IberSPEECH 2021
24-25 March 2021, Valladolid, Spain

Convolutional Recurrent Neural Networks for Speech Activity Detection in Naturalistic Audio from Apollo Missions

Pablo Gimeno, Dayana Ribas, Alfonso Ortega, Antonio Miguel, Eduardo Lleida

ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain
{pablogj, dribas, ortega, amiguel, lleida}@unizar.es

Abstract

Speech Activity Detection (SAD) aims to correctly distinguish audio segments containing human speech. Several solutions have been successfully applied to the SAD task, with deep learning approaches being specially relevant nowadays. This paper describes a SAD solution based on Convolutional Recurrent Neural Networks (CRNN) presented as the ViVoLab submission to the 2020 Fearless steps challenge. The dataset used comes from the audio of Apollo space missions, presenting a challenging domain with strong degradation and several transmission noises. First, we explore the performance of 1D and 2D convolutional processing stages. Then we propose a novel architecture that executes the fusion of two convolutional feature maps by combining the information captured with 1D and 2D filters. Obtained results largely outperform the baseline provided by the organisation. They were able to achieve a detection cost function below 2% on the development set for all configurations. Best results were reported on the presented fusion architecture, with a DCF metric of 1.78% on the evaluation set and ranking fourth among all the participant teams in the challenge SAD task.

Index Terms: speech activity detection, convolutional recurrent neural networks, Fearless steps challenge, naturalistic audio

1. Introduction

Speech activity detection (SAD) aims to determine whether an audio signal contains speech or not, and its exact location in the signal. This constitutes an essential preprocessing step in several speech-related applications such as speech and speaker recognition, as well as speech enhancement. In many cases, the SAD is used as a preliminary block to separate the segments of the signal that contain speech from those that are only noise. This way, enabling the overall system to, for instance, performing speaker recognition only on speech segments.

A large number of approaches have been proposed for the SAD task. Starting with unsupervised approaches, some examples can be cited: based on energy [1], or based on the estimation of the signal long-term spectral divergence [2]. Traditionally, statistical approaches have been used with relevant results under the assumption of quasi-stationary noise. Several works rely on the extraction of specific acoustic features [3] [4]. Conversely, other methods are model-based [5] [6], aiming to estimate a statistical model for the noisy signal. Recently, deep learning approaches are becoming more and more relevant in the SAD task. The research presented in [7] implements a SAD system based on a multilayer perceptron with energy efficiency as the main concern. A deep neural network approach is used in [8] to perform SAD in a multi-room environment. In [9], new optimisation techniques based on the area under the ROC curve are explored in the framework of a deep learning SAD system.

Recurrent Neural Networks (RNN) are specially relevant in order to deal with temporal sequences of information. The Long Short Term Memory (LSTM) networks [10] are a kind of RNN that introduces the concept of the memory cell in order to learn, retain, and forget information in long dependencies. Some research has already proposed the use of LSTM networks to solve the SAD task. Authors in [11] presented an LSTM network to classify speech and non-speech segments in a noisy speech from Hollywood movies. A similar system is used in [12] to implement a noise-robust vowel based SAD. In this context, we have been able to obtain competitive results in the framework of diarisation tasks [13] [14] based on the properties of the Bidirectional LSTM (BiLSTM) classifier.

Convolutional Recurrent Neural Networks (CRNN) combine the capability of convolutional networks to capture frequency and time dependencies simultaneously seeking to extract discriminative features, and the capability of recurrent networks to deal with temporal series. Several examples of the use of CRNN in audio processing can be found in the literature [15] [16] [17]. Recently, CRNN have been proposed in the SAD task with relevant results. The approach presented in [18], based on the use of 2D convolutional layers, ranked first among all submissions in the 2019 Fearless steps challenge SAD task[1].

In this paper, we present our submission to the SAD task proposed for the Fearless steps challenge 2020. We introduce a supervised deep learning solution based on a CRNN that is fed with Mel filterbank energies as input. We explore alternatives for the convolutional layers, namely 1D and 2D filters. Then, we present a novel approach based on the fusion of two convolutional layers that combines the information of 1D and 2D filters to be processed by the RNN.

The remainder of the paper is organised as follows: a brief description of the Fearless steps challenge is given in section 2. Our CRNN based system proposal is described in section 3. The experimental setup for the challenge is introduced in section 4. Section 5 presents and discusses the results obtained. Finally, a summary and the conclusions are presented in section 6.

2. Fearless Steps challenge

The Fearless steps initiative has resulted in the digitisation of the original analog recordings from the Apollo space missions. Part of these data has been made available through the Fearless steps corpus, consisting of a cumulative 19,000 hours of conversational speech coming from the Apollo 11 mission [20]. Audio data belongs to 30 different communication channels, with multiple speakers in different locations. Most channels show a strong degradation with transmission noise or noise due to tape ageing. Furthermore, the signal-to-noise ratio (SNR) has a

[1] Results are no longer available online, but a summary of the best submissions can be found in [19]

10.21437/IberSPEECH.2021-6

strong variance, with levels ranging from 0 to 20 dB.

Aiming to motivate the research effort on this challenging domain, a series of annual challenges is being held proposing different speech related tasks. The inaugural Fearless steps challenge [19] took place in 2019, proposing the SAD task among other 4 different tasks. The focus on this first challenge was made on the development of unsupervised or semi-supervised systems. Only 20 hours of in-domain manually transcribed audio were available for the participants to use.

This new version of the challenge released in 2020 [21] changes its focus to the development of supervised systems, releasing around 80 hours of human labelled data through the training and development datasets. The SAD task is proposed again among other 5 different tasks. The fact that a larger amount of in-domain annotated data is available in this version opens a new possibility for supervised approaches such as the one proposed by this paper. Note that the use of out-of-domain data in these specific conditions, namely naturalistic audio and strongly degraded channels, could lead to poor results.

3. Proposed SAD system

3.1. Feature extraction

As input features for our proposed SAD system, we consider log Mel-filter bank energies. Namely, we use 64 log Mel-coefficients concatenated with the log energy of the frame. Note that as the input audio is sampled at f_s = 8 kHz, Mel filters span across the frequency range between 64 Hz and $f_s/2$. Features are computed every 10ms using a 25 ms Hamming window. As a final step, the mean and variance at feature level are used to normalise the corresponding file. All the alternatives developed to the SAD system proposal share the same set of features.

3.2. Neural architectures

In our submission to 2020 Fearless steps challenge for the SAD task we experimented with different neural architectures. In the following lines we briefly describe each of them.

As our baseline model, we choose a solution that is inspired by the SAD system proposed in our previous work in the diarisation framework [13]. It consists of an RNN block generated by stacking three BiLSTM layers with 128 neurons each. This block is then followed by a linear layer that generates the speech class score as a single neuron output.

The following architectures proposed are built on top of the RNN block from the baseline system, incorporating a set of convolutional layers working as a processing stage previous to the RNN block. The schematic representation of the proposed alternatives for the CRNN model is described in Figure 1. Note that the RNN block followed by a linear layer is shared by the three architectures. Then, the difference comes from the convolutional stage, that is implemented in three different ways:

- **Architecture A**: This model uses three 2D convolutional blocks processing the input features. Each of these blocks is integrated by a 2D convolutional layer with 3x3 or 5x5 kernel size and 64 filters. Then it is followed by a batch normalisation [22] and the application of a rectified linear unit (ReLU) [23] activation function. Finally, a max-pooling mechanism is applied considering a 4x1 stride, so that only the frequency axis is downsampled.
- **Architecture B**: This model similarly uses three 1D convolutional blocks. Even though, in this case, we experiment with different variations for the 1D convolutional layer. The first approach uses a kernel size of 3 in all the convolutions with no dilation. In the second implementation, each of the three layers uses kernel sizes of 5, 3 and 3 with dilations 1, 2 and 4 respectively. For the third approach, we experiment with the concept of group convolution, which has recently demonstrated its effectiveness in models such as ResNeXt [24]. This alternative employs a kernel size of 3 in all the convolutions but, in this case, these are implemented as 5 independent groups. This change does not affect the dimension of the input and the output feature maps but it reduces computational complexity and the number of model parameters. Finally, to obtain a comparable representation to the 2D setup, the convolutional layers have 256 output filters and a max-pooling mechanism is applied on the frequency axis using a 4x1 stride after the batch normalisation layer and a ReLU activation function.
- **Architecture C**: Some previous work has already shown the combined capabilities of 1D and 2D convolutions applying system level fusion techniques [25]. This alternative proposes a novel approach to the CRNN model in the SAD task where we aim to combine the information extracted by two different convolutional branches. One consisting of three Conv2D blocks and other consisting of three Conv1D blocks, both implemented as described in the previous architectures. This fusion is done in an intermediate feature space, where both branches are then combined to be processed by the RNN block. The fusion block (depicted in blue) could be implemented in many different ways. In our experiments we test three different options: 1) a bilinear layer combining both convolutional branches, 2) the sum of the output of both branches, and 3) the concatenation of the output of both branches.

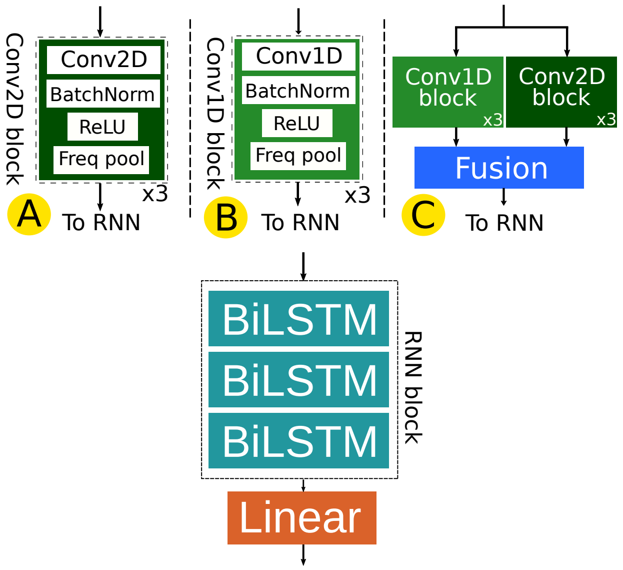

Figure 1: *Schematic representation of the different variations on the proposed convolutional recurrent neural network used for the SAD task.*

A common characteristic among all the models evaluated in this paper is that training and evaluation are performed using finite-length sequences. The input audio is separated in overlapping fragments of 3 second length and 2.5 second advance in order to limit the delay of the dependencies that the network may take into account. The final prediction is generated by taking the first half of the overlapped part from the previous window, and the second half from the next window. This way the labels corresponding to the boundaries of each fragment are discarded as they may not be reliable. It must be noted too that, in all cases, the neural networks emit a SAD label for each frame processed at the input, one each 10ms in this case.

4. Experimental setup

4.1. Data description

The Fearless steps challenge follows open training conditions. Participants can use any available data in addition to the provided challenge data to train and tune their systems. However, in this work we have not used any additional datasets. Considering the specific domain of the audio, namely quite degraded channels and several kinds of transmission noises, we opted to use for training and development only the labelled data provided by the organisation. These data consist of 3 different partitions. In the following lines, we describe them and explain how they have been used in our submission:

- **Train**: Training subset is made of 125 files of around 30-minutes duration each. This makes a total of around 62.5 hours of audio. In our experiments we used 10% of these data for training validation. This way, all the proposed systems were trained with around 56 hours of audio from the train partition.
- **Development**: There are available 30 files of 30 minutes length for development purposes, resulting in around 15 hours of audio. This subset was used to obtain an empirical threshold, in order to minimise the detection cost function (DCF) metric. We also report our results on this subset.
- **Evaluation**: There are available 40 files of 30 minutes, which become 20 hours of audio for evaluation. We report our results on this subset as provided by the challenge organisation. The DCF metric obtained in the evaluation subset is the one used to rank the participants.

4.2. Training strategies

Models in this work are trained using Adam optimiser [26] with a learning rate that decays exponentially from 10^{-3} to 10^{-4} during the 20 epochs that data is presented to the neural network, with a minibatch size of 64. Cross entropy criterion is chosen as loss function, as usually done in classification tasks. Model selection is done choosing the best performing model in terms of frame classification accuracy using the validation subset. All the models in this paper have been developed using the PyTorch toolkit [27].

4.3. Evaluation metric

Two different errors can be considered when dealing with a SAD system: a false positive (FP), this is the identification of speech in a segment where the reference identifies non-speech, and a false negative (FN), this is the missed identification of speech in a segment where the reference identifies speech. With these two errors, we can define the probability of a false positive and the probability of a false negative according to the following equations:

$$P_{FP} = \frac{\mathrm{T}_{FP}}{\mathrm{T}_{\text{ref non-speech}}}, \qquad P_{FN} = \frac{\mathrm{T}_{FN}}{\mathrm{T}_{\text{ref speech}}}, \qquad (1, 2)$$

Table 1: *SAD results in terms of DCF metric on the development and evaluation partition, and number of trainable parameters for different systems considered for submission.*

System	# Param	DCF(%)	
		Dev	Eval
Organisation baseline [28]	-	12.50	13.60
RNN baseline	266K	2.02	2.54
A1 - CRNN 2D (3x3)	340K	1.65	2.07
A2 - CRNN 2D (5x5)	473K	1.67	2.28
B1 - CRNN 1D	421K	1.76	2.33
B2 - CRNN 1D dilation	455K	1.86	2.30
B3 - CRNN 1D groups	300K	1.76	2.46
C1 - CRNN fusion bilinear	641K	1.46	1.78
C2 - CRNN fusion sum	377K	1.60	1.89
C3 - CRNN fusion concat	411K	1.43	1.82

where T_{FP} and T_{FN} are, respectively, the total false positive time and total false negative time , $\mathrm{T}_{\text{ref non-speech}}$ represents the total annotated non-speech time in the reference, and $\mathrm{T}_{\text{ref speech}}$ represents the total annotated speech time in the reference.

In the SAD task of the Fearless steps challenge false negative errors are considered more important than false positive errors. This is shown in the primary evaluation metric for the challenge, the DCF, which is calculated as follows:

$$\text{DCF}(\theta) = 0.75 P_{FN}(\theta) + 0.25 P_{FP}(\theta)\,, \qquad (3)$$

where P_{FN} is the probability for a false negative and P_{FP} is the probability for a false positive. Participants are responsible to choose a threshold (θ) that minimises the DCF.

5. Results

Table 1 presents the obtained results for the different systems submitted. We compare our RNN baseline system and the three proposed architectures to the baseline provided by the organisation [28]. Note that the organisation's baseline is based on a statistical approach. Concerning the fusion architectures, they use the best configurations achieved with the development set: A1 for the 2D setup, and B3 for the 1D setup, as it obtains similar results to B1 with a significantly smaller number of parameters. Results are reported in terms of DCF metric for both, development and evaluation partitions. To measure the level of complexity of the models, we present the number of trainable parameters for all submissions.

Regarding the results reported on the development partition, all our presented systems significantly outperform the baseline algorithm proposed by the organisation. Furthermore, all the systems that include a convolutional processing stage improve the performance compared to the RNN baseline. Our experimental findings are in line with the ones presented in [18], where 2D CRNN models provided better performance than 1D CRNN based SAD systems. In our case, using the 3x3 filter configuration we were able to obtain a DCF of 1.65%, which is better than all the 1D based systems evaluated. For the 1D convolution setup, it is interesting to mention the configuration using groups. A significant relative improvement of 12.77% compared to the RNN baseline is obtained being the CRNN model with the lowest number of parameters. These experimental results indicate that the combination of 1D and 2D feature maps is beneficial for our SAD system. Best overall results are obtained

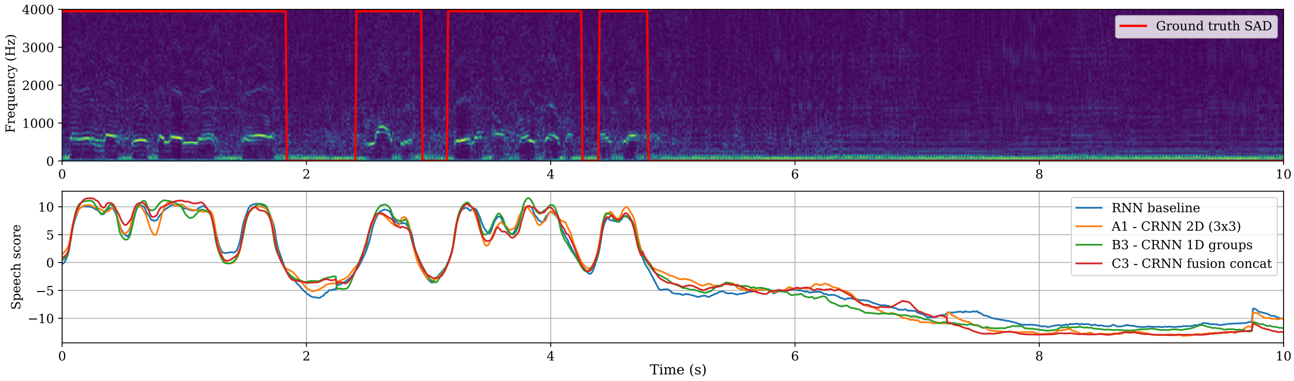

Figure 2: *Qualitative visualisation of SAD scores for the model alternatives described in the paper in a 10 seconds audio fragment extracted from file "`FS02_dev_001`". From top to bottom: audio spectrogram with the SAD ground truth overlapped in red colour and speech score for different SAD systems proposed.*

using the proposed fusion architecture. The most relevant result is the one that concatenates both convolutional outputs, obtaining the best result in the development set while keeping a lower number of parameters when compared to the fusion using a bilinear layer. With this setup we were able to achieve a competitive result for the development set with a 1.43% DCF metric.

Concerning the evaluation partition results, similar trends to the ones described in the development partition can be observed. In general terms, the behaviour for the three different kinds of architectures is consistent. Again, the 2D convolution setup outperforms its 1D equivalent. However, a difference can be observed in the 1D setup between development and evaluation results. While the groups configuration is the best performing in the development set, in the evaluation set this is done by the dilation setup. The fusion architectures show the best overall results, with a DCF metric below 2% for all the variations proposed. This solution allows our best submission to achieve a DCF metric of 1.78% using a bilinear layer as fusion method. This result was ranked in seventh position among the 28 challenge submissions to the SAD task, and fourth among the 7 participant teams[2]. Note that, unlike it was observed in the development set, the fusion based on concatenation offers a slight degradation in performance compared to the bilinear fusion, while keeping the number of parameters significantly smaller.

Additionally, Figure 2 presents a qualitative visualisation of the SAD performance for the best performing architectures in the development partition. It can be observed that, as it was expected, a high positive value in the neural network speech score is associated with a strong evidence of speech in the audio signal. In general terms, we can see that all the systems shown in Figure 2 can accurately capture the speech and non-speech segments in the audio fragment with the empirical threshold minimising the DCF being $\theta = -2$. Anyway, some inconsistencies can be observed between the ground truth and the speech scores on some points. This is probably due to labelling conditions, where a few non-speech segments in between two speech segments are labelled as speech. Proposed systems are able to capture this effect by showing a local minimum in the speech score for the mentioned fragments (see the first two seconds).

Focusing on the individual performance of the proposed systems, we can observe that the 2D and fusion systems show a lower score when a long fragment of non-speech is processed. On the other hand, the system based on 1D tends to output a higher score for speech fragments. In the case of transitions, no significant difference is observed among the systems presented, indicating a similar response between speech and non-speech fragments and vice-versa. It must be noted that all the systems presented in this paper achieve competitive results without introducing post-processing or smoothing techniques on the neural network output. As it can be observed from the depicted scores of Figure 2, the BiLSTM layers are able to impose a certain amount of inertia on the output so that the speech class score is smooth enough to be used by itself.

6. Conclusions

In this paper, we presented the ViVoLab submission to the SAD task of the Fearless Steps Challenge 2020. In this Challenge, we processed audio with degraded channels and several kinds of transmission noises from Apollo space missions. For our submission, we explored different CRNN models using 1D and 2D filters in the convolutional layers. We proposed a novel architecture that combines information coming from 1D and 2D filters in an intermediate feature space, which then is processed by the recurrent neural network. Obtained results largely outperform the baseline provided by the Challenge organisation. Our experimental achievements are in line with previous publications where 2D convolutions obtained better performance than equivalent 1D convolutions. Additionally, we showed that the combination of the information provided by 1D and 2D filters is beneficial for the SAD system, performing with the best results in the development and evaluation sets. Our best submission achieved a DCF metric of 1.46% and 1.78% respectively in the development and evaluation sets, ranking seventh among the 28 submissions to the challenge SAD task, and fourth among the 7 participant teams.

7. Acknowledgements

This work has been supported by the Spanish Ministry of Economy and Competitiveness and the European Social Fund through the project TIN2017-85854-C4-1-R, Government of Aragón (Reference Group T36_20R) and co-financed with Feder 2014-2020 "Building Europe from Aragón".

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

[2] `https://fearless-steps.github.io/ChallengePhase2/Final.html`

8. References

[1] K.-H. Woo, T.-Y. Yang, K.-J. Park, and C. Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electronics Letters*, vol. 36, no. 2, pp. 180–181, 2000.

[2] J. Ramırez, J. C. Segura, C. Benıtez, A. De La Torre, and A. Rubio, "Efficient voice activity detection algorithms using long-term speech information," *Speech communication*, vol. 42, no. 3-4, pp. 271–287, 2004.

[3] S. V. Gerven and F. Xie, "A comparative study of speech detection methods," in *Fifth European Conference on Speech Communication and Technology*, 1997.

[4] J.-C. Junqua, B. Mak, and B. Reaves, "A robust algorithm for word boundary detection in the presence of noise," *IEEE Transactions on speech and audio processing*, vol. 2, no. 3, pp. 406–412, 1994.

[5] J.-H. Chang, N. S. 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.

[6] T. Ng, B. Zhang, L. Nguyen, S. Matsoukas, X. Zhou, N. Mesgarani, K. Veselỳ, and P. Matějka, "Developing a speech activity detection system for the DARPA RATS program," in *Proc. Interspeech*, 2012, pp. 1969–1972.

[7] 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.

[8] 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 *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2016, pp. 3391–3398.

[9] Z.-C. Fan, Z. Bai, X.-L. Zhang, S. Rahardja, and J. Chen, "AUC optimization for deep learning based voice activity detection," in *Proc. IEEE ICASSP*, 2019, pp. 6760–6764.

[10] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] 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 ICASSP*, 2013, pp. 483–487.

[12] J. Kim, J. Kim, S. Lee, J. Park, and M. Hahn, "Vowel based voice activity detection with LSTM recurrent neural network," in *Proc. 8th International Conference on Signal Processing Systems*, 2016, pp. 134–137.

[13] I. Viñals, P. Gimeno, A. Ortega, A. Miguel, and E. Lleida, "Estimation of the number of speakers with variational bayesian PLDA in the DIHARD diarization challenge," in *Proc. Interspeech*, 2018, pp. 2803–2807.

[14] ——, "In-domain adaptation solutions for the RTVE 2018 diarization challenge," in *Proc. Iberspeech*, 2018, pp. 220–223.

[15] T. N. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *Proc. IEEE ICASSP*, 2015, pp. 4580–4584.

[16] F. Vesperini, L. Romeo, E. Principi, A. Monteriù, and S. Squartini, "Convolutional recurrent neural networks and acoustic data augmentation for snore detection," in *Neural Approaches to Dynamics of Signal Exchanges*. Springer, 2020, pp. 35–46.

[17] X. Huang, L. Qiao, W. Yu, J. Li, and Y. Ma, "End-to-end sequence labeling via convolutional recurrent neural network with a connectionist temporal classification layer," *International Journal of Computational Intelligence Systems*, pp. 341–351, 2020.

[18] 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," *Proc. Interspeech 2019*, pp. 2045–2049, 2019.

[19] J. H. Hansen, A. Joglekar, M. C. Shekhar, V. Kothapally, C. Yu, L. Kaushik, and A. Sangwan, "The 2019 Inaugural Fearless Steps Challenge: A Giant Leap for Naturalistic Audio," in *Proc. Interspeech*, 2019, pp. 1851–1855.

[20] J. H. Hansen, A. Sangwan, A. Joglekar, A. E. Bulut, L. Kaushik, and C. Yu, "Fearless steps: Apollo-11 corpus advancements for speech technologies from earth to the moon," in *Proc. Interspeech*, 2018, pp. 2758–2762.

[21] A. Joglekar, J. H. Hansen, M. C. Shekar, and A. Sangwan, "FEARLESS STEPS challenge (FS-2): Supervised learning with massive naturalistic apollo data," *Proc. Interspeech 2020*, pp. 2617–2621, 2020.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv preprint arXiv:1502.03167*, 2015.

[23] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[24] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[25] H. Zeinali, L. Burget, and H. Cernocky, "Acoustic scene classification using fusion of attentive convolutional neural networks for DCASE2019 challenge," DCASE2019 Challenge, Tech. Rep., 2019.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[27] A. Paszke, S. Gross, F. Massa, A. Lerer *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, 2019, pp. 8024–8035.

[28] A. Ziaei, L. Kaushik, A. Sangwan, J. H. Hansen, and D. W. Oard, "Speech activity detection for NASA apollo space missions: Challenges and solutions," in *Proc. Interspeech*, 2014, pp. 1544–1548.