Statistical and Neural Network Based Speech Activity Detection in Non-Stationary Acoustic Environments

Jens Heitkaemper, Joerg Schmalenstroeer, Reinhold Haeb-Umbach

Paderborn University, Department of Communications Engineering, Paderborn, Germany

{heitkaemper, schmalen, haeb}@nt.upb.de

Abstract

Speech activity detection (SAD), which often rests on the fact that the noise is "more" stationary than speech, is particularly challenging in non-stationary environments, because the time variance of the acoustic scene makes it difficult to discriminate speech from noise. We propose two approaches to SAD, where one is based on statistical signal processing, while the other utilizes neural networks. The former employs sophisticated signal processing to track the noise and speech energies and is meant to support the case for a resource efficient, unsupervised signal processing approach. The latter introduces a recurrent network layer that operates on short segments of the input speech to do temporal smoothing in the presence of non-stationary noise. The systems are tested on the Fearless Steps challenge database, which consists of the transmission data from the Apollo-11 space mission. The statistical SAD achieves comparable detection performance to earlier proposed neural network based SADs, while the neural network based approach leads to a decision cost function of $1.07\,\%$ on the evaluation set of the 2020 Fearless Steps Challenge, which sets a new state of the art.

Index Terms: voice activity detection, speech activity detection, neural network, statistical speech processing

1. Introduction

Speech activity detection (SAD) is an integral part of many speech processing pipelines. For example, it is used to define speech on/offsets for diarization [1, 2], to reduce the computational effort of speech recognition systems by specifying the temporal regions for automatic speech recognition (ASR) decoding [3, 4], or to support noise power estimation in speech enhancement algorithms [5]. Indeed, SAD has been the focus of research efforts for years [6, 7, 8, 9].

Traditionally, SAD is formulated as a statistical hypothesis test employing probabilistic models, such as Gaussians, mixtures of Gaussians, or Laplacian distributions [6, 10, 11, 12]. During the last decade, however, deep neural networks (DNNs) have achieved impressive results on some of the more taxing SAD tasks, outperforming the traditional approaches [8, 13, 14]. Here, SAD is formulated as a supervised learning problem by presenting the speech signal at the network's input and the class labels (speech / no speech) as training targets at the output.

One of these challenging tasks is the SAD on the transmission data from the Apollo-11 mission which is one of the objectives of the Fearless Steps challenge [15]. The signals are degraded due to high channel noise, system noise, attenuated signal bandwidth, analog tape ageing, etc.. Furthermore, the noise conditions and signal-to-noise ratios change rapidly over time and channels. Many characteristics of these signals can also be observed in analog speech transmission over High Frequency (HF) radio bands.

In the 2019 edition of the challenge, the top performing neural network-based system achieved a $66.4\,\%$ improvement over the baseline system [14]. The latter consisted of two Gaussian mixture models (GMMs) applied to the one-dimensional principal component analysis (PCA) of a concatenation of multiple high dimensional noise robust features [12]. The neural network, on the contrary, consisted of several convolutional neural network (CNN) layers with subsequent recursive neural network (RNN) layers to exploit temporal information, and employed majority voting on the output of multiple networks. Additionally, a post filter was used to smooth unwanted oscillations in the network's decisions over time.

One should note, however, that the baseline GMM system is rather simplistic. If one used more sophisticated signal processing techniques one should be able to come closer to the performance of the neural network while still requiring considerably less computational and memory resources. To show that this is indeed possible, we present a statistical SAD, which combines multi-layer minimum statistics-based noise estimation and Wiener filter-based enhancement, followed by an energy-based SAD. We show that the presented statistical SAD achieves competitive results compared to those published previously on the Fearless Steps dataset [16, 17], thereby closing the gap between statistical [12, 16] and neural network based SAD [14, 17].

Additionally, we propose an improved neural network based SAD which achieves even better results. It is a CNN-based system inspired by the latest advances in sound event detection [18]. The network topology is similar to the one in [14]. But unlike that system, we introduce a segment RNN which conducts temporal smoothing inside the network rather than by a postfilter. The segment RNN operates on a fixed segmentation of the input signal to control the context observed by the RNN. This is different from other layer types like the hierarchical multiscale RNN [19] and the RNN-based approach presented in [20], which have to learn the segmentation in addition to the segment labeling.

Furthermore, the network calculates multiple predictions per time frame, which are subsequently aggregated. However, instead of using different temporal context lengths as in [21], we aggregate information from different segments with overlapping input frames after the RNN layer for an automatic smoothing. We show that the segment RNN outperforms all previously published results on the fearless dataset w.r.t. the decision cost function (DCF) measure.

The remainder of this paper is structured as follows. In Section 2 the statistical SAD, and in Section 3 the neural network-based SAD system is introduced. Section 4 includes an evaluation of the two systems on the Fearless Steps challenge dataset.

2. Statistical SAD

The proposed statistical SAD is following the idea of [22] to conduct a two-stage processing: In the first stage denoising is
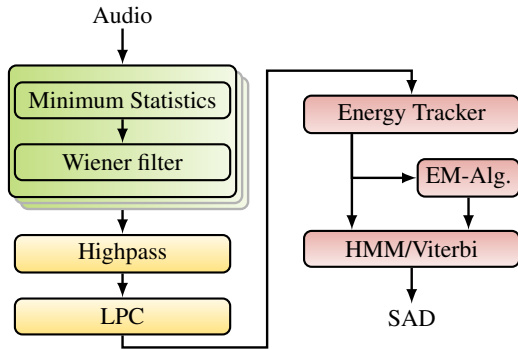
Figure 1: *Overview on statistical SAD components and signal processing queue.*

Figure 2: *Input signal (left) and processed signal (right).*

carried out, and in the second stage a time domain energy criterion is applied to decide on speech activity. Although this SAD works reliably for signals with medium ($\geq 5\,\mathrm{dB}$) signal to noise ratios (SNRs) and stationary noise conditions, it fails for the highly non-stationary noise of the Fearless dataset for multiple reasons: The recordings exhibit a large variety of different and changing noise types, and they have much lower SNR values. Additionally, the signal magnitude of the active speakers varies by several orders of magnitude, even in the same recording, making it difficult to decide whether the noise floor or a speaker with low energy signals is observed.

Our approach is depicted in fig. 1: It consists of repeated application of a denoising stage (fig. 1, green blocks), highpass and linear predictive coding (LPC) filtering (fig. 1, yellow blocks), and the statistical SAD (fig. 1, orange blocks). Each block is described in the following.

As input to the statistical SAD we choose the short time Fourier transform (STFT)-coefficients of the observed signal $X_{t,f}$ with $t \in [0, T]$ as the frame index and $f \in [0, F]$ as the frequency bin index. From these coefficients, the minimum statistics based noise power spectral density (PSD) estimate $\overline{|V(t,f)|^2}$, and $|X(t,f)|^2$, i.e., the PSD estimate of the current analysis window, are calculated and the corresponding Wiener filter $W(t,f)$ is given by:

$$W(t,f) = \max\left(1 - \gamma\frac{\overline{|V(t,f)|^2}}{|X(t,f)|^2}, G_{\min}\right) \qquad (1)$$

The oversubtraction factor γ is chosen relatively high, $\gamma > 20$, to compensate for the bias of underestimating the noise level via minimum statistics and to force the Wiener Filter to aggressively apply noise reduction. Furthermore, $W(t,f)$ is lower bounded by $G_{\min}$ to prevent $W(t,f)$ from becoming negative in case of low noise levels co-occuring with speech absence in the same bin. Noise tracking and Wiener filtering iterate multiple times over the audio signal, decreasing the noise level in each iteration and at the same time keeping the maximum peaks corresponding to speech untouched. Thereby, the SNR is improved with each stage, however, at the cost of deteriorating the audio quality. Since neither a low number of acoustic artifacts (e.g., musical tones), nor superior speech quality are of interest here, this loss is acceptable.

After the denoising stage, a linear highpass filter is applied to remove low frequency noise. Furthermore, a simple 1st-order LPC filter is employed to enhance the well predictable speech signals and to suppress the unpredictable noise.

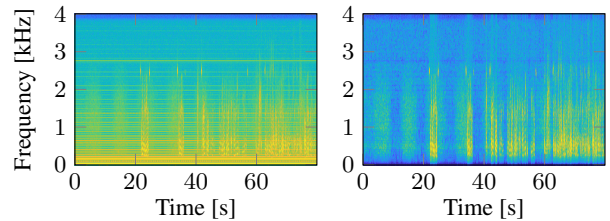From the enhanced audio signal the energy per sub-band is calculated, where each sub-band has a bandwidth of $1\,\mathrm{kHz}$.

Temporal smoothing with an averaging window of size $0.48\,\mathrm{s}$ reduces the sub-band energy variance. Afterwards, the smoothed sub-band energies are weighted with an exponential decay factor ($1/s$ for the sth sub-band) and accumulated to a single value per frame, called combined sub-band energy (CSBE)(t).

2.1. Adaptive threshold

The CSBE values are the basic information source for deciding whether an active speaker or just noise is observed. However, finding an optimal decision threshold is a nontrivial task. Furthermore, the resulting threshold may be dependent on the dataset. In order to circumvent a fixed threshold, minimum statistics is applied once more. This time, the CSBE values are tracked to find the floor values (F-CSBE), which belong to the non-speech parts of the recordings. Additionally, the mean of all F-CSBE values for each recording can be calculated to get the average CSBE floor value (A-CSBE), i.e., an estimate for the average noise level of the recording. A frame is marked as speech if CSBE exceeds the sum of F-CSBE and A-CSBE by a certain factor.

The described minimum tracking approach delivers an initial estimate of speech activity, but it shows a high sensitivity towards the chosen threshold. To overcome this issue, a statistical model in the log-domain is established. As depicted in fig. 3, the logarithm of the CSBE values is taken and all values smaller than the noise threshold are considered to be caused by noise and thus used to estimate a GMM representing noise components. Similarly, values larger than the chosen speech threshold are considered speech and used to estimate a GMM for speech. The thresholds for speech and noise are derived from the A-CSBE value, adding some additional safety margins.

The final decision of the statistical SAD is derived with a Viterbi decoder operating on an hidden Markov model (HMM) which consists of 5 consecutive states for noise and 5 states for speech, each with probability of 0.9 for staying in the state and 0.1 for state switching. The HMM emission probabilities are given by the aforementioned GMMs.

3. Neural Network-Based SAD

The neural network architecture is adapted from [18] and consists of multiple CNN blocks with subsequent layers for temporal
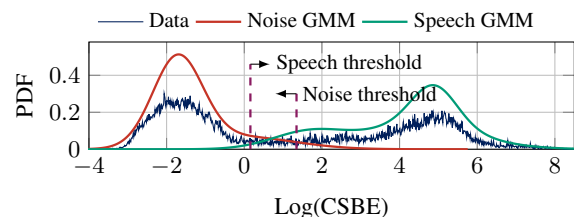

Figure 3: *Histogram of logarithmic CSBE values with noise and speech GMM models.*
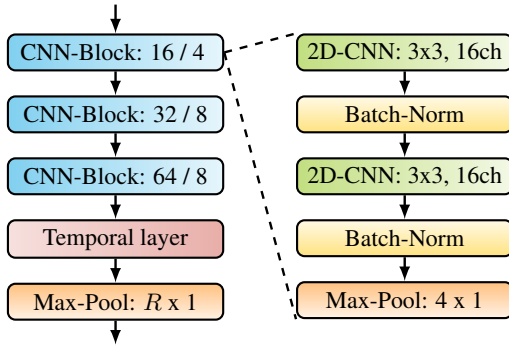
Figure 4: *Block diagram of the DNN model used for SAD, where R symbolizes the output size of the temporal layer.*

smoothing, an output activation and max pooling over the feature dimension. Each CNN block consist of two 2D-CNN layers with a 3x3 kernel and a subsequent batch normalization, followed by a single max pooling layer of stride 4 which operates along the feature dimension. In contrast to [18], no pooling along the time dimension is applied to ensure frame-wise outputs. To exploit temporal information, we either use a 1D-CNN structure as in [18] or an RNN layer with a bidirectional gated recurrent unit (GRU) [23] and a subsequent feed-forward (FF) layer as classifier to enable the network to use a larger context. The described network structure is shown in fig. 4 and table 1 compares the two options for gathering temporal information.

To enforce temporal smoothing, the common approach is to use Viterbi decoding on an HMM with GMMs emission distributions to model speech and noise statistics, as described in the previous section. However, the smoothing can also be done by the DNN or, more specifically, by the temporal layer. Therefore, we propose a segment RNN as replacement for the temporal layer where each input utterance is segmented into M overlapping chunks of length L with a shift S between segments as outlined in fig. 5. Each segment is processed by an RNN with subsequent classifier layer as specified in table 1. The parameters are shared between the layers to ensure that all segments are processed equally. For each segment only the last output frame is chosen as the prediction for the whole segment. Speech activity is assumed for a segment i if the estimated output y_i exceeds a certain threshold α: $d_i = \begin{cases} 1 \text{ if } y_i > \alpha \\ 0 \text{ if } y_i \leq \alpha \end{cases}$. Finally, speech presence is declared for a given frame if at least one segment containing that frame indicates speech presence. Thereby, the occurrence of oscillations in the decision signal is reduced at the cost of overestimating the speech activity. Since the segment length L and shift S are fixed, the segment RNN approach may lead to a higher hit rate while also increasing the false alarm rate. However, in most applications a high true positive rate (TPR) is more important than a low false positive rate (FPR). For example, in case of ASR an overestimation of the length of speech activity is not as harmful as missing part of an utterance. The shift S

Table 1: *Comparison of two possible layer structures for the temporal layer where FF represents a feed forward layer.*

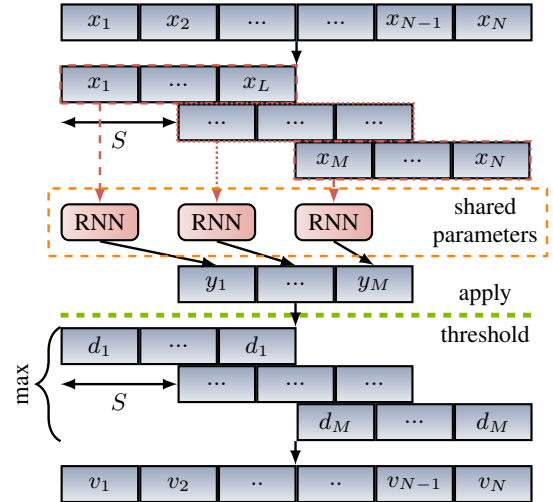| Layer type	#Layer	Params	Classifier
1D-CNN	2	3x1 Kernel / (128,10)ch	–
RNN	1	BI-GRU: 64x256	FF: 256x10

Figure 5: *Example of the segment RNN*

allows for a complexity reduction since a larger value reduces the number of segments and thus the number of chunks to be processed by the RNN. However, increasing S also reduces the overlap between segments and thereby the number of segments contributing to the activity estimation for each frame. Changes in the segment length L, control the temporal context seen by the RNN layer.

For training, the 30 min streams are divided randomly into 4 s segments which are independently sent through the network, This prevents overfitting since it ensures that all possible speech-silence-ratios are observed during training. As cost function the binary cross entropy is chosen.

4. Evaluation

The presented SAD systems are tuned on the development set of the fearless dataset [15]. As metrics we use precision (P), recall (R), F_1-score$= 2\frac{P \cdot R}{P+R}$ and DCF with

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}}, \qquad (2) \qquad R = \frac{\text{TP}}{\text{TP} + \text{FN}}, \qquad (3)$$

$$\text{DCF} = 0.75 \cdot \frac{\text{FN}}{\text{TP} + \text{FN}} + 0.25 \cdot \frac{\text{FP}}{\text{TN} + \text{FP}}, \qquad (4)$$

where TP, FP, TN and FN are the number of true positive, false positive, true negative and false negative predictions. The scoring is done with the openSAD evaluation tool [24].

4.1. Fearless Steps Dataset

The Fearless Steps dataset [15] consists of 8 kHz recordings from the Apollo 11 mission. The part of the Fearless Steps dataset used during this challenge for training and development consists of 290 speakers with an SNR between 0 dB and 20 dB. Note, that the development set includes 34 unique speakers not seen during training. All examples are 30 min long, and the training set consists of 29.56 % of speech on average, whereas the development set includes an average of 32.87 % of speech. The main challenges of the dataset are, first, that speech activity is typically very short, consisting of one or two words only, and, second, that the noise is highly non-stationary with varying SNR. If not stated otherwise, all experiments are carried out on the development set.
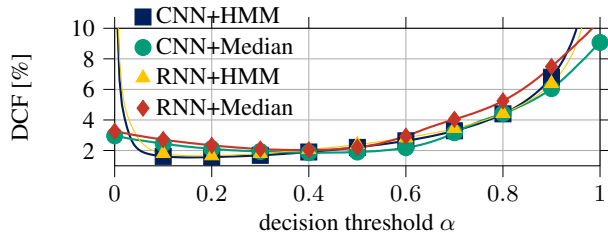
Figure 6: *Comparison of the CNN and RNN temporal layer with median filter or HMM-based smoothing*

4.2. DNN tuning

The input to the DNN is the magnitude spectrum obtained from an STFT with an FFT size of 512 samples, a window length of $50\,\mathrm{ms}$ and a frame shift of $10\,\mathrm{ms}$. All networks were trained for 50000 iterations with a batch size of 24 and a learning rate of 0.001 using the Adam optimizer [25].

In fig. 6, the DCF is displayed as a function of the decision threshold α for the first two temporal layer variants described in section 3. According to the figure, RNN and the 1D-CNN achieve similar results with a small edge for the 1D-CNN. In both cases, the threshold can be chosen in a fairly large range without a substantial impact on the performance.

Additionally, two types of temporal smoothing are compared: On one hand, HMM-based smoothing described in section 2 and, on the other hand, a median filter with a fixed window length. It can be observed that the HMM-based approach outperforms the median filter for both the RNN and CNN layer.

In fig. 7, the DCF values achieved with the segment RNN as temporal layer are plotted for different shifts S and lengths L. For each S and L the individual optimal threshold α is chosen. The results are compared to the previously best presented system, the CNN+HMM. It is apparent that the segment RNN clearly outperforms the CNN+HMM SAD estimation for short segment length L. All results with $L \leq 250\,\mathrm{ms}$ and $S \leq 150\,\mathrm{ms}$ achieve at least a small gain compared to the CNN temporal layer with HMM smoothing. Reducing the shift and length in the segment RNN to $10\,\mathrm{ms}$ and $50\,\mathrm{ms}$ which equals 1 and 5 frames respectively, results in the lowest DCF. A possible explanation is the non-stationarity of the distortions and that the high overlap between segments due to the small shift successfully smooths the network output.

4.3. System comparison

In table 2 the results for all systems tuned to their optimal threshold are shown. The presented statistical SAD achieves a DCF of $2.98\,\%$ and thus an improvement of $9.52\,\%$ over the challenge baseline in terms of absolute numbers. In comparison, the
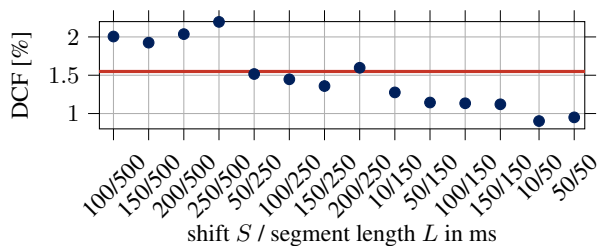

Figure 7: *DCF results for the segment RNN with different shifts S and segment length L for the respective optimal threshold. The red line symbolizes the DCF value for the CNN+HMM system.*

Table 2: *Results for all presented systems for different metrics in % on the Dev and Eval set for the Fearless Steps SAD Challenge.*

| System | DEV | | | | EVAL |
	F1	P	R	DCF	DCF
Stat. SAD	94.32	90.94	97.97	2.98	4.60
CNN + HMM	97.91	96.83	99.02	1.42	2.45
RNN + HMM	97.84	96.85	98.85	1.54	2.17
Seg. RNN	98.62	97.42	99.85	0.81	1.19
Baseline	–	–	–	12.50	13.60

neural network-based approaches achieve a DCF of $1.42\,\%$ and $1.54\,\%$ with a CNN and RNN as temporal layer, respectively. However, the segment RNN temporal layer outperforms all other approaches achieving a DCF of $0.81\,\%$. The high precision P indicates that a small segment length L and shift S allows to increase the hit rate without causing an higher false alarm rate. The results achieved on the evaluation dataset are similar to the ones observed on the development data for both the DNN-based and statistical SAD. Indicating, that the systems are robust to small changes in the data. Using majority voting on the output of all neural networks in table 2 we get a DCF of $1.07\,\%$ which is the best submitted SAD result during the 2020 Fearless Steps Challenge but only a slight improvement over the single model with a segment RNN temporal layer.

The two proposed approaches differ in many aspects, e.g. their type of signal processing or their implementations. In table 3 a comparison between the systems in terms of processing time is stated. Please note that the table shows results for systems which are implemented on different tool chains, and that further optimizations on the code may improve the realtime factors. The table shows that all systems allow real time processing.

Although the statistical SAD achieves worse detection rates than the DNN-based method presented here, the approach is interesting because it has a lower real-time factor and because it is an unsupervised learning approach not requiring labeled training data. Arguably, this makes it easier to adapt the system to other data sets.

5. Conclusions

This paper proposes a new statistical SAD which achieves competitive results compared to other DNN-based systems. Furthermore, a new DNN-based SAD with a segment RNN-based smoothing is presented which allows to define the context observed by the RNN. This single system approach achieves a DCF value of $1.19\,\%$ which is the second best result submitted to the 2020 Fearless Steps Challenge and is only slightly outperformed by a majority voting between a combination of the proposed DNN architectures (DCF: $1.07\,\%$). Hereby, the majority voting results are the best submitted to the 2020 Fearless Steps Challenge, outperforming the baseline by more than $10\,\%$ in absolute values.

6. Acknowledgement

Table 3: *Realtime factor of SAD on an [Intel®Xeon®CPU E3-1240 v6 @ 3.70GHz, 8GB RAM].*

System	Stat. SAD	CNN + HMM	Seg. RNN
Realtime factor	0.004 72	0.0119	0.0445
Tool chain	C++/Matlab	Pytorch [26]	Pytorch [26]

7. References

[1] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 2, pp. 356–370, 2012.

[2] R. Mertens, H. Po-Sen, L. Gottlieb, G. Friedland, A. Divakaran, and M. Hasegawa-Johnson, "On the applicability of speaker diarization to audio indexing of non-speech and mixed non-speech/speech video soundtracks," *IJMDEM*, pp. 1–19, 2012.

[3] S. H. Dumpala and K. N. R. K. R. Alluri, "An algorithm for detection of breath sounds in spontaneous speech with application to speaker recognition," in *Speech and Computer*, 2017, pp. 98–108.

[4] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on YouTube using deep neural networks," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013.

[5] R. Martin and I. Cohen, *Single-Channel Speech Presence Probability Estimation and Noise Tracking*. John Wiley & Sons, Ltd, 2018, ch. 6, pp. 87–106.

[6] Jongseo Sohn, Nam Soo Kim, and Wonyong Sung, "A statistical model-based voice activity detection," *IEEE Signal Processing Letters*, vol. 6, no. 1, pp. 1–3, Jan 1999.

[7] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *Trans. Audio, Speech and Lang. Proc.*, vol. 16, no. 5, p. 910–919, 2008.

[8] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7378–7382.

[9] B. Liu, Z. Wang, S. Guo, H. Yu, Y. Gong, J. Yang, and L. Shi, "An energy-efficient voice activity detector using deep neural networks and approximate computing," *Microelectronics Journal*, vol. 87, pp. 12 – 21, 2019.

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

[11] M. Graciarena, A. Alwan, D. Ellis, H. Franco, L. Ferrer, J. H. L. Hansen, A. Janin, B. suk Lee, Y. Lei, V. Mitra, N. Morgan, S. O. Sadjadi, T. Tsai, N. Scheffer, L. N. Tan, and B. Williams, "All for one: Feature combination for highly channel-degraded speech activity detection," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013, pp. 709–713.

[12] A. Ziaei, L. Kaushik, A. Sangwan, J. Hansen, and D. Oard, "Speech Activity Detection for NASA Apollo Space Missions: Challenges and Solutions," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2014.

[13] G. Saon, S. Thomas, H. Soltau, S. Ganapathy, and B. Kingsbury, "The IBM speech activity detection system for the DARPA RATS program," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 3497–3501, 2013.

[14] A. Vafeiadis, E. Fanioudakis, I. Potamitis, K. Votis, D. Giakoumis, D. Tzovaras, L. Chen, and R. Hamzaoui, "Two-dimensional convolutional recurrent neural networks for speech activity detection," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 2045–2049.

[15] J. Hansen, A. Sangwan, A. Joglekar, A. Bulut, L. Kaushik, and C. yu, "Fearless steps: Apollo-11 corpus advancements for speech technologies from earth to the moon," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2018, pp. 2758–2762.

[16] L. Kaushik, A. Sangwan, and J. H. L. Hansen, "Speech activity detection in naturalistic audio environments: Fearless steps apollo corpus," *IEEE Signal Processing Letters*, vol. 25, no. 9, pp. 1290–1294, Sep. 2018.

[17] B. Sharma, R. Das, and H. Li, "Multi-level adaptive speech activity detector for speech in naturalistic environments," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019, pp. 2015–2019.

[18] J. Ebbers, L. Drude, A. Brendel, W. Kellermann, and H.-U. Reinhold, "Weakly supervised sound activity detection and event classification in acoustic sensor networks," in *IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, Guadeloupe, West Indies, 2019.

[19] J. Chung, S. Ahn, and Y. Bengio, "Hierarchical multiscale recurrent neural networks," *CoRR*, vol. abs/1609.01704, 2016.

[20] L. Kong, C. Dyer, and N. Smith, "Segmental recurrent neural networks," *CoRR*, vol. abs/1511.06018, 11 2015.

[21] X. Zhang and D. Wang, "Boosting contextual information for deep neural network based voice activity detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 2, pp. 252–264, 2016.

[22] E. Warsitz, R. Haeb-Umbach, and J. Schmalenstroeer, "Zweistufige sprache/pause-detektion in stark gestoerter umgebung," in *33. Deutsche Jahrestagung fuer Akustik (DAGA 2007)*, 2007.

[23] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *CoRR*, vol. abs/1406.1078, 2014.

[24] NIST U.S. Department of Commerce, *NIST Open Speech-Activity-Detection Evaluation*, 2016 (accessed April 9, 2020). [Online]. Available: https://www.nist.gov/itl/iad/mig/nist-open-speech-activity-detection-evaluation

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[26] 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.