A Semi-Automatic Approach to Create Large Gender-and Age-Balanced Speaker Corpora: Usefulness of Speaker Diarization & Identification

Rémi Uro, David Doukhan, Albert Rilliard, Laëtitia Larcher, Anissa-Claire Adgharouamane, Marie Tahon, Antoine Laurent

Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3271–3280
Marseille, 20-25 June 2022

A Semi-Automatic Approach to Create Large Gender- and Age-Balanced Speaker Corpora: Usefulness of Speaker Diarization & Identification.

Rémi Uro[1,2], David Doukhan[1], Albert Rilliard[2,3], Laëtitia Larcher[1], Anissa-Claire Adgharouamane[1], Marie Tahon[4], Antoine Laurent[4]
[1]French National Institute of Audiovisual (INA), Paris, France. [2]Université Paris Saclay, CNRS, LISN.
[3]Universidade Federal do Rio de Janeiro, Brasil. [4]LIUM, Le Mans Université, France.
{ruro,ddoukhan,llarcher,aadgharouamane}@ina.fr
albert.rilliard@lisn.fr, {marie.tahon,antoine.laurent}@univ-lemans.fr

Abstract

This paper presents a semi-automatic approach to create a diachronic corpus of voices balanced for speaker's age, gender, and recording period, according to 32 categories (2 genders, 4 age ranges and 4 recording periods). Corpora were selected at French National Institute of Audiovisual (INA) to obtain at least 30 speakers per category (a total of 960 speakers; only 874 have be found yet). For each speaker, speech excerpts were extracted from audiovisual documents using an automatic pipeline consisting of speech detection, background music and overlapped speech removal and speaker diarization, used to present clean speaker segments to human annotators identifying target speakers. This pipeline proved highly effective, cutting down manual processing by a factor of ten. Evaluation of the quality of the automatic processing and of the final output is provided. It shows the automatic processing compare to up-to-date process, and that the output provides high quality speech for most of the selected excerpts. This method shows promise for creating large corpora of known target speakers.

Keywords: semi-automatic processing, corpus creation, diarization, speaker identification, gender-balanced, age-balanced, speaker corpus, diachrony

1. Introduction

This paper describes a semi-automatic method for speeding-up speaker corpora construction. There is a growing need for reference speech corpora featuring reliable and known speaker characteristics. Manual annotations of high level features such as speaker segmentation is a time consuming process (Broux et al., 2018). We propose to use up-to-date diarization and speaker identification to reduce human intervention, and enable the creation of large spoken corpora at a minimum cost.

This work is part of the Gender Equality Monitor (GEM) project, that aims at describing male and female representation differences in the French broadcast media at scale, using automatic information extraction methods. A main institution involved in the project is the French National Institute of Audiovisual (INA), a public institution in charge of archiving French audiovisual heritage (so called legal deposit). INA's collections include 23 million hours of TV and radio programs broadcasted since the 1930's.

First, the targeted characteristics of the corpus are presented, with a discussion of comparable datasets, and the potential use of this resource. A literature review on diarization and speaker identification systems is then presented. The rest of the paper presents: (1) the methodological steps for the creation of this corpus, (2) evaluations of these steps in terms of performance on available data with comparable characteristics, (3) a subjective evaluation of the quality of voice of the final output, and (4) an estimation of processing time reduction compared to manual processing.

2. Targeted corpus and existing resources

2.1. Voices & gender over time

For the needs of the GEM project, the capacity to describe the acoustic characteristics of a given speaker's voice is important, to compare to its socially recognized characteristics, that predominantly includes the speaker's gender and age. Having a reliable statistical representation (i.e., based on representative data for several age and gender categories) of the acoustic variation of voices presented in the public arena allows sociological analysis of gender representation, gender stereotypical characteristics, potential changes in voice as a device to present one's self public image, etc. The voice is an important aspect of the construction of individual social persona or character (Podesva, 2007; Sadanobu, 2015); it changes according to our role in society — e.g., while talking as an employee to a superior, as a professor to students, as a parent to children, as a friend to sporting mates etc. Our voice is developed during childhood and adolescence, varying because of our physiological development (Vorperian et al., 2011), but also as a culturally shaped representation of our social projection as an individual (Sergeant and Welch, 2009; Guzman et al., 2014; Scott, 2022). Gender is a key aspect of the construction of voices, and is culturally shaped: speakers have been shown to have different mean pitch in different cultures, and to change their pitch for reaching culture-matching expectations, notably on gender (van Bezooijen, 1995; Ohara, 1992;

Ohara, 2001). Changes in gender-related vocal cues have effects on many aspects of social interaction, pitch having been related to a series of characteristics like credibility, charisma, cuteness, sexual attractiveness, etc. (Geiselman and Crawley, 1983; Niebuhr et al., 2016; Jang, 2021; Gussenhoven, 2016).

To better document the aspects of self that our voices may voluntarily or not display (Scott, 2022), a reference corpus of voices is a great tool, allowing studies on voice presentation in the public arena via broadcast media in France (the GEM project is centered on this country). This corpus would ideally contain voices used in comparable dialogue situations – i.e., not expressive situations where voice may be especially loud, or express emotions, two aspects which have important effects on vocal characteristics (Titze and Sundberg, 1992; Liénard and Di Benedetto, 1999; Traunmüller and Eriksson, 2000; Liénard, 2019; Goudbeek and Scherer, 2010). Although such phenomena may be interesting to describe, they add variability that would require more data to be controlled for. Thus broadcast programs featuring interviews or discussions with invited people were selected, mostly those recorded in studio situations. Note that audiovisual program metadata does not allow a full control of the recording settings, especially as it is common to find reporting on a given topic, that may happens to feature the target speaker. Details on the audiovisual document selection process are given in the next section.

To achieve accurate representation of voices, through their long-term acoustic qualities, it is mandatory to obtain a duration that allows acoustic characteristics to stabilize over articulation and other dynamic aspects. Löfqvist and Mandersson (1987) showed Long Term Average Spectrum is stabilized at about 20 seconds of continuously voiced speech – so about twice this duration for raw speech; see also Arantes and Eriksson (2014) on prosody. We set a lower limit of three minutes of diarized speech per speaker. It was thought important so to limit the influence of noises and potential backchannels.

2.2. Available resources

As far as we know, there is no available speech corpora providing balanced distribution of speaker gender, age and recording date for French media.

For resources in French, various corpora are available via the cocoon website[1] – that mostly proposes dialectological, sociological or ethnological resources. The quality, type of speech, and information available about each speaker is highly variable. One of these resources, the ESLO corpus (Baude, 2019), proposes interviews of many residents of Orleans city, made at two periods: beginning of 1970's and of 2010's (Eshkol-Taravella et al., 2011). If an interesting resource, it is strictly restricted to one city and two time periods; it is also based on field recordings of variable recording qualities.

[1] https://cocoon.huma-num.fr/exist/crdo

Another set of corpora was developed during evaluation campaigns for NLP tools. For example, the ESTER corpus provides sounds and transcriptions for one hundred hours of broadcast news from French media (Galliano et al., 2012). The ETAPE corpus is a follow-on development with less prepared speech from other types of radio or TV programs (Gravier et al., 2012). These corpora contain information on speaker gender, but not directly on speaker age. Unfortunately, none of them present a diachronic dimension. One resource, the Eurodelphes database (Barras et al., 2002), proposes a set of broadcast documents spread from the 1940's up to the 1990's; it is nonetheless relatively limited in size for the oldest decades, and is highly unbalanced for gender and age (Boula de Mareüil et al., 2012), thus not suited for the targeted use.

The corpus presented in Suire and Barkat-Defradas (2020) is the closest to what we are trying to build. It is based on media programs with about thirty speakers of each gender selected by period of ten years, over seven decades, with about 5 seconds of speech for each speakers; the age of speakers was not informed. These two limitations (short extracts, and no information on speaker's age) make it unsuitable for the study.

2.3. Existing (semi-)automatic speaker corpora

Building audiovisual speaker databases is costly, since it requires finding speakers in audiovisual documents and obtaining the time codes for speakers' speech turns. To that aim, few fully or partially automated approaches were proposed for building large speaker databases with minimal human involvement.

INA's speaker dictionary was prepared using a semi-automatic procedure based on unsupervised speaker segmentation (diarization) and Optical Character Recognition (OCR) (Salmon and Vallet, 2014; Vallet et al., 2016). OCR decoded embedded text presenting people in TV news programs, and filtered characters corresponding to the first twenty thousand most referenced people in INA's audiovisual databases. Speech segments corresponding to decoded embedded text were on a second stage presented to human annotators in charge of stating if speech segments were corresponding to the decoded person name, resulting in an average involvement time of 22,8 seconds per segment.

VoxCeleb was built from YouTube videos using a fully automatic procedure (Nagrani et al., 2017). YouTube video queries were applied using target speakers' names and the word 'interview'. The set of target speakers was a subset of celebrities known by a preexisting face verification classifier. Active speaker verification models were used to detect the portions of video with facial lip movements synchronized with the audio track. Face candidates were in a latter stage presented to the face verification classifier using a high threshold.

While these approaches allowed to build large speaker collections, they suffer from several limitations. Both

of them require video material, and cannot be used for processing radio collections. INA's speaker dictionary's strategy requires embedded text to be displayed during target's speech turn, while VoxCeleb's strategy requires target speaker's face to be already known by a face verification classifier. Moreover, YouTube's queries do not allow search by speakers' age.

The corpus-building process aims at creating a large corpus (more than 960 speakers) of gender- and age-balanced speakers that may serve as a reference of voice qualities (linked to gender representation) presented in broadcast media, from the 1950's until now. As a manual gathering is out of reach, we aimed at limiting to a minimum manual intervention so to speed the process. This paper present the details of the approach, with an evaluation of the automatized process, and of the quality of the obtained dataset, as well as an estimation of the time gained through the semi-automated method, compared to a fully handmade process.

3. Methods

Speaker selection guidelines were defined to obtain a diachronic speaker corpus with balanced gender, age, and recording period. These guidelines were provided to INA's archivists to constitute a corpus of audiovisual documents containing a balanced amount of target speakers. Figure 1 show the semi-automatic pipeline designed to extract target speaker excerpts from this huge collection of audiovisual documents with a minimal amount of human involvement. A *Clean Speech Detection* process was defined to discard speech segments with acoustic properties that may interfere with acoustic parameter extraction. An unsupervised speaker segmentation and clustering procedure (diarization) was used on the resulting clean speech segments in order to assign numeric identifiers to each speaker found in the recording. Resulting speaker segmentations were then presented to human annotators in charge of the manual identification of the target speakers. If the found speaker excerpts in the manually processed documents are less than three minutes, these excerpts were used to perform automatic cross-document speaker retrieval, to complete the speaker sample in the corpus.

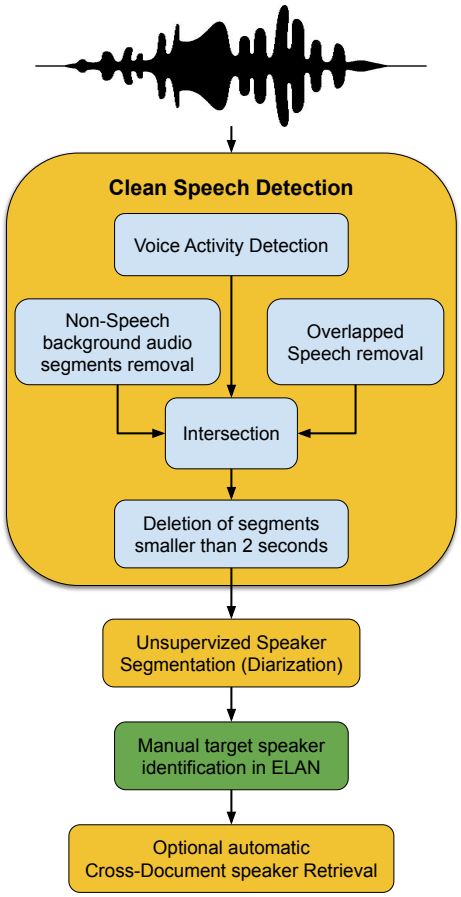

Figure 1: Semi-automatic pipeline proposed for the extraction of *clean* target speaker segments

3.1. Balanced diachronic speaker corpus definition

To avoid environmental noise and vocal changes linked to stylistic variations, broadcast programs featuring dialog and interviews, recorded indoors, are prioritised.

A total of 30 distinct speakers is required for each of **32 adult speaker categories** (adding up to a total of at least 960 different speakers), based on the combination of 3 parameters: **gender** (male and female), **age** (4 groups: 20 to 35, 35 to 50, 51 to 65 and over 65 years old), and **periods** of time (4 periods: 1955-1956, 1975-1976, 1995-1996, 2015-2016). The age groups were based on known changes in voice linked to age and gender (Sataloff et al., 2017; Yamauchi et al., 2015). The periods of time span from the 1950's to the 2010's with 20 years intervals: this is somewhat arbitrary, but these periods were chosen because it is harder to find archives before the 1950's, especially featuring female speakers, and four periods were thought sufficient, and limit the amount of target speakers. Each speaker was selected only in one of the 32 categories to avoid statistical dependency; we are thinking about preparing a section of the corpus with longitudinal recordings of speakers available in three to four time periods.

3.2. Audiovisual document selection

Document selection based on the corpus definition was realized by INA's archivists, and required 3 weeks of work. The list of participants and the date of diffusion was extracted from TV and radio archives meta-data, and linked to INA's personality thesaurus to obtain date of birth and gender information. This allowed to assign each unique speaker (our "target speakers") to one of the 32 gender, age, and period categories. A manual selection of 450 TV and radio shows was realized, that usually feature reasonably long studio-recorded interviews, well-known personalities, low amounts of background music and noise, and a low amount of conflict-

Period	# TV docs	# Radio docs	# Unique speakers	Duration (days)
1955-56	133	508	594	11
1975-76	849	642	1220	46
1995-96	1565	4686	2393	176
2015-16	933	7991	1845	160

Table 1: Characteristics of the speaker corpus collected by INA's archivists: number of documents (docs) from TV and radio, of speakers, and total duration.

ing interactions between participants.
While the number of male found for each category turned out to be greater than thirty, several categories of female had to benefit from additional manual retrieval in INA's collections to reach the minimum amount of speakers required to meet corpus specifications. Female stand out more rarely in broadcast media (Doukhan et al., 2018b), and their number was particularly scarce within the older age groups and in the oldest periods in INA thesaurus, possibly reflecting the gender bias of media presentation and representation (Tuchman, 2000). Research for complementary speakers was necessary and resulted in the enrichment of INA's personality thesaurus, with the inclusion of the characteristics of 182 media personalities.
Table 1 presents the characteristics of the corpus constituted by INA's archivists, with 25092 entries for 6051 distinct speakers, distributed between 17307 audiovisual documents, for a total duration of about 393 days of content (9400 hours). This is clearly too large to consider a complete manual exploration, and requires the semi-automatic procedure described below.

3.3. Clean Speech Detection (CSD)

A *Clean Speech Detection* procedure was proposed to detect the cleanest speech excerpts, suited to prosodic parameter extraction. Figure 1 describes the components of our proposal, which allows to obtain speech segments with low amount of overlapped speech, background music or noise. Speech segments shorter than 2 seconds were rejected, as they are likely to be only small parts of sentences and hence be of little interest to achieve an accurate representation of the voices.

3.3.1. Voice Activity Detection (VAD) and Overlapped Speech Detection (OVL)

Voice activity detection was performed using `InaSpeechSegmenter` (Doukhan et al., 2018a). This system is based on a CNN architecture trained to distinguish speech from music and noise. `InaSpeechSegmenter` was ranked in first position for the VAD task of MIREX 2018 challenge, containing TV and radio corpora which are representative of our target material[2].
In order to isolate the segments corresponding to only one speaker, `pyannote-audio`'s overlap speech detection (Bredin et al., 2020; Bullock et al., 2020) was also used. The detected overlapping speech segments were then cut from the initial VAD.

[2] `https://www.music-ir.org/mirex/wiki/`

Level	bgvl	bg	similar	fg	music
Recall (%)	65	89.9	97.0	97.2	99

Table 2: Music detection recall obtained on OpenBMAT for varying music levels: bgvl (hard-to-hear background music), bg (background music), similar (music and other signals mixed at similar levels), fg (foreground music), music (music only)

3.3.2. Non-Speech audio event detection

Audiovisual documents may contain non-speech audio events overlapped with speech, such as music or noise. Such events may interfere with vocal feature estimation and excerpts with these events should be discarded.
A non-speech audio event detection model is proposed, based on `spleeter` source separation framework (Hennequin et al., 2020). We used `spleeter` vocals and instrumental accompaniment separation model. The potential presence of non-speech audio events was linked to the energy of the extracted instrumental accompaniment track, estimated using the root mean square of signal with 200 ms window size and 100 ms hop size. The energy is filtered using a median filter of size 11, and a threshold set at 5% was used.
The evaluation of the non-speech audio detection model is difficult, due to a lack of annotated resources containing overlapping speech, music and noise annotations. Table 2 presents the evaluation of our proposal on OpenBMAT, a database of audio streams with annotated music levels (Meléndez-Catalán et al., 2019). With respect to our use-case (obtaining clean speech, even with a low document coverage rate) and to the availability of annotated resources (no dataset with speech, music and noise annotations), we used `sed_eval` segment-based detection recall for estimating the performance of our proposal, using time tolerance (collar) of 1 second (Mesaros et al., 2016). The results show detection performance above 90% for audible music, and 65% for hard-to-hear background music, which shall less affect acoustic analysis.

3.3.3. Clean Speech Detection pipeline coverage

We tested our pre-processing pipeline on the DIHARD II Development set (Ryant et al., 2019). Table 3 shows the duration of detected speech at different stages of the pre-processing.DIHARD II focuses on *hard* diarization, i.e. with lots of low volume background speech, *in the wild* speech with music, noise and overlapping speech. Note that we target clean speech: low level noisy speech is not of interest for our intended prosodic analyses. Our CSD system eliminates more than half of the total speech time. Since we value a better precision than recall, considering that only about 40% of a given corpus is usable seems enough. Moreover, one can assume that the TV and radio broadcast documents that

Method	Duration (s)	Coverage
Reference	72311	100%
Ref+OVL	63781	88.2%
Ref+NSE	32013	44.3%
VAD	69247	89.1%
VAD+OVL	56953	78.8%
VAD+NSE	30166	41.7%
VAD+OVL+NSE	28360	39.2%
CSD	23980	**33.2%**

Table 3: Duration of detected speech on the DIHARD II Dev set for the different pre-processing steps and coverage relatively to the reference. (VAD: only VAD; OVL: overlapped speech removal; NSE: non-speech events removal; CSD: clean speech detection –VAD+OVL+NSE + removal of segments less than 2 seconds)

Input	DER (%)
Reference	23.8
VAD	24.5
VAD+OVL	21.3
VAD+NSE	16.5
CSD	**14.7**

Table 4: Performance of the diarization system (measured by DER) on DIHARD II dev set for the different stages of pre-processing. (collar=0.25)

this system targets should contain more clean speech than DIHARD II documents. So the process shall remove a smaller part of the content on such documents.

3.4. Diarization with VBx

We used our CSD as an input VAD for diarization, meaning that *clean* speech is considered as *speech* and *non-clean* speech as *non-speech*. We use the x-vector based diarization system VBx (Landini et al., 2022) with the `ResNet101_16kHz` model, pretrained on VoxCeleb1 (Nagrani et al., 2017), VoxCeleb2 (Chung et al., 2018) and CN-CELEB (Fan et al., 2020). The diarization step outputs clusters id corresponding to a unique speaker.

We have evaluated the VBx model using our CSD on the DIHARD II Development set (Ryant et al., 2019). Table 4 shows the Diariation Error Rate (DER) for the different stages of pre-processing. The DER is computed by removing *non-clean* segments from the reference. As expected, we observe a better DER when *non-clean* speech segments are removed, the diarization task being easier. We obtained a DER of 14.7 with a 0.25s collar using our pre-processing pipeline, which is comparable to the DER of 12.23% obtained by (Landini et al., 2022) with oracle VAD.

3.5. Manual speaker identification

Each audiovisual document of the source corpus was associated to a list of target speakers with known age, gender and role (anchor, participant) provided by INA's archivists (see section 3.2), to be manually identified. The *clean speech diarization* described above was exported to ELAN video annotation tool and presented to human annotators, together with the list of target speakers (Sloetjes and Wittenburg, 2008).

For each document, annotators had to map diarization cluster id's to target speakers' identities.

The complexity of this task varies a lot depending on the type of document and the role of the target. For instance a recent TV interview with only two speakers can be processed in a few seconds, whereas an old radio show with multiple characters, mostly unknown nowadays, may require to listen almost all the document, and sometimes the use of internet to find photos or details to spot the target.

3.6. Automatic cross-show speaker identification

The corpus aims at presenting at least three minutes of speech by speaker. For most documents, the manual identification described in 3.5 was enough because the documents were chosen to maximise the speaking time of the target speakers. However, it is not the case for all documents. Then, the segments linked to the target speaker in the manually annotated document were used as a reference to automatically identify this speaker in other documents.

Using VBx x-vector extraction model, we retrieve one x-vector per segment corresponding to the target speaker in the annotated document, giving us a matrix x_{known}. The non-annotated document was then pre-processed and automatically diarized in the same way, and one x-vector per segment of this document was extracted, giving us a matrix x_{target}. Cosine similarity was computed between all the vectors in x_{known} and x_{target}. For each vector in x_{target}, the mean similarity to the x_{known} vectors gives the probability score of the vector corresponding to the target speaker. If this score exceeds a given threshold, we considered that the segment corresponds to the target speaker. If no segment had a score above the threshold, we considered the target speaker absent from the document. We chose to focus on segment-level identification in order to maximize the precision, even though it may increase the false negative rate.

We evaluated our speaker identification system on INA speaker dictionary (Vallet et al., 2016) which contains materials extracted from French TV archives and is similar to the contents our system was designed for. This dataset contains about 1300 speakers extracted from French TV broadcasts. We analyse the similarity score between three different types of recording pairs: the same speaker in two different recording sessions, two different speakers of the same gender and two different speakers of different gender. This was evaluated on a gender-balanced subset of 718 speakers allowing for each speaker to be in at least two recording sessions,

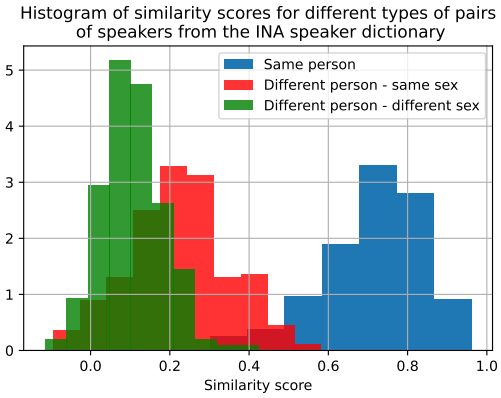

Figure 2: Histogram of similarity scores for different types of pairs of speakers from the INA speaker dictionary

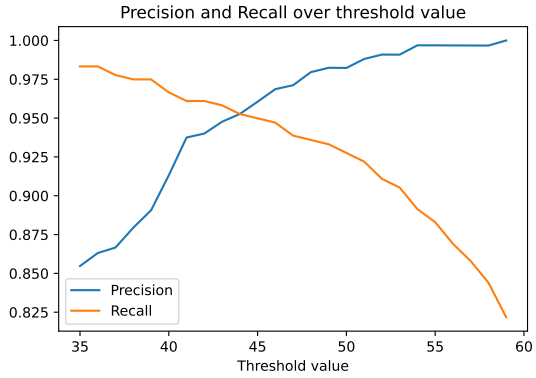

Figure 3: Precision and Recall over threshold values on the INA speaker dictionary

ending up with 359 pairs for each type. The mean duration of the used segments was 14s. Figure 2 shows the distribution of similarity scores for the different types of pairs. There is few overlap between the scores corresponding to the same speaker and those corresponding to different speakers. There is very little (4%) overlap between scores for the same speaker and scores for speakers of different genders.

Speaker identification was evaluated according to the Equal Error Rate (EER). We obtain an EER of 3.9%[3] on the INA speaker dictionary for a similarity threshold equal to 0.40. However, because precision is more important than recall here, we decided to use a threshold equal to 0.52, at which we obtain a precision of 0.99 and a recall of 0.91 (see Figure 3).

3.7. Subjective quality evaluation

At the end of this selection process, for each target speaker, series of speech segments were available amounting to at least three minutes by speakers (i.e. for the complete corpus, that would amount to about 50 hours of speech). In order to subjectively evaluate the quality of the automatic process, and as it is hardly feasible to listen to the complete corpus, we opted for applying a perceptual annotation on a subset of the available extracts. The subset was composed by a random selection of one segment for each target speaker; these segments were annotated for the presence of the following potential problems: backchannel, more than one person speaking, musical background, and audible noise. Backchannel was defined as up to two syllables produced by another speaker than the target; if two speakers spoke more than two syllables, it was annotated as more than one person speaking. Audible background music or noise, while listening the extracts with headphones, were annotated as such.

The available extracts were divided in three parts, assigned to three different annotators, with a 309-large subset annotated by the three annotators. The extracts of the common part were selected so as to propose up to 10 extracts per category of period, gender and age (let's recall we aimed at collecting 30 speakers per category); this shall amount to 320 extracts if a sufficient number of speakers were available in each category, which was not the case. The females of 51 to 65 year-old in the 1955-1956 period, and over 65 year-old in the 1975-1976 period were only 5 and 4 in the corpus currently available, so the final number of 309 (see details in the result section for available speakers).

A python notebook was used to randomly load one extract, play it, and request the annotator to answer a series of letters indicating the presence of the potential problems. A field for free comments was available, before hearing the next extract.

4. Results

4.1. Obtained target speakers

Each target speaker was manually identified for one archive. This manual identification work took about 140 hours along 20 days. The available speakers are detailed in table 5. Note that it corresponds to more than 40 hours of final speech (more than 3 minutes for 874 speakers). A total of 533 male and 341 female targets have been identified – thus 874 different speakers. Sixteen categories of speakers out of 32 did not reach the aim of at least thirty speakers. Four groups from the 1990's and 2010's (3 females groups) were almost complete, only missing a few speakers. For the other groups, ten have between 10 to 20 speakers (four of them are male groups), and two female groups have only 4 and 5 speakers. A total of 211 speakers (22% of our target) are missing, so to complete all groups with at least 30 speakers. The sixteen groups with at least 30 speakers present 125 extra speakers.

For each category, more than thirty targets have been searched for, but for a series of reasons, in some cases the requirements were not met. These problems overwhelmingly arise for female targets, known to be

[3]Previous work on the same corpus(Vallet et al., 2016) obtained an EER of 7.3% using an equivalent evaluation protocol.

	20-35	36-50	51-65	>65
1955-56	**13**/34	**17**/61	**5/19**	**17/10**
1975-76	**16/14**	**18**/37	**11**/31	**4/11**
1995-96	30/**27**	32/47	**29**/48	**29**/35
2015-16	31/30	**29**/51	30/48	30/30

Table 5: Number of speakers with sufficient data extracted for each period (row) and age group (columns), for gender (Female/Male). Categories with less than 30 speakers shown in boldface.

under-represented in media (Doukhan et al., 2018b), and for archives from the 1950's and 1970's – for which documentation and quality is worst. These missing speakers are mostly linked to the following factors: (i) the target appears in the notice, but was the topic of a program without actually appearing, or the target may not speak (or not sufficiently), or spoke in a foreign language and was interpreted. (ii) the target may be interviewed in a noisy place (and was not detected as clean speech), or the target voice may appear during a movie trailer (and thus do not fit with our criterion of conversational speech).

A main reason more male targets were found is linked to the fact that most documents have several targets (female and male) so working on a document that features a female target generally led to the identification of one or more male targets, while the reverse is not true. Moreover, female targets are more prone to be presented by male speakers, without actually speaking. A typical case is programs about cinema in the 1970's, that interviewed the director (generally a male), but just present the female actress during extracts of the movie. For these reasons, the male categories generally present well above 30 target speakers; note that age also introduces bias in terms of presence in the media.

4.2. Perceptual evaluation of speech quality

On a subset of 309 extracts, the three annotators did evaluate the four scales. The number of extracts detected by each annotator with each type of potential problem is reported on the left part of table 6. The corresponding inter-annotator agreement, measured with an exact Fleiss' kappa (Fleiss, 1971; Gamer et al., 2019), equals 0.629 for backchannel, 0.569 for more than one speaker, 0.855 for music, 0.448 for noise – this amount to a kappa of 0.649 for finding a potential problem in a given extract. These kappa values shows the relative reliability of the annotation, especially for music and backchannel. The comparatively low kappa for noise show that noise is a more complex concept than music, and that the three annotators have somehow different views on what is a noisy extract.

From this common ground of 309 extracts, the results on the complete set of 874 utterances was grouped. For the utterance evaluated by three annotators, the presence of a problem was considered only if at least two of them reported it. The number of each error on 874 utterances are reported in the right part of table 6. The percentages of these four types of potential problems, on the complete corpus, and for each category of period, gender, and age, are reported in table 7.

	Common sub-part			Total
Annotator	A1	A2	A3	
Backchannel	57	82	44	148
Several Spk.	5	15	21	33
Music	19	17	15	33
Noise	30	51	37	72

Table 6: Number of problems, for each category, spotted by the three annotators on the 309 common sentences of the perceptual evaluation (left part), or on the complete set (right part).

The amount of observed backchannels is stable across gender and age (at about 17%), but is higher for the 1970's, and clearly lower for the 1950's. Presence of more than one speaker in one extract is much lower (mean below 4%), and does not seem to be correlated to gender, while the same pattern on periods is observed (clearly higher for the 70's, lower for the 50's). A pattern for age appears: the older the target, the less several speakers were found.

	Bac	SSp	Mus	Noi	Any
Globally	16.9	3.8	3.8	8.2	29.7
1955-56	9.1	1.7	0.6	9.1	19.3
1975-76	21.8	9.2	12.7	26.1	55.6
1995-96	17.7	2.5	3.2	3.2	26.4
2015-16	18.6	3.6	1.8	3.6	26.5
Female	18.5	3.5	4.4	10.0	32.6
Male	15.9	3.9	3.4	7.1	28.0
20-35	17.9	5.1	2.6	9.2	29.7
36-50	15.1	5.1	3.8	8.9	30.5
51-65	17.2	2.7	5.4	6.3	28.5
Over 65	18.7	1.2	3.0	8.4	30.1

Table 7: Percentage of utterance detected with each category of potential problem (Bac: Backchannel, SSp: Several speakers, Mus: Music, Noi: Noise, Any: presence of at least one error in the extract), globally, and for each category of period, gender, and age.

The presence of music varies mostly with the period, the 1970's having about 13% of its extracts with perceivable musical background, while the other periods have lower percentages. Only reduced changes on the presence of music are observed with age and gender.

Noise is about twice more frequent than music. Like music, noise is particularly frequent in the 1970's (26%) but, unlike music, is also relatively frequent in the 1950's, while low levels of noise are observed in the two more recent periods. Slight changes of noise pres-

ence are observed across gender and age categories.
The percentage of extracts that do present at least one potential problem of 30% globally. This percentage varies mostly across periods, more than half of the extracts from the 1970's being annotated with potential problems, while one on five extracts from the 1950's has a potential problem.

5. Discussion & Conclusions

We proposed a semi-automatic method to help selecting speakers with known characteristics (here in term of age and gender) in large media archives, avoiding silence, noise and musical background.
About 140 hours of work was invested to manually identify 915 target speakers. Among these speakers, 41 were either not found in the documents or do not speak sufficiently to build a model of their voices (32 of these speakers had less than 20 seconds of annotated speech). The work required for the application of the automatic processing scripts, and for processing the files (time spent by human, not by machine processing) was estimated to 20 hours for the corpus presented here. The complete set of archives used to obtain the current set of speakers had a total duration of 453 hours.
Bazillon et al. (2008) measured the time required to manually transcribe spontaneous speech at eight times the duration of the target speech; manual diarization process is more simple than a full transcription, an estimate of four times the archive duration for manual processing seems reasonable. Due to the complexity of the target speaker identification process only (that may require an almost complete listening of the archive), this estimation does not seem unreasonable.
Given these estimations, we can assume the manual extraction of the target speech data from 453 hours of archives would have required at least more than the viewing time and up to 1800 hours of human labor. Using the proposed method, it took about 160 hours – i.e. four to ten times less than a manual annotation, which seems to be a fairly efficient method. By comparison, but not on the same task, the semi-automatic transcription method proposed by Bazillon et al. (2008) cut by half the manual processing time.
The perceptual evaluation of potential problems shows that one extract in three has at least one potential problem. This may seem a relatively high rate; meanwhile, annotation of backchannel amounts for about half this number (see table 6 and shall not be a problem for the targeted analyses, as backchannels are very short regarding to the duration of extracts (about 0.1s vs. 10s for the mean duration of annotated extracts): this shall not affect the voice's spectral characteristics or mean pitch values. Presence of music, thanks to the music detection process, was limited, compared to the frequent use of mixed music in radio and TV shows (see the evaluation part). Moreover, even if audible wearing a headset, the levels of music that were annotated are low compared to the levels of voices. Noise is a more complex question: we have seen its presence is difficult to annotate. This is certainly linked with the complexity of defining noise, that may be any audible sound added to the soundtrack but speech and music (e.g. street noise, steps, natural noises), but also sounds linked to the recording place (echo from the room), from the recording equipment, from the many hardware used to archive the media (disk, tape), or from compression used to store audio files. The fact noises are more present in the 1970 cue for high presence of added noise, because the selected programs happen to have more outdoor recordings, for example (observation made during the manual target spotting). On the contrary, the lower levels of noise in the 1950's compared to the 70's (somehow counterintuitive) shows archive processing at INA shall not introduce major bias – even if more recent media have better quality.
Noise and music potential effects on the targeted measures will be evaluated once the corpus is completed, but this is outside the scope of this paper. The presented methodology may apply for the construction of corpora dedicated to e.g., sociological work, that do not necessarily require high sound quality.
Presence of several speakers in one extract is an unwanted feature, and ideally these extracts shall be removed. It'll be important to screen the extracts labelled as such to estimate the relative ratio of extracts with completely different speakers, compared to extracts featuring speakers with comparable voices. The second case is less problematic than the former. During informal testing to set up the perceptual evaluation, one extract was labelled as featuring two different speakers only by one of the three annotators: the other two hadn't noticed – only the dialogue's semantics allowed judging there were two speakers. The fact the kappa for this measure was 0.57, shows in a good deal of the cases, the difference in voices was not spotted by all annotators, which pleads for similarity between the voices. The evaluation of the speaker identification system, with higher similarity within than across genders shows it is a probable outcome.
Identification work will continue until having a complete set of speakers. Then, evaluations of the output quality will be applied, with estimation of signal-to-noise ratios, potential distortion of acoustic measurements due to music background, etc. This corpus has a vocation to be shared via INA's online resource management system[4], once the project will be over. The software developed to apply the processing described here shall also be made public in the future, after consolidation of their use on other corpus building.

6. Acknowledgements

This work has been partially funded by the French National Research Agency (project Gender Equality Monitor - ANR-19-CE38-0012).

[4] https://dataset.ina.fr/

7. Bibliographical References

Arantes, P. and Eriksson, A. (2014). Temporal stability of long term measures of fundamental frequency. In *Speech Prosody 2014*, pages 1149–1152. ISCA, May.

Barras, C., Allauzen, A., Lamel, L., and Gauvain, J.-L. (2002). Transcribing audio-video archives. In *ICASSP'02*, pages I–13–I–16, Orlando, FL, USA, May. IEEE.

Bazillon, T., Estève, Y., and Luzzati, D. (2008). Manual vs assisted transcription of prepared and spontaneous speech. In *LREC'08*.

Boula de Mareüil, P., Rilliard, A., and Allauzen, A. (2012). A Diachronic Study of Initial Stress and other Prosodic Features in the French News Announcer Style: Corpus-based Measurements and Perceptual Experiments. *Language and Speech*, 55(2):263–293, June.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., and Gill, M.-P. (2020). pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 7124 – 7128, Barcelona, Spain, May.

Broux, P.-A., Doukhan, D., Petitrenaud, S., Meignier, S., and Carrive, J. (2018). Computer-assisted speaker diarization: How to evaluate human corrections. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Bullock, L., Bredin, H., and Garcia-Perera, L. P. (2020). Overlap-aware diarization: resegmentation using neural end-to-end overlapped speech detection. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 7114 – 7118, Barcelona, Spain, May.

Chung, J. S., Nagrani, A., and Zisserman, A. (2018). Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, pages 1086–1090.

Doukhan, D., Carrive, J., Vallet, F., Larcher, A., and Meignier, S. (2018a). An open-source speaker gender detection framework for monitoring gender equality. In *ICASSP 2018*, pages 5214–5218. IEEE.

Doukhan, D., Poels, G., Rezgui, Z., and Carrive, J. (2018b). Describing gender equality in french audiovisual streams with a deep learning approach. *VIEW Journal of European Television History and Culture*, 7(14):103–122.

Eshkol-Taravella, I., Baude, O., Maurel, D., Hriba, L., Dugua, C., and Tellier, I. (2011). Un grand corpus oral " disponible " : le corpus d'Orléans 1 1968-2012. *Revue TAL*, 53(2):17–46. Publisher: ATALA (Association pour le Traitement Automatique des Langues).

Fan, Y., Kang, J., Li, L., Li, K., Chen, H., Cheng, S., Zhang, P., Zhou, Z., Cai, Y., and Wang, D. (2020). Cn-celeb: A challenging chinese speaker recognition dataset. In *ICASSP 2020*, pages 7604–7608.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Gamer, M., Lemon, J., and Singh, I. F. P., (2019). *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84.1.

Geiselman, R. E. and Crawley, J. M. (1983). Incidental processing of speaker characteristics: voice as connotative information. *Journal of Verbal Learning and Verbal Behavior*, 22(1):15–23, February.

Goudbeek, M. and Scherer, K. (2010). Beyond arousal: Valence and potency/control cues in the vocal expression of emotion. *The Journal of the Acoustical Society of America*, 128(3):1322.

Gussenhoven, C. (2016). Foundations of Intonational Meaning: Anatomical and Physiological Factors. *Topics in Cognitive Science*, 8(2):425–434, April.

Guzman, M., Muñoz, D., Vivero, M., Marín, N., Ramírez, M., Rivera, M. T., Vidal, C., Gerhard, J., and González, C. (2014). Acoustic markers to differentiate gender in prepubescent children's speaking and singing voice. *International Journal of Pediatric Otorhinolaryngology*, 78(10):1592–1598, October.

Hennequin, R., Khlif, A., Voituret, F., and Moussallam, M. (2020). Spleeter: a fast and efficient music source separation tool with pre-trained models. *Journal of Open Source Software*, 5(50):2154.

Jang, H. (2021). How cute do I sound to you?: gender and age effects in the use and evaluation of Korean baby-talk register, Aegyo. *Language Sciences*, 83:101289, January.

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

Liénard, J.-S. and Di Benedetto, M.-G. (1999). Effect of vocal effort on spectral properties of vowels. *The Journal of the Acoustical Society of America*, 106(1):411–422, July.

Liénard, J.-S. (2019). Quantifying vocal effort from the shape of the one-third octave long-term-average spectrum of speech. *The Journal of the Acoustical Society of America*, 146(4):EL369–EL375, October.

Löfqvist, A. and Mandersson, B. (1987). Long-Time Average Spectrum of Speech and Voice Analysis. *Folia Phoniatrica et Logopaedica*, 39(5):221–229.

Mesaros, A., Heittola, T., and Virtanen, T. (2016). Metrics for polyphonic sound event detection. *Applied Sciences*, 6(6):162.

Nagrani, A., Chung, J. S., and Zisserman, A. (2017). Voxceleb: a large-scale speaker identification dataset. In *INTERSPEECH*, pages 2616–2620.

Niebuhr, O., Voße, J., and Brem, A. (2016). What makes a charismatic speaker? A computer-based

acoustic-prosodic analysis of Steve Jobs tone of voice. *Computers in Human Behavior*, 64:366–382, November.

Ohara, Y. (1992). Gender-dependent pitch levels: A comparative study in Japanese and English. In Kira Hall, et al., editors, *Locating power. Proceedings of the Second Berkeley Women and Language Conference*, volume 2, pages 469–477. Berkeley Women and Language Group. Backup Publisher: Berkeley Women and Language Group.

Ohara, Y. (2001). Finding one's voice in Japanese: A study of the pitch levels of L2 users. In Aneta Pavlenko, et al., editors, *Multilingualism, Second Language Learning, and Gender*. DE GRUYTER MOUTON, Berlin, New York, January.

Podesva, R. J. (2007). Phonation type as a stylistic variable: The use of falsetto in constructing a persona. *Journal of Sociolinguistics*, 11(4):478–504, September.

Sadanobu, T. (2015). "Characters" in Japanese Communication and Language: An Overview. *Acta Linguistica Asiatica*, 5(2):9–28, December.

Salmon, F. and Vallet, F. (2014). An effortless way to create large-scale datasets for famous speakers. In *LREC'14*, pages 348–352.

Sataloff, R. T., Kost, K. M., and Linville, S. E. (2017). Chapter 13. The Effects of Age on the Voice. In Robert Thayer Sataloff, editor, *Clinical assessment of voice*, pages 221–240. Plural Publishing, Inc, San Diego, CA, second edition edition.

Scott, S. K. (2022). The neural control of volitional vocal production—from speech to identity, from social meaning to song. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 377(1841):20200395, January.

Sergeant, D. C. and Welch, G. F. (2009). Gender Differences in Long-Term Average Spectra of Children's Singing Voices. *Journal of Voice*, 23(3):319–336, May.

Sloetjes, H. and Wittenburg, P. (2008). Annotation by category-elan and iso dcr. In *6th international Conference on Language Resources and Evaluation (LREC 2008)*.

Suire, A. and Barkat-Defradas, M. (2020). Evolution of human pitch: Preliminary analyses in the French population using INA audiovisual archives of Vox Pops. In *2020 IASA - FIAT/IFTA Joint Conference*, Dublin, France, October.

Titze, I. R. and Sundberg, J. (1992). Vocal intensity in speakers and singers. *The Journal of the Acoustical Society of America*, 91(5):2936–2946, May.

Traunmüller, H. and Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *The Journal of the Acoustical Society of America*, 107(6):3438–3451, June.

Tuchman, G. (2000). The symbolic annihilation of women by the mass media. In *Culture and Politics*, pages 150–174. Palgrave Macmillan US.

Vallet, F., Uro, J., Andriamakaoly, J., Nabi, H., Derval, M., and Carrive, J. (2016). Speech trax: A bottom to the top approach for speaker tracking and indexing in an archiving context. In *LREC'16*, pages 2011–2016.

van Bezooijen, R. (1995). Sociocultural Aspects of Pitch Differences between Japanese and Dutch Women. *Language and Speech*, 38(3):253–265, July.

Vorperian, H. K., Wang, S., Schimek, E. M., Durtschi, R. B., Kent, R. D., Gentry, L. R., and Chung, M. K. (2011). Developmental Sexual Dimorphism of the Oral and Pharyngeal Portions of the Vocal Tract: An Imaging Study. *Journal of Speech, Language, and Hearing Research*, 54(4):995–1010, August.

Yamauchi, A., Yokonishi, H., Imagawa, H., Sakakibara, K.-I., Nito, T., Tayama, N., and Yamasoba, T. (2015). Quantitative Analysis of Digital Videokymography: A Preliminary Study on Age- and Gender-Related Difference of Vocal Fold Vibration in Normal Speakers. *Journal of Voice*, 29(1):109–119, January.

8. Language Resource References

Baude, Olivier. (2019). *Corpus d'Orléans*. Centre Orléanais de Recherche en Anthropologie et Linguistique.

Galliano, Sylvain and Gravier, Guillaume and Chaubard, Laura. (2012). *The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts*. ELRA.

Gravier, Guillaume and Adda, Gilles and Paulson, Niklas and Carré, Matthieu and Giraudel, Aude and Galibert, Olivier. (2012). *The ETAPE corpus for the evaluation of speech-based TV content processing in the French language*. ELRA.

Meléndez-Catalán, Blai and Molina, Emilio and Gómez, Emilia. (2019). *Open Broadcast Media Audio from TV (OpenBMAT)*. distributed via Zenodo: 10.5281/zenodo.3381249.

Ryant, Neville and Church, Kenneth and Cieri, Christopher and Cristia, Alejandrina and Du, Jun and Ganapathy, Sriram and Liberman, Mark. (2019). *The Second DIHARD Diarization Challenge: Dataset, Task, and Baselines*. LDC.