some problems of the classification VAD output. The output is more noisy, and the VAD predictions become quite large for some non-speech acoustic events. On the other hand, the VNR prediction is consistently low, hovering around -10 dB, indicating very unlikely speech activity at any time. It is worthwhile mentioning that the lowest equal error rate (false alarm equals miss rate) on the validation set is achieved with a VNR threshold of -7 dB.

VI. CONCLUSIONS

We have proposed an efficient real-time neural network based VAD that achieves state-of-the-art results on challenging real recordings. We showed that the segmental VNR is a more noise robust training target for VAD than the clean speech level based activity, and can be further improved by combining both targets for multi-target training. The proposed network is flexible for most applications, as the frame-level based decisions can be converted to coarser granularity by simple post-processing. The efficient network design allows straightforward integration in various speech processing tasks and different implementation platforms, as inference time on a modern CPU is around 7 ms per second of audio without any runtime optimizations.

REFERENCES

[1] E. Hänsler and G. Schmidt, *Acoustic Echo and Noise Control: A pracical Approach*. New Jersey, USA: Wiley, 2004.

[2] I. Tashev, *Sound Capture and Processing: Practical Approaches*. Wiley, July 2009.

[3] B. Schuller, A. Batliner, S. Steidl, and D. Seppi, "Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge," *Speech Communication*, vol. 53, no. 9, pp. 1062 – 1087, 2011, sensing Emotion and Affect - Facing Realism in Speech Processing.

[4] T. Bäckström, *Speech Coding with Code-Excited Linear Prediction*, 1st ed. Springer, 2017.

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

[6] T. Gerkmann and R. C. Hendriks, "Unbiased MMSE-based noise power estimation with low complexity and low tracking delay," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 4, pp. 1383 –1393, May 2012.

[7] R. Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics," *IEEE Trans. Speech Audio Process.*, vol. 9, pp. 504–512, Jul. 2001.

[8] N. Dhananjaya and B. Yegnanarayana, "Voiced/nonvoiced detection based on robustness of voiced epochs," *IEEE Signal Processing Letters*, vol. 17, no. 3, pp. 273–276, 2010.

[9] L. N. Tan, B. J. Borgstrom, and A. Alwan, "Voice activity detection using harmonic frequency components in likelihood ratio test," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2010, pp. 4466–4469.

[10] C.-C. Hsu, K.-M. Cheong, T.-S. Chi, and Y. Taso, "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, no. 10, pp. 1808–1817, 2015.

[11] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, 2013.

[12] I. Yoo, H. Lim, and D. Yook, "Formant-based robust voice activity detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 12, pp. 2238–2245, 2015.

[13] S. Graf, T. Herbig, M. Buck, and G. Schmidt, "Features for voice activity detection: a comparative analysis," *EURASIP Journal on Applied Signal Processing*, vol. 91, 2015.

[14] F. Eyben, F. Weninger, S. Squartini, and B. Schuller, "Real-life voice activity detection with LSTM recurrent neural networks and an application to hollywood movies," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, May 2013, pp. 483–487.

[15] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7378–7382.

[16] F. Vesperini, P. Vecchiotti, E. Principi, S. Squartini, and F. Piazza, "Deep neural networks for multi-room voice activity detection: Advancements and comparative evaluation," in *International Joint Conference on Neural Networks (IJCNN)*, 2016, pp. 3391–3398.

[17] I. Tashev and S. Mirsamadi, "DNN-based causal voice activity detector," in *Information Theory and Applications Workshop*. University of California - San Diego, February 2016.

[18] H. Li, D. Wang, X. Zhang, and G. Gao, "Frame-level signal-to-noise ratio estimation using deep learning," in *Proc. Interspeech*, 2020.

[19] K. Murphy, *Machine learning - a probabilistic perspective*. MIT Press, 2012.

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

[21] S. Braun, H. Gamper, C. K. A. Reddy, and I. Tashev, "Towards efficient models for real-time deep noise suppression," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[22] K. Cho, B. V. Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Proc. 8th Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST-8)*, 2014.

[23] C. K. A. Reddy, H. Dubey, V. Gopal, R. Cutler, S. Braun, H. Gamper, R. Aichner, and S. Srinivasan, "ICASSP 2021 deep noise suppression challenge," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[24] G. J. Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech?—a dataset, insights, and challenges," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1006–1010, 2015.

[25] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[26] P. Seetharaman, G. Wichern, B. Pardo, and J. L. Roux, "AutoClip: Adaptive gradient clipping for source separation networks," in *30th International Workshop on Machine Learning for Signal Processing (MLSP)*, 2020, pp. 1–6.

[27] J. Kim and M. Hahn, "Voice activity detection using an adaptive context attention model," *IEEE Signal Processing Letters*, vol. 25, no. 8, pp. 1181–1185, 2018.

[28] S. Strassel, A. Morris, J. Fiscus, C. Caruso, H. Lee, P. Over, J. Fiumara, B. Shaw, B. Antonishek, and M. Michel, "Creating HAVIC: Heterogeneous Audio Visual Internet Collection," in *8th International Conference on Language Resources and Evaluation (LREC)*, 2012.