Overlapped Speech Detection for Improved Speaker Diarization on Tamil Dataset

S.T. Jarashanth
Faculty of Engineering
University of Jaffna
Kilinochchi, Sri Lanka
jarashanth@eng.jfn.ac.lk

K. Ahilan
Faculty of Engineering
University of Jaffna
Kilinochchi, Sri Lanka
ahilan@eng.jfn.ac.lk

R. Valluvan
Faculty of Engineering
University of Jaffna
Kilinochchi, Sri Lanka
valluvan.r@eng.jfn.ac.lk

T. Thiruvaran
Faculty of Engineering
University of Jaffna
Kilinochchi, Sri Lanka
thiruvaran@eng.jfn.ac.lk

A. Kaneswaran
Faculty of Engineering
University of Jaffna
Kilinochchi, Sri Lanka
kanesh@eng.jfn.ac.lk

***Abstract*— Speaker diarization is the task of partitioning a speech signal into homogeneous segments corresponding to speaker identities. We introduce a Tamil test dataset, considering that the existing literature on speaker diarization has experimented with English to a great extent; however, none on a Tamil dataset. An overlapped speech segment is a part of an audio clip where two or more speakers speak simultaneously. Overlapped speech regions degrade the performance of a speaker diarization system proportionally due to the complexity of identifying individual speakers. This study proposes an overlapped speech detection (OSD) model by discarding the non-speech segments and feeding speech segments into a Convolutional Recurrent Neural Network model as a binary classifier: single speaker speech and overlapped speech. The OSD model is integrated into a speaker diarizer, and the performance gain on the standard VoxConverse and our Tamil datasets in terms of Diarization Error Rate are 5.6% and 13.4%, respectively.**

Keywords— Overlapped speech detection, Speaker diarization, Convolutional recurrent neural network, Binary classifier, Tamil dataset

I. INTRODUCTION

Speaker diarization, a task widely known as determining "who spoke when" in a multi-speaker conversation, labels speech signals with classes corresponding to speaker identities [1]. Typical applications of speaker diarization include meeting conversation analysis, speech recognition with speaker identification, and multi-media information retrieval [2]. With the advent of x-vectors, speaker embedding-based speaker diarization systems have produced state-of-the-art performances [3]. In such systems, speaker embedding extraction is one of the critical components where embeddings are extracted from short speech segments, typically 1.5 seconds long. Since these embeddings are directly used for speaker clustering, extracting reliable speaker embedding (i.e., an embedding that uniquely identifies a speaker) is vital. In a multi-speaker conversation in the real world, extracting embedding is challenging as the conversation may happen in an adverse noisy environment or contain overlapped speech.

In a spontaneous conversational speech, it is prevalent that multiple speakers speak simultaneously (i.e., overlapped speech). Overlapped speech regions commonly occur at speaker turn points or as backchannel utterances or interruptions. In speaker diarization, overlapped speech causes performance degradation since it provokes increases in missed speaker rates due to the difficulties in identifying concurrent speakers.

Researchers have handled overlapped speech in speaker diarization in several ways, such as removing overlapped speech from the dataset and separating individual speech signals from the mixture [4]. Unsupervised signal processing methods were used in the early stages to design suitable techniques for detecting overlapped segments. For example, Yantorno et al. used spectral autocorrelation peak valley ratio to tag overlapped segments [5]. Boakye et al. proposed a Hidden Markov Model (HMM) system as a ternary classifier (non-speech, speech, and overlapped speech) utilizing spectral audio features (e.g., MFCCs and RMS energy) to detect and exclude overlaps prior to speaker clustering [6].

With the advent of deep neural networks (DNNs), several investigations applied DNN architectures to detect overlapped speech by treating them as a sequence labelling task. A pioneering study on this domain used long short-term memory (LSTM) networks to predict frame-wise overlap scores as a regressor where a threshold is learned to differentiate between overlap and non-overlap speech [7]. Yousefi et al. proposed a convolutional neural network (CNN) architecture and studied the impact of different auditory features; spectral magnitude, Mel Filter-Banks, pyknogram, and MFCC to learn the impacts on OSD [4].

Jung et al. proposed a convolutional recurrent neural network (CRNN) model in which audio segments are classified into non-speech, single-speaker speech, and overlapped speech [8]. Unlike the conventional classification of an audio segment into two classes, overlapped speech and non-overlapped speech, the authors further segment the non-overlapped speech class as single-speaker speech and non-speech. They have proved that the fine-grained classification makes the model more generalizable. Moreover, this alleviates the class imbalance problem because only a small portion of the speech contains overlapped speech.

This study proposes a novel overlapped speech detection approach for improved speaker diarization. An overlapped speech detection (OSD) model is developed, adopting a CRNN model [8]. Our OSD model is defined as a sequence labelling problem, where the speech regions identified using a voice activity detector (VAD) are classified into two classes for each 25ms frame with a shift of 10ms: single speaker

978-1-6654-7607-2/22/$31.00

speech and overlapped speech. VAD is a tool used to identify and remove non-speech regions in an audio file [9]. Our model can learn local contextual features and sequential information effectively by exploiting the potential of both CNNs and RNNs.

The existing models on OSD perform framewise labelling for the entire audio signal. Due to the availability of more accurate VADs in recent times, we contend that an OSD model would perform better if we do the framewise labelling on the speech regions as a two-class problem: single-speaker speech and overlapped speech. Our approach may make an OSD model more generalizable since the non-speech class with more variants, such as silences, background noises, and music, are discarded by the VAD before training [10].

Our OSD model is integrated with a speaker diarizer built using ECAPA-TDNN network [11] and spectral clustering with eigen-gap-based heuristics [12]. Generally, speaker diarization systems are evaluated in English using publicly available datasets such as VoxConverse [13] and AMI meeting corpus [14], to name a few, and to our knowledge, no study has experimented with Tamil. To bridge this gap, we introduce a Tamil test dataset for speaker diarization. Like VoxConverse, our Tamil dataset is collected from YouTube videos, including interviews, talk shows, panel discussions, and political debates, but the entire annotation is done manually, whereas, for VoxConverse, it is semi-automated. This study further reports the performances of a speaker diarization system by detecting and excluding overlapped speech on VoxConverse and Tamil datasets.

Rest of the paper is organized as follows. Section II details the methodologies of our proposed OSD model and the speaker diarization system and introduces our Tamil test dataset. Experiments conducted and the results obtained with a discussion are provided in Sections III and IV, respectively. Conclusions are drawn in Section V.

II. Methodology

This section details our proposed methodologies for OSD and speaker diarization and how the OSD model is integrated with the speaker diarizer.

A. Overlapped Speech Detection

This study defines the OSD model as a sequence labelling task. The pipeline of the proposed method for the OSD is shown in Figure 1.

Each audio file is converted to a mono signal with a sampling rate of 16 kHz unless they are already. An input to the OSD model is an audio clip with 150 frames, where each frame is 25ms with a shifting size of 10ms. Each audio file is chunked into a set of 150 frames with a shift of 50 frames. The duration of the input size of a chunk to the OSD model is 1.515s (150 frames), and the chunks or audio files lesser than 1.515s are discarded from training.

The internal parts and configurations of our proposed CRNN model are displayed in Table I. The model incorporates CNNs and an RNN, accounting for the unique capabilities of each network: CNN and RNN are good at modelling local patterns and sequential data, respectively [8]. The input signals are converted to 64-dimensional log Mel-Spectrograms before feeding into the OSD model. The model includes 3 CNN blocks, where each block includes 2 CNNs. All CNNs have a filter with a size of 3x3 and a stride with a size of 1. A bidirectional Gated Recurrent Unit (Bi-GRU) network is applied to the mean pooled output of the third CNN block. Then the output is forwarded to a fully connected layer with 256 neurons and, finally, to the classification layer with two neurons to accommodate two classes: single speaker speech and overlapped speech. To standardize the inputs/outputs within the internal layers of the network, batch normalization and mean normalization are applied at the appropriate places.

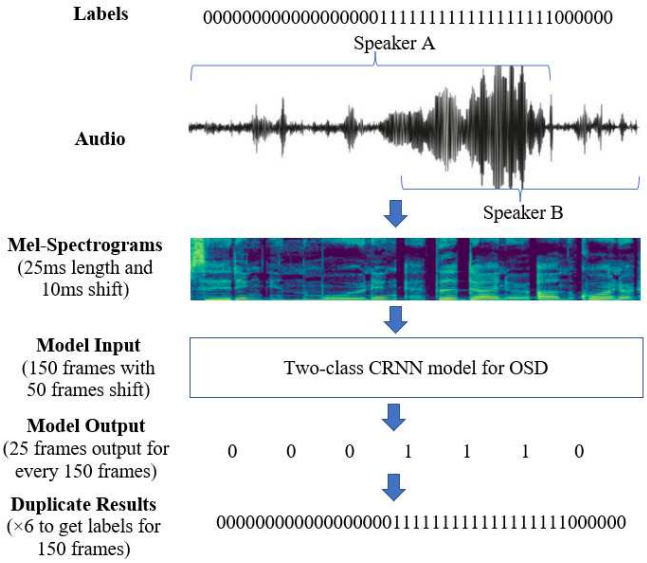

Fig. 1. Our proposed OSD approach.

TABLE I. Architecture of our proposed CRNN model for OSD. Conv (3, 1): CNN with filter size of 3x3 and stride=1. FC: Fully connected layer.

Layer	Model and Parameters	Output Shape
Input	Mel-Spectrogram	(150×64×1)
CNN blocks 1 to 3	Conv (3,1) Batch Normalization ReLU Activation Conv (3,1) Batch Normalization ReLU Activation Average Pooling	(25×32×64)
Normalization	Mean Pooling	(25×64)
RNN	Bi-GRU (256) × 2	(512,)
Linear	FC (256) LeakyReLU Linear (2)	(2,1)

The novelty in this work is how we present the input to the OSD model. Unlike previous works, we remove non-speech signals, namely silences, background noises, and music, using a VAD, known as silero-vad [15], before feeding inputs to the OSD model. The segmentations are carried out on the speech timestamps produced by the VAD. Since the VAD's timestamps contain only speech signals (i.e., single-speaker speech and overlapped speech), the model's capability to generalize improves while mitigating the class imbalance problem caused by non-speech and single-speaker speech grouped into a single class as non-overlapped regions.

Generally, a dataset for training an OSD comes with audio files in wav format, and the corresponding labels are provided in rich transcription time marked (RTTM) format [16]. An

RTTM file includes a list of timestamps for a particular audio file where the onsets and the durations of each speaker turn are mentioned. To produce labels for the inputs to the OSD model, we implement a mechanism to convert the RTTM output in a way that each frame (25ms) gets a label of either 1 (overlapped speech) or 0 (single-speaker speech).

Due to the use of CNNs and pooling layers, the model's output shrinks from 150 to 25 frames. In order to compensate for this act, each output is duplicated six times. The architecture for the evaluation is the same as the model for training, but instead of the 50 frames shift, a complete shift is performed (i.e., 150 frames shift) to avoid multiple predictions among chunks. Adjacent frames classified as 1 (i.e., overlapped speech) are merged, and an RTTM file is produced for each audio file with a list of onsets and offsets of overlapped regions if any.

B. *Speaker Diarization*

In this section, the proposed speaker diarization system, as shown in Figure 2, to extract speaker turns with speaker identities is explained. Since speaker diarization is not required for non-speech regions, a VAD (silero-vad) is applied to discard those prior to passing to the diarization system. Then the speech signals are converted to a set of features that can encode the idiosyncrasies of a speaker's voice (MFCCs).

The critical component of the pipeline is embedding extraction, where the speech segments are mapped into a multi-dimensional array so that intra-speaker variance is minimized and inter-speaker variance is maximized (i.e., same speaker embeddings are close by, whereas different speaker embeddings are far apart). We employ an off-the-shelf pre-trained ECAPA-TDNN [11] model from SpeechBrain[17], a PyTorch-based open-source speech processing toolkit, and extract 192-dimensional vectors, known as x-vectors, for each 1.5s segment with a 0.75s shifting.

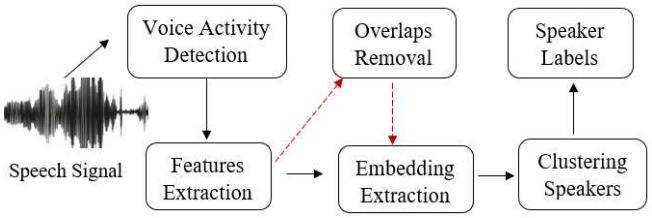

Fig. 2. Pipeline of our proposed Speaker Diarization model

The embeddings are analysed by Spectral Clustering [12], which estimates the number of speakers in an audio file using an eigengap heuristics and maps each embedding to a specific cluster (i.e., assigns a speaker identity to each embedding). Adjacent speech segments belonging to a particular speaker are merged, and the output is converted to an RTTM file which holds a list of speaker turns with speaker identities.

Extracting meaningful embedding is often hindered by factors such as background noises, frequent speaker turns, and overlapped speech, to name a few. This study also investigates the degradation in performance in speaker diarization caused by overlapped speech.

Conventional speaker diarization systems are likely to map an overlapped segment to the dominant speaker, increasing the missed speech rate from other speakers. Several studies have analysed the overlapped segments independently and identified multiple speakers. For example, Bullock et al. have assigned secondary speakers in overlapped regions using a posterior speaker matrix from VB-HMM re-segmentation [18].

Our study adopts another well-known method in which the identified overlapped regions are discarded from speaker diarization. Overlapped regions are identified using our OSD model, and those regions are excluded from speaker diarization using an un-partitioned evaluation map (UEM) proposed in the dscore toolkit [19]. UEM files are used to specify the scoring regions within each audio file.

C. *Tamil Test Dataset for Speaker Diarization*

Speaker diarization has been vastly explored in English; however, there is a dearth in other languages, especially in Sri Lankan national languages: Sinhala and Tamil. We implement a dataset for Tamil using 'in the wild' videos on YouTube using constraints similar to VoxConverse. Unlike VoxConverse, which uses a semi-automated pipeline, we annotate our dataset manually from scratch utilizing a tool called VGG Image Annotator (VIA) [20]. VIA is a lightweight, standalone, and offline software package designed to run in a web browser using HTML, Javascript, and CSS. With the help of this software, we can annotate spatial regions (e.g., the speaker turns) with ease and export the output to a CSV format. We implemented a script to convert the CSV data to RTTM format. A screenshot of the annotation process using the VIA is displayed in Figure 3.

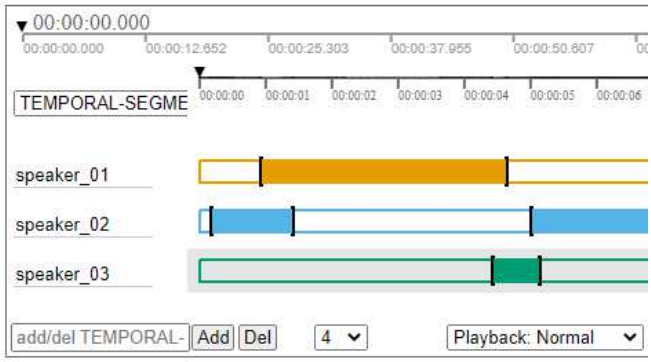

Fig. 3. A screenshot from the VGG Image Annotator while doing the manual annotation for the Tamil dataset. Temporal data of each speaker is recorded separately.

The videos are taken from multi-speaker conversations, including celebrity interviews, political debates, panel discussions, and talk shows which have taken place in challenging acoustic conditions such as environments with background noises including cross-talk, laughter, and applause. The statistics of the dataset is given in Table II.

TABLE II. STATISTICS OF OUR TAMIL TEST DATASET. ENTRIES THAT HAVE 3 VALUES ARE REPORTED AS MIN/MEAN/MAX.

Total Videos	Total Duration (min)	Number of Speakers	Duration (min)
40	367	2 / 4.4 / 9	5.05 / 9.17 / 15.15

III. EXPERIMENTS

This section provides particulars on the model configurations, experimental setups, and evaluation criteria. The OSD model is evaluated in VoxConverse (test) dataset, whereas the Speaker Diarizer is evaluated in both VoxConverse (test) and Tamil (test) datasets.

A. Configurations and Details

Both overlapped speech detection (OSD) model and speaker diarization system are implemented in PyTorch. In OSD model, the loss is calculated using categorical cross-entropy, and the gradients are updated using the Adam optimizer. StepLR learning rate scheduler from PyTorch decays the learning rate by 5% after every ten epochs. The model is trained for 60 epochs on the VoxConverse dev dataset. The batch size is set to be 512. Since a pre-trained model is used to extract embeddings for speaker diarization (as explained in II.B), we do not have any training for speaker diarization .

B. Experimental Setups

We report the performances of a speaker diarization system before and after excluding overlapped speech segments for two different scenarios: clustering based on the oracle number of speakers and eigengap-based estimation for the number of speakers. The "oracle number of speakers" is a technical term for finding the number of speakers in an audio file using the ground truth RTTM file. Moreover, the experiments are further expanded with no-collar and a standard collar size of 250ms. The collar is a forgiveness area at the onset and offset of a speaker turn due to annotation errors, and those regions are excluded from performance metrics calculations [21].

C. Evaluation Criteria

The performance of our OSD model on the VoxConverse test dataset is assessed using precision and recall as metrics. To evaluate the performances of our speaker diarization system variants, a metric known as Diarization Error Rate (DER) is used [22]. A perfect speaker diarization system would give a DER up to 0, whereas the DER can go beyond 100 if either the system is incapable of doing its job or the dataset is extremely wild. We use the NIST implementation of DER from the dscore library to calculate DER [19].

IV. Results and Discussion

As given in Table III, our overlapped speech detection (OSD) model gives precision and recall scores of 61.47 and 59.04, respectively. Since our model has been trained solely on the VoxConverse dev set, a performance hike would be expected if multiple datasets are combined or various types of data augmentation techniques are applied.

TABLE III. The performance of our overlapped speech detection (OSD) model in terms of precision and recall on the VoxConverse test set.

System	Dataset	Precision	Recall
Our Proposed OSD	VoxConverse (test)	61.47	59.04

Table IV compares the results of our speaker diarization system before and after applying OSD with no collar applied for both VoxConverse and Tamil datasets. We also report the results on clustering based on the oracle number of speakers and eigengap-based heuristics.

Before applying the OSD model, our speaker diarization system obtains diarization error rates (DERs) of 22.56 and 27.16 on the VoxConverse dataset for the oracle number of speakers and eigengap-based number of speakers, respectively, and similarly 21.25 and 25.16 on the Tamil dataset. In both datasets, there is a significant performance difference between the number of speakers estimation using oracle number of speakers and eigengap-based number of speakers. The oracle number of speakers in an audio file is calculated using the ground truth RTTM file, whereas, in real life, the number of speakers should be estimated. In our case, the optimal number of speakers is estimated using the eigengap approach [12].

TABLE IV. Performance comparision of our proposed Speaker Diarization system with zero collar before and after removing overlaps for the VoxConverse and Tamil datasets.

	Using Oracle Number of Speakers		Using Eigen-Gap Number of Speakers	
Dataset	Baseline	After OSD	Baseline	After OSD
VoxConverse (test)	22.56	21.31	27.16	25.73
Tamil	21.25	18.43	25.16	22.39

From here onwards, we only discuss the results of eigengap-based speaker estimation. From Table IV, it is evident that the performance of a speaker diarization system improves when overlapped speech are excluded from scoring since the DER of our speaker diarizer drops from 27.16 to 25.73 on the VoxConverse dataset, whereas from 25.16 to 22.39 on Tamil dataset. Since the OSD model is a deep learning model that needs a large dataset for better training, and we trained with a limited dataset, we could expect better performance when the performance of the OSD model is improved further.

A legitimate question is why there is a lower DER on Tamil Dataset than VoxConverse dataset although the embedding extractor has been only pre-trained in English, and the linguistic differences may impact the extractor to produce distinguishable embeddings. Certainly, the model can better encode speech signals according to speaker identities if the embedding extractor is trained on a Tamil dataset. However, the speaker diarizer's performance does not always depend on the embeddings but is further affected by the acoustic environments and the number of speakers present in that audio file. Although our Tamil dataset is curated using videos filmed in challenging acoustic environments to represent the real world, considering the difficulties in manual annotation, the maximum number of speakers were limited to 9, and the acoustic conditions are not extremely challenging.

The performance gain of our speaker diarizer after excluding overlapped regions is 5.3% and 11% for the VoxConverse and Tamil datasets, respectively. This could be due to our dataset's higher ratio of overlapped to single-speaker speech compared to the VoxConverse.

TABLE V. Performance comparision of our proposed Speaker Diarization system with 250ms collar before and after removing overlaps for the VoxConverse and Tamil datasets.

	Using Oracle Number of Speakers		Using Eigen-Gap Number of Speakers	
Dataset	Baseline	After OSD	Baseline	After OSD
VoxConverse (test)	16.63	14.87	21.83	20.16
Tamil	14.47	12.19	18.17	15.77

This paragraph explains the effects of a collar size of 250ms applied before and after applying the OSD model on the speaker diarizer. Each speaker turn is pruned with 250ms at the onset and offset, considering the inability of human ears to find the exact turning points while doing the manual annotation. Table V shows a DER drop from 27.16 to 21.83 and 25.16 to 18.17 for both VoxConverse and Tamil datasets, respectively. Excluding those regions, which are often the source of speaker confusion and overlapped speech, improves the performance of a speaker diarization system. After excluding overlapped regions, the DER of the diarizer drops from 21.83 to 20.61 and from 18.17 to 15.77 on VoxConverse and Tamil datasets, respectively.

V. Conclusions

This study proposed an overlapped speech detection (OSD) model that identified temporal segments where multiple speakers speak simultaneously. A novel idea has been proposed to label speech segments into single-speaker speech and overlapped speech with a CRNN to better encode local and global patterns. The model was trained on a publicly available English dataset called VoxConverse. The objective of this research was to improve the performance of a speaker diarization system by excluding overlapped regions as the performance degrades proportionally to the duration of overlapped speech. The evaluation was conducted on the VoxConverse test set and a locally developed Tamil dataset. The performance gain of the proposed speaker diarizer after incorporating the OSD model is 5.6% and 13.4% on VoxConverse test set and Tamil dataset, respectively, with a 250ms collar. Through our study, it is evident that detecting and excluding overlapped speech leads to improved performance in speaker diarization. Although the OSD model has been trained solely on English, the speaker diarization performance on the Tamil dataset after discarding overlaps is impressive. However, there is room to improve the performance further if the OSD model is fine-tuned on a Tamil dataset using transfer learning approaches to compensate for domain mismatch, which will be our future direction.

References

[1] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," Comput. Speech Lang., vol. 72, p. 101317, Mar. 2022, doi: 10.1016/j.csl.2021.101317.

[2] M. Kim, T. Ki, A. Anshu, and V. R. Apsingekar, "North America Bixby Speaker Diarization System for the VoxCeleb Speaker Recognition Challenge 2021," Sep. 2021, [Online]. Available: http://arxiv.org/abs/2109.13518

[3] F. Landini et al., "But System for the Second Dihard Speech Diarization Challenge," in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020, vol. 2020-May, no. Npu Ii, pp. 6529–6533. doi: 10.1109/ICASSP40776.2020.9054251.

[4] M. Yousefi and J. H. L. Hansen, "Frame-Based Overlapping Speech Detection Using Convolutional Neural Networks," in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020, vol. 2020-May, no. January, pp. 6744–6748. doi: 10.1109/ICASSP40776.2020.9053108.

[5] R. . Yantorno, K. R. Krishnamachari, J. . Lovekin, N. Streets, D. . Benincasa, and S. . Wenndt, "The Spectral Autocorrelation Peak Valley Ratio (SAPVR) – A Usable Speech Measure Employed as a Co-channel Detection System," IEEE Int. Work. Intell, Signal Process., no. January 2001, pp. 2–6, 2001.

[6] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, Mar. 2008, pp. 4353–4356. doi: 10.1109/ICASSP.2008.4518619.

[7] J. T. Geiger, F. Eyben, B. Schuller, and G. Rigoll, "Detecting overlapping speech with long short-term memory recurrent neural networks," in Interspeech 2013, Aug. 2013, no. August, pp. 1668–1672. doi: 10.21437/Interspeech.2013-27.

[8] J. Jung, H.-S. Heo, Y. Kwon, J. S. Chung, and B.-J. Lee, "Three-Class Overlapped Speech Detection Using a Convolutional Recurrent Neural Network," in Interspeech 2021, Aug. 2021, vol. 4, pp. 3086–3090. doi: 10.21437/Interspeech.2021-149.

[9] Z.-H. Tan, A. kr Sarkar, and N. Dehak, "rVAD: An unsupervised segment-based robust voice activity detection method," Comput. Speech Lang., vol. 59, pp. 1–21, Jan. 2020, doi: 10.1016/j.csl.2019.06.005.

[10] R. Wang, R. Tong, Y. T. Yeung, and X. Chen, "The HUAWEI Speaker Diarisation System for the VoxCeleb Speaker Diarisation Challenge," Oct. 2020, [Online]. Available: http://arxiv.org/abs/2010.11657

[11] B. Desplanques, J. Thienpondt, and K. Demuynck, "ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification," in Interspeech 2020, Oct. 2020, vol. 2020-Octob, pp. 3830–3834. doi: 10.21437/Interspeech.2020-2650.

[12] Q. Wang, C. Downey, L. Wan, P. A. Mansfield, and I. L. Moreno, "Speaker Diarization with LSTM," 2018 IEEE Int. Conf. Acoust. Speech Signal Process., vol. 2018-April, pp. 5239–5243, Oct. 2017, doi: 10.1109/ICASSP.2018.8462628.

[13] J. S. Chung, J. Huh, A. Nagrani, T. Afouras, and A. Zisserman, "Spot the Conversation: Speaker Diarisation in the Wild," in Interspeech 2020, Oct. 2020, vol. 2020-Octob, pp. 299–303. doi: 10.21437/Interspeech.2020-2337.

[14] J. Carletta et al., "The AMI Meeting Corpus: A Pre-announcement," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 3869 LNCS, 2006, pp. 28–39. doi: 10.1007/11677482_3.

[15] Silero Team, "Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier," GitHub, 2021. https://github.com/snakers4/silero-vad

[16] R. Vs, "The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan," Evaluation, vol. 2009, pp. 1–18, 2009.

[17] M. Ravanelli et al., "SpeechBrain: A General-Purpose Speech Toolkit," pp. 1–34, Jun. 2021, [Online]. Available: http://arxiv.org/abs/2106.04624

[18] L. Bullock, H. Bredin, and L. P. Garcia-Perera, "Overlap-Aware Diarization: Resegmentation Using Neural End-to-End Overlapped Speech Detection," in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020, vol. 2020-May, pp. 7114–7118. doi: 10.1109/ICASSP40776.2020.9053096.

[19] "Diarization scoring tools," GitHub, 2019. https://github.com/nryant/dscore

[20] A. Dutta and A. Zisserman, "The VIA Annotation Software for Images, Audio and Video," in Proceedings of the 27th ACM International Conference on Multimedia, Oct. 2019, pp. 2276–2279. doi: 10.1145/3343031.3350535.

[21] J. Kalda and T. Alumäe, "Collar-Aware Training for Streaming Speaker Change Detection in Broadcast Speech," in The Speaker and Language Recognition Workshop (Odyssey 2022), Jun. 2022, pp. 141–147. doi: 10.21437/Odyssey.2022-20.

[22] J. G. Fiscus, J. Ajot, and J. S. Garofolo, "The rich transcription 2007 meeting recognition evaluation," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 4625 LNCS, pp. 373–389, 2008, doi: 10.1007/978-3-540-68585-2_36.