Joint speech and overlap detection: a benchmark over multiple audio setup and speech domains

Martin Lebourdais, Théo Mariotte, Marie Tahon, Anthony Larcher, Antoine Laurent, Silvio Montresor, Sylvain Meignier, Jean-Hugh Thomas

Joint speech and overlap detection: a benchmark over multiple audio setup and speech domains

Martin Lebourdais[1⋆], *Théo Mariotte*[1,2⋆], *Marie Tahon*[1], *Anthony Larcher*[1], *Antoine Laurent*[1], *Silvio Montrésor*[2], *Sylvain Meignier*[1], *Jean-Hugh Thomas*[2]

[1]LIUM, [2]LAUM UMR 6613 IA-GS, Le Mans Université

{first name}.{last name}@univ-lemans.fr

Abstract

Voice activity and overlapped speech detection (respectively VAD and OSD) are key pre-processing tasks for speaker diarization. The final segmentation performance highly relies on the robustness of these sub-tasks. Recent studies have shown VAD and OSD can be trained jointly using a multi-class classification model. However, these works are often restricted to a specific speech domain, lacking information about the generalization capacities of the systems. This paper proposes a complete and new benchmark of different VAD and OSD models, on multiple audio setups (single/multi-channel) and speech domains (e.g. media, meeting...). Our 2/3-class systems, which combine a Temporal Convolutional Network with speech representations adapted to the setup, outperform state-of-the-art results. We show that the joint training of these two tasks offers similar performances in terms of F1-score to two dedicated VAD and OSD systems while reducing the training cost. This unique architecture can also be used for single and multi-channel speech processing.

1. Introduction

Speaker diarization answers the question *Who spoke and when?* in an audio stream. Today, this task remains difficult as shown by the numerous challenges recently organized [1, 2, 3].

Given an audio stream, speaker diarization pipelines generally address speech segmentation and speaker clustering in two distinct stages [1].

Therefore, robust speech segmentation - mainly Voice Activity Detection (VAD) and Overlapped Speech Detection (OSD) - is essential to improve speaker diarization performance as shown in previous studies [4, 5]. VAD consists in segmenting an audio signal into speech and non-speech segments.

Several approaches have been proposed in the literature such as signal processing methods [6], statistical models [7], and neural-based approaches [8]. OSD detects segments in which at least two speakers are simultaneously active. Early studies mostly focus on statistical models [9, 10] while recent approaches are mostly based on neural networks [5, 11, 12] and show promising results.

While VAD and OSD have mainly been considered as two independent binary classification tasks, they can be addressed jointly by considering three classes – non-speech, single speaker, and overlapped speech – according to the number of present speakers in each speech segment. In [13], such a 3-class problem is solved by training a recurrent convolutional network. The use of far-field microphones and a Self-Attention Channel Combinator (SACC) feature extractor [14] revealed the potential of spatial information for OSD. [15] demonstrated that Temporal Convolutional Network (TCN) is well adapted for multiple-speaker activity detection with far-field microphones.

In this paper, we propose two 2-class VAD and OSD and 3-class VAD+OSD for mono and multi-channel signals. We evaluate how beneficial is the 3-class approach in comparison to the use of two independent VAD and OSD models in terms of F1-score and training resources. Each system is trained and evaluated on four different datasets covering various speech domains including both single and multiple microphone scenarios. To the best of our knowledge, no benchmark has been conducted on these approaches across various speech domains and recording setups (multi/mono channel) in the literature.

This paper presents several contributions: It first claims a new state of the art for OSD on multiple corpora, it introduces a benchmark on 4 different datasets covering various speech domains in multi/mono channel scenarios and presents a reduction of the training cost using a 3-class approach with a detailed analysis of the benefits of this system.

2. Datasets

Our benchmark datasets combine multiple speech domains including far-field audio recordings. For each dataset, VAD and OSD labels are derived from the provided ground-truth segmentation. Table 1 summarizes corpus characteristics.

Table 1: *Corpus characteristics.* ⋆*multi-microphone data.*

Corpus	Domain	Duration	Overlap prop.
DIHARD	Multiple	34 h	11.6%
ALLIES	Media	328 h	3.2%
ALLIES-clean	Media	6 h	13.9%
AMI⋆	Meeting	100 h	24.7%
CHiME-5⋆	Dinner party	60 h	22.9%

2.1. Single Channel

Single channel experiments are conducted on 3 datasets: ALLIES [16], DIHARD [1] and AMI [17]. The ALLIES corpus is a soon-to-be-available French meta-corpus designed to gather and extend previous French data collected for diarization and transcription evaluation campaigns. It consists of 328 h of audio extracted from 1998 to 2014 in 1008 shows with 5901 different speakers. The overlap proportion (in duration) fluctuates widely between broadcast news with little to no interaction and debates (around 10% of overlaps). Despite a harmonization effort, the data collected and annotated under different protocols introduces some homogeneity problems [18]. 15 debate shows, referred to as ALLIES-clean, were selected in order to get a high overlap proportion, a manual and homogeneous speech segmen-

⋆ Both authors contributed equally.

tation, and diversity in the shows represented.

The DIHARD corpus contains data from 7 domains with various recording qualities, situations, and degrees of spontaneity from read speech to phone conversations. Since spontaneous speech naturally contains a high proportion of overlapped speech, this corpus is well-suited for OSD. This corpus is partitioned as intended for the challenge and evaluated on the official evaluation partition.

The AMI meeting corpus contains recordings of realistic meetings involving up to 5 participants in various environments. The *headset-mix* is used for single-channel experiments on this dataset. The data partition follows the protocol proposed in [19].

2.2. Multiple Channels

Multiple-channel experiments are conducted on 2 corpora: AMI [17] and CHiME-5 [20]. We select AMI audio data captured by the *Array 1* as a distant multi-microphone signal. It consists of a uniform circular array (UCA) composed of 8 omnidirectional microphones placed in the center of the table during meetings.

The CHiME-5 dataset contains 20 dinner-party sessions involving 4 participants in a real-home environment. Speakers were asked to move between 3 rooms during the party. Audio signals thus feature a strong background noise diversity with varying acoustic conditions. Audio signals are captured with 6 linear arrays composed of 4 microphones. For our experiments, only the first microphone of each array is selected. Finally, the resulting signal contains 6 channels.

3. System overview

Figure 1 depicts an overview of VAD, OSD, and VAD+OSD systems. While the feature extractor (in blue) is adapted with respect to the number of input channels, the sequence modeling network (in purple) processes the sequence of features before the frame classification. The frame classification is done at a rate of 100 Hz, while the raw waveform is sampled at 16 kHz.

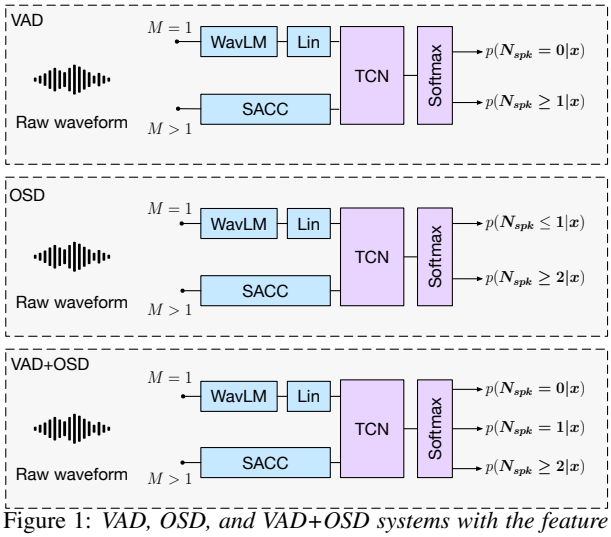

Figure 1: *VAD, OSD, and VAD+OSD systems with the feature extractor (blue) and the sequence modeling network (purple), M is the number of channels.*

3.1. Single channel features ($M = 1$)

The single channel feature extractor is based on the WavLM pre-trained model [21]. This choice is motivated by the performance obtained on the diarization task according to the SUPERB benchmark [22]. Furthermore, WavLM has been trained using simulated overlapped speech and is then more robust to this type of data. WavLM outputs speech representations every 20 ms. In order to align this representation with the target sequence, we decide to add a linear layer on top of the frozen WavLM. The linear layer aims to transform a segment of 99 features extracted with WavLM over a 2 s window, into a 200-frame vector, supposedly aligned with our target.

3.2. Multiple channel features ($M > 1$)

When multiple channels are available, feature extraction is performed using the Self-Attention Channel Combinator (SACC) [23]. This architecture has previously shown its efficiency for OSD under distant speech conditions [14]. The algorithm consists of a self-attention module [24] which computes per-channel weights from the multi-channel Short-Time Fourier Transform (STFT) of the input signal. The channels are then weighted and combined in order to get a single-channel representation. Combination weights are computed from the multi-channel STFT calculated on 25 ms segments with 10 ms shift. The attention module is composed of a single attention head of size $d = 256$. The combined representation is converted to the log-mel scale using $N_f = 64$ filters. Global Mean and Variance Normalization (MVN) is also applied before feeding the sequence modeling network.

3.3. Sequence modeling and classification

The sequence modeling network (in purple in Fig. 1) takes as input a sequence $\boldsymbol{x}$ of single or multi-channel features and assigns a class to each frame of this sequence. This task is performed using a TCN [27] since this architecture has shown noticeable results on both VAD and OSD tasks [12, 14, 15, 25]. It is composed of 5 residual convolutional blocks repeated 3 times. Classification is performed by a 1-d convolutional layer followed by a $\mathrm{softmax}$ activation function.

For each frame in the output sequence, the VAD outputs the pseudo-probability to contain at least one speaker $p(N_{spk} \geq 1|\boldsymbol{x})$. The OSD outputs the pseudo-probability to contain speech from more than one speaker $p(N_{spk} \geq 2|\boldsymbol{x})$. Both VAD and OSD are then binary classifiers. The joint VAD+OSD system outputs the pseudo-probability of either containing any speech $p(N_{spk} = 0|\boldsymbol{x})$, speech from a single speaker $p(N_{spk} = 1|\boldsymbol{x})$, or speech from more than one speaker $p(N_{spk} \geq 2|\boldsymbol{x})$. The 3-class approach is then converted to 2-class VAD and OSD by merging the relevant classes.

3.4. Training and Evaluation

In order to estimate the robustness over different speech domains, the three systems are trained and evaluated independently on the 5 datasets aforementioned. To counteract the small number of overlap segments, 50% of the training segments are augmented on-the-fly by summing them to another randomly sampled training segment. Associated labels of each segment are also combined [28]. The loss function is a cross-entropy, and we used the ADAM optimizer with a learning rate of $lr = 10^{-3}$. Single-channel audio data is augmented with noise extracted from MUSAN [29] and additional reverberation using simulated room impulse responses. Preliminary experiments have shown that data augmentation did not bring significant improvement in the far-field scenario.

Following the DIHARD evaluation plan, we use the F1-score obtained on the evaluation set as a performance metric. In

Table 2: *Overview of the F1-score (%) for each system on the evaluation set of several corpora covering various domains, ⋆ indicates multi-microphone data, † indicates that the results are taken from the original article.*

		VAD					OSD				
		DIHARD	ALLIES	AMI	AMI⋆	CHiME⋆	DIHARD	ALLIES	AMI	AMI⋆	CHiME⋆
2-class	VAD (ours)	97.0	99.8	97.4	96.4	99.8	-	-	-	-	-
	OSD (ours)	-	-	-	-	-	**66.2**	**71.6**	**79.6**	**72.2**	**75.9**
	Mel+CRNN [15]	-	-	-	-	-	51.3	-	66.0	57.2	-
	Mel+TCN [25]	-	-	-	-	-	54.7	-	73.4	65.8	-
3-class	VAD+OSD (ours)	97.0	89.2	97.2	96.6	99.3	**66.8**	**75.4**	**80.4**	**71.8**	**75.5**
	Mel+CRNN [15]	-	-	-	-	-	50.8	-	69.6	61.2	-
	Mel+TCN [25]	-	-	-	-	-	54.5	-	73.8	67.9	-
	SincNet+BLSTM [26]†	-	-	-	-	-	59.9	-	75.3	-	-

the 2-class approach, only the positive class output ($N_{spk} \geq 1$ for VAD, and $N_{spk} \geq 2$ for OSD) is used for prediction and two detection thresholds are applied to predict binary labels [28]. In the 3-class approach, the class associated with the maximum $\mathrm{softmax}$ output is selected at the frame level. A working version of the code will soon be released[1].

4. Results

OSD and VAD results obtained on 5 single (DIHARD, ALLIES, AMI) and multi-channels (AMI⋆, CHIME-5⋆) datasets are presented in Table 2.

4.1. Single Channel

So far, ALLIES corpus has only been studied for speaker diarization while discarding overlapping speech obtained in the manual reference [16]. We provide the first evaluation of OSD for ALLIES data with a 71.6% F1-score using the 2-class approach.

VAD performances are similar between the 2- and 3-class approaches except for the ALLIES data for which a strong F1-score degradation of 10.6% is noticeable. The 3-class approach improves OSD results in all single-channel datasets, particularly on ALLIES (+5.3%). Results on ALLIES should be treated cautiously as the average proportion of overlap is rather low, and we identified some issues in the manual segmentation. In summary, except for ALLIES, the joint VAD+OSD system offers better performance than the two dedicated systems. It even outperforms the previous state-of-the-art results on DIHARD and AMI data with a new F1-score at 66.8% and 80.4% respectively.

4.2. Domain adaptation

Table 3: *VAD and OSD F1-score (%) obtained on the ALLIES evaluation set. The model trained on DIHARD is fine-tuned on the subset ALLIES-clean.*

Model	Task	DIHARD	ALLIES
Fine-tuning		*ALLIES-clean*	*No*
2-class	VAD	99.7	**99.8**
	OSD	75.3	71.6
3-class	VAD	**99.8**	89.2
	OSD	75.0	**75.4**

[1]Hidden link for anonymous submission

The presence of errors in the reference segmentation of ALLIES introduces some noise during the training stage, and thus, degrades the performance of the 3-class approach especially regarding VAD. To cope with this issue, we propose to use the model trained on DIHARD and fine-tune it with the clean subset ALLIES-clean. Table 3 shows that fine-tuning on ALLIES-clean brings a similar OSD performance (75.0%) as a model trained with ALLIES data only (75.4%). More interestingly, fine-tuning significantly improves the VAD performance with a relative +11.9% gain on the F1-score, with only 6 h of in-domain speech. This gain can be explained by the diversity and quality of the annotations in DIHARD. We conclude that it is better to train the model on clean and diverse data and apply fine-tuning on in-domain data.

4.3. Multiple channels

On the AMI meeting corpus, we notice lower performances on multi-channel data AMI⋆ in comparison to the close-talk recordings of AMI. Two factors can explain this degradation. First, multi-channel signals are recorded under distant speech conditions. This leads to lower quality recordings and thus performance degradation [14]. Moreover, unlike single channels, the multi-channel feature extraction algorithm does not rely on pre-trained features. Therefore, the SACC features are less optimized compared to WavLM features. On AMI⋆, the joint VAD+OSD system offers similar VAD performance as the 2-class approach. The same behavior is observed on the OSD task where the 3-class system degrades with a 0.5% relative F1-score degradation. A single 3-class VAD+OSD system thus offers similar performance as two dedicated VAD and OSD systems on multi-channel audio from AMI⋆. VAD and OSD performance are also evaluated on audio data recorded during dinner parties with the CHiME-5 dataset. Again, the 3-class VAD+OSD system offers similar VAD and OSD performance as two dedicated VAD and OSD systems with about 0.5% relative F1-score degradation on each task. Results on these two multi-microphone datasets show that joint VAD+OSD is also adapted to the distant speech scenario with SACC features.

5. Analysis

This section evaluates the benefits of such an approach in terms of training time, speech domains, and spatial information in the multi-channel scenario.

5.1. Training time

In order to assess the value of training a joint VAD+OSD system against two dedicated models, we compare the training time required for each approach. Each system is trained on an RTX6000 GPU card until it reaches its best F1-score on the validation set. Figure 2 presents the elapsed time to obtain the best-performing model.

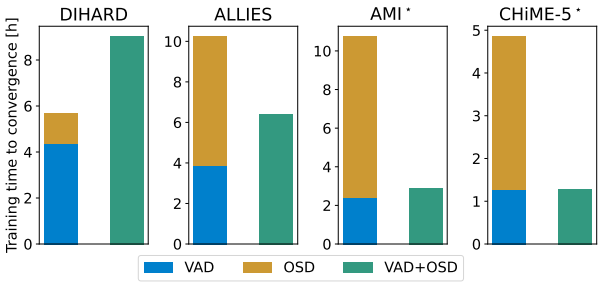

Figure 2: *Training time for each system to converge on two single-channel and two multi-channel datasets.*

2-class OSD task clearly requires more resources than VAD. Indeed the discrimination of the spectral information between the presence of one speaker or several speakers is more difficult than between speech and non-speech signals. Multi-channel VAD+OSD system converges as fast as the 2-class VAD system, as observed on the AMI* and CHiME-5* datasets. In the single channel scenario, the gain is less significant (and no gain at all with DIHARD), probably because the spatial information helps to detect multiple speakers.

5.2. Influence of the speech domain on performance

In order to study the influence of the speech domain on OSD, we analyze the OSD F1-score distributions for each of the DIHARD evaluation files, manually separated into 7 domains (see Fig. 3). *Clinical* contains conversations between a clinician and a child, *facetoface* contain interviews, *phone* contains phone conversations, *map task* contains a game in which someone guides a person remotely on a map, *group chat* contains spontaneous conversations, *court* contains court recordings and *audiobook* contains read speech.

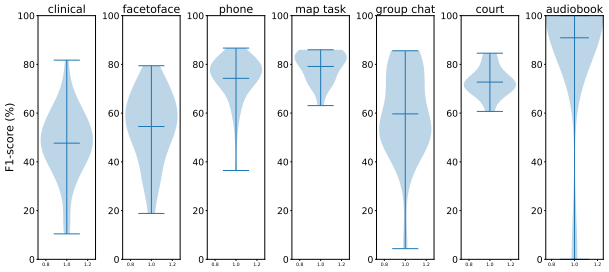

Figure 3: *Distribution of F1-scores on DIHARD speech domains*

Fig. 3 shows that the F1-score is globally better for *phone* conversations than for *clinical* and *face-to-face* conversations, despite the fact that the three domains are dyadic interactions. We can then hypothesize that the absence of visual cues in phone conversations limits the diversity of overlaps contained in the audio files. Another difference between domains is the quality of the recordings. For example, *group chat* and *face-to-face* files feature strong background noise and low-quality recordings, which could explain the low performance obtained in these domains. This analysis concludes that the speech domain is of major importance for OSD. The presence of noise, the diversity of overlaps, and the differences in turn-taking driven by the speech domain is clearly a major issue for OSD.

5.3. Spatialisation

In the CHiME-5 dataset, the rooms where participants are located are annotated for each utterance in the evaluation set. We can thus study which microphone the SACC feature extractor activates as a function of the speakers' positions. Since the VAD+OSD system is trained using one microphone per array in the CHiME data, we can visualize the combination weights for each array in each room. Two arrays are located in the kitchen, two are located in the dining area and two are in the living room. Figure 4 shows the SACC combination weights of each channel depending on the location of the speakers. On these utterances, the SACC system mostly activates the channels placed in the areas where speakers are located. The system seems able to select microphones with the most information for the VAD+OSD task. An in-depth study should however be conducted to better assess the information used by the system in the multiple-channel scenario.

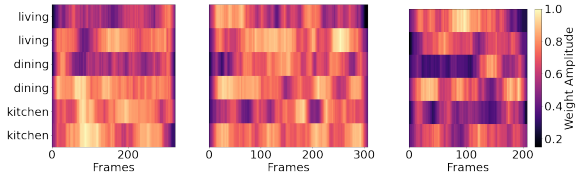

Figure 4: *Combination-weights applied to each channel depending on the room where speakers are staying on 3 CHiME-5 utterances. (Left) Kitchen, (Middle) Dining room, (Right) Living room. Utterance-wise normalization is applied for better visualization by dividing weights by their maximum value.*

6. Conclusion

This article presents a benchmark on two speech segmentation tasks – Voice Activity Detection and Overlapped Speech Detection – over multi/mono channel and various domains in 5 datasets. Two approaches are compared by solving jointly or independently VAD and OSD. The VAD+OSD joint training offers similar performance as the traditional 2-class OSD or VAD approaches on both single-channel audio data and distant multi-microphone signals. The proposed system reaches a new state-of-the-art for OSD on DIHARD (66.8%) and AMI (80.4%) data. Particularly in the case of ALLIES data with domain adaptation, joint training brings an improvement of +11.8% for VAD and +5.3% for OSD. Furthermore, joint training requires fewer resources as it reduces the training time on most of the datasets, especially in the case of multi-channel data.

A deeper analysis demonstrates that background noise and face-to-face conversations are clearly hard to segment. We also visualize how the combination weights obtained with the SACC multi-channel feature extractor are prone to locate active speakers within a session.

Since VAD and OSD performances on multi-microphone data highly depend on the number of microphones during training, we intend to evaluate our system in a cross-domain scenario with different types of adaptation to go towards a robust multi-corpus segmentation model. The impact of the proposed VAD+OSD system on diarization will also be evaluated.

7. Acknowledgments

This work was performed using HPC resources from GENCI–IDRIS (Grant 2022-AD011012565), the French ANR GEM (ANR-19-CE38-0012), and LMAC grant from Région Pays de la Loire.

8. References

[1] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, and M. Liberman, "The Second DIHARD Diarization Challenge: Dataset, Task, and Baselines," in *Proc. Interspeech 2019*, 2019, pp. 978–982.

[2] F. Yu, S. Zhang, Y. Fu, L. Xie *et al.*, "M2met: The icassp 2022 multi-channel multi-party meeting transcription challenge," in *ICASSP*. IEEE, 2022, pp. 6167–6171.

[3] M. Zelenák, H. Schulz, and J. Hernando, "Speaker diarization of broadcast news in albayzin 2010 evaluation campaign," *EURASIP Journal on Audio Speech and Music Processing*, vol. 19, Jul 2012.

[4] L. P. Garcia Perera, J. Villalba *et al.*, "Speaker Detection in the Wild: Lessons Learned from JSALT 2019," in *The Speaker and Language Recognition Workshop (Odyssey 2020)*, 2020, pp. 415–422.

[5] L. Bullock, H. Bredin, and L. P. Garcia-Perera, "Overlap-Aware Diarization: Resegmentation Using Neural End-to-End Overlapped Speech Detection," in *ICASSP*, 2020, pp. 7114–7118.

[6] H. Ghaemmaghami, B. Baker, R. Vogt, and S. Sridharan, "Noise robust voice activity detection using features extracted from the time-domain autocorrelation function," in *11th Annual Conference of the ISCA*, 2010, pp. 3118–3121.

[7] E. Nemer, R. Goubran, and S. Mahmoud, "Robust voice activity detection using higher-order statistics in the lpc residual domain," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 217–231, 2001.

[8] M. Lavechin, M.-P. Gill, R. Bousbib, H. Bredin, and L. P. Garcia-Perera, "End-to-End Domain-Adversarial Voice Activity Detection," in *Interspeech 2020*. ISCA, Oct. 2020, pp. 3685–3689.

[9] K. Boakye, O. Vinyals, and G. Friedland, "Improved overlapped speech handling for speaker diarization," in *Interspeech*, 2011, pp. 941–944.

[10] S. H. Yella and H. Bourlard, "Overlapping speech detection using long-term conversational features for speaker diarization in meeting room conversations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1688–1700, 2014.

[11] V. Andrei, H. Cucu, and C. Burileanu, "Detecting overlapped speech on short timeframes using deep learning." in *Interspeech*, 2017, pp. 1198–1202.

[12] M. Lebourdais, M. Tahon, A. Laurent, and S. Meignier, "Overlapped speech and gender detection with WavLM pre-trained features," in *Proc. Interspeech 2022*, 2022, pp. 5010–5014.

[13] J.-w. 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, pp. 3086–3090.

[14] T. Mariotte, A. Larcher, S. Montrésor, and J.-H. Thomas, "Microphone Array Channel Combination Algorithms for Overlapped Speech Detection," in *Proc. Interspeech 2022*, 2022, pp. 4636–4640.

[15] S. Cornell, M. Omologo, S. Squartini, and E. Vincent, "Overlapped speech detection and speaker counting using distant microphone arrays," *Computer Speech & Language*, vol. 72, p. 101306, 2022.

[16] A. Larcher, A. Mehrish, M. Tahon, S. Meignier, J. Carrive, D. Doukhan, O. Galibert, and N. Evans, "Speaker Embedding For Diarization Of Broadcast Data In The ALLIES Challenge," in *ICASSP*, Toronto, Canada, 2021, pp. 5799–5803.

[17] I. Mccowan, G. Lathoud, M. Lincoln, A. Lisowska *et al.*, "The ami meeting corpus," in *In: Proceedings Measuring Behavior 2005, 5th International Conference on Methods and Techniques in Behavioral Research.*, 2005.

[18] M. Lebourdais, M. Tahon, A. Laurent, S. Meignier, and A. Larcher, "Overlaps and Gender Analysis in the Context of Broadcast Media," in *LREC 2022*, Marseille, France, Jun. 2022.

[19] F. Landini, S. Wang, M. Diez, L. Burget *et al.*, "But system for the second dihard speech diarization challenge," in *ICASSP*, 2020, pp. 6529–6533.

[20] J. Barker, S. Watanabe, E. Vincent, and J. Trmal, "The fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, task and baselines," in *Interspeech 2018 - 19th Annual Conference of the International Speech Communication Association*, Hyderabad, India, Sep. 2018.

[21] S. Chen, C. Wang *et al.*, "WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing," *IEEE Journal of Selected Topics in Signal Processing*, pp. 1–14, 2022.

[22] S.-W. Yang, P.-H. Chi *et al.*, "SUPERB: Speech Processing Universal PERformance Benchmark," in *Interspeech 2021*, 2021, pp. 1194–1198.

[23] R. Gong, C. Quillen, D. Sharma, A. Goderre *et al.*, "Self-Attention Channel Combinator Frontend for End-to-End Multichannel Far-Field Speech Recognition," in *Interspeech*, 2021, pp. 3840–3844.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit *et al.*, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17, 2017, p. 6000–6010.

[25] S. Cornell, M. Omologo, S. Squartini, and E. Vincent, "Detecting and Counting Overlapping Speakers in Distant Speech Scenarios," in *Interspeech*, 2020, pp. 3107–3111.

[26] H. Bredin and A. Laurent, "End-to-end speaker segmentation for overlap-aware resegmentation," in *Interspeech*, Brno, Czech Republic, 2021.

[27] S. Bai, J. Z. Kolter, and V. Koltun, "An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling," *arXiv:1803.01271 [cs]*, 2018.

[28] H. Bredin, R. Yin, J. M. Coria, G. Gelly *et al.*, "Pyannote.Audio: Neural Building Blocks for Speaker Diarization," in *ICASSP*, 2020, pp. 7124–7128.

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