# Data Selection Effects on Self-Supervised Learning of Audio Representations for French Audiovisual Broadcasts

**Valentin Pelloin**[*], **Lina Bekkali**[*†], **Reda Dehak**[‡], **David Doukhan**[*]
[*]Institut National de l'Audiovisuel (INA), France,
[†]École nationale des ponts et chaussées (ENPC), France,
[‡]EPITA Research Laboratory (LRE), France
{vpelloin, ddoukhan}@ina.fr, lina.bekkali@eleves.enpc.fr, reda.dehak@epita.fr

**Abstract**

Audio and speech self-supervised encoder models are now widely used for a lot of different tasks. Many of these models are often trained on clean segmented speech content such as LibriSpeech. In this paper, we look into how the pretraining datasets of such SSL (Self-Supervised Learning) models impact their downstream results. We build a large pretraining corpus of highly diverse TV and Radio broadcast audio content, which we describe with automatic tools. We use these annotations to build smaller subsets, which we use to train audio SSL models. Then, we evaluate the models on multiple downstream tasks such as automatic speech recognition, voice activity and music detection, or speaker recognition. The results show the potential of pretraining SSL models on diverse audio content without restricting it to speech. We also perform a membership inference attack to evaluate the encoder ability to memorize their training datasets, which highlight the importance of data deduplication. This unified training could bridge speech and music machine learning communities.

**Keywords:** self-supervised learning, pretraining dataset, audio encoders, speech, music

## 1. Introduction

Self-Supervised Learning (SSL) consists in pretraining models on unsupervised data, without using labeled data. In the context of audio and speech SSL models, an encoder model is pretrained on a large corpus of audio content. This model then generates embeddings that can be finetuned and used as input of downstream models to perform various tasks: music information retrieval, automatic speech recognition, speaker recognition, etc. However, speech encoders are trained on clean segmented speech (Parcollet et al., 2024; Zanon Boito et al., 2024), with many of them using read or audiobooks datasets as LibriSpeech (Baevski et al., 2020, 2023). Having access to large corpora of audio content without any clean annotation or segmentation of speech, one might be tempted to pretrain an audio encoder on this content. However, there remain multiple questions on the viability of such approaches. If this model is pretrained on content which includes music, noises, and speech, will it obtain good performances on downstream tasks such as speech recognition? Will the generated features be useful to perform voice activity or music detection?

Gender biases are commonly present in speech models (Adda-Decker and Lamel, 2005; Attanasio et al., 2024; Garnerin et al., 2019), including in SSL models (Fuckner et al., 2023; Zanon Boito et al., 2022). Does balancing the pretraining data across speaker genders reduce this bias for downstream tasks such as ASR or speaker recognition?

Another question arises from the quantity of duplicated content inside the pretraining dataset. For Natural Language Processing, others (Lee et al., 2022; Carlini et al., 2021) have demonstrated the negative impact it has on performances and sensitive data extraction. However, this impact has yet to be demonstrated on speech and audio tasks. As we release the pretrained and downstream models, we want to hinder the ability to extract pretraining data information from the model.

In this paper, we introduce a new 100,000 hours audio corpus derived from TV and Radio broadcasts from *Institut National de l'Audiovisuel* (INA, the French National Audiovisual Institute). This corpus is deduplicated using an audio deduplication tool, and segments are automatically described. With these pieces of information, we construct and pretrain 6 different audio SSL models, each on a subsample of 1,000 hours of content. To answer the questions highlighted above, we construct each subsample in order to evaluate the consequences of data selection during pretraining on the downstream evaluations. We evaluate our models on multiple downstream tasks: (gendered) automatic speech recognition, voice activity detection, music detection, speaker recognition, and a membership inference attack.

Pretrained and downstream models are published on HuggingFace[1]. Due to obvious copyright concerns, the training datasets are not released to

[1]https://hf.co/spaces/ina-foss/LREC-2026-Data-Selection-Effects



*Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 10221–10232
11–16 May 2026. 

the public, however, researchers seeking to obtain audiovisual archives may address their requests to *Le Lab*, an entity dedicated to researchers willing to access French audiovisual archives (Lezer, 2022).

## 2. Related works

Introduced by Baevski et al. (2023), data2vec2 is an efficient and multimodal architecture to train SSL encoders. This architecture is composed of a teacher-student encoder, and can be trained with similar objectives for text, image or speech. Using an equivalent architecture, Li et al. (2022) published music2vec, a model pretrained on music (1,000h). It obtains SOTA results on multiple Music Information Extraction (MIR) tasks. However, there does not exist a single unified encoder model suitable for both French speech and non speech tasks.

Parcollet et al. (2024) released pretrained speech encoders models for French following the architecture of Wav2vec 2.0 (Baevski et al., 2020). This architecture was also employed by Zanon Boito et al. (2022), where they evaluated the impact of pretraining gender biases for the downstream ASR task, suggesting that gender-balanced pretraining might provide a better initialization for the finetuning process.

Voice Activity Detection (VAD) using self-supervised speech representations has been experimented with success by Gimeno et al. (2021); Kunešová and Zajíc (2023); Karan et al. (2024). In particular, Karan et al. (2024) used a Wav2vec 2.0 encoder pretrained on 436k hours of speech and finetuned it alongside their downstream model to obtain SOTA performances, while keeping the throughput speed reasonable.

Although VAD and Music detection are closely related, unified architectures and techniques have not been developed between the two communities. While Doukhan et al. (2018) presented a model that detects voice and music, it cannot do both at the same time, for example when someone speaks over background music. This limitation might arise from available corpora labelled with both information. To the best of our knowledge, only AVA-Speech (Chaudhuri et al., 2018) is labelled for both speech and music detection.

The most recent Speaker Recognition (SR) models are based on large SSL models (Chen et al., 2022a; Novoselov et al., 2023; Chen et al., 2022b; Peng et al., 2023, 2024). These models use a backend model to aggregate and map hidden and temporal representations onto an embedding representation for speaker recognition purposes. The cosine distance between the embeddings is then used as a scoring method. Various approaches have been proposed. In simple methods, representations from all hidden layers of the pretrained SSL model are averaged with learnable weights and then fed as input features to a standard speaker recognition model such as ECAPA-TDNN (Chen et al., 2022b). Novoselov et al. (2023) suggests using a TDNN-based backend to directly aggregate hidden Wav2vec 2.0 representations. Peng et al. (2023) proposed an attention-based backend that uses key and value flow; the embeddings are then obtained via a weighted average. While these models demonstrate the importance of initial layers in defining the speaker embedding space (Chen et al., 2022a), the best performance relies on finetuning the pretrained model.

The ability of SSL models to remember (memorize) their training data has been studied by many. For example, for text models, Carlini et al. (2021) managed to generate URLs seen during the training of GPT-2 XL. For speech SSL models, Tseng et al. (2022) successfully set up Membership Inference Attacks (MIA) where they probe the model representations for the memorization of speaker and utterance information.

## 3. Audio datasets for SSL

Our objective is to build audio SSL models as general as possible. These models could work for both speech analysis tasks such as speech recognition, speech understanding, speaker diarization or verification; and also for Music Information Retrieval (MIR) tasks: music and singing voice detection. We aim at applying these models on audiovisual archives to extract audio embeddings, and use these as input of multiple downstream classifiers which could describe content at scale automatically.

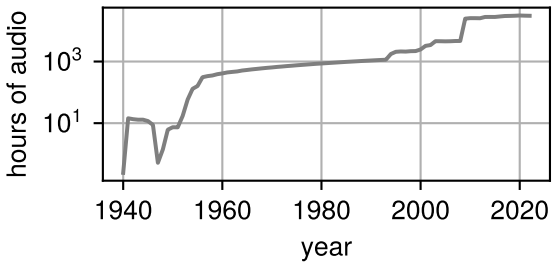


Figure 1: Number of hours of audio content in the source INA dataset per year.

The *Institut National de l'Audiovisuel* (INA) is in charge since 1975 of collecting and archiving TV and Radio content in France. In partnership with them, we obtained a randomly sampled dataset of 473k hours of content, broadcast on 113 French TV and Radio channels, from 1940 to 2022. Thus, this dataset covers various kinds of audiovisual content: news, adverts, documentaries, game shows, movies, musics, cartoons, sports, etc. The dataset is composed of mostly 1h long audio files, each file

corresponding to an unsegmented chunk of broadcast content on one channel on a particular date. In Figure 1 the number of audio content in this dataset each year is presented. Since 1995, new legislations are enforcing archival of TV and Radio (legal deposit). We notice a large increment starting in the recent years, thanks to the legal deposit[2].

We present in Figure 2 an overview of the data preprocessing pipeline, that we describe in the following sections.

### 3.1. Audio deduplication

Chenot and Daigneault (2014) showed that there was on average only about 8h of fresh content every day across 12 French TV channels, with some of them having as low as 2.3h of fresh content on each day. Many noted that deduplication is an important preprocessing step when preparing datasets for machine learning: it allows for faster training and better generalisation (Mikolov et al., 2018; Lee et al., 2022). A more serious issue regarding data privacy was raised by others (Carlini et al., 2021; Kandpal et al., 2022; Yan et al., 2024). Carlini et al. (2021) shows that the more sensitive information is repeated, the more it is at risk for memorization. According to Kandpal et al. (2022), text sequences present 10 times in training data of Language Models are on average generated 1000x more often that sequences present only once.

In order to mitigate this risk, we decided to deduplicate our training dataset using the repeated content detection tool described by Chenot and Daigneault (2014). The tool extracts lightweight audio fingerprints. A database of all fingerprints is constructed, copies are detected when at least 4 similar consecutive fingerprints are found, and we discard all copies of a content once it has already been found. The tool is described to be robust to many signal alterations, such as low pass filtering or temporal splits into short extracts (98% recall with 24s chunks). 154k hours of audio content were removed from our corpus with this deduplication step, representing 32.6% of the original corpus.

### 3.2. Removal of evaluation corpora

We intend to evaluate our models on corpora containing French audiovisual contents. Therefore, we want to avoid pretraining on content found in evaluation datasets. As in section 3.1, we use Chenot and Daigneault's tool, this time to remove existing datasets from our training corpus. We first aggregate the fingerprints of multiple datasets, including: ESTER1 (Gravier et al., 2004), ESTER2 (Galliano et al., 2009), EPAC (Estève et al., 2010), QUAERO (Boudahmane et al., 2011), ETAPE (Gravier et al., 2012), REPERE (Giraudel et al., 2012), Rhapsodie (Lacheret et al., 2014), Orféo (Benzitoun et al., 2016), InaGVAD (Doukhan et al., 2024), is24_news_topic (Pelloin et al., 2024). We then remove all audio chunks that matches with these fingerprints. 623h were removed from the 473k hours dataset in this step.

### 3.3. Chunking

Finally, we randomly sample 12M audio chunks of 30s, in order to create a corpus of 100,000 hours from the remaining content available. This step is necessary to restrict the required processing time for the automatic content description tools presented in the next section, while also ensuring data diversity. As a result of audio deduplication and chunking, this corpus of 100k hours represents 21% of the original corpus provided by INA.

### 3.4. Automatic description of audio chunks

We use different tools to automatically describe the 12M audio chunks. These data are then used in sec. 4 to obtain controlled pretraining corpora.

We first use Whisper (*whisper-large-v3-turbo*, Radford et al., 2023) to transcribe audio chunks into text. We do not set the content language and let Whisper perform the language identification.

We use InaSpeechSegmenter (Doukhan et al., 2018), a Voice Activity Detection (VAD) and Speaker Gender Segmentation (SGS) tool built for TV and radio audiovisual content. It has already been used by others as a dataset curation tool for SSL (Zanon Boito et al., 2024), to obtain clean speech segments. It allows us to predict a segmentation with active speech along with gender information. InaSpeechSegmenter also predicts a label "music" and "noise" but unfortunately, it cannot predict speech and music separately: it cannot tell if there is music in the background behind speech.

Although many open-source VAD systems exist, we did not find music detection tools suitable for our needs. Instead, we bootstrap a small MLP model using embeddings generated by music2vec (Li et al., 2022) and finetuned on OpenBMAT (Meléndez-Catalán et al., 2019) to predict the proportion of the *no-music*, *background-music* and *foreground music* classes. It obtains a Mean Absolute Error of 12.54% globally across OpenBMAT (13.30% on our test split).

### 3.5. Dataset statistics

In Table 1 we show the global statistics of the 100,000 hours dataset we prepared. We obtain the

[2]Legal deposit content between 1995 and 2009 was not broadly available due to storage format constraints.

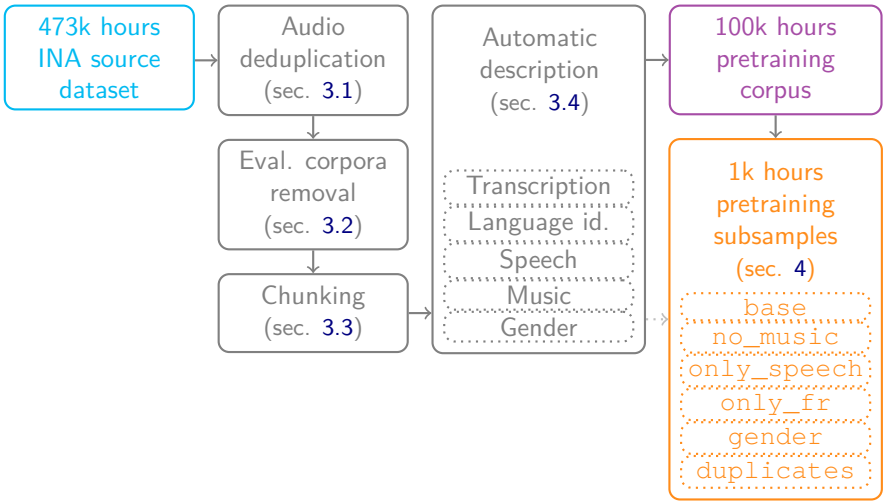


Figure 2: Overview of the data preprocessing pipeline.

| | |
|---|---|
| **Global:** | |
| Segment duration | 30s |
| Audio segments | 12,000,000 |
| Total duration | 100,000h |
| Period | [01/01/1940 - 31/12/2022] |
| Channels (TV+Radio) | 113 |
| **Content type*:** | |
| Segments with speech | 72.51% |
| Segments with music | 55.23% |
| **Gender balance*:** | |
| Women speaking time | 29.95% |
| Men speaking time | 70.05% |
| **Language*:** | |
| French segments | 91.69% |
| ↪ among "speech" segments | 99.50% |
| ↪ among "music" segments | 85.45% |
| English segments | 7.22% |
| ↪ among "speech" segments | 0.30% |
| ↪ among "music" segments | 12.75% |

Table 1: Global statistics on the 100,000 hours audio dataset. *Statistics obtained through heuristics and/or automatic tools.

speech and music segments counts with the following heuristics: segments are considered to have music if our music tool predicted less than 85% of *no-music*, and considered to be speech if more there is more than 20s according to InaSpeechSegmenter and less than 30% of *foreground-music*.

The language is determined by Whisper. French segments represent 91.69% of all segments according to Whisper. When filtering out segments not characterized as speech segments, French language represents 99.50% of the dataset. On the other hand, speech segments with English only represent 0.30% of them, while 12.75% for music segments.

The gender balance is obtained with InaSpeechSegmenter, and is plotted per year in Figure 3. The global speaking time for women in the whole 100,000h dataset is at 29.95%.

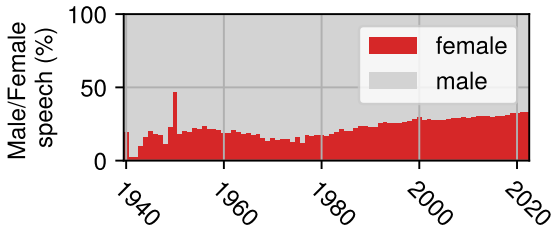


Figure 3: The per-gender speech ratio per year.

## 4. Self-Supervised Learning

We train audio SSL models following data2vec2 architecture presented by Baevski et al. (2023). The architecture follows a teacher-student encoder setting, where the teacher corresponds to the exponentially moving average from the student weights. The student has to predict the masked audio sequence representation of the teacher. The encoder contains a CNN network, followed by 12 transformers layers. Our models contains 93.2M trainable parameters, similar to other speech models considered as *base* models (Baevski et al., 2023; Parcollet et al., 2024). Models are trained for 100k steps with a cosine learning rate scheduler configured with a warmup of 8k steps to $7.5 \times 10^{-4}$. The full training configuration and metrics are released along with the models. Each model is trained for around 70 GPU-hours on NVIDIA H100 80GB cards.

Derived from the 100,000 hours training corpus presented in section 3, we define 6 pretraining datasets with controlled acoustic properties, in order to investigate the effects of data selection on downstream tasks. Each one of them is composed of 120,000 audio segments of 30s, for a total of 1,000h of speech. We train one model on each of these pretraining dataset. The six models are:

**base:** A model trained on a random sample of 1,000 hours.

**no_music:** Trained on a subsample composed of segments detected as not containing music with our described heuristics.

**only_speech:** Trained on a subsample with only segments containing speech, as described by our heuristics.

**only_fr:** Trained on a subsample on only French segments identified by Whisper.

**gender:** Trained on a subsample with a balanced proportion of male and female speech, as identified by InaSpeechSegmenter. With this subsample, we aim to see if these bias could be reduced through the pretraining dataset.

**duplicates:** Trained on a subsample sourced from **base**, where 1% of the segments were duplicated 10 times. To have the same total duration as base, 10,800 other segments were randomly removed from this sample. The objective of this subsample is to create a comparable version to **base**, but without a deduplicating step, as presented in section 3.1.

## 5. Evaluation on downstream tasks

In this section we benchmark our audio encoders with multiple downstream tasks: automatic speech recognition, voice activity detection, music detection and speaker recognition. We also assess the ability of our models to recall their pretraining dataset with a membership inference attack. We compare our audio encoders with speech encoder baselines also trained on 1,000h of content.

Unless otherwise stated, for all downstream tasks presented below, we train the downstream model and eventually finetune the audio encoder itself on either the official training subset of corpora used, or on our own split if it was not provided. We optimize hyperparameters on the development sets, and test on the remaining unseen data. We compute confidence intervals at 97.5% using the bootstrap sampling strategy with $n = 1000$.

### 5.1. Automatic Speech Recognition

We benchmark the different audio encoder models on Automatic Speech Recognition (ASR), i.e. the task of transcribing speech. We feed the last transformer layer of the audio encoder into a linear projection layer. The model is trained with a Connectionist Temporal Classification (CTC) loss to predict character-level outputs. We first initialize this added layer by training it for 1k steps with the encoder freezed, and then we continue training with the rest of the model unfreezed for up to 30k steps. Models are evaluated with a greedy decoder and without a language model.

ASR models are trained on the combination of Antract (Carrive et al., 2021), QUAERO (Boudahmane et al., 2011), EPAC (Estève et al., 2010), ESTER1 (Gravier et al., 2004) and REPERE (Giraudel et al., 2012) for a total of 258h, and evaluated on their test sets. We present in Table 2 the results in Word Error Rate (WER) of the different audio encoders on this task. We compare our audio encoders with *LB/wav2vec2-FR-1K-base* (Parcollet et al., 2024), also trained on 1,000 hours of French. Unlike our models, which are pretrained on spontaneous and diverse audio content, this model was pretrained on an audiobook dataset (MLS French Pratap et al., 2020). We notice our `base` obtains much better results that this baseline model, with absolute an improvement of 16.1% WER. Comparing our models trained on different subsamples, we can see that training without music (`no_music`) or with speech content only (`only_speech`) improves the results compared to the standard `base` setting. The model pretrained on data containing `duplicates` seems to perform worse than it's `base` counterpart. This confirms results previously observed by others (Lee et al., 2022).

In Table 3, we present the per-gender WER on the datasets with speaker gender annotation: QUAERO, EPAC, ESTER1 and REPERE. Similar to Zanon Boito et al. (2022), we compute the relative difference of WERs between male and female ($\Delta_{\text{rel}}$) as in Eq. 1. If the $\Delta_{\text{rel}}$ is greater than 0, the model is biased towards male as it performs better for male than for female, and better for female if the value is less than 0. Our gender analysis compares `base` and `gender` models as the two were trained with the same kind of content. During pretraining, 70% of the speech for the `base` model was from male, while it accounted to 54% in the `gender` model.

$$\Delta_{\text{rel}} = 100 \times \frac{\text{WER}_{\text{F}} - \text{WER}_{\text{M}}}{0.5 \times (\text{WER}_{\text{F}} + \text{WER}_{\text{M}})} \quad (1)$$

Globally, we can see EPAC women transcriptions seem more difficult than men's. On the other hand, we notice QUAERO and ESTER seems easier for female speech. This result could be explained by biased speaker roles inside broadcast content corpora (Adda-Decker and Lamel, 2005;

| Model | *Global* | Antract | QUAERO | EPAC | ESTER1 | REPERE |
|---|---|---|---|---|---|---|
| *LB/wav2vec2-FR-1K-base* | 31.4 ±0.1 | 28.4 ±0.1 | 36.8 ±0.5 | 29.8 ±0.2 | 32.4 ±0.2 | 34.0 ±0.2 |
| `base` | 15.3 ±0.1 | 14.2 ±0.1 | 18.5 ±0.4 | 14.2 ±0.2 | 15.8 ±0.2 | 16.4 ±0.2 |
| `no_music` | **14.4** ±0.1 | 13.5 ±0.1 | **17.4** ±0.4 | **13.1** ±0.2 | **14.9** ±0.2 | **15.4** ±0.2 |
| `only_speech` | 14.5 ±0.1 | **13.3** ±0.1 | 18.0 ±0.4 | 13.4 ±0.2 | 15.1 ±0.2 | 15.5 ±0.2 |
| `only_fr` | 15.0 ±0.1 | 13.7 ±0.1 | 18.2 ±0.3 | 14.0 ±0.2 | 15.5 ±0.2 | 15.9 ±0.2 |
| `gender` | 15.1 ±0.1 | 14.0 ±0.1 | 18.5 ±0.4 | 13.9 ±0.2 | 15.7 ±0.2 | 16.1 ±0.2 |
| `duplicates` | 15.6 ±0.1 | 14.5 ±0.1 | 18.9 ±0.4 | 14.4 ±0.2 | 16.2 ±0.2 | 16.5 ±0.2 |

Table 2: ASR results in WER (↓) of the different models on the test sets of Antract, QUAERO, EPAC, ESTER1 and REPERE and globally across all corpora. Best results in **bold**, results within the confidence interval of the best model underlined.

| Dataset | | `base` | `gender` |
|---|---|---|---|
| **QUAERO** | male | 18.8 ±0.4 | **18.7** ±0.4 |
| | female | **17.9** ±0.6 | 18.2 ±0.6 |
| | $\Delta_{rel}$ | -4.9 | **-2.7** |
| **EPAC** | male | 14.0 ±0.2 | **13.7** ±0.2 |
| | female | 15.8 ±0.4 | **15.2** ±0.4 |
| | $\Delta_{rel}$ | 12.1 | **10.4** |
| **ESTER1** | male | 16.2 ±0.2 | **16.1** ±0.2 |
| | female | 14.9 ±0.3 | **14.7** ±0.3 |
| | $\Delta_{rel}$ | **-8.4** | -9.1 |
| **REPERE** | male | 16.4 ±0.2 | **16.1** ±0.2 |
| | female | 16.9 ±0.5 | **16.5** ±0.4 |
| | $\Delta_{rel}$ | 3.0 | **2.5** |

Table 3: Gendered ASR results in WER, and relative difference between male and female ($\Delta_{rel}$).

Garnerin et al., 2019). Regarding the model gender bias, measured with the relative difference of WERs ($\Delta_{rel}$), we notice that `gender` obtains less bias than `base` for all datasets except ESTER1, which indicates training on balanced gender corpora could be beneficial. However, it should be noted that the overall differences between `base` and `gender` results are not statically different.

## 5.2. Voice Activity Detection

We evaluate the impact of the pretraining set on a simple Voice Activity Detection downstream task. VAD aims to identify and segment portions of an input audio recording that contain human speech, separating them from silence, breathing, background noise or music. For our experiments, it is treated as a binary frame classification problem with two mutually exclusive classes (speech and non speech).

Our downstream models uses the concatenation of the representations extracted from the CNN and the first transformer block of the SSL model. Based on empirical results, using only these two layers provide a good tradeoff between accuracy and inference speed. The concatenated hidden representations are frozen and then fed to a simple MLP classifier, described in Figure 4. We predict classes on frames sampled at 50Hz, aligned with the hidden representations frequency, and use the Viterbi algorithm for smoothing transition probabilities of 1% for *speech* to *non speech* and inversely.

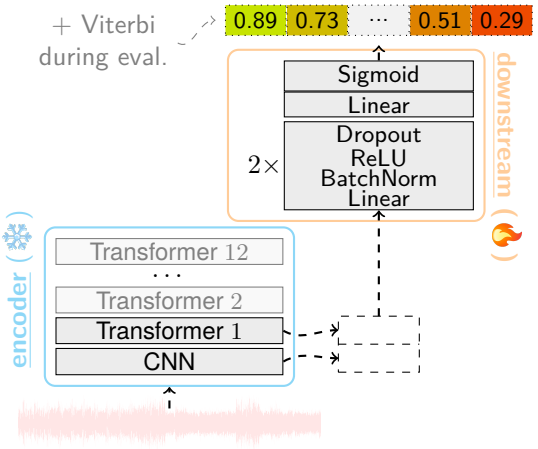


Figure 4: Overview of architecture used of the Voice Activity Detection and Music Detection models.

Models are trained on the *dev*[3] subset of InaG-VAD (Doukhan et al., 2024), a corpus designed to evaluate VAD systems on French audiovisual broadcast content.

The global and per-category results on InaG-VAD are presented in Table 4. We compare our downstream models with previous state-of-the-art VAD baselines: InaSpeechSegmenter (Doukhan et al., 2018) and pyannote.audio (Bredin and Laurent, 2021). While earlier systems struggled with generalist TV and music radio content, our trained models show a clear improvement and strong performances across all categories. Previous systems were typically trained on separate speech and mu-

[3]We sample it into training (80%) and validation sets.

| Model | *Global* | | generalist radio | generalist tv | music radio | news tv |
|---|---|---|---|---|---|---|
| | Acc | F1 | Acc | Acc | Acc | Acc |
| *InaSpeechSegmenter* | 93.0 ±2.4 | 94.3 ±2.0 | 95.4 ±3.9 | 88.3 ±4.1 | 98.1 ±3.1 | 94.9 ±2.8 |
| *pyannote* | 88.7 ±4.8 | 91.3 ±4.2 | 96.2 ±3.0 | 89.6 ±5.0 | 75.5 ±16.8 | 96.1 ±1.7 |
| *MFCC* | 89.9 ±2.7 | 91.4 ±2.9 | 91.3 ±5.8 | 86.3 ±4.6 | 92.2 ±5.9 | 94.1 ±2.9 |
| *m-a-p/music2vec-v1* | 96.4 ±1.1 | 97.0 ±1.0 | 96.5 ±2.7 | 94.6 ±2.1 | **98.5** ±1.7 | **97.7** ±1.2 |
| *FB/data2vec-audio-base* | 95.2 ±1.8 | 95.9 ±1.7 | 96.3 ±3.3 | 93.2 ±2.9 | 96.6 ±4.0 | 97.0 ±2.6 |
| `base` | **96.8** ±1.1 | **97.3** ±1.0 | **97.6** ±2.1 | **95.6** ±1.7 | 97.8 ±2.5 | **97.7** ±1.1 |
| `no_music` | 96.0 ±1.3 | 96.6 ±1.2 | 96.6 ±2.3 | 94.4 ±2.4 | 97.6 ±2.5 | 97.1 ±1.7 |
| `only_speech` | 96.2 ±1.2 | 96.8 ±1.2 | 96.7 ±2.3 | 94.9 ±2.0 | 97.2 ±3.2 | 97.4 ±1.1 |
| `only_fr` | 96.3 ±1.2 | 96.9 ±1.1 | 96.5 ±2.3 | 95.1 ±2.0 | 97.6 ±3.0 | 97.4 ±1.2 |
| `gender` | 96.5 ±1.3 | 97.0 ±1.2 | 97.1 ±2.1 | 95.3 ±1.9 | 97.2 ±3.5 | 97.6 ±1.1 |
| `duplicates` | 96.7 ±1.0 | 97.2 ±1.0 | 97.0 ±2.3 | 95.4 ±2.0 | **98.5** ±1.6 | 97.4 ±1.3 |

Table 4: Accuracy, and F1-Score of the different embedding representations models for the VAD task on the test set of InaGVAD (both globally and per channel category). Best results in **bold**, results within the confidence interval of the best model underlined.

sic audio files, whereas incorporating a more challenging broadcast type during training proved beneficial for our models. Overall, the `base` model performs best, although the music2vec encoder obtains good results as well.

### 5.3. Music Detection

Next, we evaluate our models on a music segmentation task. Given an input audio file, the goal of this task is to determine timestamps of segments containing music. As with VAD, we configure our downstream model to use the concatenation CNN representations as well as hidden representations of the 1st transformer block. This concatenation is fed into a MLP classifier, similar to the one used for VAD, described in Figure 4. We predict a binary class (music/no-music) at a 50Hz frequency. We use the Viterbi algorithm with transition probabilities of $5$% for *music* to *no-music* and vice-versa. To manage the memory consumption during training and evaluation, we slice audio files into chunks of 30s, with no overlap.

Our downstream models are trained on OpenBMAT (Meléndez-Catalán et al., 2019), the Music/Speech detection dataset of the Mirex 2015 competition (MIREX, 2015), and the music detection corpus presented by Seyerlehner et al. (2007), denoted as Seyerlehner.

We present in Table 5 our results. We can see that the music2vec-v1 pretrained model, presented by Li et al. (2022) achives the best overall results. This model has been pretrained on 1,000 hours of music audio content, and is therefore only suitable for MIR tasks. Next we observe that our models obtain the best results when pretrained on content which includes music and non-speech: `base`, `gender` and `duplicates`. The model using `no_music`, pretrained on content without music has a F1-Score 2.2% lower than it's `base` counterpart.

### 5.4. Speaker Recognition

We evaluate the representations extracted by our models (`base`, `only_speech`, and `gender`) and *MS/WavLM-base* (Chen et al., 2022a) using the VoxCeleb speaker recognition benchmark (Nagrani et al., 2017; Chung et al., 2018). To aggregate the representation layers, we use the attention-based MHFA backend described in Peng et al. (2023). The 12 transformer layers and the CNN layer are combined with the attention-based model to produce 256-dimensional embedding. The cosine distance is used to compute the test score between the test and the target embeddings. Only the backend model parameters were finetuned during 50 epochs using a subset of 1,000 speakers derived from the VoxCeleb2 development set with AAM-softmax, a margin of 0.2, and a scaling of 30. Both Equal Error Rate (EER) and the minimum Detection Cost Function (minDCF) are used to evaluate the speaker verification systems. The target probability $P_{tar}$ is set to $0.01$ or $0.05$, for DCF1 and DCF5, respectively. $C_{fa}$ and $C_{fr}$ have an equal weight of $1.0$.

As shown in Table 6, the performances of all evaluated models are consistent. The `gender` model achieves the best overall scores. This suggests that speaker identity information can be inferred from the hidden layer representations of our SSL model. WavLM achieves slightly better performance, likely due to its pretraining on English speech, which matches the language used in the VoxCeleb dataset. However, these scores remain below those usually reported by finetuned SSL models, indicating that finetuning SSL models is neces-

| Model | *Global* | | | Mirex2015 | OpenBMAT | Seyerlehner |
|---|---|---|---|---|---|---|
| | F1 ±0.1 | P ±0.1 | R ±0.1 | F1 ±0.1 | F1 ±0.1 | F1 ±0.2 |
| *MFCC* | 82.8 | 76.3 | 90.5 | 95.4 | 79.8 | 83.4 |
| *m-a-p/music2vec-v1* | **91.2** | **90.6** | **91.8** | 96.7 | **89.7** | **92.0** |
| *FB/data2vec-audio-base* | 87.5 | 84.6 | 90.6 | 97.0 | 85.4 | 87.7 |
| `base` | 89.4 | 89.4 | 89.3 | **97.3** | 87.1 | 91.0 |
| `no_music` | 87.1 | 86.5 | 87.7 | 96.9 | 84.5 | 88.4 |
| `only_speech` | 87.5 | 87.2 | 87.9 | 97.0 | 85.0 | 88.9 |
| `only_fr` | 87.7 | 84.2 | 91.6 | 97.2 | 85.1 | 89.8 |
| `gender` | 88.3 | 86.2 | 90.5 | **97.3** | 85.6 | 90.7 |
| `duplicates` | 88.9 | 88.4 | 89.4 | 97.2 | 86.5 | 90.8 |

Table 5: Frame-level F1-Score (F1), Precision (P) and Recall (R) for the music detection task globally and individually on each dataset. The confidence interval is the maximum of all models for a given column.

| Model | VoxCeleb1-O | | | VoxCeleb1-E | | | VoxCeleb1-H | | |
|---|---|---|---|---|---|---|---|---|---|
| | EER | minDCF1 | minDCF5 | EER | minDCF1 | minDCF5 | EER | minDCF1 | minDCF5 |
| *MS/WavLM-base* | **6.32** | **0.540** | **0.397** | **6.98** | **0.575** | **0.400** | **11.45** | **0.707** | **0.553** |
| `base` | 8.20 | 0.591 | 0.450 | 8.02 | 0.615 | 0.455 | 11.97 | 0.722 | 0.572 |
| `only_speech` | 8.00 | 0.634 | 0.454 | 7.84 | 0.612 | 0.440 | 11.50 | 0.710 | 0.561 |
| `gender` | 7.35 | 0.576 | 0.409 | 7.83 | 0.610 | 0.439 | 11.64 | 0.712 | 0.567 |

Table 6: EER(%) and minDCF of pretrained SSL models on the VoxCeleb Speaker Recognition benchmark.

sary to achieve higher performance, as finetuned models adapt their representations more effectively to the target application. During training, analysis of the learned weights of the backend model reveals higher weights for the first layers, a pattern that aligns with previous work (Chen et al., 2022a). This indicates that speaker identity features are represented in the first layers.

### 5.5. Membership Inference Attack

We evaluate the ability of the audio encoder to remember its pretraining dataset by performing a Membership Inference Attack (MIA). Our intuition behind this attack is to assess whether the encoder is able to memorize some data examples seen during the pretraining. While a model remembering it pretraining dataset would not necessarily enable its extraction or generation, a model not even able to remember its pretraining data would lower this risk.

First, we construct a 22-hour long training dataset, composed of 1320 segments seen once during the pretraining of `base` and `duplicates`, and 1320 segments never seen in any of them.

The MIA downstream model is described in Figure 5. It is composed of a weighted sum of hidden states (CNN+transformers layers), followed by a MLP. The weights of the weighted sum are learned along with the linear layers. We freeze the audio encoder during the training of the downstream model.

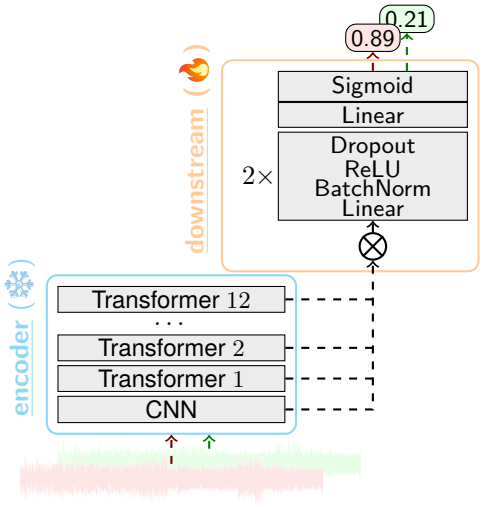


Figure 5: Overview of the proposed Membership Inference Attack (MIA) downstream model.

During training, we notice the downstream MIA model struggles to converge on the development set, exhibiting the difficulty of the task.

Next, for evaluating the MIA, we construct three different test sets of 10 hours (1,200 segments each): *unseen* where neither of the model has seen the examples during their pretraining ; *once* where both models saw the examples only once in the pretraining dataset ; *duplicated* where the `duplicates` model saw the examples 10 times in its pretraining

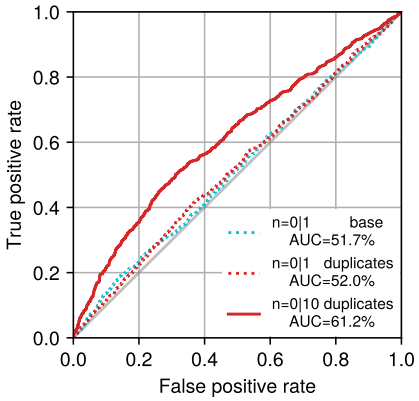


Figure 6: ROC curves for the membership inference attack.

dataset. Neither of the three sets overlaps with each other nor the 22h downstream training set.

We present in Figure 6 the Receiver Operating Characteristic (ROC) curves[4] for the `base` and `duplicates` models on *unseen* vs *once* ($n = 0|1$) and *unseen* vs *duplicated* ($n = 0|10$).

We can see that when testing on examples seen multiple times during pretraining ($n = 0|10$ for `duplicates` model), the attack is able to succeed better than by chance, with an Area Under the Curve (AUC) of 61.2%. The membership inference attack on `base` fail, with an AUC of 51.7% suggesting the model was not able to remember its pretraining dataset. We notice that the `duplicates` model was not able to remember examples seen once ($n = 0|1$) with an AUC of 52.0%. Overall, this suggests that pretraining with duplicates elements do not make unduplicated segments more recoverable, contrary to duplicated elements, which become recoverable with an AUC of 61.2%.

## 6. Conclusion

In this paper, we construct a 100,000 hours pretraining corpus of audiovisual TV and Radio content. We build 6 pretrained audio SSL models that we benchmark on various downstream evaluations.

Our observations shows that for speech recognition, pretraining on content without music improves the results compared to more diverse content. Gender-wise, pretraining on as much male than female speech seems to reduce the gender bias between male and female WER. On voice activity detection, pretraining on diverse audiovisual data obtains the best results. Regarding music detection, we observe that pretraining purely on music achieves the best results, although training on various audiovisual content obtains close results. For speaker recognition, our models achieved performance comparable to another speech SSL encoder, with the Gender-wise model achieving the best results. Overall, across all tasks, we can say that the results between downstream models are close enough to the specialized models that we can consider pretraining a general purpose audio model on the whole 100,000 hours corpus, suitable for both music and speech tasks. This experiment is left for future works. Lastly, the membership attack highlighted the importance of data deduplication, with a great reduction of training data memorization.

[4]For the sake of clarity, the ROC curve for the `base` model on $n = 0|10$ is not shown. It has an AUC of 50.5%.

## 7. Acknowledgments

The authors would like to thank Jean-Hugues Chenot, Nicolas Hervé and Sandrine Depoix for their help during the construction of the INA dataset and with the audio deduplication tool. They also thank Aude Formagne regarding the legal challenges of publishing the pretrained models.

This research has been funded by the French National Research Agency (ANR), project “Pantagruel” (ANR-23-IAS1-0001). This work was performed using HPC resources from GENCI at IDRIS and CINES under the allocations 2022-A0131013801, 2023-A0151013801, 2024-A0171013801, 2024-A0161015074, and 2025-A0191013801 on the Jean Zay and Adastra supercomputers.

## 8. Bibliographical References

Martine Adda-Decker and Lori Lamel. 2005. Do speech recognizers prefer female speakers? In *Interspeech 2005*, pages 2205–2208.

Giuseppe Attanasio, Beatrice Savoldi, Dennis Fucci, and Dirk Hovy. 2024. Twists, humps, and pebbles: Multilingual speech recognition models exhibit gender performance gaps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21318–21340, Miami, Florida, USA. Association for Computational Linguistics.

Alexei Baevski, Arun Babu, Wei-Ning Hsu, and Michael Auli. 2023. Efficient self-supervised learning with contextualized target representations for vision, speech and language. In *International Conference on Machine Learning*, pages 1416–1429. PMLR.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In *Interspeech 2021*, pages 3111–3115.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022a. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng. 2022b. Large-scale self-supervised speech representation learning for automatic speaker verification. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6147–6151.

Jean-Hugues Chenot and Gilles Daigneault. 2014. A Large-Scale Audio and Video Fingerprints-Generated Database of TV Repeated Contents. In *12th International Workshop on Content-Based Multimedia Indexing (CBMI2014)*, pages 1–6, Klagenfurt, Austria.

J. S. Chung, A. Nagrani, and A. Zisserman. 2018. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*.

David Doukhan, Jean Carrive, Felicien Vallet, Anthony Larcher, and Sylvain Meignier. 2018. An open-source speaker gender detection framework for monitoring gender equality. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5214–5218.

Marcio Fuckner, Sophie Horsman, Pascal Wiggers, and Iskaj Janssen. 2023. Uncovering bias in asr systems: Evaluating wav2vec2 and whisper for dutch speakers. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 146–151.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2019. Gender representation in french broadcast corpora and its impact on asr performance. In *Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery*, AI4TV '19, page 3–9, New York, NY, USA. Association for Computing Machinery.

Pablo Gimeno, Alfonso Ortega, Antonio Miguel, and Eduardo Lleida. 2021. Unsupervised representation learning for speech activity detection in the fearless steps challenge 2021. In *Interspeech 2021*, pages 4359–4363.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 10697–10707. PMLR.

Biswajit Karan, Joshua Jansen van Vüren, Febe de Wet, and Thomas Niesler. 2024. A Transformer-Based Voice Activity Detector. In *Interspeech 2024*, pages 3819–3823.

Marie Kunešová and Zbyněk Zajíc. 2023. Multitask detection of speaker changes, overlapping speech and voice activity using wav2vec 2.0. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Y. Li, R. Yuan, G. Zhang, Y. MA, C. Lin, X. Chen, A. Ragni, H. Yin, Z. Hu, H. He, E. Benetos, N. Gyenge, R. Liu, and J. Fu. 2022. Lv-49: Map-music2vec: A simple and effective baseline for self-supervised music audio representation learning. In *23rd International Society for Music Information Retrieval Conference (ISMIR 2022)*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

A. Nagrani, J. S. Chung, and A. Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*.

Sergey Novoselov, Galina Lavrentyeva, Anastasia Avdeeva, Vladimir Volokhov, Nikita Khmelev, Artem Akulov, and Polina Leonteva. 2023. On the robustness of wav2vec 2.0 based speaker recognition systems. In *Interspeech 2023*, pages 3177–3181.

Titouan Parcollet, Ha Nguyen, Solène Evain, Marcely Zanon Boito, Adrien Pupier, Salima Mdhaffar, Hang Le, Sina Alisamir, Natalia Tomashenko, Marco Dinarelli, Shucong Zhang, Alexandre Allauzen, Maximin Coavoux, Yannick Estève, Mickael Rouvier, Jerôme Goulian, Benjamin Lecouteux, François Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, and Laurent Besacier. 2024. Lebenchmark 2.0: A standardized, replicable and enhanced framework for self-supervised representations of french speech. *Computer Speech & Language*, 86:101622.

Junyi Peng, Oldřich Plchot, Themos Stafylakis, Ladislav Mošner, Lukáš Burget, and Jan Černocký. 2023. An attention-based backend allowing efficient fine-tuning of transformer models for speaker verification. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 555–562.

Shengyu Peng, Wu Guo, Haochen Wu, Zuoliang Li, and Jie Zhang. 2024. Fine-tune Pre-Trained Models with Multi-Level Feature Fusion for Speaker Verification. In *Interspeech 2024*, pages 2110–2114.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Wei-Cheng Tseng, Wei-Tsung Kao, and Hung yi Lee. 2022. Membership Inference Attacks Against Self-supervised Speech Models. In *Interspeech 2022*, pages 5040–5044.

Biwei Yan, Kun Li, Minghui Xu, Yueyan Dong, Yue Zhang, Zhaochun Ren, and Xiuzhen Cheng. 2024. On protecting the data privacy of large language models (llms): A survey.

Marcely Zanon Boito, Laurent Besacier, Natalia Tomashenko, and Yannick Estève. 2022. A Study of Gender Impact in Self-supervised Models for Speech-to-Text Systems. In *Interspeech 2022*, pages 1278–1282.

Marcely Zanon Boito, Vivek Iyer, Nikolaos Lagos, Laurent Besacier, and Ioan Calapodescu. 2024. mHuBERT-147: A Compact Multilingual HuBERT Model. In *Interspeech 2024*, pages 3939–3943.

## 9. Language Resource References

Christophe Benzitoun, Jeanne-Marie Debaisieux, and Henri-José Deulofeu. 2016. Le projet orféo: un corpus d'étude pour le français contemporain. *Corpus*, (15).

Karim Boudahmane, Bianka Buschbeck, Eunah Cho, Josep Maria Crego, Markus Freitag, Thomas Lavergne, Hermann Ney, Jan Niehues, Stephan Peitz, Jean Senellart, Artem Sokolov, Alex Waibel, Tonio Wandmacher, Joern Wuebker, and François Yvon. 2011. Advances on spoken language translation in the quaero program. In *Proceedings of the 8th International Workshop on Spoken Language Translation: Evaluation Campaign*, pages 114–120, San Francisco, California.

Jean Carrive, Abdelkrim Beloued, Pascale Goetschel, Serge Heiden, Antoine Laurent, Pasquale Lisena, Franck Mazuet, Sylvain Meignier, Bénédicte Pincemin, Géraldine Poels, and Raphaël Troncy. 2021. Transdisciplinary Analysis of a Corpus of French Newsreels: The ANTRACT Project. *Digital Humanities Quarterly*, 15(1). Editors: Taylor Arnold, Jasmijn van Gorp, Stefania Scagliola, and Lauren Tilton.

Sourish Chaudhuri, Joseph Roth, Daniel P. W. Ellis, Andrew Gallagher, Liat Kaver, Radhika Marvin, Caroline Pantofaru, Nathan Reale, Loretta Guarino Reid, Kevin Wilson, and Zhonghua Xi. 2018. Ava-speech: A densely labeled dataset of speech activity in movies. In *Interspeech 2018*, pages 1239–1243.

David Doukhan, Christine Maertens, William Le Personnic, Ludovic Speroni, and Reda Dehak. 2024. InaGVAD : A challenging French TV and radio corpus annotated for speech activity detection and speaker gender segmentation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8963–8974, Torino, Italia. ELRA and ICCL.

Yannick Estève, Thierry Bazillon, Jean-Yves Antoine, Frédéric Béchet, and Jérôme Farinas. 2010. The EPAC corpus: Manual and automatic annotations of conversational speech in French broadcast news. In *Proceedings of the*

*Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. 2009. The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Interspeech 2009*, pages 2583–2586.

Aude Giraudel, Matthieu Carré, Valérie Mapelli, Juliette Kahn, Olivier Galibert, and Ludovic Quintard. 2012. The REPERE corpus : a multimodal corpus for person recognition. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1102–1107, Istanbul, Turkey. European Language Resources Association (ELRA).

G. Gravier, J-F. Bonastre, E. Geoffrois, S. Galliano, K. McTait, and K. Choukri. 2004. The ESTER evaluation campaign for the rich transcription of French broadcast news. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Guillaume Gravier, Gilles Adda, Niklas Paulsson, 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 *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 114–118, Istanbul, Turkey. European Language Resources Association (ELRA).

Anne Lacheret, Sylvain Kahane, Julie Beliao, Anne Dister, Kim Gerdes, Jean-Philippe Goldman, Nicolas Obin, Paola Pietrandrea, and Atanas Tchobanov. 2014. Rhapsodie: a prosodic-syntactic treebank for spoken French. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 295–301, Reykjavik, Iceland. European Language Resources Association (ELRA).

Arthur Lezer. 2022. L'analyse des médias au lab INA.

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):43–51.

MIREX. 2015. Music/Speech classification and detection dataset. https://www.music-ir.org/mirex/wiki/2015:Music/Speech_Classification_and_Detection.

Valentin Pelloin, Léna Dodson, Émile Chapuis, Nicolas Hervé, and David Doukhan. 2024. Automatic Classification of News Subjects in Broadcast News: Application to a Gender Bias Representation Analysis. In *Interspeech 2024*, pages 3055–3059.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411.

Klaus Seyerlehner, Tim Pohle, Markus Schedl, and Gerhard Widmer. 2007. Automatic music detection in television productions. In *Proc. of the 10th International Conference on Digital Audio Effects (DAFx'07)*. SCRIME/LaBRI Bordeaux.