A Transformer-Based Voice Activity Detector

Biswajit Karan[1], *Joshua Jansen van Vüren*[1], *Febe de Wet*[1], *Thomas Niesler*[1]

[1]Department of Electrical and Electronic Engineering,
Stellenbosch University, Stellenbosch, South Africa

bkaran@sun.ac.za, jjvanvueren@sun.ac.za, fdw@sun.ac.za, trn@sun.ac.za

Abstract

Voice activity detection (VAD) is the task of distinguishing speech from other types of audio signals, such as music or background noise. We introduce a novel end-to-end VAD architecture which incorporates a pre-trained transformer model (Wav2Vec2-XLS-R). We evaluate the proposed architecture on an established VAD dataset, AVA-Speech, and a manually-segmented corpus of under-resourced multilingual speech. As benchmarks, we include a hybrid CNN-BiLSTM system and an off-the-shelf enterprise VAD. On the AVA-Speech test set, our proposed VAD achieves an area under the curve (AUC) of 96.2% while the benchmarks achieve 94.8% and 81.9% respectively. On the multilingual dataset, the gap widens to 92.2% for the transformer-based VAD and 80.8% and 74.6% for the two baselines. Therefore, the proposed VAD offers improved performance in all cases, with an absolute increase of more than 11% for our target domain. We conclude that the proposed end-to-end architecture improves VAD performance.

Index Terms: VAD, transformer, multilingual speech

1. Introduction

The process of determining which parts of an audio stream are speech and which are not is known as voice activity detection (VAD). Real-world speech signals are often embedded in background noise or other interference, such as music. Therefore, several practical voice processing tasks, such as automatic speech recognition (ASR), speaker identification and speech enhancement, require reliable and accurate VAD as a preprocessing step. Our own eventual objective is to automatically and accurately determine segments of clean speech, for subsequent manual annotation, within a multilingual audio stream in under-resourced languages. Currently this segmentation is achieved manually, a process which has shown itself to be both slow and cumbersome.

Early approaches to VAD simply used an energy threshold to identify speech [1]. The addition of a hangover period helped to reduce speech truncation, while the application of adaptive thresholds to different temporal and spectral aspects of the signal enhanced the accuracy of this simple approach [2, 3]. For example, the discrete Fourier transform (DFT) coefficients of speech and noise can be modelled as asymptotically independent Gaussian random variables [4]. VAD is then accomplished using a likelihood ratio test in conjunction with an HMM hangover model. This work has served as the foundation for several related techniques that employ different signal characteristics [5, 6] or distributions [7, 8] or decision procedures [9]. However, these model-based approaches were not effective when faced with challenging non-speech interference, such as music.

Recently, machine learning-based systems have achieved state-of-the-art VAD performance. Classifiers can be trained to distinguish between speech and non-speech frames by approaching VAD as a frame-based classification problem. Support vector machine classifiers have for example been utilised successfully in this way [10, 11, 12, 13]. More recently, neural network architectures have been used instead [14, 15, 16, 17].

Our requirement for VAD is as a preprocessing step in the development of ASR and word spotting in a highly resource-constrained environment [18]. The speech of interest is typically spontaneous, multilingual and in languages for which very few or even no resources exist. This scenario was addressed in recent work which proposes a hybrid CNN-BiLSTM structure and demonstrates state-of-the-art VAD performance [19]. However, when applied to our under-resourced target domain, we found that this architecture is not able to identify clean speech with sufficient accuracy.

In this paper, we propose a VAD architecture that incorporates a large pre-trained transformer model, namely Wav2Vec2-XLS-R [20]. In this way, advantage is taken of the very large multilingual dataset used to train this neural network model. We will show that the resulting VAD offers improved performance on two datasets relative to two benchmarks, a hybrid CNN-BiLSTM and SileroVAD, an off-the-shelf enterprise VAD implementation [21].

The remainder of this paper is structured as follows. Section 2 provides information on the datasets that we utilise. Our methodology is presented in Section 3. The resulting analysis of the proposed VAD framework is discussed in Section 4, and in Section 5 we present our concluding remarks.

2. Datasets

We will evaluate VAD performance on two datasets. AVA-Speech is a publicly-accessible dataset developed specifically for VAD benchmarking. In addition, we consider a corpus of manually segmented speech from 48 episodes of South African soap operas, this under-resourced dataset resembles the target application for a voice activity detection system, from which we would aim to automatically obtain segments of speech.

2.1. AVA Speech dataset

This dataset comprises 40 hours of labelled audio data extracted from 160 movies found on YouTube. Approximately 15 minutes audio from each movie is labelled using the following four mutually-exclusive classes to identify sequential segments: "Clean-Speech", "Speech+Music", "Speech+Noise", and "Non-Speech" [22]. AVA-Speech includes a wide range of languages, acoustic settings and speakers. Moreover, film audio can be viewed as representative of broadcast media, set-

ting it apart from more carefully curated and therefore artificial datasets that are often used for VAD testing and development. AVA-Speech contains approximately equal amounts of speech and non-speech data, with the majority of the speech data being noisy. Of the 40 hours of data, we will utilise 4.25 hours for validation, 4.25 hours for testing, and the remainder for training. Table 1 summarises the statistics of the AVA-Speech dataset.

Table 1: *Statistics of the AVA-Speech dataset*

Label	**Average duration (sec)**	**Segments** (%)	**Time** (%)
Clean-Speech	2.97	16.68	14.55
Speech+Noise	3.28	25.41	24.32
Speech+Music	3.43	13.33	13.46
Non-Speech	3.68	44.57	47.68

2.2. South African soap opera dataset

As a second dataset, we use a collection of 48 manually-segmented episodes of South African soap operas, also collected from YouTube. The speech in this data has been reported to be highly multilingual and spontaneous, making it a challenging and representative test scenario for our VAD, since our objective is to apply it to under-resourced languages in Africa [17, 23]. Our data consists of entire episodes, each on average 24 minutes long. We utilise 22 episodes for training, six for development, and 20 for testing, constituting 8.86h, 2.44h and 8.15h of audio respectively. The statistics for this dataset are summarised in Table 2.

Our over-arching objective is to automatically identify segments of clean speech within a multilingual under-resourced audio stream that can be used for subsequent manual annotation, Therefore, the audio in this corpus has been delineated into two mutually exclusive classes, namely "Clean-Speech" and "Other". This means that the classes "Non-Speech", "Speech+Music" and "Speech+Noise" used in the AVA-Speech dataset all correspond to "Other" in this dataset.

By comparing Tables 1 and 2 it is clear that the soap opera dataset contains a similar ratio of clean speech to other audio as the AVA-Speech dataset. However, the average duration's of these segments are substantially different. This is a consequence of the segmentation process, in which short clean speech segments are typically surrounded by much longer segments of non-clean speech.

We note that the VAD task for AVA-Speech requires the processing of 15 to 30 minute long segments extracted from movies to be processed, the soap opera data requires entire episodes to be processed without any pre-processing. This, for example, includes the title and credit sequences.

Table 2: *Statistics of South African soap opera corpus*

Label	**Average duration (sec)**	**Segments** (%)	**Time** (%)
Clean-Speech	4.16	49.38	11.27
Other	31.95	50.62	88.73

3. Methodology

Our proposed VAD architecture is shown in Figure 1. The end-to-end transformer model, Wav2Vec2-XLS-R, has been pre-trained on 436k hours of unlabelled speech drawn from the

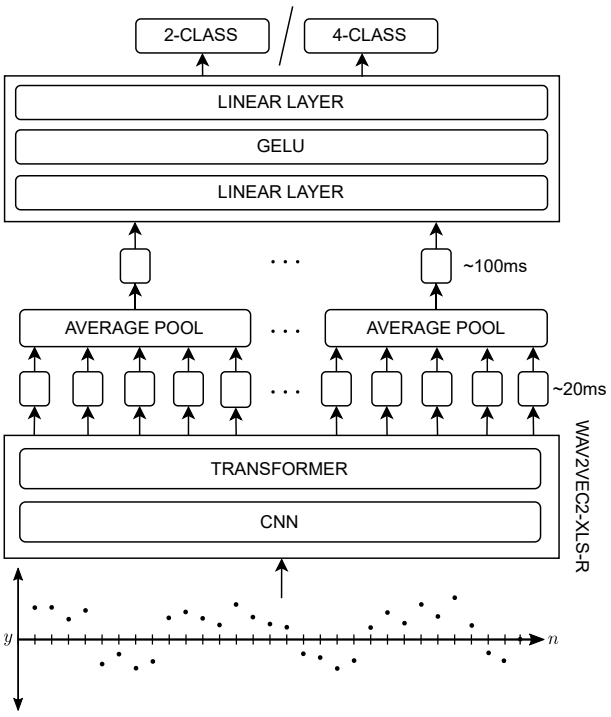

Figure 1: *Structure of Wav2Vec2 voice activity detection system. Feature vectors from Wav2Vec2 are passed to average pooling layer in groups of five, which is then passed to a two layer linear model with a GeLU activation function, the final layer has dimensionality of one or four.*

VoxPopuli, MLS, CommonVoice, BABEL, and VoxLingua107 speech corpora, which encompass 128 languages [20]. Except for English, only one of the target languages for our South African dataset forms part of the XLS-R training dataset, namely isiZulu, and constitutes only 56 hours of data.

The transformer model receives the raw audio samples as input. Because the computational complexity of the transformer model is proportional to the square of the length of the input sequence, we divide the input audio into fixed 20 second long windows with no overlap. Given a 20 second sequence of audio samples, the transformer model produces, as output, a sequence of hidden states, each corresponding to 25ms of audio, with a stride of approximately 20ms. For our application, we regard a speech boundary placed within 105ms of the reference boundary as satisfactorily accurate. Therefore, we pool the output hidden states produced by the transformer into groups of five, producing one 768-dimensional vector approximately every 105ms. These are passed though a dense layer with GELU activations and an output dimensionality half of its input (384). Finally, the GELU outputs are passed through a softmax with either two outputs (corresponding to the classes "Clean-Speech," and "Other") or four outputs (corresponding to "Clean-Speech", "Speech+Music", "Speech+Noise" and "Non-Speech"). Target training labels are derived from the labels in the AVA-Speech and soap opera datasets by determining the class that describes the largest proportion of the frame being classified.

This architecture is fine-tuned utilising cross-entropy. We use a batch size of 16 and a peak learning rate of $5 \cdot 10^{-5}$ which warms up linearly for the first 10% of training steps and then linearly cools to zero by the end of training. Each model is trained for 16 epochs, during which the CNN feature-extractor layers

are frozen, while the transformer layers and final classification layers are trained.

3.1. Baselines

We employ two baseline VAD architectures as benchmarks. First, we include a hybrid CNN-BiLSTM system which provides state-of-the-art performance on the AVA-Speech dataset [19]. Second, we include the Silero VAD, an off-the-shelf enterprise-grade system [21]. This model has been trained on a large undisclosed corpus containing over 100 languages, and is reported to achieve a 90% AUC on four hours of speech sampled from the AVA- speech dataset.

The baseline CNN-BiLSTM was trained from a random initialisation for 24 epochs with a batch size of 64. The structure of the model is the same as the best architecture presented in [19]. This model produces a frame level output every 320ms.

4. Results

The results for two-class and four-class classification are presented separately below.

4.1. Two-class VAD

This section presents and analysis of classifier performance when these are trained to distinguish between the two classes "Clean-Speech" and "Other". In the case of AVA-Speech, this means that "Speech+Music", "Speech+Noise", and "Non+Speech" are all regarded as "Other".

4.1.1. AVA-Speech data

Figure 2 presents the ROC curve for the fine-tuned transformer VAD, the CNN-BiLSTM VAD, and the Silero VAD, for the AVA-Speech development and test sets. Note that only the CNN-BiLSTM and transformer VADs are trained on the AVA-Speech training set, while the Silero VAD is used as-is. In contrast to the BiLSTM VAD trained in [19], here we select only clean speech as the positive class, as it better represents for our eventual goal of isolating speech for manual transcription. The CNN-BiLSTM model achieves an area under the ROC curve of 93.7% and 94.8% on the development and test sets respectively. The transformer model improves on this, achieving an AUCs of 95.0% and 96.2% respectively. Both the CNN-BiLSTM and the transformer models outperform the Silero VAD, which achieves AUCs of 83.9% and 81.9% on the development and test sets respectively. We hypothesise that this model performs slightly worse than the reported 90% because here our positive class is defined to only include clean-speech.

Table 3: *AVA speech voice activity detection results for the development (Dev) and test sets. AUC: Area under ROC curve, TPR: True positive rate, FPR: False positive rate.*

Model	AUC		TPR @ FPR 10%	
	Dev	Test	Dev	Test
Wav2Vec2-XLS-R	**95.0%**	**96.2%**	**88.0%**	**88.9%**
CNN-BiLSTM	93.7%	94.8%	76.7&	84.4%
Silero	83.9%	81.9%	48.4%	26.5%

Our aim is to use the VAD to isolate segments of clean multilingual speech that can be passed to a human transcriber. The goal is to minimise human effort, and therefore false positives, where non-speech or noisy speech are labelled as clean speech, are to be avoided because they would require manual re-adjustment of the segment boundaries. The fine-tuned Wav2Vec2 model, as seen in Table 3, shows better performance than both the LSTM model and the off-the-shelf VAD, offering a 4.5% and 62.4% absolute test set improvement in true positive rate at a false positive rate of 10% respectively.

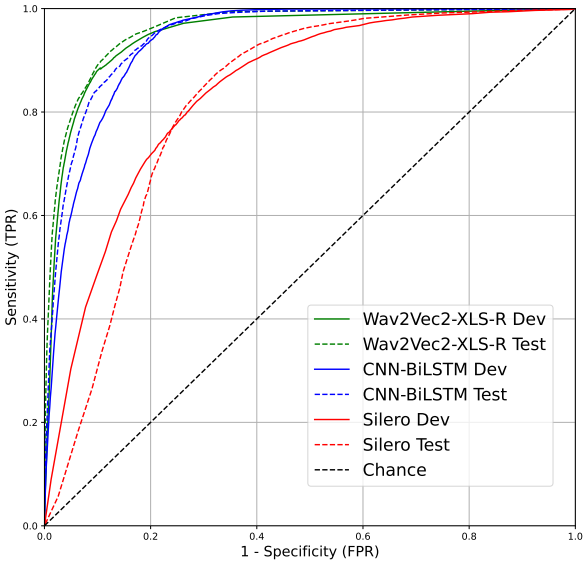

Figure 2: *AVA-Speech corpus development and test set ROC curves for Wav2Vec2, CNN-BiLSTM and Silero VAD.*

4.1.2. Soap opera speech data

Table 4: *Soap opera voice activity detection results for the development (Dev) and test sets. AUC: Area under ROC curve, TPR: True positive rate, FPR: False positive rate.*

Model	AUC		TPR @ FPR 10%	
	Dev	Test	Dev	Test
Wav2Vec2-XLS-R	**93.3%**	**92.2%**	**75.6%**	**71.6%**
CNN-BiLSTM	82.3%	80.8%	46.0%	39.8%
Silero	74.1%	74.6%	32.3%	28.0%

In Figure 3 and Table 4 we present the performance of the same three models when applied to the soap opera dataset. We see that, again, the transformer VAD outperforms the two baselines, this time by a larger margin. Although the AUC of CNN-BiLSTM model deteriorates by 14.0% absolute on the test set for the under-resourced soap opera dataset, the Wav2Vec2 model only deteriorates by 4.0%. It would appear, therefore, that the Wav2Vec2 model is more robust to the detection of clean speech in multilingual low-resource datasets.

Again, our over-arching objective is to isolate segments of clean multilingual speech that can be passed to a human transcriber. When considering the true positive rate (TPR) at a false positive rate of 10% the performance gap between the Wav2Vec2 model and the two baseline again widens. The Wav2Vec2 model is able to identify 71.6% of the clean speech from the test set while 10% of the negative class is incorrectly classified as clean speech. In contrast, at the same FPR the CNN-BiLSTM model is only able to identify 39.8% of the clean-speech from the test set, and the Silero VAD system only 28.0%. Thus the transformer model affords an absolute improvement over the CNN-BiLSTM model of 29.6% absolute.

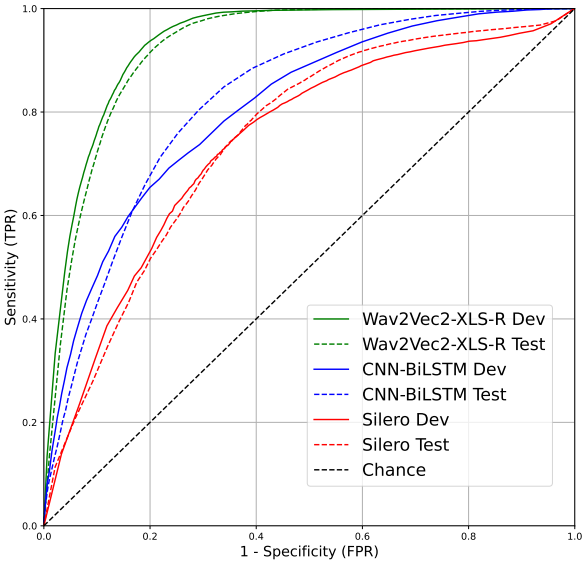

Figure 3: *Soap opera corpus development and test set ROC curves for Wav2Vec2, CNN-BiLSTM and Silero VAD.*

Table 5: *Comparison of number of model parameters and throughput (Seconds of audio processed per second).*

Model	Throughput (s/s)	Parameters (Millions)
Wav2Vec2-XLS-R	113.1	316
CNN-BiLSTM	164.6	0.552
Silero	37.5	0.180

In Table 5 we compare the runtime performance of the three considered models. This performance is denoted as throughput, which is the seconds of input audio processed per second and is the inverse of real-time factor (RTF). We estimate this value by measuring the wall-time during the segmentation of the entire Soap Opera test set (8.15h). We find that the CNN-BiLSTM model achieves the highest throughput. However, although Wav2Vec2 contains 572x more parameters, the reduction in throughput is only 31.3% slower relative to the BiLSTM model. This is because the transformer architecture can process entire audio segments simultaneously, while the LSTM architecture must process the input in a recurrent fashion. We find that the Silero model performance is the worst of the three models, which is consistent with the performance reported by the authors (36 s/s).[1].

4.2. Four-class VAD

To further investigate the effectiveness of our proposed VAD architecture, we train both the transformer as well as the CNN-BiLSTM models to classify all four classes labelled in the AVA-Speech dataset, namely: "Clean-Speech", "Speech+Music", "Speech+Noise" and "Non-Speech". Figure 4 presents the confusion matrices for the two VAD approaches, computed on the test set. The figure shows that the transformer VAD is better able to separate all four classes than the CNN-BiLSTM model. For example, it is able to correctly identify 70.1% of the clean speech, while the CNN-BiLSTM experiences a higher confusion between the clean speech and especially speech with music. In addition to this, the Wav2Vec2 model more effectively distinguishes Speech+Noise from the other three classes, while the CNN-BiLSTM misclassifies a large proportion (31.0%) of speech with noise as speech with music. Overall the transformer VAD more robustly delineates the four environmental conditions.

[1]However, utilising third-party libraries such as ONNX a considerably improved throughput of 158s/s can be achieved.

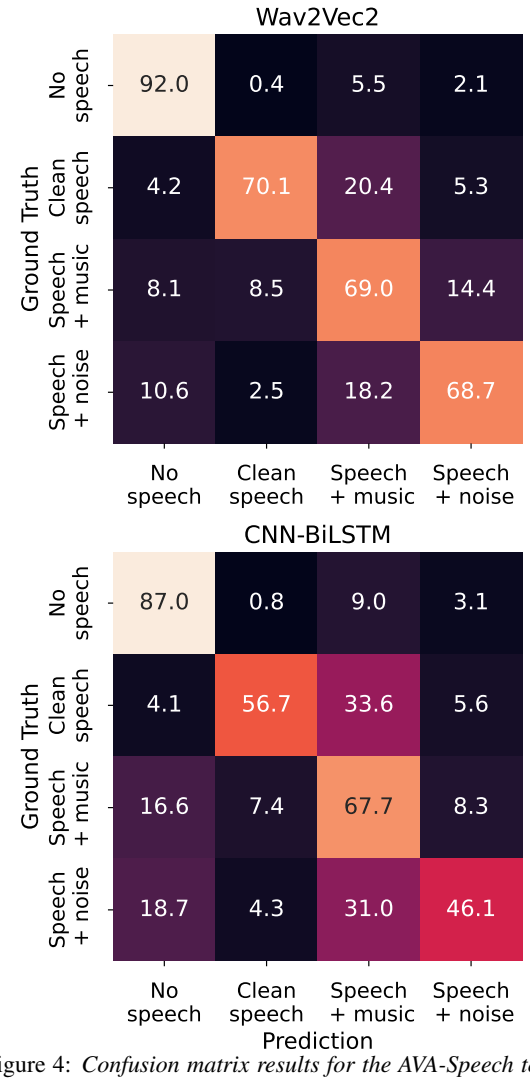

Figure 4: *Confusion matrix results for the AVA-Speech test set when the proposed transformer-based VAD and the baseline CNN-BiLSTM VAD are trained to classify all four classes. Results are given as percentages (%) and are normalised row-wise.*

5. Conclusions

In this work we have introduced a novel transformer architecture for voice activity detection (VAD) that incorporates the pre-trained Wav2Vec2-XLS-R model. The approach is demonstrated to offer consistent performance over two baselines on for both the AVA speech corpus and a dataset composed of complete multilingual soap opera episodes. The proposed architecture achieves test set AUCs of 96.2% and 92.2% on the AVA and soap opera datasets respectively, in both cases but especially for the latter corpus improving on both baselines. Furthermore, it is shown to substantially improve the true positive rate at a fixed low false positive rate relative to both baselines. This leads to a reduction in the required manual adjustment of the clean speech segment boundaries, while maintaining a high yield of clean speech, indicating that the proposed VAD is well-suited to the task of isolating speech for subsequent manual transcription.

6. References

[1] K. Bullington and J. Fraser, "Engineering aspects of TASI," *Transactions of the American Institute of Electrical Engineers, Part I: Communication and Electronics*, vol. 78, no. 3, pp. 256–260, 1959.

[2] T. ITU, "Annex b: A silence compression scheme for g. 729 optimized for terminals conforming to recommendation v. 70," *International Telecommunication Union*, 1996.

[3] J. Ramirez, J. M. Górriz, and J. C. Segura, "Voice activity detection. fundamentals and speech recognition system robustness," *Robust speech recognition and understanding*, vol. 6, no. 9, pp. 1–22, 2007.

[4] J. Sohn, N. S. Kim, and W. Sung, "A statistical model-based voice activity detection," *IEEE signal processing letters*, vol. 6, no. 1, pp. 1–3, 1999.

[5] S. Gazor and W. Zhang, "A soft voice activity detector based on a laplacian-gaussian model," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 5, pp. 498–505, 2003.

[6] J. W. Shin, J.-H. Chang, H. S. Yun, and N. S. Kim, "Voice activity detection based on generalized gamma distribution," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Philadelphia, USA, 2005.

[7] J. Ramírez, J. C. Segura, C. Benítez, L. García, and A. Rubio, "Statistical voice activity detection using a multiple observation likelihood ratio test," *IEEE Signal Processing Letters*, vol. 12, no. 10, pp. 689–692, 2005.

[8] 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.

[9] D. Enqing, L. Guizhong, Z. Yatong, and Z. Xiaodi, "Applying support vector machines to voice activity detection," in *Proceedings of the 6th International Conference on Signal Processing*, Beijing, China, 2002.

[10] J. Ramírez, P. Yélamos, J. M. Górriz, J. C. Segura, and L. García, "Speech/non-speech discrimination combining advanced feature extraction and SVM learning," in *Proceedings of the Ninth International Conference on Spoken Language Processing (ICSLP)*, Pittsburgh, USA, 2006.

[11] Q.-H. Jo, J.-H. Chang, J. Shin, and N. Kim, "Statistical model-based voice activity detection using support vector machine," *IET Signal Processing*, vol. 3, no. 3, pp. 205–210, 2009.

[12] J. Wu and X.-L. Zhang, "Efficient multiple kernel support vector machine based voice activity detection," *IEEE Signal Processing Letters*, vol. 18, no. 8, pp. 466–469, 2011.

[13] X.-L. Zhang and J. Wu, "Deep belief networks based voice activity detection," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 4, pp. 697–710, 2012.

[14] N. Ryant, M. Liberman, and J. Yuan, "Speech activity detection on YouTube using deep neural networks." in *Proceedings of Interspeech*. Lyon, France, 2013.

[15] T. Hughes and K. Mierle, "Recurrent neural networks for voice activity detection," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, 2013.

[16] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "CNN architectures for large-scale audio classification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, USA, 2017.

[17] N. Wilkinson, A. Biswas, E. Yilmaz, F. De Wet, E. Van der westhuizen, and T. Niesler, "Semi-supervised acoustic modelling for five-lingual code-switched ASR using automatically-segmented soap opera speech," in *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, Marseille, France, 2020.

[18] A. Biswas, R. Menon, E. van der Westhuizen, and T. Niesler, "Improved low-resource Somali speech recognition by semi-supervised acoustic and language model training," *arXiv preprint arXiv:1907.03064*, 2019.

[19] N. Wilkinson and T. Niesler, "A hybrid CNN-BiLSTM voice activity detector," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, Canada, 2021.

[20] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," in *Proceedings of Interspeech*, Incheon, Korea, 2022.

[21] Silero Team, "Silero VAD: pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier," https://github.com/snakers4/silero-vad, 2021.

[22] S. Chaudhuri, J. Roth, D. Ellis, A. C. Gallagher, L. Kaver, R. Marvin, C. Pantofaru, N. C. Reale, L. G. Reid, K. Wilson, and Z. Xi, "AVA-speech: A densely labeled dataset of speech activity in movies," in *Proceedings of Interspeech*, Hyderabad, India, 2018.

[23] E. van der Westhuizen and T. Niesler, "A first South African corpus of multilingual code-switched soap opera speech," in *Proceedings of the 11th Language Resources and Evaluation Conference (LREC)*, Miyazaki, Japan, 2018.