# Unsupervised Representation Learning for Speech Activity Detection in the Fearless Steps Challenge 2021

Pablo Gimeno, Alfonso Ortega, Antonio Miguel, Eduardo Lleida

# Unsupervised Representation Learning for Speech Activity Detection in the Fearless Steps Challenge 2021

*Pablo Gimeno, Alfonso Ortega, Antonio Miguel, Eduardo Lleida*

ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain

{pablogj, ortega, amiguel, lleida}@unizar.es

## Abstract

In this paper, we describe the ViVoLab speech activity detection (SAD) system submitted to the Fearless steps Challenge - phase III. This series of challenges have proposed a number of speech processing task dealing with audio from Apollo space missions over the last few years. The focus in this edition is set on the generalisation capabilities of the systems, with new evaluation data from different channels. Our proposed submission is based on the use of the unsupervised representation learning paradigm, seeking to obtain a new and more discriminative audio representation than traditional perceptual features such as log Mel-filterbank energies. These new features are used to train different variations of a convolutional recurrent neural network (CRNN). Experimental results show that features learned via unsupervised learning provide a much more robust representation, significantly reducing the mismatch observed between development and evaluation partition results. Obtained results largely outperform the organisation baseline, achieving a DCF metric of 2.98% on the evaluation set and ranking third among all the participant teams.

**Index Terms**: unsupervised representation learning, speech activity detection, fearless steps challenge

## 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.

While supervised learning algorithms have been essential for the development of deep learning applications, labelled data is not always easy to obtain. This problem is becoming more and more relevant as models grow faster in size and computational requirements. In this context, unsupervised learning solutions [10] have emerged in order to alleviate the need for labels. These methods expose the model to huge amounts of data, with the objective of understanding the data source, by learning to make predictions related to it. This idea was already presented for discrete sources such as text, forcing the network to predict the next items [11]. When dealing with real valued signals, the idea was initially approached by minimising the reconstruction of the signal [12]. Some evolution on this idea were proposed, such as the reconstructions of missing or corrupted fragments [13]. However, greater gains have been obtained by constructing pretext tasks, where the objective of the system is to solve a prediction as a classification. In many works, this objective is to select an unseen fragment of the signal among other randomly selected distractor fragments [10]. This self supervised mechanism was successfully implemented for large scale tasks with good results for image [14] and speech recognition [15] [16].

In this paper, we present ViVoLab submission to the SAD task proposed by the Fearless Steps challenge 2021. We build our work upon the neural architectures evaluated in previous editions of this challenge [17], where convolutional recurrent neural networks (CRNNs) yielded competitive results. Our focus in this new edition is set on the input features fed to the neural network. We propose the introduction of the unsupervised learning paradigm to obtain a new representation of audio signals more discriminative than traditional perceptual features such as log Mel-filterbank energies, seeking to improve the performance of our SAD system using thousand of hours of unlabelled audio.

The remainder of the paper is organised as follows: a brief description of the Fearless steps challenge and its context is given in section 2. Our proposed system for the SAD task, our approach for unsupervised feature learning and the SAD neural networks are described in section 3. The experimental setup for the challenge, stating the data and metrics considered, 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 [18]. 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

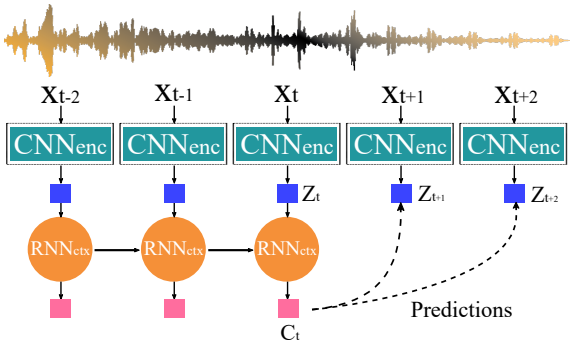


Figure 1: *Schematic overview of the system used for unsupervised representation learning.*

tape ageing. Some channels even have the presence of babble noise depending on the location of the personnel in the mission control centre. Furthermore, the signal-to-noise ratio (SNR) has a strong variance, with levels ranging from 0 to 20 dB. Most recordings were made with head-mounted microphones, but some in-spacecraft recordings were made using fixed far-field microphones which also picked up the presence of environmental noise. All those characteristics are likely to degrade the performance of speech technology applications.

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. As a progression, following the previous versions of the challenge [19] [20], the 3rd phase of the challenge (FSC P3), held in 2021, focuses on the development of single-channel supervised learning strategies with an aim to test system generalisation to varying channel and mission data. Besides 20 hours of Apollo-11 evaluation data, 5 hours of unseen channel evaluation data and an additional 5 hours of blind-set Apollo-13 mission evaluation data have been included in the evaluation dataset for the challenge. This fact indicates that a larger mismatch between development and evaluation results could be observed.

# 3. Proposed SAD system

## 3.1. Unsupervised representation learning

Our proposed representation learning approach is inspired by the one presented in [10], with some variations. As shown in Figure 1, two stages are combined to learn a feature extractor: First, an strided convolutional neural network (CNN) encoder runs directly on the 8 kHz waveform mapping the input sequence $X_t$ to a latent space $Z_t$. The total downsampling factor of the network is 80, resulting in a feature vector every 10 ms of audio. The second part is implemented as a GRU recurrent neural network [21] with 512 dimensional hidden state. The output of the GRU at every timestep is used as the context $C_t$ from which we predict 8 timesteps using a contrastive loss. Similarly to Decoar [22] or BERT text models [23], the left and right embedding contexts of a bidirectional GRU are used to predict future and past timesteps respectively, providing a loss term from future and past frame prediction tasks. The final loss function is the sum of both directional losses. This results in a final representation of 1024 dimensions being used as the context embedding. This array is extracted after training to be used as our proposed learned features.

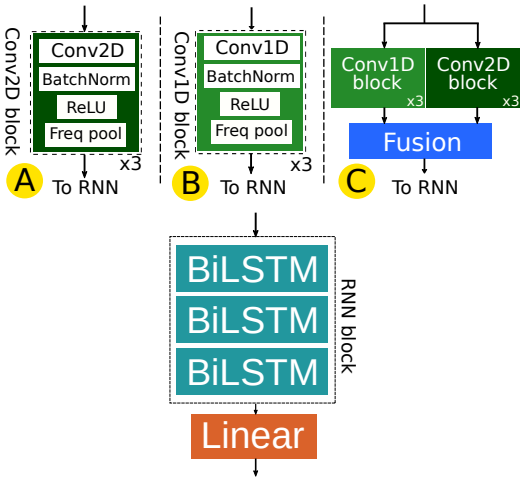


Figure 2: *Schematic representation of the different variations on the proposed convolutional recurrent neural network used for the SAD task.*

Concerning the prediction process of our system, unlike [15], our approach uses a single head for predicting future and past timesteps respectively, with an architecture consisting of a single hidden layer [24]. Furthermore, predictions for a clean reference $Z_t$ are obtained using augmented data from context $C_t$. Noises from MUSAN database [25] are added with a signal to noise ratio (SNR) that is sampled from an uniform distribution in the range (3, 15) dB. We also simulated different room impulse responses (RIR) using the gpuRIR toolkit [26] to incorporate reverberated conditions in training time.

These new features obtained through unsupervised learning are compared with a traditional set of perceptual features, considering log Mel-filterbank energies. Namely, we use the same configuration as in our previous participation in the Fearless Steps challenge, 64 log Mel-filterbank energies concatenated with the log energy of the frame.

## 3.2. SAD neural network

The neural architectures used for the SAD task are taken from our previous experience in this challenge [17]. All the architectures proposed are built on top of a RNN block, incorporating a set of convolutional layers working as a processing stage previous to it. The schematic representation of the proposed alternatives for the CRNN model is described in Figure 2. 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 uses three 2D convolutional blocks. Each of these blocks is integrated by a 2D convolutional layer with 3x3 kernel size and 64 filters. Then it is followed by a batch normalisation [27] and the application of a rectified linear unit (ReLU) [28] activation function. Finally, a max-pooling mechanism is applied considering a 4x1 stride, so that only the frequency axis is downsampled. Architecture B 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, implementing convolu-

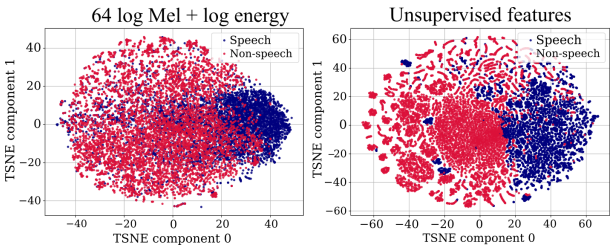


Figure 3: *TSNE 2D projection for the validation subset of both set of features considered in this work: 64 log Mel-filterbank energies + log energy (left) and features obtained through unsupervised learning (right).*

tions as 5 independent groups. 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 combines 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. In all cases, the SAD neural network is trained using a cross entropy loss function optimised using Adam algorithm, with a learning rate decaying exponentially from $10^{-3}$ to $10^{-4}$ during 20 epochs, and a minibatch size of 32.

The neural network output may result in a noisy estimation of class boundaries. Aiming to avoid high-frequency transitions, speech score is smoothed using an L order averaging filter. This filter is implemented as a zero-phase FIR filter to avoid phase distortions in the output signal. The optimal value for L was empirically found to be in the range between 50 and 60, which is equivalent to considering a moving average of 500 to 600 ms.

# 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. Our proposed unsupervised learning representation system can benefit from huge amounts of unlabelled training data. In order to train it, we include data from several English datasets: Librispeech, RSR2015, Tedlium release 1, Voxforge, Librilight, Voxceleb 1 and 2, Commonvoice (English only) and MLS (English only). Note that the full dataset is included unless explicitly noted. This results in a total of around 130 thousands hours of unlabelled audio used in training for the unsupervised feature learning system.

Given the specific audio domain considered for this challenge, with quite degraded channels and several kinds of transmission noises, the SAD neural network has been trained using only data provided by the challenge 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. Training subset 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 SAD systems were trained with around 56 hours of audio from the train partition. Development subset comprises 15 hours of audio. This subset was used to obtain an empirical threshold, in order to minimise the detection cost function (DCF) metric. Also, results are reported on this subset. Evaluation subset contains 34 hours of audio. 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. 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:

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

where $\mathrm{T}_{FP}$ and $\mathrm{T}_{FN}$ 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:

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

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 ($\theta$) that minimises the DCF.

# 5. Results

Seeking to obtain a deeper understanding of the information provided by the new set of learned features, we compare them to perceptual log Mel-filterbank energies through a t-SNE projection [29]. Figure 3 shows the t-SNE projection on a 2D plane of the validation susbet for both 64 log Mel-filterbank energies and the features obtained through unsupervised learning. It can be observed that log Mel-filterbank energies show a significant overlap between the speech and non-speech classes. On the other hand, unsupervised features provide a much more separable representation of both classes, with a minimum amount of overlap between speech and non-speech when compared to log Mel-filterbank energies. It is also to interesting to observe how the unsupervised features tend to assemble on small subclusters. This fact may come motivated by the sequentiality introduced by the RNN from our unsupervised representation learning system, grouping together features that are temporally close.

Once we have experimentally validated the separability provided by the new unsupervised features, we evaluate them on our SAD system. Table 1 presents the obtained results for the different systems submitted. We compare the performance of all the presented CRNN architectures using log Mel-filterbank energies and the representation obtained by unsupervised learning as input. Results are reported in terms of DCF metric for

Table 1: *SAD results in terms of DCF metric on the development and evaluation partition for the CRNN trained using log Mel-filterbank energies and the proposed unsupervised features.*

| System | Mel | | Unsup | |
|---|---|---|---|---|
| | DCF(%) | | DCF(%) | |
| | Dev | Eval | Dev | Eval |
| A1 - CRNN 2D (3x3) | 1.27 | 7.82 | 0.92 | 3.63 |
| B1 - CRNN 1D | 1.44 | 8.49 | 0.65 | 2.98 |
| B2 - CRNN 1D dilation | 1.59 | 7.13 | 0.91 | 3.66 |
| B3 - CRNN 1D groups | 1.37 | 8.46 | 0.96 | 3.13 |
| C1 - CRNN fusion bilinear | 1.30 | 7.55 | 0.87 | 3.86 |
| C2 - CRNN fusion sum | 1.28 | 8.64 | 0.97 | 3.11 |
| C3 - CRNN fusion concat | 1.48 | 9.06 | 0.84 | 3.36 |

both, development and evaluation partitions. As a comparison, we also report the baseline provided by the organisation, which is based on a statistical approach [30]. This system yielded a DCF value of 12.50% and 15.61% respectively for the development and evaluation partitions. As it can be observed, all our submissions largely outperform the baseline provided by the organisation. Concerning the results using log Mel-filterbank energies as input, competitive results are obtained on the development partition, with a DCF metric of 1.27% obtained in the best case with the 2D convolutional setup. However, a strong degradation in results can be observed when comparing to the metric reported in the evaluation partition, with a best case DCF of 7.13% for the 1D convolutional setup using dilation. This fact could be expected as new data from unseen channels has been included in this new edition of the challenge, so unlike our results in [17], a bigger mismatch is observed between development and evaluation results.

An overall improvement can be observed on all the neural architectures evaluated when using the new learned features. Best results are obtained with the 1D convolutional setup, yielding a DCF metric of 0.65% and 2.98% on the development and evaluation partitions respectively. Even though the boost in performance observed using the unsupervised features is significant in the development and evaluation partitions, it should be noted that this improvement is more relevant in the case of the evaluation dataset. While the average relative improvement observed comparing log Mel-filterbank energies and unsupervised features among all the architectures evaluated is around 37% in the development data, this percentage increases to 58% in the case of the evaluation data. Given the composition of the evaluation data in this year's edition of the challenge, with new unseen channels being present, these results suggest that learned features show a robust behaviour discriminating speech and non-speech classes even in possibly shifted acoustic conditions.

Additionally, Figure 4 shows a qualitative visualisation of our SAD system performance on an audio excerpt from development partition, comparing the use of log Mel-filterbank energies and unsupervised features as input for the Conv1D SAD neural network. As expected, a high speech score is associated with a strong evidence of speech presence in the audio spectrogram. Generally, it can be observed that both solutions can capture accurately the speech and non-speech fragments (Note that DCF obtained for file `fsc_p3_dev_001` is 0.02% for log Mel-filterbank energies and 0.01% for unsupervised features). It is also interesting to mention that the system using unsupervised features as input consistently provides a higher score than the setup using log Mel-filterbank energies for speech fragments,

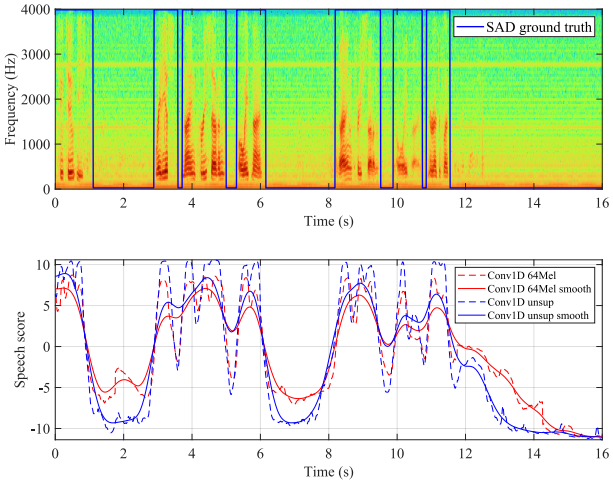


Figure 4: *Qualitative visualisation of SAD scores in a 16 seconds audio fragment extracted from file "`fsc_p3_dev_001`". From top to bottom: spectrogram with SAD ground truth overlapped, and speech score for different SAD systems proposed.*

and lower scores for the non-speech fragments. Concerning the behaviour of the averaging filter, it can be seen that some of the high frequency occurrences in the speech score are eliminated, resulting in a smoother signal being used in the thresholding process.

## 6. Conclusions

In this paper, we presented the ViVoLab submission to the SAD task of the Fearless Steps Challenge 2021 dealing with audio from Apollo space missions. For this edition, the focus was set on testing the systems generalisation capabilities, with an evaluation dataset that includes new unseen channel data. Seeking to obtain a new and more robust audio representation, for our submission we explored the unsupervised representation learning paradigm. Our system uses a contrastive loss to learn a feature representation combining an strided CNN encoder with a RNN that provides a context embedding, which is extracted after training to be used by the SAD neural network as input. The obtained features are then used to train different variants of a CRNN architecture.

Experimental results suggest that features learned via unsupervised learning provide a much more robust representation, significantly reducing the mismatch observed between development and evaluation partition results when compared to a set of traditional perceptual features such as log Mel-filterbank energies. Obtained results largely outperform the baseline provided by the challenge organisation. Our best submission achieved a DCF metric of 0.65% and 2.98% respectively in the development and evaluation sets, ranking third among the 6 participant teams in the SAD task.

## 7. Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 101007666, by the Spanish Ministry of Economy and Competitiveness and the European Social Fund (TIN2017-85854-C4-1-R) and the Government of Aragón (Reference Group T36_20R), and by Nuance Communications, Inc. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

# 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] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

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

[12] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[13] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," *arXiv preprint arXiv:1904.03240*, 2019.

[14] T. H. Trinh, M.-T. Luong, and Q. V. Le, "Selfie: Self-supervised pretraining for image embedding," *arXiv preprint arXiv:1906.02940*, 2019.

[15] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *arXiv preprint arXiv:1904.05862*, 2019.

[16] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

[17] P. Gimeno, D. Ribas, A. Ortega, A. Miguel, and E. Lleida, "Convolutional Recurrent Neural Networks for Speech Activity Detection in Naturalistic Audio from Apollo Missions," in *Proc. IberSPEECH 2021*, 2021, pp. 26–30. [Online]. Available: http://dx.doi.org/10.21437/IberSPEECH.2021-6

[18] 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.

[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] 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.

[21] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[22] S. Ling, Y. Liu, J. Salazar, and K. Kirchhoff, "Deep contextualized acoustic representations for semi-supervised speech recognition," in *Proc. IEEE ICASSP*, 2020, pp. 6429–6433.

[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[24] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[25] D. Snyder, G. Chen, and D. Povey, "Musan: A music, speech, and noise corpus," *arXiv preprint arXiv:1510.08484*, 2015.

[26] D. Diaz-Guerra, A. Miguel, and J. R. Beltran, "gpuRIR: A python library for room impulse response simulation with gpu acceleration," *Multimedia Tools and Applications*, vol. 80, no. 4, pp. 5653–5671, 2021.

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

[28] 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.

[29] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[30] 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.