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

LREC 2022

ANR-19-CE38-0012

Aim of this work

- Present a semi-automatic method for balanced speaker corpus building
- Evaluate the process in terms of
  - quality of the automatic processing steps
  - quality of the output
- Present the data selected using this semi-automatic process
- Estimate the time saved

The case corpus

Gender, age, and period-balanced corpus of voices

- Gender representation is highly biased in media[1]
  - More men than women
  - Difference of roles
  - Other phenomena (manterrupting)
- Voices also change with
  - Age: older female have lower voices; older male have higher voices[2]
  - Time: can we observe a shift in voice usage with time?

⇒ Need to have representative dataset to base ***acoustic*** measures on

[1] Global Media Monitoring Project https://whomakesthenews.org/wp-content/uploads/2021/10/Rapport-national-France.pdf
[2]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.

Target composition of the corpus

- 2 Gender groups *male* | *female*
- 4 Age groups *20-35* | *36-50* | *51-65* | *>65*
- 4 Time periods *1950's* | *1970's* | *1990's* | *2010's*
- = 32 categories of speakers

- Acoustic measurements:
  - long term measure: > 40s ⇒ target: > 3 minutes

- ≥ 30 speakers by categories
  - Target of 960 speakers

- Clean speech
  - Low level of noise / music background
  - No speech overlap, relatively long (>2s) continuous speech segments.

Diving into INA’s archives

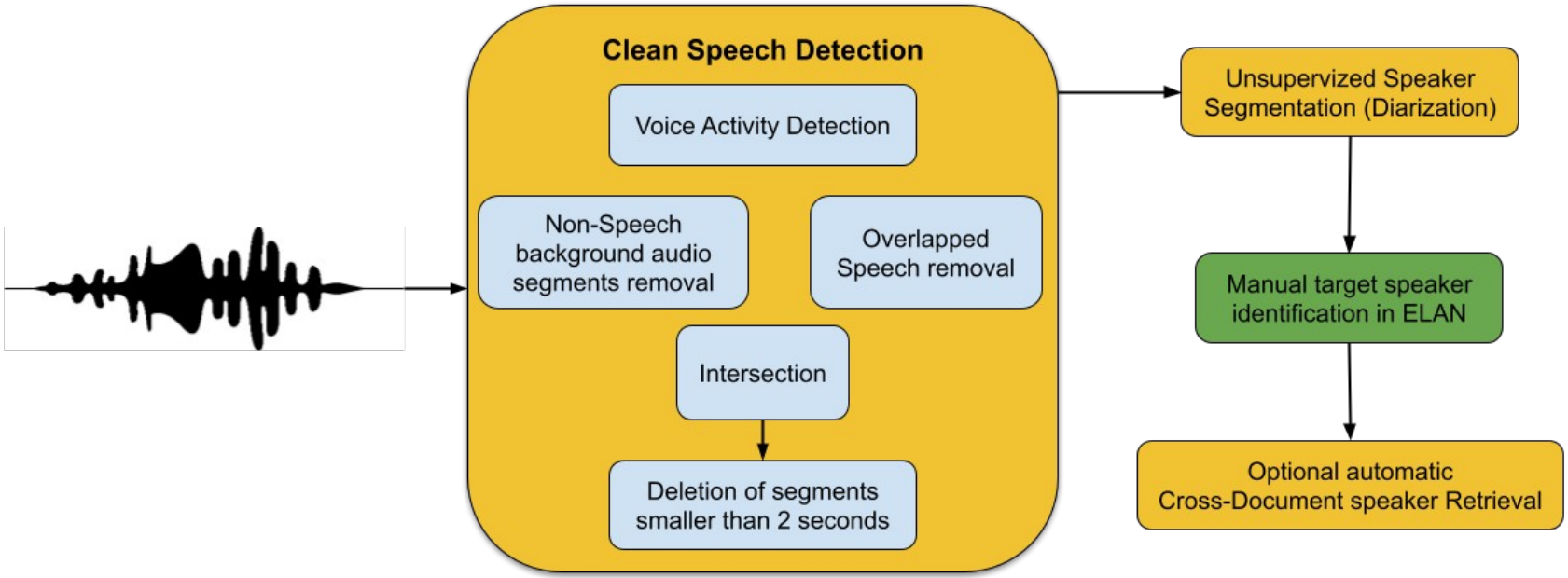
Institut National de l’Audiovisuel :
Audiovisual archives from French TV and radio

ID Notice PHD88025680
Titre propre Le monde est un spectacle du 16 octobre 1955
Titre collection Le monde est un spectacle
Chaîne de diffusion Programme Parisien
Date de diffusion 16/10/1955
Date d'enregistrement 16/10/1955
Statut de diffusion Première diffusion
Heure de diffusion 20:40:00
Heure de fin de diffusion 20:52:05
Thématique Variétés ; Spectacle
Genre Interview ; Interprétation
Générique PRO,Monestier Marianne ; PRO,Goupillières Roger ; PRE,Ascain Jacqueline ; PRE,Michel Jean Claude ; INT,Mania Rose ; INT,Louvier Nicole ; PAR,Louvier Nicole ; PAR,Antoine André Paul
Résumé documentaire "- Rose MANIA : chante ""Le Danseur de Charleston""
- Nicole LOUVIER : interviewée et chante ""Ou te trouverai-je""
- Interview de André-Paul ANTOINE sur sa pièce ""La Tueuse""
- Interview de Janine FAVIER sur le rôle qu'elle joue dans cette pièce"
Œuvres "- Le Danseur de Charleston / Interprétation : MANIA Rose
- Ou te trouverai-je / Interprétation : LOUVIER Nicole"
Société de programmes RTF
Extension géographique Régional
Base Archives Radio Pro

- Prefer studio interviews
- Natural language description
- Using INA’s thesaurus to get date of birth

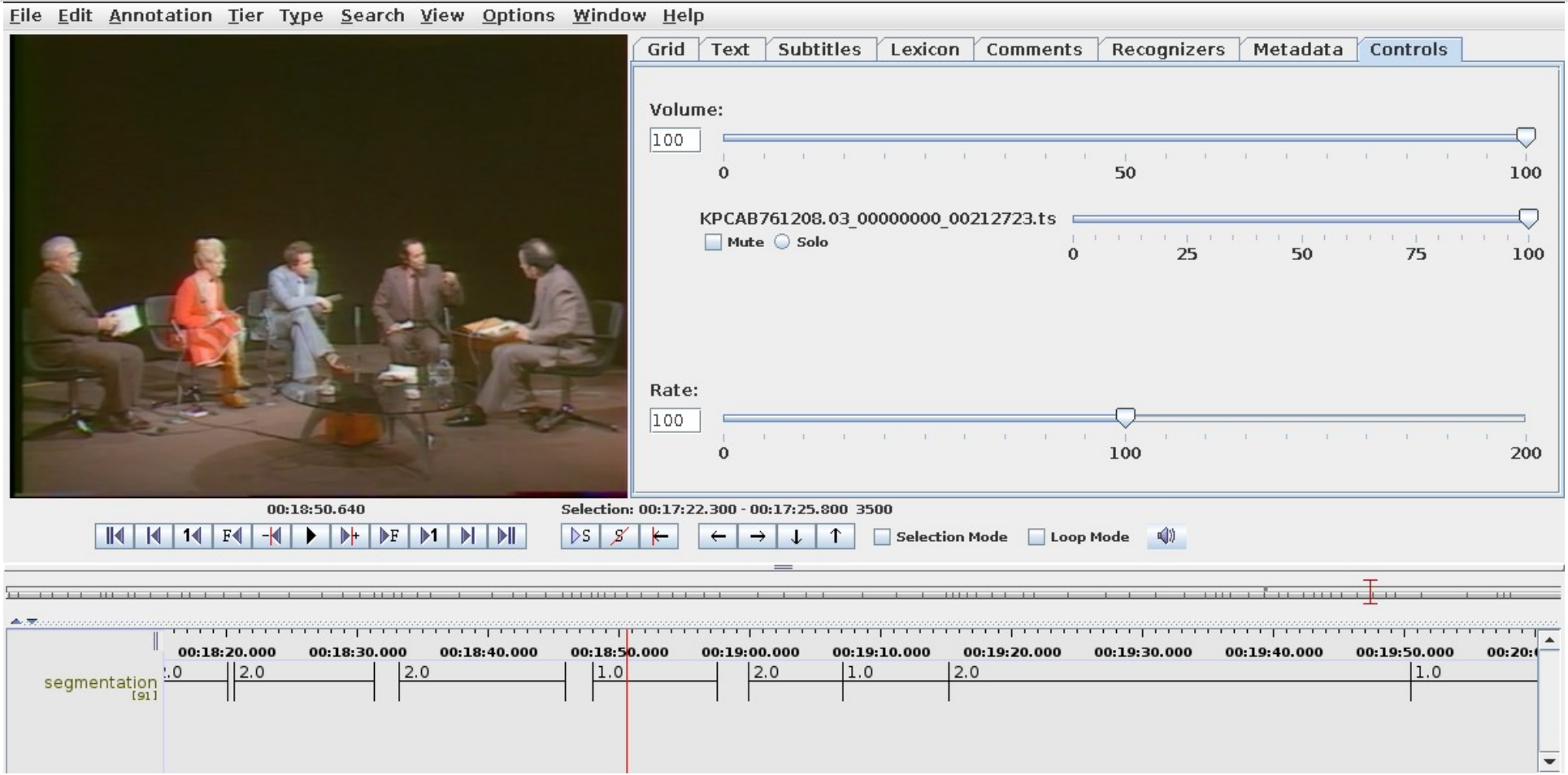
- Example:
  - Louvier, Nicole
  - F
  - 1933 - 2003
  - Chanteuse, compositrice, écrivain, productrice.
  - Age: 22
  - Categorie 20/35_F_1955/56

Workflow



Automatic processes to minimize manual intervention

Annotation


File Edit Annotation Tier Type Search View Options Window Help
Grid Text Subtitles Lexicon Comments Recognizers Metadata Controls
Volume:
100
0 50 100
KPCAB761208.03_00000000_00212723.ts
Mute Solo
0 25 50 75 100
Rate:
100
0 100 200
00:18:50.640
Selection: 00:17:22.300 - 00:17:25.800 3500
Selection Mode
Loop Mode
00:18:20.000 00:18:30.000 00:18:40.000 00:18:50.000 00:19:00.000 00:19:10.000 00:19:20.000 00:19:30.000 00:19:40.000 00:19:50.000
segmentation
[91]
2.0 2.0 2.0 1.0 2.0 1.0 2.0 1.0

Results – VAD

Voice Activity Detection with InaSpeechSegmenter[1]

Dataset	Precision	Recall	F-Measure
Mirex 2018[2] - D1	90.9	92.9	91.9
Mirex 2018 - D2	96.0	95.6	95.8
DiHARD 2[3] - Dev	90.1	84.6	87.3

[1] 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

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

[3] Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., and Liberman, M. (2019). The second DIHARD diarization challenge: Dataset, task, and baselines. In Interspeech, pages 978–982.

Results – Non-speech detection

- Speech Vs Musical instrument source separation with Spleeter[1]
- Estimation of non-speech events activity based on the energy of the musical instrument source obtained
- Evaluation on OpenBMAT[2] with 1 sec collar
- No annotated database including noise, music and speech

Sound Level	hard-to-hear background music	background music	music and other signals mixed at similar levels	foreground music	music only
Music Recall (%)	65	89.9	97.0	97.2	99

[1] 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.
[2] Meléndez-Catalán, Blai and Molina, Emilio and Gómez, Emilia. (2019). Open Broadcast Media Audio from TV (OpenBMAT).

Results – Clean speech coverage

Clean detected speech coverage on DiHard2 Dev[1]

Method	Speech Coverage (%)
Reference VAD	100
Reference VAD – paynnote Speech Overlap[2]	88.2
Reference VAD - Non Speech Audio Events	44.3
Automatic VAD	**89.1**
Auto VAD - Speech Overlap	78.8
Auto VAD - Non Speech Events	41.7
Auto Vad - Speech Overlap - Non Speech Audio Events	39.2
All & removal of segments < 2 sec	**33.2**

[1] Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., and Liberman, M. (2019). The second DIHARD diarization challenge: Dataset, task, and baselines. In Interspeech, pages 978–982

[2] Bredin, H. et al. (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

Results – Clean speech coverage

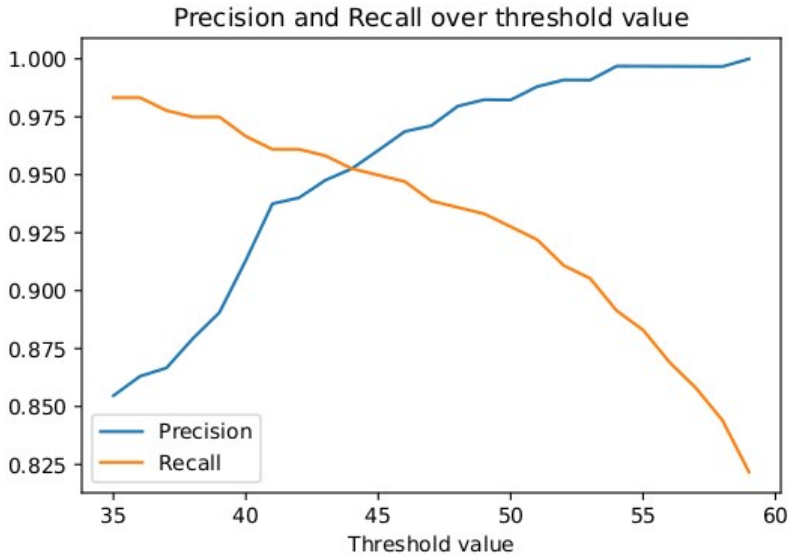
Part of initial VAD remaining after clean speech detection

- 1950’s more noisy
- Other periods : ≈1/3 removed

1955-56	19%
1975-76	65%
1995-95	69%
2015-16	67%

Results – Cross show speaker indentification

- Evaluated on INA's speaker dictionary[1] - gender balanced subset of 718 speakers
- EER: **3.9%** with threshold at 0.4
- Final Threshold used: 0.52
  - Precision = 0.99
  - Recall = 0.9



[1] 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

Subjective evaluation of the outcome

- Manual annotation of the following events:

- Backchannel
- ≤ 2 syllables by another speaker
- More than one speaker
- > 2 syllables by two speakers
- Music
- Audible music background?
- Noise
- Audible noise?

- Three annotators for a total of 565 extracted segments
  - 309 subset annotated by all annotators

Subjective evaluation of the outcome

- Inter annotator agreement (common subpart; exact Fleiss' kappa):
  - backchannel: **0.629**
  - Several speakers: **0.569**
  - Music: **0.855**
  - *Noise:* ***0.448***
- Global: **0.649**

Notes :

- Definition of “noise” is problematic
- Music is reliable
- Several speakers → similar voices?

	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

Subjective evaluation: problems by category

- The 1970's have much more problems
  - Program choices?
- Relative good quality for the 50's
- Slight effect of gender
  - Bias of automatic systems?
  - Bias of shows / roles?
- Little effect of age

	Bachannel	Several Speakers	Music	Noise	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

Outcome: What we gathered so far

- 874 speakers out of 960:
  - 341 female / 533 male
  - 915 target identified; 41 missing (do not speak)
- 16 incomplete categories on 32
  - 4 from recent periods almost complete, mostly female missing
  - 10 have from 10 to 20 speakers (6 female groups)
  - 2 female groups have 4 and 5 speaker only (older females)
- Missing 211 targets; 125 extra targets

	20-25		36-50		51-65		>65	
	F	M	F	M	F	M	F	M
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

- Work done
  - 20 to run scripts and deal with files
  - 140 hours for the manual identification process (915 speakers)
  - Volume of archives used: 453 hours
  - estimation of manual time: between one and four times realtime ⇒ up to 1600 hours

References

Global Media Monitoring Project https://whomakesthenews.org/wp-content/uploads/2021/10/Rapport-national-France.pdf

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.

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

MIREX https://www.music-ir.org/mirex/wiki/

Ryant, N., Church, K., Cieri, C., Cristia, A., Du, J., Ganapathy, S., and Liberman, M. (2019). The second DIHARD diarization challenge: Dataset, task, and baselines. In Interspeech, pages 978–982.

Bredin, H. et al. (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

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.

Meléndez-Catalán, Blai and Molina, Emilio and Gómez, Emilia. (2019). Open Broadcast Media Audio from TV (OpenBMAT).

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