ALLIES: a Speech Corpus for Segmentation, Speaker Diarization Speech Recognition and Speaker Change detection

Marie Tahon, Anthony Larcher, Martin Lebourdais, Bougares Fethi, Ana Silnova, Pablo Gimeno

ALLIES: a Speech Corpus for Segmentation, Speaker Diarization, Speech Recognition and Speaker Change detection

Marie Tahon[1], Anthony Larcher[1], Martin Lebourdais[1], Fethi Bougares[2], Anna Silnova[3], Pablo Gimeno[4],
[1]LIUM, Le Mans Université, France, [2]Elyadata,
[3]Brno University of Technology, Speech@FIT, Brno, Czechia
[4]VivoLab, Universidad de Zaragoza, Spain
{marie.tahon; anthony.larcher}@univ-lemans.fr

Abstract

This paper presents a new release of ALLIES, a meta corpus which gathers and extends existing French corpora collected from radio and TV shows. The corpus contains 1048 audio files for about 500 hours of speech. Agglomeration of data is always a difficult issue, as the guidelines used to collect, annotate and transcribe speech are generally different from one corpus to another. ALLIES intends to homogenize and correct speaker labels among the different files by integrated human feedback within a speaker verification system. The main contribution of this article is the design of a protocol in order to evaluate properly speech segmentation (including music and overlap detection), speaker diarization, speech transcription and speaker change detection. As part of it, a test partition has been carefully manually 1) segmented and annotated according to speech, music, noise, speaker labels with specific guidelines for overlap speech, 2) orthographically transcribed. This article also provides as a second contribution baseline results for several speech processing tasks.

Keywords: segmentation, music, noise, overlap, speaker diarization, transcription, media

1. Introduction

In the last few years, multimedia data collection has shown a noticeable increase. In this context, archivists and media producers require storing and indexing these large-scale databases with various characteristics from the segmentation of the audio signal to speaker labeling. Automatic annotation is one of the most straightforward approaches to reduce the costs of manual annotation. More specifically, automatic speaker diarization task answers the question "who speaks when ?". It is a key component for many speech technologies such as automatic speech recognition (ASR) (Mao et al., 2020), speaker identification, and dialog monitoring in different multi-speaker scenarios, including TV/radio, meetings, and medical conversations. Some recent works also explore the joint prediction of the linguistic transcription and speaker diarization (Kanda et al., 2022) or at least speaker change detection (Anidjar et al., 2023). However in order to evaluate all these systems, there is a need for large scaled speech databases segmented and annotated with speaker labels and transcribed.

ALLIES corpus[1] was originally designed for lifelong human-assisted speaker diarization (Shamsi et al., 2022). The corpus has already been registered in ELRA catalogue[2] It is a French meta-corpus that gathers and extends previous French TV/radio data collected for diarization and transcription evaluation campaigns. This collection covers a wide range of shows, *i.e.* TV or radio programs. Collected between 1998 and 2014 by INA, who is in charge of the Legal Deposit for national TV and radio channels, this data is of relatively high quality (16kHz 16bit).

However, the initial version faces strong issues regarding annotations. The data originally collected has been annotated with dedicated guidelines specific to each evaluation campaign. Despite a harmonization effort, the differences in these guidelines introduce some heterogeneity problems, especially regarding speech segmentation and overlaps (Lebourdais et al., 2022). In addition, the orthography of speaker names is not consistent across corpora and shows. Consequently diarization accros shows, using speaker IDs shared among the different shows, cannot be investigated on this corpus, similarly to most of diarization corpora. Finally, segmentation is limited to the presence/absence of speech in the recording. Recent developments in audio segmentation (Gimeno et al., 2020) suggest enlarging to other sound events such as noise and music.

Even though ALLIES corpus has already been presented for lifelong diarization, the present work differs in various aspects. While most available speech databases are designed for ASR, including some of the original datasets present in ALLIES,

[1] https://lium.univ-lemans.fr/corpus-allies/
[2] https://www.islrn.org/resources/397-116-696-859-2/

Partition	Speech (h)	Trans. (h)	# files
Train	282.90	162.46	546
DiarTest-SeenShows	106.26	0	181
DiarTest-UnseenShows	88.87	5.30	286
FullTest-CleanAnnotations	17.95	10.34	35
total	495.90	178.50	1048

Table 1: ALLIES partitions for segmentation, diarization and transcription.

speaker verification, or diarization, the present work proposes a unified partition for all these tasks. Speaker labels have been unified with unique IDs, thus enabling speaker linkage across the entire collection. This should allow cross-show diarization in the future. A subset of 35 recordings (files) covering a wide range of shows including debates, has been carefully transcribed and segmented in terms of speech, overlap, noise, and music. This cleaning process allows evaluation of a wide range of technologies on the same set of audio and textual data. This article presents a new release of ALLIES, an audio and textual speech corpus designed to set up speaker diarization and speech recognition systems. The data preparation has been done for JSALT 2023, and ALLIES corpus was used in XDiar project to develop interpretable models for speaker diarization performed on both audio and textual transcription.

2. Data collection

ALLIES corpus is a French broadcast media corpus that extends previous evaluation campaigns (Larcher et al., 2016), (Shamsi et al., 2022). It gathers ESTER1&2 (Galliano et al., 2006) which includes EPAC (Esteve et al., 2010), REPERE (Giraudel et al., 2012), ETAPE (Gravier et al., 2012) and new data.

2.1. Content

The entire corpus consists of almost 500 hours of speech extracted from 1998 to 2020 in 1048 recordings (files) with 7188 different speakers. In total, 7 radio stations and 4 TV channels were collected.
Six radio channels namely France Inter, France Info, Radio France International (RDI), Radio Télévision Marocaine (RTM), France Culture and Radio Classique come from ESTER/EPAC databases (1998-2004). The last one being Africa1 comes from another dataset. Such data mainly consists of broadcast news, but also debates (France Inter *Le Téléphone sonne*, street interviews, music (Radio Classique), advertisements and jingles. BFM and LCP TV channels from the REPERE database (2012-2014) are multimodal data with a wide range of difficulties in both audio and video modalities. These recordings include political debates, spontaneous speech. ETAPE data (2010-2011) focuses on real-life TV material from LCP and TV8 channels, plus radio debates (France Inter) with various degree of spontaneity and overlapping speech with a wide range of background noises.
Additional data still collected by INA from more recent episodes of existing shows until 2014 has been included in ALLIES corpus. Speakers are identified with their names when possible, and else with a unique ID related with the following naming: `speaker#4348` This new data have been precisely manually annotated for overlapping speech along with the speaker segmentation. In the new set, speech segments where two speakers are active were labeled with the IDs of these two speakers, while segments involving three or more are labeled with *+3*.

2.2. Partitions

The initial version of ALLIES addressed lifelong human-assisted speaker diarization (Shamsi et al., 2022). Consequently a specific evaluation protocol has been proposed in the eponymous challenge in which the train/dev/test sets were set chronologically into three disjoint parts. In this work, we propose a new partition for segmentation, diarization and transcription described in Table 1. This partition cares about shows, but neither chronology, nor original databases from where the data comes. In the current version, we impose that the train (+ development) and test sets can be used for both speaker diarization and transcription.

Train The train set contains the 546 files partially transcribed. Almost 60% comes with a manual transcription. This train set can be used to train either a French ASR system, or speaker verification models, or to train or tune some parts of the diarization system.

FullTest-CleanAnnotations We first designed carefully the test set in order to reach a high level of diversity regarding shows, speakers, debates, music, and noise. Therefore, we selected 35 files in which there are debates, interviews, jingles, live translations, etc. Speech segmentation and transcription – with a special care regarding overlapping speech – of these 35 files have been manually verified and completed The guidelines are described in the next section. This test partition

can be used for ASR, segmentation and diarization evaluations.

DiarTest-UnseenShows A specific partition has been designed for diarization on unseen shows. The shows from this partition differs from the ones in the train set. This enables to evaluate the robustness of the segmentation, diarization and speaker verification on unseen domains : new scenarios, new speakers (different presenters and interviewees), new acoustic environment. Files are partially transcribed, and no manual correction have been applied to this subset. This partition can be included in the train set for ASR training and validation purposes.

DiarTest-SeenShows This partition has been designed for cross-file diarization. The shows in the train and this partition are the same. Of course, many speakers are unseen, *i.e.* street interviews, experts that appear only once. None the files of this subset has been transcribed.

3. Homogenization and annotation

3.1. Human assisted corrections

The agglomeration of corpora required a homogenization of annotations. Performance of recent speaker recognition systems has reached a level such that many "errors" of the system often appear to be instead errors in the initial manual annotations (i.e., the reference). In order to correct the reference (speaker labels only), we perform a within file speaker recognition task to compare pairs of speech segments, detect inconsistencies between the system output and the initial annotation and manually correct the reference when needed. The work described in this section is a first step and other corrections are still ongoing. For this reason, we only corrected the speaker labeling without questioning the quality of the segmentation. For each file, we removed all segments including more than one active speaker (overlapped speech) as well as all segments shorter than 3 seconds according to the reference. For each remaining speech segment, a speaker embedding has been extracted. All pairs of embeddings which belong to the same (resp. different) speaker(s) according to the reference, and which obtained a cosine similarity below 0.3 (resp. above 0.7) were sent for manual verification. Those pairs of speech segments present a strong mismatch between the reference and the system output. A total of 7,022 couples of segments (almost 194 hours of speech, 678 different speakers) from 267 files have been manually checked. 99% of the pairs which belong to the same speaker according to the reference but obtained an automatic score lower than 0.3 happened to be wrongly labeled and belong to different speakers. On the other side, 69% of the pairs

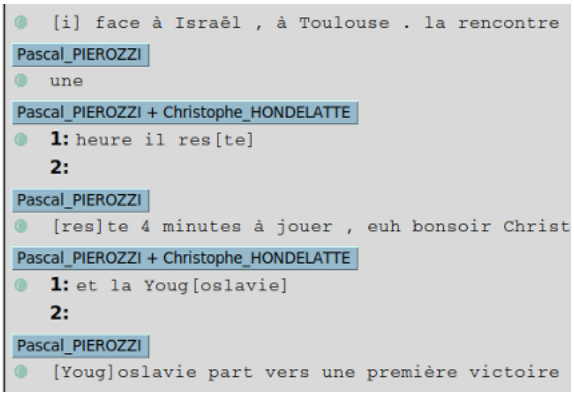

Figure 1: Example of an overlap speech segment with appropriate transcription.

labeled with different speaker ID in the reference but with scores above 0.7 actually belonged to a same speaker. We also found during the process a number of speaker labeling errors due to a bad segmentation or to overlapped speech that was labeled with a unique speaker ID (a good example of this case are segments of non-French speakers overlapped with the voice of a translator but that have been labeled with the ID of the original speaker).
The automatic speaker recognition system was an ECAPA-TDNN network trained with AAM-Loss (Desplanques et al., 2020) on VoxCeleb1&2 datasets (Nagrani et al., 2020) on top of a large WavLM pre-trained model. The comparison of speaker embeddings is performed via cosine similarity.

3.2. Segmentation and transcription annotation of the FullTest-CleanAnnot set

ALLIES test data have been designed in order to be able to set up segmentation, speaker diarization and speech recognition systems. Within the constraint of having this evaluation set ready for JSALT 2023, we decided to focus on different tasks such as speech, noise, music and overlap speech segmentation, automatic transcription, speaker labels. Further annotations could also be useful for interpretability such as the emotional states of the speakers, the linguistic intent, named entities, etc...
The manual annotation of this subset consists in 1) the addition of the transcription where it was missing, 2) the correction of existing transcription and segment boundaries, 3) the addition of missing overlap speech segments, 4) noise and music segmentation.
One difficulty regarding overlapping speech is that the second speaker is likely to take the floor within a words pronounced by the first speaker. The guidelines regarding overlapping speech have been carefully designed in order to satisfy both the tran-

scription (where the words need to be completely written) and get a correct segmentation. Therefore we decided to duplicate the word in two speech segments and to precise the part of the word pronounced in each segment as shown on the example Figure 1.
All annotations have been done with Transcriber. Steps 1 to 3 were realized in a single phase, while music and noise segmentation have been done afterwards. It is worth noting that the selected 35 files contain a high proportion of debates and political interviews, therefore overlapping speech occurs frequently which significantly increase the annotation time.

4. Characteristics of the corpus

ALLIES data contains 1048 audio files, each corresponding to a single show at a specific date. Each audio file has its corresponding transcription file in TRS format, whether a manual or automatic transcription has been realized or not. RTTM files have been generated from these transcription files for segmentation based evaluation. Different characteristics of the corpus are summarized in Table 4. The relative number of different shows highly differs between the DiarTest and the FullTest partitions. This result comes from the choice of having a diverse representative selection of shows in the full test partition. The average number of speaker per file is slightly lower in the Train and Full Test partitions. The Train partition contains many broadcast news with only one single presenter especially from radio channels. We can also notice that the number of unnamed speakers is relatively low in the FullTest set, because the manual correction of speaker names allows to name unidentified speakers.
The overlap proportion (in duration) fluctuates widely between broadcast news with little to no interaction and debates with around 10% of overlaps.

5. Tasks and Evaluation

Existing corpora for speech processing generally address only a single task such as diarization, sound event detection, or speech recognition. The goal of this section is to provide a benchmark which shows the high potential of ALLIES in addressing several tasks in speech processing with the same data in French. We provide the baseline results regarding several standard speech processing tasks on the FullTest-CleanAnnotation set.

5.1. Speech segmentation

We first target a voice activity detection (VAD). The addition of the presence of music and noise at the frame level is also investigated. We then face a multi-class (speech, overlap, music, and noise) problem that we address with two different approaches : a multilabel segmentation (all labels can occur simultaneously), and a 3-classes speech and overlap (one single label at a frame). Performances are given in terms of F-score.

The multiclass system (3cl) is composed of a WavLM large as a feature extractor and a TCN for the classification using the same architecture as (Lebourdais et al., 2024). This system has been trained on the train partition of ALLIES. The multilabel system (ml) uses an architecture similar to the multiclass, but with 4 outputs instead of 3, each one using a separated cross-entropy loss. This type of architecture allows predicting the presence of multiple classes simultaneously. To get samples from all the classes, this system has been trained on DIHARDIII (Ryant et al., 2021), AragonRadio (Ortega et al., 2012), OpenBMAT (Meléndez-Catalán et al., 2019), and ALLIES.

To increase the proportion of music, noise, and overlap in the training data, two augmentation methods are used. The first takes music and noise segments from other corpora (MUSAN (Snyder et al., 2015), and a proprietary noise dataset), and adds them to existing segments in the training data. The second method is to generate the training segments by summing two randomly selected original segments from the training set. This way we increase the proportion of the examples with noise, music, and overlap without using any additional data source.

Table 3 presents the results obtained on the different test partitions of ALLIES for both 3-classes (3cl) and multilabel (ml) systems. All metrics are calculated at the time level following standard evaluation challenges in speech diarization. The noise class is undergoing an annotation process and thus has not been evaluated yet. The performances in voice activity detection (speech/non speech) are similar between both multi-class and multi-label systems. The overlapped speech and music detection results obtained with the multi-label model are not as good as expected, probably because there is a domain mismatch between training and testing. As there is not enough annotated music data in ALLIES Train, the music information comes from other training data. The overlap information was also extracted in majority from another corpus, thus preventing good results on these tasks.

5.2. Speaker diarization

For the diarization experiments, we opted for two commonly used clustering-based diarization approaches: Agglomerative Hierarchical Clustering (AHC) and VBHMM xvector-based diarization (VBx) (Landini et al., 2022). We use an

Partition	Train	DiarTest-SeenShows	DiarTest-UnseenShows	FullTest-CleanAnnot
# files	546	181	286	35
# different shows	23	4	8	12
# speakers	4065	1489	2020	299
# named spks	2308	1011	1272	231
% of unseen spks from the train	na	87	85	64
avg. nb of spk per file	9.46	14.28	14.17	9.97
avg. nb of unnamed spk per file	3.22	2.61	2.56	1.94
% overlap	6.45	8.06	4.46	6.73

Table 2: ALLIES characteristics.

Model	FullTest	DiarTest Seen	DiarTest Unseen
Speech-3cl	.978	.987	.983
Overlap-3cl	.647	.782	.689
Speech-ml	.988	.986	.982
Overlap-ml	.687	.652	.538
Music-ml	.384	NA	NA
Noise-ml	.355	NA	NA

Table 3: Segmentation results (F1-score %) on the 3 test partitions of ALLIES.

open source implementation of VBx[3], which includes pre-trained ResNet101 embedding extractor (trained on VoxCeleb 1&2), AHC (since the result of AHC is used as the initialization for VBx), and VBx itself. The authors of the implementation provide recipes for several datasets; in our experiments, we use the same pre-trained models and hyperparameter settings as in the DIHARD2 recipe. The only difference is that we modify the threshold for AHC when it is used by itself and not as an initialization for VBHMM clustering. The reason is that, in the original recipe, the AHC threshold is set to overestimate the number of speakers and let VBx drop the unnecessary ones later. While in our experiments, we are interested to see the performance of AHC alone. The threshold in this case was set to 0.

Table 4 shows the results achieved with the described diarization approaches on several ALLIES conditions. The performance is reported in terms of Diarization Error Rate (DER) computed with dscore[4] tool, collar was set to 0. Both of the selected diarization methods are unable to deal with overlapped speech (in the best-case scenario, they can assign one of the speakers correctly). However, when presenting the results, the overlapped speech regions are not excluded from the evaluation. For both AHC and VBx, we show the results with the "oracle" VAD labels as well as with the automatic VAD presented above in Section 5.1.

Notice that in our baseline diarization experiments, we did not utilize the training set of ALLIES. The reason is that we wanted to provide the results of easily-available and easily reproducible systems while training a diarization system from scratch would not satisfy these criteria.

From the results, we can conclude that VBx is competitive towards a simple AHC. The performances on the DiarTest-SeenShows set are the lowest. An hypothesis for this result is that this set contains a high proportion of overlap speech which is not tackle by the evaluated systems. So far, the diarization results are good even on the FullTest for which the selected shows contain debates with a high number of speakers.

Model	FullTest	DiarTest Seen	DiarTest Unseen
Oracle VAD + AHC	14.23	19.94	16.41
Pred. VAD + AHC	19.46	22.19	18.75
Oracle VAD + VBx	8.45	14.54	13.91
Pred. VAD + VBx	13.73	18.39	16.46

Table 4: Diarization results (DER %) on ALLIES data.

5.3. Speech Recognition and Speaker Change Detection

As regards the Automatic Speech Recognition (ASR) experiments, we explored an encoder-decoder architecture, where the encoder transforms the speech feature sequence into hidden representations and the decoder generates an output sequence. We used SpeechBrain toolkit (Ravanelli et al., 2021) which supports sequence-to-sequence ASR models including recurrent and transformer structures and joint CTC-attention training. In order to increase the number of transcribed data to train the model, a new partitions were designed for ASR. The Train and DiarTest-UnseenShows sets presented in Section 2.2 are merged and re-

[3] https://github.com/BUTSpeechFIT/VBx

[4] https://github.com/nryant/dscore

split into a training and development sets as described in Table 5 which preserves the FullTest-CleanAnnot subset as the same test split as the one used in other tasks.

Set	Trans. (h)	# files	# seg
Train	157.16	533	120313
Dev	4.59	13	3988
FullTest	10.34	35	8279

Table 5: ALLIES partitions for ASR experiments

We kept only segments of a duration higher than 1 second to train our ASR systems. The results of the ASR systems are reported in terms of Word Error Rates (WER) in Table 6. The first row of the table corresponds to the WER obtained on dev and test sets. The baseline ASR systems achieve good recognition results with a WER of 14.83 and 14.99 on the dev and test set respectively.

Model	Dev	FullTest
Baseline ASR	14.83	14.99
SC-ASR	14.77	15.07

Table 6: Transcription results for the baseline ASR (WER %) and the Speaker Change ASR (WER including the # token) on ALLIES data.

As part of this work, we also trained a Speaker Change aware ASR system (SC-ASR row in table 6). Unlike conventional ASR training process, our SC-ASR system is trained using speech segments from more than one speaker. In fact, we pre-processed the data in such a way that all the consecutive segments (start time of segment $n+1$ equal to end time of segment n) from two different speakers are merged into one segment. This process led to merging 10% of the training data. As shown in figure 2, we also added a special token (#)in order to denote the speaker change position in the corresponding transcription. The SC-ASR system was trained with this special token to indicate the two-speakers segments and the word position of change.

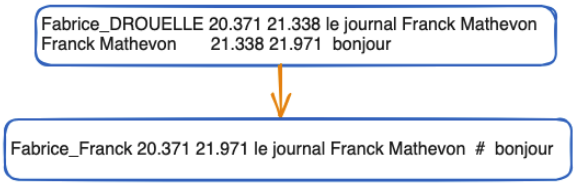

Figure 2: Example of consecutive segments merged to train an SC-ASR system.

As regards the evaluation of the SC-ASR system, we used WER to evaluate the transcription quality with this training configuration. This new WER includes the special token # as a word. As we can see, the transcription quality is almost the same as with the baseline ASR, while we have added some tokens.

In addition, we also evaluated the speaker change detection using the purity and the F1-score metrics at the time level.

F1-score is also calculated at the segment level and measure the ability of the SC-ASR model to predict the speaker change for two-speakers segments. A true positive meaning that the speaker change present in the reference has been predicted in the transcription eventually on the wrong position. To evaluate the position of the speaker change token we also include the purity, defined as the intersection (number of common frames) between reference and predicted speaker intervals. This metric is calculated at the time level and measures the capacity of the SC-ASR model to correctly identify the two speakers segments.

	purity	prec.	rec.	F1	support
Dev	0.86	0.92	0.91	0.91	545
FullTest	0.87	0.93	0.84	0.88	1189

Table 7: Time-based evaluation of Speaker Change detection on ALLIES data : purity, precision, recall and F1 . Support is the number of 2-speakers segments per set.

As we can see from Table 7, the Speaker-Change aware ASR system is able to detect the presence of a speaker change in a given segment with a precision of 86% and 87% on Dev and FullTest respectively. This is also reflected by the high precision and recall of detecting a speaker change at the segment level.

6. Conclusion

This article presents ALLIES corpus along with protocols. ALLIES is a French meta-corpus of almost 500 hours of speech and 1048 files. In addition to the Train set, 3 Test sets have been designed to evaluate segmentation tasks. A first contribution is the homogenization of speaker names in the whole corpus and the complete segmentation and transcription work done on the FullTest partition regarding overlap, music, noise, speakers, segmentation and transcription. This corpus will be included in ELRA catalogue in 2024. As part of the work done during JSALT 2023, this article also provides as a second contribution, several baseline results for speech segmentation, speaker diarization, speech recognition and speaker change detection from the transcription.

7. Acknowledgements

This work has been partially funded by the French National Research Agency (project Gender Equality Monitor - ANR-19-CE38-0012). 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. This work was performed using HPC resources from GENCI–IDRIS (Grant 2022-AD011012565). The research reported here was conducted at the 2023 Frederick Jelinek Memorial Summer Workshop on Speech and Language Technologies, hosted at Le Mans University (France) and sponsored by Johns Hopkins University.

8. Bibliographical References

Or Haim Anidjar, Yannick Estève, Chen Hajaj, Amit Dvir, and Itshak Lapidot. 2023. Speech and multilingual natural language framework for speaker change detection and diarization. *Expert Systems with Applications*, 213:119238.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. 2020. ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification. In *Proc. Interspeech 2020*, pages 3830–3834.

Pablo Gimeno, Ignacio Viñals, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida. 2020. Multiclass audio segmentation based on recurrent neural networks for broadcast domain data. *EURASIP Journal on Audio, Speech, and Music Processing*, 2020(1).

Naoyuki Kanda, Xiong Xiao, Yashesh Gaur, Xiaofei Wang, Zhong Meng, Zhuo Chen, and Takuya Yoshioka. 2022. Transcribe-to-Diarize: Neural Speaker Diarization for Unlimited Number of Speakers using End-to-End Speaker-Attributed ASR. ArXiv:2110.03151 [cs, eess].

Federico Landini, Ján Profant, Mireia Diez, and Lukáš Burget. 2022. Bayesian hmm clustering of x-vector sequences (vbx) in speaker diarization: theory, implementation and analysis on standard tasks. *Computer Speech & Language*, 71:101254.

Anthony Larcher, Kong Aik Lee, and Sylvain Meignier. 2016. An Extensible Speaker Identification SIDEKIT in Python. In *ICASSP)*, pages 5095–5099, Shanghai, China.

Martin Lebourdais, Pablo Gimeno, Théo Mariotte, Marie Tahon, Alfonso Ortega, and Anthony Larcher. 2024. 3MAS: a multitask, multilabel, multidataset semi-supervised audio segmentation model. In *The Speaker and Language Recognition Workshop, Odyssey*.

Martin Lebourdais, Marie Tahon, Antoine Laurent, Sylvain Meignier, and Anthony Larcher. 2022. Overlaps and Gender Analysis in the Context of Broadcast Media. In *LREC 2022*, Marseille, France.

Huanru Henry Mao, Shuyang Li, Julian McAuley, and Garrison W. Cottrell. 2020. Speech Recognition and Multi-Speaker Diarization of Long Conversations. In *Proc. Interspeech 2020*, pages 691–695.

Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. 2020. Voxceleb: Large-scale speaker verification in the wild. *Computer Speech & Language*, 60:101027.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. ArXiv:2106.04624.

Meysam Shamsi, Anthony Larcher, Loïc Barrault, Sylvain Meignier, Yevhenii Prokopalo, Marie Tahon, Ambuj Mehrish, Simon Petitrenaud, Olivier Galibert, Samuel Gaist, Andre Anjos, Sébastien Marcel, and Marta R. Costa-Jussà. 2022. Towards Lifelong Human Assisted Speaker Diarization. *Computer Speech and Language*.

9. Language Resource References

Yannick Esteve, Thierry Bazillon, Jean-Yves Antoine, Frederic Bechet, and Jerome Farinas. 2010. The EPAC corpus: manual and automatic annotations of conversational speech in French broadcast news. In *LREC*, page 4, Valleta, Malta.

S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukri. 2006. Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news. In *LREC*, pages 139–142, Genoa, Italy.

Aude Giraudel, Matthieu Carre, Valerie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard. 2012. The REPERE Corpus : a multimodal corpus for person recognition. In *LREC*, page 7, Istanbul, Turkey.

Guillaume Gravier, Gilles Adda, Niklas Paulson, Matthieu Carré, Aude Giraudel, and Olivier

Galibert. 2012. The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *LREC - Eighth international conference on Language Resources and Evaluation*, Turkey.

Blai Meléndez-Catalán, Emilio Molina, and Emilia Gómez. 2019. Open broadcast media audio from TV: A dataset of TV broadcast audio with relative music loudness annotations. *Transactions of the International Society for Music Information Retrieval*, 2(1).

Alfonso Ortega, Diego Castan, Antonio Miguel, and Eduardo Lleida. 2012. The Albayzín 2012 audio segmentation evaluation. In *Proc. IberSPEECH*, pages 283–289.

Neville Ryant, Prachi Singh, Venkat Krishnamohan, Rajat Varma, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman. 2021. The third dihard diarization challenge. In *Proc. ISCA Interspeech*, pages 3570–3574, Brno, Czechia.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. MUSAN: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.