Speaker diarization with overlapped speech

Justine Diliberto, Cindy Pereira, Anna Nikiforovskaja
Université de Lorraine, IDMC

Supervised by Md Sahidullah

Introduction

Speaker diarization: identify speakers and when they talk (*who spoke when*).
Applications: speaker-attributed speech-to-text, handling audio archives, improving automatic speech recognition, spotting speakers in voice assistant technology.
Overlapped speech: when at least two speakers talk at the same time; major and recurrent cause for diarization errors.

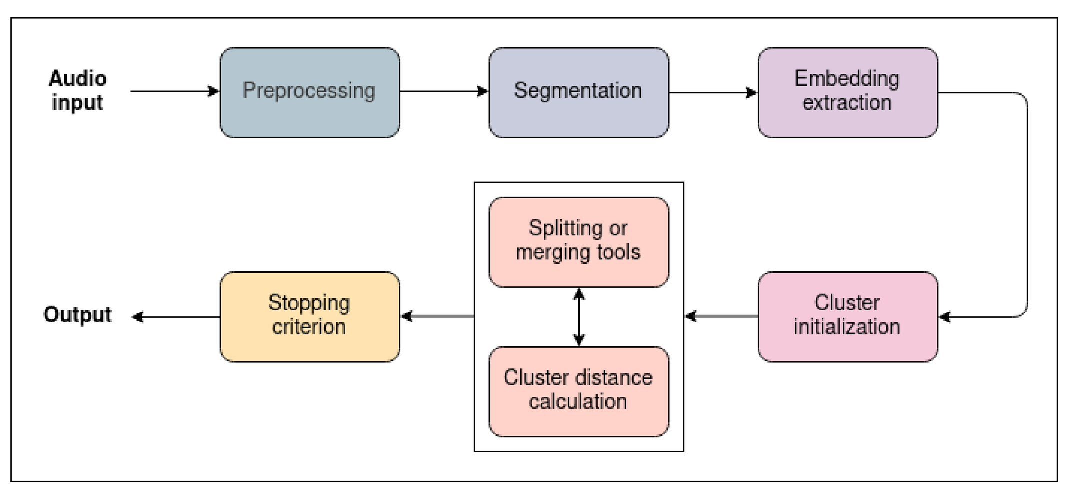

Figure 1: Components of a speaker diarization system

Experimental setup

Dataset source: *Second DIHARD Diarization Challenge*

- Single channel condition (one voice channel)
- Reference speaker activity detection (SAD – ground truth)
- 11 domains: audiobooks, broadcast interviews, child language, clinical, courtroom, map task, meeting, restaurant, socio-linguistic field and lab, and web videos

Related Works

Diliberto, J., Pereira, C. & Nikiforovskaja, A. (2021) Speaker diarization with overlapped speech; Realization report.
Diliberto, J., Pereira, C. & Nikiforovskaja, A. (2021) Speaker diarization with overlapped speech; Bibliographical report.

Performance impact

Performance: measured in terms of Diarization Error Rate (DER):

$$\mathrm{DER} = \frac{\mathrm{FA} + \mathrm{MISS} + \mathrm{ERROR}}{\mathrm{TOTAL}}$$

FA: False alarm (speech falsely identified)
MISS: Missed speech (speech not identified)
ERROR: Speaker error (speech attributed to the wrong speaker)

Assumption: if overlap worsens the performance, removing it should improve the results

Category	Average DER original data	Average DER overlap removed	Average % of overlap
Audiobooks	4	1.3	0
Broadcast interview	9	14.1	0.9
Child	31.7	37.5	7.5
Clinical	18.5	40.5	2.4
Court	16.3	29.3	1.6
Maptask	6.7	28.2	2
Meeting	34.1	49	21.3
Restaurant	50.5	59	21.4
Socio field	14.7	35.4	5.7
Socio lab	10.4	29.7	3.7
Webvideo	38.1	35.3	17.7

Figure 2: Results after running the baseline on the dataset with overlap segments removed

Conclusion of the experiment:

- Unexpected results, DER worsened after removing overlap
- No correlation between the DER difference and the number of seconds removed from the files
- No correlation between the DER difference and the average percentage of overlap per category
- Performance can be altered by other factors (eg background noise)
- Overlap impacts the whole audio

Acoustic impact

Assumption: overlapped speech may be identified through acoustic features (eg pitch, formants, voicing)

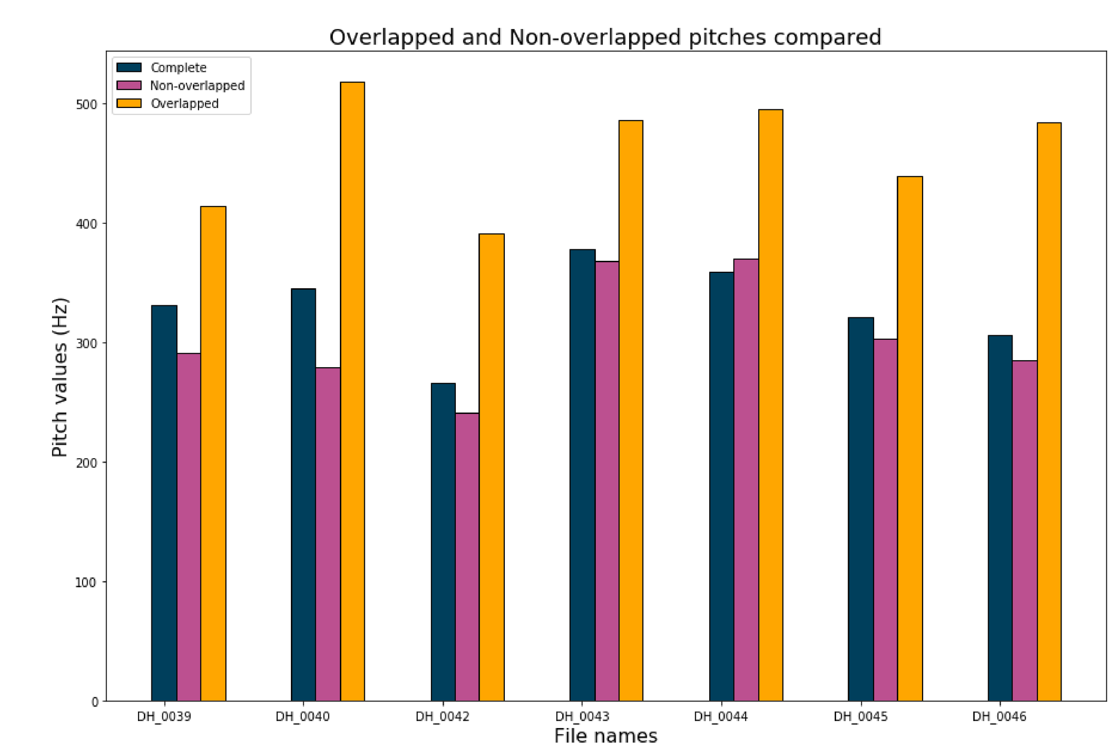

Figure 3: Pitch values for overlapped and non-overlapped speech samples in the category 'restaurant'

Pitch always obtains higher scores in the case of overlapped speech

Feature	Mean			Median			Std Dev	
	NOV	OV	Ratio	NOV	OV	Ratio	NOV	OV
Pitch	439	641	1.46	367	628	1.71	218	233
SpectralFlux: amean	0.32	0.47	1.46	0.23	0.36	1.60	0.22	0.30
F0: meanFallingSlope	98	132	1.35	88	127	1.44	43	56
Loudness: amean	0.73	0.99	1.34	0.59	0.82	1.39	0.46	0.56
SlopeV0-500: amean	0.015	0.02	1.39	0.008	0.01	1.65	0.03	0.03
Unvoiced seg len: mean	0.41	0.24	0.59	0.31	0.19	0.61	0.25	0.13
Voiced seg/sec	1.93	2.78	1.44	2.01	2.62	1.32	0.56	0.76
F2 frequency: amean	1683	1692	1.00	1656	1675	1.01	114	114
F3 frequency: amean	2703	2707	1.00	2697	2711	1.00	97.28	92.83

Figure 4: Statistical results of some features based on the study of 26 files

NOV: files without any overlapped speech
OV: files containing only overlapped speech

Similar acoustic values will have a ratio closer to 1

Conclusion of the experiment:

- Some features have distinctive values when computed on overlapped speech
- Other features (eg formants 2 and 3) have similar values

Overlap detectors

Assumption: X-vectors can be used to detect segments with overlap to further improve the performance of speaker diarization.

X-vector: trained embeddings for speech segments.

Method	UAR big context	UAR small context
RidgeClassifier	0.26	0.23
SVC	0.20	0.20
SGDClassifier	0.24	0.23
DecisionTreeClassifier	0.22	0.22
LinearNet	0.24	0.24
TDNNBasedModel	0.25	0.23
BLSTMBasedModel	0.23	0.20
GRUBasedModel	0.22	0.19

Figure 5: Evaluation results for classification methods

Method	R2 big context	R2 small context
Lasso	0.006	-0.051
SVR	0.070	0.052
SGDRegressor	-2e27	-1.5e27
DecisionTreeRegressor	-0.79	-0.808
LinearNet	-0.34	-0.333
TDNNBasedModel	-0.065	-0.124
BLSTMBasedModel	-0.498	-0.252
GRUBasedModel	-1.207	-0.551

Figure 6: Evaluation results for regression methods

Big context: 3 segments before and 1 after
Small context: 2 segments before
UAR: unweighted average recall
R2: coefficient of determination

Conclusion of the experiment:

- Classification-based: better for overlap prediction
- TDDN-based: best deep learning method; improves with larger context
- X-vectors contain some information which can be used for overlap detection