3MAS: a multitask, multilabel, multidataset semi-supervised audio segmentation model

Martin Lebourdais, Pablo Gimeno, Théo Mariotte, Marie Tahon, Alfonso Ortega, Anthony Larcher

3MAS: a multitask, multilabel, multidataset semi-supervised audio segmentation model

Martin Lebourdais[1]*, Pablo Gimeno*[2]*, Théo Mariotte*[1]*, Marie Tahon*[1]*,*
Alfonso Ortega[2]*, Anthony Larcher*[1]*.*

[1]LIUM / Le Mans Université, France
[2] ViVoLab, Aragón Institute for Engineering Research (I3A), University of Zaragoza, Spain
marie.tahon@univ-lemans.fr

Abstract

When processing audio data, multiple challenges arise, one of them being the diversity of information present in the audio signal. Various audio segmentation subtasks appeared including voice activity detection (VAD), overlapped speech detection (OSD), music or noise detection. These tasks are often completed by separate models trained on different datasets, thus increasing computational costs and limiting the usage to specific datasets. We first show that a multiclass VAD and OSD model outperforms state of the art models. Then, we propose 3MAS, a novel deep learning-based audio segmentation model capable of handling multiple datasets, and assessing multiple simultaneously as a multilabel segmentation problem. 3MAS provides similar performances as specialized models with a similar architecture and can be trained using partial and unbalanced annotations on different datasets. 3MAS is a gain in computational time, and opens new opportunities to include new labels.

1. Introduction

Audio segmentation systems aim to divide an audio signal into shorter fragments according to a predefined set of rules so that each fragment contains only information from a specific audio typology. In the context of raw audio recordings such as broadcast data, podcasts, or voice assistants, a large taxonomy of sound events can occur. In general terms, we can consider at least the speech, music, or noise categories. An initial segmentation of speech fragments can be used as a guide for automatic systems that work on speech-only fragments. Indeed, this preprocessing is an essential step before applying automatic speech recognition (ASR) or speaker recognition systems. In addition to speech detection, being able to recognize overlapping speech –speech fragments in which at least two speakers are simultaneously active– is also required in several audio processing tasks. Most ASR systems assume that they are fed with single-speaker utterances, consequently the presence of overlapped speech is usually an important source of error in ASR [1]. In addition, the presence of untreated overlapped speech degrades the performance of speaker diarization [2]. The distinction of musical content may also hold significance from the standpoint of document information retrieval. In broadcast content, music detection plays a key role in order to monitor copyright infringement [3]. The accurate detection of noisy events could be relevant in to remove it, or at least to reduce with a speech enhancement algorithm on the audio signal [4]. The definition of what is considered as noise is clearly not consensual. It can be background noise (brouhaha, urban, in-car, etc.), but also isolated occurrences of specific sound events (dog barking, laugh, mouth noise, etc.). In the present work, we define noise as a background sound coming from a non-vocal source, hindering the comprehension of the message.

While Voice Activity Detection (VAD) and Overlapped Speech Detection (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 [5]. Thus, two main segmentation by classification approaches, *i.e.*, the segmentation obtained by classifying each time frame, can be observed: the multiclass framework, where only a single class can be active at a given time, and the multilabel framework, where several classes can be active at a given time. One of the main benefit of considering a multilabel solution is scalability. While the problem complexity grows in a factorial way for a multiclass classification task that considers all possible combinations of classes [6] (i.e., speech, speech+overlap, speech+music, speech + noise ...), the same problem observes a linear growth in complexity for the multilabel paradigm. When dealing with multiple labels, one of the main issues concerns the available labels. Indeed, training a multipurpose model requires to have a homogeneous distribution of the labels in the data. However, databases annotated with a large number of labels are generally small due to annotation costs. Therefore, one option is to merge different databases annotated according to different sets of labels.

In this study, we first compare the performances in terms of segmentation (F1-score) and training time of multiclass models trained on databases in which both the presence of speech and overlapped speech is annotated (DIHARD, AMI). Then we propose a new methodology which extends the approach to new sound classes under a multilabel framework. This methodology relies on a modified loss able to take into account missing labels, and an adapted data augmentation process. As we include non-speech classes, we also investigate new acoustic features that are supposed to generalize to a large set of sound events.

The remainder of the paper is organized as follows: Section 2 introduces relevant previous work in the field of audio segmentation. The description of all data used and the details of data processing are shown in Section 3. Experimental set up and evaluations for the multiclass, respectively the multilabel, are presented in Section 4, respectively Section 5. The conclusions are drawn in the last section 6.

M. Lebourdais is now with IRIT, Toulouse, France; T. Mariotte is now with Telecom Paris, Palaiseau, France.

2. Related work

Depending on the final task, the diversity of sound events predicted by an audio segmentation system is quite wide. Here, we only focus on the models that predict the exact time location when occurs the event. We will not discuss classification approaches where a label is assigned to a pre-segmented audio fragment. One of the most well-known segmentation tasks is voice activity detection (VAD), which determines the exact location of speech samples in an audio stream. Historically, VAD systems have used energy measures [7] or statistical models [8]. Nowadays, most VAD models rely on deep learning methods, mainly bi-directional recurrent or convolutional models [9, 10]. Another relevant task is overlapped speech detection (OSD), whose goal is to extract audio fragments where at least two speakers are simultaneously active. Similarly to VAD, current state-of-the-art for OSD is also based on recurrent or convolutional deep learning approaches [11]. The use of Temporal Convolutional Network (TCN) has proven to be very efficient for OSD [12, 13]. Additionally, some works have combined both VAD and OSD in a single segmentation system. The 3 class OSD convolutional model presented in [5] deals with this problem from a multiclass perspective, while a modified version of the end-to-end diarization (EEND) approach [14] is based on the multilabel paradigm. This paradigm also has shown to be usable in broader speech-related tasks, as a voice type classifier [15].

Concerning audio segmentation solutions related to music information, several systems deal with binary tasks such as music detection [16], or the separation of speech and music [17], but the multiclass paradigm has also been investigated [18]. Other approaches go one step beyond, considering additional audio typologies, such as noise. For instance, the task proposed in the Albayzín audio segmentation evaluation campaigns [19, 20] aims at segmenting broadcast data according to three classes: speech, music, and noise.

As it can be inferred from previous explanations, audio segmentation is a heterogeneous discipline. This characteristic is also expressed in audio segmentation corpora, with different datasets containing different kinds of annotations. The annotation process with detailed taxonomies and several options is significantly more complex than using a binary labeling schema. That is the reason the amount of labeled data for audio segmentation tasks with multiple classes is still limited.

While most of the speech databases, the presence of speech is annotated, overlapped segments are not homogeneously annotated [21]. For instance, in some databases overlap is considered as noise. In such databases, noise and music are considered as non speech sounds, and are not precisely identified. Semi-supervised learning approaches deal with partially labeled data. In [22], both labeled and unlabeled data are present within a batch and the loss is a combination of a MSE obtained from the labeled part and a consistency cost of the pseudo labels predicted on the unlabeled part. Another option is to iteratively train the model on pseudo-labels [23]. In the multilabel paradigm, the complete loss can not be estimated if one label is missing, therefore we propose a very simple approach which consists in ignoring the unlabeled part in the calculation of the global loss.

To the best of our knowledge, the system presented in this paper is the first audio segmentation model able to provide automatic annotations for speech, overlap, music, and noise simultaneously. Furthermore, by following our training approach, the system could theoretically be extended to any additional class, even with training data that is partially annotated.

Corpus	Annotated hours per class			
	Speech	Music	Noise	Overlap
AMI*	80.90	-	-	10.97
DIHARD*,†	53.29	-	-	5.37
ALLIES†	368.26	-	-	25.10
Albayzín 2010†	19.92	13.76	7.36	-
Albayzín 2012†	51.2	12.77	22.7	-
OpenBMAT†	-	10.85	-	-
Train 3MAS†	224.78	22.45	12.45	14.06
Test 3MAS†	234.02	11.17	5.46	15.69
Total 3MAS†	492.67	37.38	30.06	30.47

Table 1: Number of annotated hours per class for each of the considered datasets. A speech segment can be annotated with several classes. * are used to train the multiclass models, † are used to train the multilabel models.

3. Data description

To be able to train and evaluate a system that can detect speech, overlap, music, and noise, several datasets with different annotations are used in our experiments. Namely, datasets considered in our work can be categorized into two big groups: those annotated for speech and overlap, and those annotated for speech, music, and/or noise. The first group (DIHARD and AMI) is used to validate the multilabel approach against the use of two independent binary models (section 4). In a second step, we gather all data, keeping the train/test partitions, to investigate how the performance degrades when adding two new classes: music and noise.

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 DIHARD III challenge with the *full* version, 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 *full-corpus-ASR* proposed in [24].

The ALLIES corpus[1] is a French meta-corpus designed to gather and extend previous French data collected for speaker diarization and transcription evaluation campaigns. The overlap proportion (in duration) fluctuates widely between broadcast news with little to no interaction and debates (around 10% of overlaps). Music and noise are also present, but not annotated. Despite a harmonization effort, the data collected and annotated under different protocols introduces some homogeneity problems [21]. The testing partition has been split in three parts, a DiarTest-SeenShows (181 files) composed of shows seen during training (but different files), a DiarTest-UnseenShows (286 files) composed of shows unseen during training, and a partition named FullTest-CleanAnnot (35 files) with carefully corrected speaker segmentation.

The two dataset of Albayzín evaluation campaigns are

[1] https://lium.univ-lemans.fr/corpus-allies/

broadcast news data in Catalan for Albayzín 2010 [19] and Spanish for Albayzín 2012 [25]. These contain annotations in speaker, music, and noise segmentation. The speech class is inferred from the speaker segmentation, but the overlapped areas are not annotated as such, thus removing this class possibility.

The OpenBMAT dataset [26] contains television broadcast audio from different countries labelled for the music detection task under two different scenarios: a binary condition where audio is annotated as containing music or not containing music, and a multiclass setup where audio is separated into foreground, background and no music fragments. In our experiments, we only consider the binary annotations.

Our test set is generated by combining standard test partitions from ALLIES, AMI, DIHARD and Albayzín 2012 datasets. This allows to obtain evaluation results on all possible classes provided by our system. Table 1 provides a comprehensive overview of the datasets used throughout the research by presenting the quantities of annotated audio for each class, as well as the distribution into train and test sets.

4. Independent vs. joint speech and overlap multiclass detection

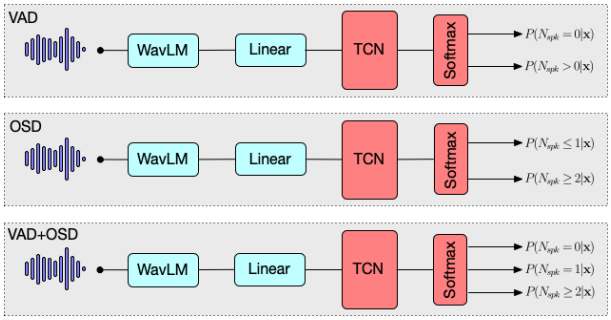

Figure 1: Independent VAD, OSD, and joint VAD+OSD systems with the feature extractor (blue) and the sequence modeling network (red). WavLM is frozen.

In this section, we address the multiclass paradigm, and investigate the advantages of a joint VAD+OSD model in comparison to two independent models. Figure 1 depicts an overview of the two independent VAD, OSD, and the joint VAD+OSD systems. While the feature extractor (in blue) encodes the input channel, the sequence modeling network (in red) processes the sequence of features before the frame classification.

4.1. Sequence modeling and classification

The frame classification is done at a rate of 100 Hz, while the raw waveform is sampled at 16 kHz. The feature extractor (blue) is based on the WavLM pre-trained model [27]. WavLM is a self-supervised system built with transformer blocks trained on Mix94k, a corpus of 94k hours drawn from LibriLight, VoxPopuli, and GigaSpeech. It learns to represent speech by masking a part of the signal and trying to predict the hidden part. Two versions are available, *large* and *base* + that differ by the number of transformer blocks (respectively 24 and 12) and the number of output dimensions (respectively 1024 and 768). In this first experiment, we use only the *large* model. This choice is motivated by the performance obtained by this model on the diarization task according to the SUPERB benchmark [28]. Furthermore, WavLM has been trained using simulated overlapped speech and is then more robust to this type of data, as demonstrated in previous works [13]. A trainable linear layer is added on top of the frozen WavLM to align this representation (one vector every 20 ms) with the target sequence (one label every 10 ms). More precisely, the linear layer transforms a segment of 99 features extracted with WavLM over a 2 s window of raw audio, into a 200-frame vector, aligned with our target.

The sequence modeling network (in red) takes as input a sequence $\boldsymbol{x}$ of features and assigns a class to each frame of this sequence. This task is performed using a TCN [29] since this architecture has shown noticeable results on both VAD and OSD tasks [12, 13, 30, 31]. This kind of architecture consists of stacked dilated 1D convolutional layers that exploit long temporal contexts from input sequences. 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 independent VAD (top) outputs the pseudo-probability of presence of at least one speaker $p(N_{spk} > 0|\boldsymbol{x})$. The independent OSD (middle) outputs the pseudo-probability to contain speech from more than one speaker $p(N_{spk} \geq 2|\boldsymbol{x})$. Both independent VAD and OSD are then binary classifiers, denoted as 2-class systems. The joint VAD+OSD system outputs the pseudo-probability of presence of either no speaker at all (non-speech) $p(N_{spk} = 0|\boldsymbol{x})$, a unique speaker $p(N_{spk} = 1|\boldsymbol{x})$, or more than one speaker $p(N_{spk} \geq 2|\boldsymbol{x})$. The 3-class approach is then converted for evaluation to 2-class VAD and OSD by merging the relevant classes.

4.2. Experimental protocol

In order to estimate the robustness over different speech domains, the three systems are trained and evaluated independently on the 2 datasets DIHARD and AMI. To counteract the small number of overlap segments as stated in Table 1, 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 following the method described in [32]. The loss function is a cross-entropy, and we used the ADAM optimizer with a learning rate of $lr = 10^{-3}$. Again, to balance the training data in favor of noisy labels, audio data is augmented with noise extracted from MUSAN [33] and additional reverberation using simulated room impulse responses.

Following DIHARD III evaluation plan, we use the F1-score obtained on the evaluation set as a performance metric. In the 2-class approach, only the positive class output ($N_{spk} > 0$ for VAD, and $N_{spk} \geq 2$ for OSD) is used for prediction and two detection thresholds (in and out) are applied to predict binary labels [32]. In the 3-class approach, the labels can not occur simultaneously. Then, the class associated with the maximum $\mathrm{softmax}$ output is selected at the frame level. The VAD is inferred by combining $N_{spk} = 1$ and $N_{spk} \geq 2$ outputs of the system. OSD relies on the $N_{spk} \geq 2$ prediction only.

4.3. Detection results

OSD and VAD results obtained on DIHARD and AMI datasets with independent or joint models are presented in Table 2. VAD performances are similar between the 2- and 3-class approaches on both datasets. The high detection scores (over 97%) confirm the ease of this task. However, we need to keep in mind that the remaining 3% could heavily penalize speaker diarization or ASR.

Table 2: Multiclass VAD and OSD F1-score (%) on AMI and DIHARD datasets for Mel+TCN [12], Mel+CRNN [31], SincNet+BLSTM [14] on the evaluation set of data covering various domains. † indicates that the results are taken from the original article.

		DIHARD				AMI			
	Models	Mel+TCN	Mel+CRNN	SincNet+BLSTM†	Ours	Mel+TCN	Mel+CRNN	SincNet+BLSTM†	Ours
2-class	VAD	-	-	-	97.0	-	-	-	97.4
	OSD	54.7	51.3	-	66.2	73.4	66.0	-	79.6
3-class	VAD	-	-	-	97.0	-	-	-	97.2
	OSD	54.5	50.8	59.9	66.8	73.8	69.6	75.3	80.4

Regarding overlap detection, we retrained and evaluated both models on DIHARD and AMI data to be able to compare our results [12, 31]. Regarding state-of-the-art results, the 3-class SincNet + BLSTM from [32] reaches the best performances on both datasets. The OSD results obtained with our approach, being with the independent or the joint approach, are above 66%.

The 3-class approach shows a small increase in OSD results for both datasets. In summary, the joint VAD+OSD system offers similar to slightly better performances than 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.

4.4. Training time

To assess the value of training a joint VAD+OSD system against two dedicated models, we compare the training time required for each approach to converge. 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.

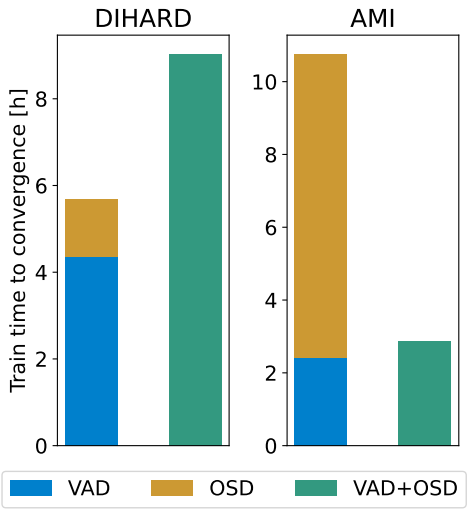

Figure 2: Training time (in hours) required for two separated VAD and OSD systems, and the joint VAD+OSD system to converge on both AMI and DIHARD datasets.

Regarding the DIHARD corpus, the VAD requires more resources than OSD. This can be explained by the fact that the data contains a lot of real-life domains, such as *group chat*, *clinical* (conversations between a clinician and a child) or *phone*, with a high amount of background noises. We observe the exact opposite with AMI data, which can be explained by the highly controlled recording conditions. In terms of training time, the 3-class model is beneficial with AMI, where the model converges as fast as a single VAD. This conclusion is not true for DIHARD where no gain is noticeable when using a 3-class approach.

4.5. Influence of the speech domain on performance

To study the influence of the speech domain on OSD performances obtained with the 3-class model, 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* contains 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.

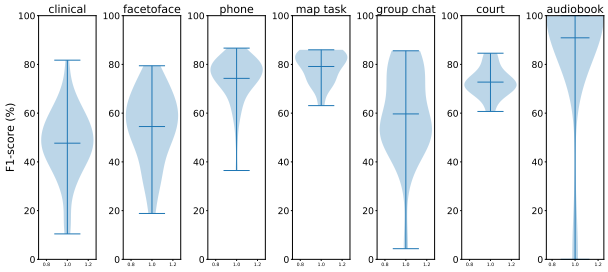

Figure 3: Distribution of F1-scores on DIHARD speech domains

We expect similar results for dyadic conversations such as *phone*, *clinical* and *face-to-face* conversations. Fig. 3 shows that the F1-score is globally better for *phone* conversations than the other. We hypothesize that the absence of visual cues in phone conversations limits the diversity of overlaps contained in the audio files, thus making overlap detection easier. 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 are major issues for OSD.

In summary, the joint VAD+OSD system offers similar to slightly better performances than 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. We also shown that such an approach can drastically reduce the training time in certain scenario. OSD performance remains limited by the audio quality and the speech domains. This opens new directions to improve the robustness of such systems.

5. Semi-supervised multilabel segmentation model (3MAS)

In this section, we extend our joint speech and overlap detection model to two new classes: music and noise. To do so, we investigate a new semi-supervised approach to train our model on missing labels and heterogeneous data. We also move towards a multilabel approach, meaning that all labels can occur at the same time.

5.1. Features description

In the previous section, we demonstrated that WavLM representations were very accurate in detecting speech and overlapped speech segments. However, this model built on transformer encoder layers has been trained on (noisy) speech data. Therefore, the features extracted with such a model might not be relevant for music detection. That is why, in addition to WavLM, we investigate other acoustic features while extending our 3-class model to music and noise labels.

As a baseline, we use a combination of 20 MFCCs with first and second derivatives (Δ and $\Delta\Delta$) and 12 chromas that correspond to a projection of a time-frequency representation onto a 12-tone scale. These characteristics are extracted on 30ms windows each 20ms to match the step of WavLM.

Leaf [34] is a trainable feature extractor that aims to provide a Mel-filterbank-like representation from an audio signal. It is similar to SincNet [35] in its composition but contains Gabor convolutions instead of sinc, a learnable Gaussian filter, and a per-channel energy normalization. This framework is presented as having better performances than mel-filterbank overall and than SincNet on music-related tasks.

5.2. Semi-supervised learning

The multilabel segmentation model is similar to the joint VAD+OSD presented in the previous section (bottom in Fig.1). It consists of a feature extractor (WavLM, MFCC+chromas, Leaf + linear layer) and a TCN. Classification is performed by a 1-d convolutional layer followed by a $\mathrm{sigmoid}$ activation function.

The loss function of the multilabel segmentation (see eq. 1) is a binary cross-entropy (BCE) computed on average at the sequence level between predicted $\hat{y}_c$ and reference y_c for each label $c \in \{\text{speech, overlap, music, noise}\}$.

$$L_{global} = \sum_{c \in C} w_c \cdot \text{BCE}(\hat{y}_c, y_c) \tag{1}$$

As shown in Table 1, the 5 datasets (ALLIES, DIHARD, Albyzìn, openBMAT) used for training are not annotated with all classes. When the annotation is available in the reference, the presence of the class is denoted by $y_c = 1$ (respectively $y_c = 0$) when absent). For instance, the overlapped speech class is present in Albyzìn data, but has not been annotated and no reference y_{ov} is available. To cope with this issue, an option is to neutralize this class by setting the weight $w_{ov} = 0$ for samples coming from this database. Otherwise the weight of the class c is $w_c = 1$

In practice, we mask the frames corresponding to missing labels by adding a -1 label (instead of 0 or 1) as illustrated in Figure 4. The loss function management has been implemented in `pyannote 2.1` [36]. The training and evaluation processes also rely heavily on `pyannote` and thus can be reproduced easily. For encouraging reproducible research, the codebase for training and evaluating our model is available on GitHub[2]

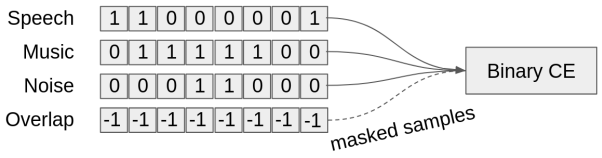

Figure 4: Masking process to ignore unannotated classes in a dataset

5.3. Data augmentation

Furthermore, the classes considered in this work are strongly unbalanced. More precisely, music, noise, and overlap classes are under-represented (see Table 1). To cope with this issue, one option is to artificially augment data. Two distinct kinds of augmentation are used. First music and noise augmentation consists of adding an audio segment from an external source, in our case, MUSAN [33] and ESPINETE, an internal dataset of noise collected from YouTube, to add noise and music to the audio signal and the associated label. This method increases the proportion of music and noise classes during the training stage.

The second augmentation adapts the method described in section 4.2 to the multilabel paradigm. Two segments from the same batch are summed, thus augmenting simultaneously the 4 classes. This augmentation brings new issues in merging the existing labels. Adding two segments of speech creates an overlap segment, but adding a segment with annotated speech ($y_{sp} = 1$) and a segment without speech annotation ($y_{sp} = -1$) should keep the speech information ($y_{sp} = 1$), without overlap information ($y_{ov} = -1$).

5.4. Experimental protocol

For these experiments, we use the ADAM optimizer with torch default parameters, a non-weighted binary cross-entropy as loss, and a batch size of 128. All segmentation models are trained and evaluated on the databases (except AMI) described in section 3. The exact amount of data for training and testing is given by the *3MAS* lines of Table 1. We decided to discard AMI from this experiment to limit the number of data with only speech and overlap annotations. Note that test partitions from all databases are merged into a single one, and we provide only global results regarding each class. Therefore, we cannot compare the results of this section to the previous one.

[2] `https://github.com/Lebourdais/3MAS`

Table 3: Multilabel precision, recall, and F1-score obtained for the different classes with different input features on Test 3MAS. The ablation of the augmentation process is done using the WavLM model.

	Speech			**Overlap**			**Music**			**Noise**		
Features	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
Merged baselines	98.1	98.4	98.2	58.6	68.7	63.2	90.2	92.1	91.1	81.1	77.2	**79.1**
MFCC+chroma	97.8	98.3	98.1	44.5	55.6	49.4	90.7	70.9	79.6	45.2	92.3	60.7
Leaf	97.1	98.0	97.6	32.2	50.3	39.3	77.0	89.1	82.6	65.8	59.9	62.7
WavLM	98.7	97.9	**98.3**	59.5	68.7	**63.8**	90.0	96.6	**93.2**	73.0	85.1	78.6
No Music+noise aug	98.3	98.1	98.2	55.8	75.1	64.0	90.3	94.4	92.3	61.3	85.2	71.3
No Overlap aug	98.5	97.6	98.0	67.7	57.8	62.4	89.6	95.6	**92.5**	65.5	90.2	**75.9**
No augmentation	98.5	98.2	**98.3**	67.7	61.0	**64.2**	92.8	89.1	90.9	81.5	70.5	75.6

Two independent baselines are combined to provide results on the four target classes, and then compare the results with the semi-supervised multilabel segmentation. The first baseline is the 3-class model presented in the previous section, in which the final classification layer is adapted towards a multilabel framework (by using $\mathrm{sigmoid}$ instead of $\mathrm{softmax}$, and by modifying the output dimensions from 1 to 2). While the 3-class outputs a single prediction, the multilabel outputs two binary predictions: speech and overlap. This baseline is evaluated in terms of speech and overlap only.

The second baseline is inspired by the work presented in [18] but considers a multilabel approach. Using the set of MFCC+chroma acoustic features presented in section 5.1, a 2-layer bidirectional LSTM network is trained to predict speech, music, and noise. Predictions obtained from both baselines are merged. For the speech class, we keep only the best detection among the two systems. This approach is referred to as *Merged baselines* in the following.

The multilabel segmentation model extended to the four classes is evaluated with three different feature sets : MFCC+chromas, Leaf and WavLM.

5.5. Detection results

The experiment is conducted on the full corpus (except AMI) presented earlier for acoustic features, leaf features, and the two mentioned WavLM models, the base plus and the large version. The Table 3 presents the precision, recall, and f1-score on the test set. Our model based on WavLM is the best we trained, achieving similar results as baselines separate models. We can notice a small gain from the baseline on music. The multilabel segmentation can benefit from the different classes to slightly improve the overlap detection in comparison to the baseline. We hypothesize that the information provided by the presence of music within the cost function improves the precision of the overlap segmentation.

Regarding the audio features, all of them get comparable performances on speech detection (VAD). We confirm that this task is not discriminant enough to assess the performance of a model. We can notice that music is better detected with Leaf features than MFCC+chroma, while it is the contrary for overlap detection. It seems that Leaf features do not generalize enough to non-music classes to be used in a multilabel segmentation approach. We also conclude that WavLM features provide high detection scores whatever the class is. It shows that this pre-trained model generalizes to non-speech signals.

5.6. Influence of the augmentation

To validate the usage of our data augmentation techniques, we propose an ablation study on our best system using WavLM. In this experiment, we remove first the augmentation by noise and music addition, keeping the mixing of segments. We then remove this second augmentation, keeping the first, and finally, both augmentations are removed.

The obtained results are summarized in the Table 3 (bottom). From this experiment, we can separate two groups of classes: the ones that need augmentation and the ones that don't. Without any augmentation, we obtain the best performance on speech and overlap detection. WavLM is heavily optimized for speech-related tasks and thus can segment speech and overlap without issues. The internal representation of these classes is sufficiently accurate (thanks to the multiclass loss function) to achieve good performances without augmentation by reaching F1-scores of 98.3% on speech and 64.2% on overlap detection. The absence of overlap augmentation when music and noise are augmented seems to penalize the model. The reason could be that the addition of music and noise strengthens the underrepresentation of the overlap class.

Contrary to speech and overlap, noise, and in a less extent music, detection require data augmentation to achieve the best results: 92.5% on music and 75.9% on noise. Indeed these classes are not well represented in the training data of WavLM and thus need to be represented somewhere else. The variety and amount of training data are important to achieve good results. This is why any augmentation that adds music or noise information is improving the results.

5.7. Output logits correlations

To better understand the behaviour of our best model, we investigate how correlated are the predicted logits on the full test set (44×10^6 frames). To do so, we compute the correlation matrix (depicted on Fig. 5) between the 4 logits obtained at the frame level. Time frames are considered as independent in this analysis. We remind that thresholds are applied on each logit to get the final labels.

In general terms, classes are not to be strongly correlated. This is expected because we set up a multilabel framework, meaning that classes are considered independent from a statistical perspective. We were expecting speech and overlap classes to be highly correlated, as the presence of overlap is conditioned by the presence of speech, however, counter-intuitively the correlation is only moderate ($\rho(sp, ov) = 0.14$). We notice a weak inverse correlation ($\rho(no, ov) = -0.55$) between noise and

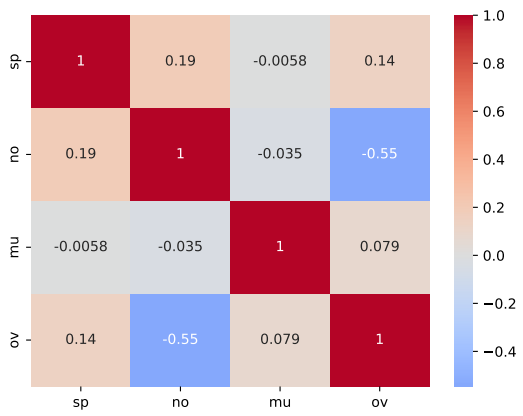

Figure 5: Correlation matrix obtained from logits for speech, overlap, music and noise labels on the full test set.

overlap. One reason behind this observation is the annotation heterogeneity issue. In some data, overlap is annotated as noise as it follows the definition of an event hindering the comprehension of the main speech. A better definition of the overlap class is necessary to ensure coherence in segmentation. To conclude, we confirm that our segmentation multilabel framework operates classes independently at the frame level. However further investigations at the sequence level should clarify the influence of the temporal aspects.

6. Conclusions

In this paper, we have investigated multiclass and multilabel approaches for audio segmentation. We have first demonstrated that a joint multiclass VAD+OSD (3-class model) reaches similar performances than two independent models (2-class). The joint approach has the advantage of drastically reducing the training time for some data. We also highlighted the fact that overlapped speech detection still remains limited by the audio quality and the speech domains. These first conclusions motivated a multilabel approach in which the model has more freedom to discriminate the different classes.

The main contribution of this paper is the introduction of the 3MAS audio segmentation model, the first of its kind to jointly provide segmentation labels for speech, overlap, music, and noise. This single system has been proven to be as effective as multiple specialized previous systems. We also assessed different features possibilities to confirm the efficiency of pretrained self-supervised models on task further from their original goal. In addition to this, we propose a solution based on loss masking to train a semi-supervised model on a fusion of different datasets partially annotated.

To tackle the lack of data for some classes, we propose two augmentation methods that can be used together or separately for music, noise, and overlap classes. An ablation study was performed, confirming that both methods can benefit self-supervised representation when dealing with music and noise classes. The proposed system relies on the multilabel classification paradigm and, thus, shows potential to be extended easily to more classes without a major change of the architecture. We finally studied the correlation between predicted classes and outlined the major flaw of the multilabel paradigm, which considers each class as an independent variable while this fact might not be true. An interesting lead to follow would be to inject knowledge of class correlation into the system to help interaction between multilabel classes without losing the modularity.

7. Aknowledgments

This work was performed using HPC resources from GENCI–IDRIS (Grants 2022-AD011012565 and AD011012527), the French ANR GEM (ANR-19-CE38-0012), and LMAC grant from Région Pays de la Loire. This project has also received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 101007666. 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. References

[1] O Çetin and E. Shriberg, "Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: insights for automatic speech recognition," in *Interspeech*, 2006, pp. paper 1915–Mon2A2O.6.

[2] Leibny Paola Garcia Perera et al., "Speaker Detection in the Wild: Lessons Learned from JSALT 2019," in *Speaker Odyssey*, 2020, pp. 415–422.

[3] Tomonori Izumitani, Ryo Mukai, and Kunio Kashino, "A background music detection method based on robust feature extraction," in *ICASSP*, 2008, pp. 13–16.

[4] Harishchandra Dubey, Vishak Gopal, Ross Cutler, Ashkan Aazami, Sergiy Matusevych, Sebastian Braun, Sefik Emre Eskimez, Manthan Thakker, Takuya Yoshioka, Hannes Gamper, and Robert Aichner, "Icassp 2022 deep noise suppression challenge," in *ICASSP*, 2022, pp. 9271–9275.

[5] J-W. Jung, H-S. Heo, Y. Kwon, J. Son Chung, and B-J. Lee, "Three-Class Overlapped Speech Detection Using a Convolutional Recurrent Neural Network," in *Interspeech*, 2021, pp. 3086–3090.

[6] Alexis Plaquet and Hervé Bredin, "Powerset multi-class cross entropy loss for neural speaker diarization," in *Interspeech*, 2023.

[7] Kyoung-Ho Woo, Tae-Young Yang, Kun-Jung Park, and Chungyong Lee, "Robust voice activity detection algorithm for estimating noise spectrum," *Electronics Letters*, vol. 36, no. 2, pp. 180–181, 2000.

[8] Joon-Hyuk Chang, Nam Soo Kim, and Sanjit K Mitra, "Voice activity detection based on multiple statistical models," *IEEE Transactions on Signal Processing*, vol. 54, no. 6, pp. 1965–1976, 2006.

[9] Fabio Vesperini, Paolo Vecchiotti, Emanuele Principi, Stefano Squartini, and Francesco Piazza, "Deep neural networks for multi-room voice activity detection: Advancements and comparative evaluation," in *Proc. IJCNN*, 2016, pp. 3391–3398.

[10] Taesoo Kim, Jiho Chang, and Jong Hwan Ko, "Ada-vad: Unpaired adversarial domain adaptation for noise-robust voice activity detection," in *ICASSP*, 2022, pp. 7327–7331.

[11] Midia Yousefi and John H. L. Hansen, "Block-based high performance cnn architectures for frame-level overlapping speech detection," *IEEE Trans. Audio Speech and Language Processing*, vol. 29, pp. 28–40, 2021.

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

[13] Martin Lebourdais, Marie Tahon, Antoine Laurent, and Sylvain Meignier, "Overlapped speech and gender detection with WavLM pre-trained features," in *Interspeech*, 2022, pp. 5010–5014.

[14] H. Bredin and A. Laurent, "End-To-End Speaker Segmentation for Overlap-Aware Resegmentation," in *Interspeech*, 2021, pp. 3111–3115.

[15] Marvin Lavechin, Ruben Bousbib, Hervé Bredin, Emmanuel Dupoux, and Alejandrina Cristia, "An open-source voice type classifier for child-centered daylong recordings," in *Interspeech*, 2020.

[16] Pablo Gimeno, Victoria Mingote, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida, "Partial AUC optimisation using recurrent neural networks for music detection with limited training data," in *Interspeech*, 2020, pp. 3067–3071.

[17] Diego de Benito-Gorron, Alicia Lozano-Diez, Doroteo T Toledano, and Joaquin Gonzalez-Rodriguez, "Exploring convolutional, recurrent, and hybrid deep neural networks for speech and music detection in a large audio dataset," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1, pp. 1–18, 2019.

[18] Pablo Gimeno, Ignacio Viñals, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida, "Multiclass audio segmentation based on recurrent neural networks for broadcast domain data," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2020, pp. 1–19, 2020.

[19] T. Butko and C. Nadeu, "Audio segmentation of broadcast news in the Albayzín-2010 evaluation: overview, results, and discussion," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2011, no. 1, pp. 1–10, 2011.

[20] Diego Castán, David Tavarez, Paula Lopez-Otero, Javier Franco-Pedroso, Héctor Delgado, Eva Navas, Laura Docio-Fernández, Daniel Ramos, Javier Serrano, Alfonso Ortega, and Eduardo Lleida, "Albayzín-2014 evaluation: audio segmentation and classification in broadcast news domains," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, pp. 33, Dec. 2015.

[21] Martin Lebourdais, Marie Tahon, Antoine Laurent, Sylvain Meignier, and Anthony Larcher, "Overlaps and Gender Analysis in the Context of Broadcast Media," in *LREC*, 2022.

[22] Léo Cances, Etienne Labbé, and Thomas Pellegrini, "Comparison of semi-supervised deep learning algorithms for audio classification," *EURASIP J. Audio Speech Music Process.*, vol. 2022, no. 1, sep 2022.

[23] Yuki Takashima, Yusuke Fujita, Shota Horiguchi, Shinji Watanabe, Paola García, and Kenji Nagamatsu, "Semi-supervised training with pseudo-labeling for end-to-end neural diarization," in *Interspeech*, 2021.

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

[25] A. Ortega, D. Castan, A. Miguel, and E. Lleida, "The Albayzín 2012 audio segmentation evaluation," in *iberspeech*, 2012, pp. 283–289.

[26] Blai Meléndez-Catalán, Emilio Molina, and Emilia Gómez, "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*, vol. 2, no. 1, 2019.

[27] Sanyuan Chen et al., "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 6, pp. 1505–1518, 2022.

[28] Shu-Wen Yang, Po-Han Chi, et al., "SUPERB: Speech Processing Universal PERformance Benchmark," in *Interspeech 2021*, 2021, pp. 1194–1198.

[29] 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.

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

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

[32] Herve Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill, "Pyannote.Audio: Neural Building Blocks for Speaker Diarization," in *ICASSP*, 2020, pp. 7124–7128.

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

[34] Neil Zeghidour, Olivier Teboul, Félix de Chaumont Quitry, and Marco Tagliasacchi, "LEAF: A learnable frontend for audio classification," in *ICLR*, 2021.

[35] Mirco Ravanelli and Yoshua Bengio, "Speaker recognition from raw waveform with sincnet," in *SLT*, 2018, pp. 1021–1028.

[36] Hervé Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *Interspeech*. Aug. 2023, pp. 1983–1987, ISCA.