# An Effortless Way to Create Large-scale Datasets for Famous Speakers

**François Salmon**[1], **Félicien Vallet**[2]

[1]ENSEA, 6 avenue de Ponceau, 95014 Cergy-Pontoise Cedex, France
[2]Institut National de l'Audiovisuel, 4 avenue de l'Europe, 94366 Bry-sur-Marne cedex, France
*francois.salmon3@gmail.com, fvallet@ina.fr*

### Abstract

The creation of large-scale multimedia datasets has become a scientific matter in itself. Indeed, the fully-manual annotation of hundreds or thousands of hours of video and/or audio turns out to be practically infeasible. In this paper, we propose an extremly handy approach to automatically construct a database of famous speakers from TV broadcast news material. We then run a user experiment with a correctly designed tool that demonstrates that very reliable results can be obtained with this method. In particular, a thorough error analysis demonstrates the value of the approach and provides hints for the improvement of the quality of the dataset.

**Keywords:** Corpus (creation, annotation, etc.), Speech resource/database, Person Identification

## 1. Introduction

With the increasing amount of audiovisual and digital data deriving from televisual and radiophonic productions but also from amateur sources (online sharing platforms like YouTube or Dailymotion), professional archives such as the French National Audiovisual Institute (Ina[1]), acknowledge a growing need for efficient indexing tools. Indeed, the missions of these archives being to store but also describe and enhance content, the need for automatic structuring methods is of utmost importance. If storage of these content (documents) is not the main issue anymore, indexing methods remain ill suited for the retrieval of information in multimedia databases. Thus, an important human effort for description is required to enhance the stored data.

In this context, many research works aim at automatically organizing and structuring large quantities of audiovisual documents. Among these, speaker tracking — the task of finding spoken segments of a particular speaker for which some training material is given — is of great interest. As stated in (Kinnunen and Li, 2010), speaker tracking is derived from the larger field of the speaker recognition technologies and has been a very hot research topic for several decades. However, the challenge of dealing with "big data" has only been tackled recently. For instance in (Jeon et al., 2012), the authors propose a new statistical utterance comparison that, coupled with kernalized locality-sensitive hashing (KLSH), they use in (Jeon and Cheng, 2012) to retrieve very large population of speakers (about 10 000 speakers). It has to be noted that in that case, speech segments are issued from the SPEECON database especially designed for the research on consumer devices (Siemund et al., 2000).

Evaluation campaigns such as NIST SRE[2] or the French ESTER (Galliano et al., 2009) and ETAPE (Gravier et al., 2012) do not focus on large-scale speaker tracking on broadcast material and the subject has been, to our knowledge, very scarcely studied due to the lack of sufficient data. Indeed the former concentrate on telephone and microphone speech while for the latter, dealing with TV and radio data (∼50 hours), no speaker recognition task is even defined. It is however worth mentioning the framework proposed in (Huijbregts and van Leeuwen, 2010) where, based on speaker diarization techniques, the authors propose to link speakers in an unsupervised fashion for about 1800 hours of Dutch television broadcasts. Similarly, the prototype developped at BBC[3] gathers 3 years of continuous audio amounting to 70 000 programmes (Raimond and Lowis, 2012). In this case, the speaker diarization of each radio programme is realised. Then, a pool of users provides relevance feedback for the identities of the speakers. Corresponding models are build and finally the recognition step is performed against the whole database. To cop with the very large amount of data a locality sensitive hashing-based speaker index is added on top of the LIUM_SpkDiarization toolkit (Meignier and Merlin, 2010) used for the diarization and recognition steps.

In this paper we propose a framework to efficiently gather speaker segments from very large TV broadcast news archives (∼3600 hours) without resorting to relevance feedback. We prove that very reliable datasets can be constructed without much effort and can thus become great resources for the creation of voice models in the speaker tracking task.

## 2. Pre-processing of the Data

To perform the extraction of audiovisual segments corresponding to a list of targeted speakers and thus automatically construct speaker datasets, a simple hypothesis is made. It states that during a newscast, when the name of a person appears on the screen, that person is presently speaking. The excerpts constitutive of the final dataset come from two newscast corpora :

- The first brings together 365 hours of news broadcasted in 2007 at 8pm on the French national channel France 2.

[1] Institut National de l'Audiovisuel: www.ina.fr

[2] NIST Speaker Recognition Evaluation: www.itl.nist.gov/iad/mig/tests/spk/

[3] The World Service Radio Archive: www.bbc.co.uk/rd/projects/worldservice-archive-proto

- The second, consists of 3423 hours of news broadcasted on 6 different French channels: TF1, France 2, France 24, LCI, BFM TV and ITélé between June 2011 and January 2013.

### 2.1. Optical Character Recognition (OCR)

Our system uses an OCR software to detect the names of personality appearing on a banner at the bottom of the screen. Our choice fell on the LOOV (LIG Overlaid OCR Text) system which easily handles the video stream through a detection of significant text regions (Poignant et al., 2012). LOOV analyzes the image at regular intervals and gathers detected strings through ”bounding boxes” defined in time and space.

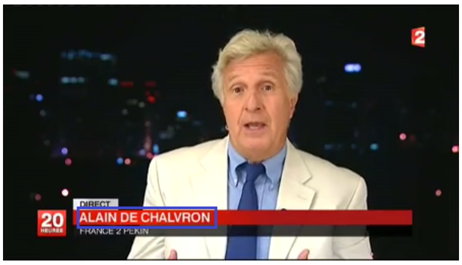


Figure 1: The key string *Alain de Chalvron* is to be detected by the LOOV software.

As illustrated in Figure 1, this tool, already used for the speaker identification task, is particularly suited to newscast processing.

### 2.2. Speaker-Turn Segmentation

This step consists in cutting the audio stream into speech turns, *i.e.* successive speakers' interventions, without any *a priori*. To do so, the diarization tool developed by the LIUM (Meignier and Merlin, 2010) is used. This set of Java methods provides basic tools for audio stream analysis that can be used for various applications. The LIUM teams took part in several national and international evaluation campaigns (such as NIST, ESTER, ETAPE or REPERE[4] (Kahn et al., 2012)) demonstrating the effectiveness of their system which is generally ranked first or second.

In this work, the toolkit is used for segmentation purposes only. Indeed, one option considered was to extract the clusters computed from the speaker diarization process but it did not provide satisfactory results due to numerous wrong associations. Indeed, it seems that during a newscast, the personalities' interventions are too short to expect homogeneous and exhaustive clusters.

## 3. Association

### 3.1. Named Entity List

For achiving purposes, Ina uses a thesaurus of 10 000 common names, 43 000 geographical regions (cities, countries, *etc.*), 750 000 individuals and 140 000 legal entities (companies, bands, associations, *etc.*). We chose to use a sublist extracted from the thesaurus that gathers the first twenty thousand most referenced people in Ina's audiovisual documentation. This list allows us to set the name of the personalities to be recognized by the OCR system.

The link to be made between the detected strings and the different personalities of the thesaurus is not always easy. Indeed, the association must be robust towards the varying quality of detection, possible spelling differences and potential titles such as Dr., Lord, General, *etc*. The decision to keep or reject the detected strings is taken by computing the Levenshtein distance with the thesaurus' terms.

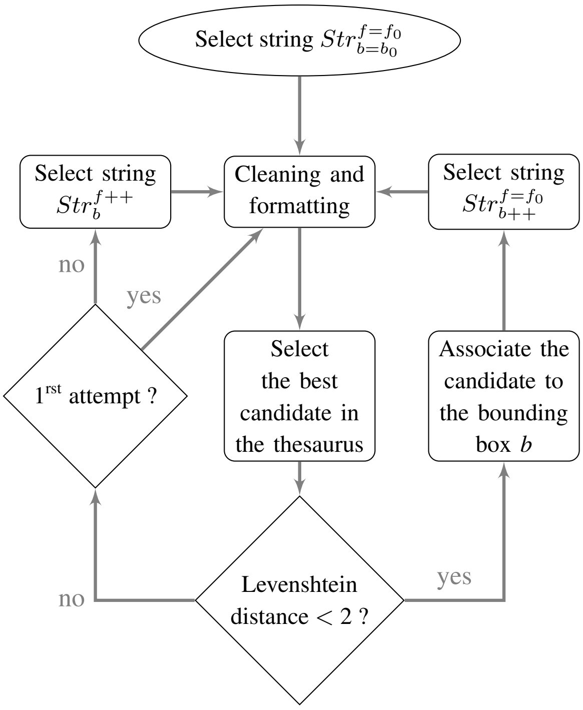

Select string $Str_{b=b_0}^{f=f_0}$
Select string $Str_b^{f++}$
Cleaning and formatting
Select string $Str_{b++}^{f=f_0}$
no
yes
1rst attempt ?
Select the best candidate in the thesaurus
Associate the candidate to the bounding box $b$
Levenshtein distance < 2 ?
no
yes

Figure 2: Algorithm associating the string detected in a banner by LOOV with a named entity from the thesaurus.

The association algorithm presented in Figure 2 has been adopted, with $b$ the index of the current bounding box (defined for a certain amount of time) and $f$ the index of one of its associated frame. $b_0$ is the index of the first bounding box of the series to be analyzed and $f_0$ the index of the first frame of a given bounding box $b$.

The selection of the best candidate in the thesaurus is performed through the use of the Lucene search engine[5].

### 3.2. Extraction

Following the assumption stating that during a newscast, when the name of a person appears on the screen, that person is presently speaking, we can finally automatically label speech segments. A speech turn is matching the name detected at the previous step when it is overlapping with the banner for at least half of it's length. Under this condition, the corresponding video clip is transferred in the dataset and the operation is repeated for each detection of a personality present in the thesaurus.

[4] Défi REPERE: http://www.defi-repere.fr/

[5] Lucene: http://lucene.apache.org/core/

## 4. Description of the Dataset

The creation of such a dataset provides annotated content, that can be used to create voice models (or with a few tweaks, face models). Processing the 3 788 hours of newscasts allowed to identify 5 403 individuals, sharing 47 267 video clips. Figure 3 represents the speakers' distribution based on their speaking time.

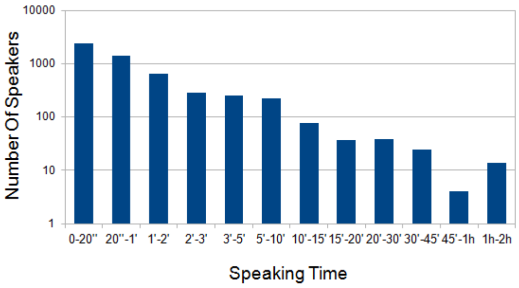


Figure 3: Speakers' distribution according to their total speaking time.

This curve shows the disparities between speakers with respect to their amount of data. A large number of speakers possess only one or two extracted segments, which explains the sharp decrease of the distribution. 1 210 people over the 5 403 have more than 1.5 minute of speech, which appears to be a critical amount of time to compute a voice model for speaker tracking.
The average duration of a speech turn on television seems to be close to 12.6 seconds. The dispersion of these durations is fairly consequent (from 2 to 20 seconds), which is mainly due to bad segmentation and detection errors.

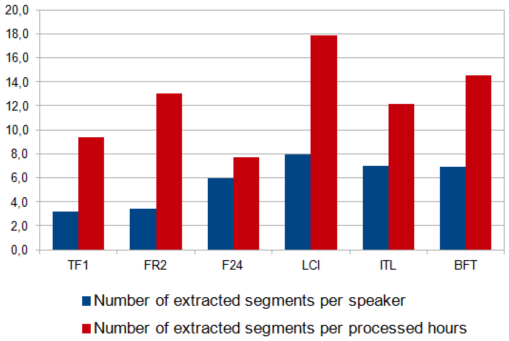


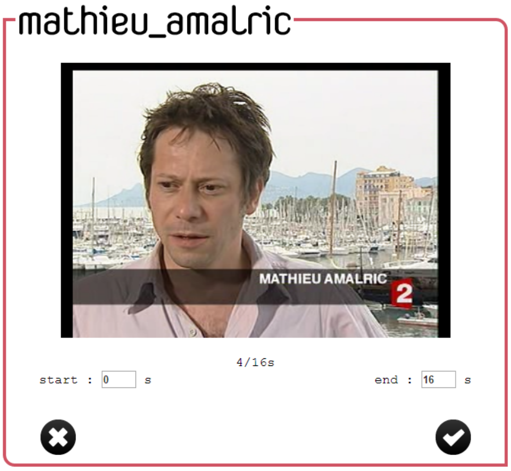
Figure 4: Number of detected speakers and extracted segments per processed hour over the different channels.

Figure 4 provides information on the system's performance over the six processed channels. It allows to notice that the four news channels LCI, France 24, ITélé and BFMTV, have the largest speakers' redundancy with an average of 7 segments extracted per personality. That was to be expected since these channels broadcast the same news reports over and over. With many segments extracted per processed hour and less excerpts per speaker, the more traditional TV channels France 2 and TF1 show more diversity and less redundancy (due to the lesser proportion of duplicates).

These results have to be analysed in regard to the various bias introduced by the system such as the quality of the turn segmentation and name detection, the recording conditions, the TV news screen layouts, *etc.*
Measuring the quality of the dataset remains problematic. The direct access to errors such as wrong associations, detections or segmentations is possible but laborious.

## 5. Creation of a Validation Tool

Due to the amount of data processed, it is practically unfeasible to ask users to manually validate and correct the entirety of the dataset previously created. However, we propose a validation tool[6] aiming at analyzing the quality of the dataset. As displayed in Figure 5, this tool enables to check randomly, the extracted segments' accuracy (meaning the adequation with the identity of the person talking) and adjust manually their temporal boundaries.



4/16s
start : 0 s
end : 16 s

Figure 5: Validation tool screenshot.

Practically, once a validation session is started, user-related information is recorded such as the number of checked, modified and validated segments, the average processing time per segment, the session duration, *etc.*
In order to evaluate the proposed validation tool, we asked volunteers to use it, and extracted usage information to provide a thorough analysis in terms of processing time and accuracy.

## 6. Results of the User Experiment

Fifteen volunteers from Ina took part in the evaluation of the dataset. Thanks to their participation, 1 710 segments were validated, rejected or modified. The overall processing time represents 10 hours and 49 minutes, corresponding to an average of 22,8 seconds per segment. The standard deviation reaches 11.1 seconds, showing great variations between users. With an average segment length of 12.6 seconds, those values appear to be quite important. They are mainly due to a large part of the database segments which had to be modified (19% of the hand-checked segments), as displayed on Figure 6. A recurrent modification turns out to be to shorten excerpts that are a few seconds too long.

[6] Validation tool:
https://github.com/FSalmon/ina_database_validation_tool/

Consequently, a better tuning of the speaker-turn segmentation software has to be applied to increase the quality of the dataset.

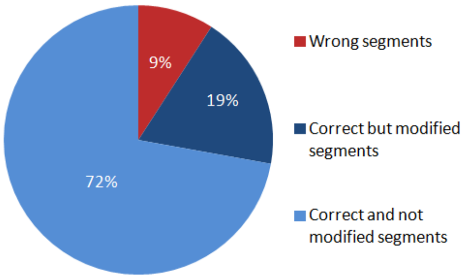


Figure 6: Proportion of segments which were validated, rejected or modified during the user experiment.

The rejected segments (9%) are the consequence of four differents types of errors listed in Figure 7 and detailed in the following sections.

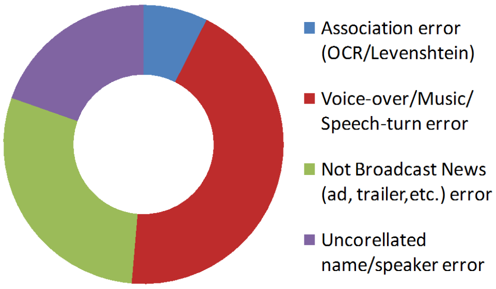


Figure 7: Analysis of the observed errors.

### 6.1. Association Error (OCR/Levenshtein)

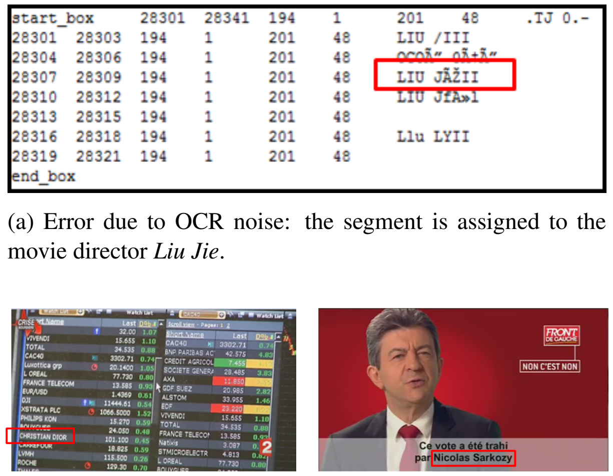
The first type of error concerns wrong associations (8% of the total errors). They can be due to OCR noise (3%). In this case, it causes the attribution of segments to personalities having few letters in their name, as illustrated in Figure 8a. Short names could eventually be discarded from the thesaurus to avoid this from happening.

Some association errors due to homonyms or spelling differences remain inevitable with our method (4%). For instance, for *Alain Prost*, among the four excerpts retrieved, two belong to the former Formula One champion while the remaining two are associated with the CEO of the French company *Lejaby*. Only homonyms concerning places such as airports, libraries or hospitals named after famous people (*Charles De Gaulle*, *Marguerite Duras*, *etc.*) could be reduced by completing the thesaurus (1%).

### 6.2. Voice-over/Music/Speech-Turn Error

Errors related to the presence of polluted speech turns represent the major cause of segment rejection (44% of the total errors). This is mostly due to the presence of voice-over (38%). The voice-over phenomenon happens almost exclusively for foreign personalities such as *David Cameron*, *Serena Williams* or *Pedro Almodovar* when their words are directly translated. These errors could be avoided by exploiting information regarding to the personnalities' nationality in the thesaurus. Indeed, this seems a much easier solution than solving the problem of the detection of overlapping speech (Geiger et al., 2012).

Bad speech-turn segmentation, often occurring during heated exchanges, also generates corrupted segments with more than one speaker (5%). Finally, the presence of music in speech turns (1%) could be reduced by introducing a music detector. However, this does not represent a very important number of errors.


```
start_box     28301  28341  194   1       201    48     .TJ 0.-
28301  28303  194    1      201   48      LIU /III
28304  28306  194    1      201   48      OCOÃ" OÃ+Ã"
28307  28309  194    1      201   48      LIU JÃŽII
28310  28312  194    1      201   48      LIU JfA»1
28313  28315  194    1      201   48
28316  28318  194    1      201   48      Llu LYII
28319  28321  194    1      201   48
end_box
```

(a) Error due to OCR noise: the segment is assigned to the movie director *Liu Jie*.

(b) Scene text detection example: the segment from which this frame is extracted is assigned to *Christian Dior*.

(c) Example of a personality detection who does not correspond to the speaker (*Nicolas Sarkozy* for *Jean-Luc Mélenchon*).

Figure 8: Various errors associating a wrong personality to the current speech turn.

### 6.3. Not Broadcast News Error

A great part of errors also concerns segments that do not belong to broadcast news such as advertisements or trailers (29% of the total errors). Program schedules could be used to limit the processing to the exact newscasts' length. Note that in the corpus pre-processing, newscasts were considered to last one hour, which is rarely the case. Another way of restricting the analysis to broadcast news material could be to use jingle/ad detectors. This approach could easily lessen the number of wrong associations due to credits appearance at the end of programs (5%), leading to the mis-attribution of excerpts to journalists or TV directors.

### 6.4. Uncorellated Name/Speaker Error

The last type of error happens when the name of a personnality is correctly detected (during a newscast) but doesn't match the identity of the current speaker (19% of the total errors). It can be due to the detection of scene-text, meaning some text which is not overlaid but part of the image as shown in Figure 8b. It is also related to the detection of a personality who does not correspond to the main speaker (see Figure 8c). This kind of error is bound to happen on occasion, for instance when a celebrity dies (*Amy Winehouse*, *Lou Reed*, *Michael Jackson*, *etc.*).

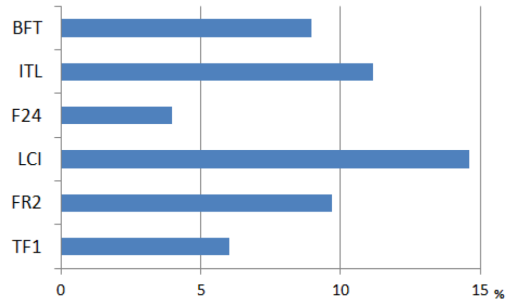


Figure 9: Percentage of channels' wrong segments among their respective checked segments.

Considering the various TV news screen layout, results differ significantly from one channel to another. For instance, given the percentages represented in Figure 9, our system seems to be more suited to France 24 than to LCI. Channel related strategies could be introduced to lower those differences. For instance, learning the exact position of banners displaying interviewees' name for each channels would enhance the system efficiency.

## 7. Conclusion and Perspectives

The challenge of dealing with "big data" makes the manual annotation of hundreds or thousands of hours of video unfeasible. Therefore, we propose in this paper a novel approach to automatically create large-scale datasets for famous speakers. With the joint use of an OCR system and a speaker-turn segmentation tool, excerpts are automatically extracted for people belonging to a named entity list.

Almost 4 000 hours of newscast were processed this way, allowing to extract more than 47 000 segments distributed over 5 000 speakers and totaling 170 hours. A thorough error analysis showed that very good results could be obtained and that the error could drop with a few adjustments.

It remains however to be checked if the error rate is low enough to allow good results for the speaker tracking task, which will be studied in the future. Besides, an additional feature allowing the user to check segments more likely to contain errors is to be introduced. This measure will be provided by tracking, for a given personality, the less similar excerpts through a voice recognition scoring of the considered segments. In particular, this feature would be very handful to allow the disambiguation of homonyms.

Finally, it has to be noted that the approach described here to extract famous speakers' excerpts may be derived to create large-scale face datasets for famous people by replacing the speaker-turn segmentation by a cut and a face detector.

### ACKNOWLEDGMENT

The authors would like to thank Dr. Sylvain Meignier from LIUM and Dr. Jean Carrive from Ina for their precious help, the ENSEA for its support, as well as the volunteers who took part in the user experiment.

## 8. References

Galliano, S., Gravier, G., and Chaubard, L. (2009). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. In *Annual Conference of the International Speech Communication Association*, Brighton, United Kingdom, septembre.

Geiger, J., Vipperla, R., Evans, N., Schuller, B., and Rigoll, G. (2012). Speech overlap detection using convolutive non-negative sparse coding: New improvements and insights. In *European Signal Processing Conference*, Bucharest, Romania, august.

Gravier, G., Adda, G., Paulson, N., Carré, M., Giraudel, A., and Galibert, O. (2012). The etape corpus for the evaluation of speech-based tv content processing in the french language. In *International Conference on Language Resources and Evaluation*, Istanbul, Turkey, may.

Huijbregts, M. and van Leeuwen, D. (2010). Towards automatic speaker retrieval for large multimedia archives. In *International Workshop on Automated Information Extraction in Media Production*, Florence, Italy, october.

Jeon, W. and Cheng, Y.-M. (2012). Efficient speaker search over large populations using kernelized locality-sensitive hashing. In *International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, march.

Jeon, W., Ma, C., and Macho, D. (2012). Statistical utterance comparison techniques for speaker clustering using factor analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(9):2482 – 2491.

Kahn, J., Giraudel, A., Carré, M., Galibert, O., and Quintard, L. (2012). Repere : premiers résultats dun défi autour de la reconnaissance multimodale des personnes. In *Journées d'Étude de la Parole*, Grenoble, France, june.

Kinnunen, T. and Li, H. (2010). An overview of text-independent speaker recognition: from features to supervectors. *Speech Communication*, 52(1):12 – 40.

Meignier, S. and Merlin, T. (2010). Lium spk_diarization: An open source toolkit for diarization. In *CMU SPUD Workshop*, Dallas, TX, USA, march.

Poignant, J., Besacier, L., Quénot, G., and Thollard, F. (2012). From text detection in video to person identification. In *International Conference on Multimedia and Expo*, Melbourne, Australia, july.

Raimond, Y. and Lowis, C. (2012). Automated interlinking of speech radio archives. In *Linked Data on the Web, WWW*, Lyon, France, april.

Siemund, R., Höge, H., Kunzmann, S., and Marasek, K. (2000). Speecon - speech data for consumer devices. In *International Conference on Language Resources and Evaluation*, Athens, Greece, may.