Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3264–3270
Marseille, 20-25 June 2022

Overlaps and Gender Analysis in the Context of Broadcast Media

Martin Lebourdais, Marie Tahon, Antoine Laurent, Sylvain Meignier, Anthony Larcher
LIUM
Avenue Olivier Messiaen, 72000 Le Mans, France
{martin.lebourdais, marie.tahon, antoine.laurent, sylvain.meignier, anthony.larcher}@univ-lemans.fr

Abstract

Our main goal is to study the interactions between speakers according to their gender and role in broadcast media. In this paper, we propose an extensive study of gender and overlap annotations in various speech corpora mainly dedicated to diarisation or transcription tasks. We point out the issue of the heterogeneity of the annotation guidelines for both overlapping speech and gender categories. On top of that, we analyse how the speech content (casual speech, meetings, debate, interviews, etc.) impacts the distribution of overlapping speech segments. On a small dataset of 93 recordings from LCP French channel, we intend to characterise the interactions between speakers according to their gender. Finally, we propose a method which aims to highlight active speech areas in terms of interactions between speakers. Such a visualisation tool could improve the efficiency of qualitative studies conducted by researchers in human sciences.

Keywords: Speech, Overlap, Gender, Corpus, Medias

1. Introduction

The analysis of conversational speech allows characterising communication phenomena and specifing the relationships between the involved participants. Such sociological study finds applications in the field of political speech, television debates or journalistic speech (Beattie, 1982). Whether in an interview or a debate, we are facing asymmetric interactions between at least two kinds of participants of different status, such as a journalist or a guest, with complementary roles. In this context, the characterisation of interruptions can help to identify gender-related phenomena such as *manterrupting* (unnecessary interruption of a woman by a man (Bennett, 2015). The Gender Equality Monitoring (GEM) project aims at exploring gender representations in French broadcast media. In this framework, one of the main challenges is to bring face to face social sciences representations of gender interactions with constrained models suitable for automatic speech processing. For instance, in this work, gender is not considered as a social construct but as a physiological voice characteristic.

The characterisation of interruptions in speech can help to analyse the relationships between speakers according to their gender and role. When a participant interrupts another one, what usually happens when journalists contradict the interviewees (Adda-Decker et al., 2008), it favours the presence of overlapping speech segments (overlaps), defined by the presence of two or more speakers speaking simultaneously. Overlaps structures the conversation and gives clues about how and when the different participants take their speech turn according to their roles. The notion of interruption implies a cognitive action which can not be treated as such by machines. In automatic speech processing, oral communication between several people is usually roughly characterised by a succession of speech turns, where the definition of an interruption is simplified using speaker changes and overlaps. The activist concept of *manterrupting* does not raise a consensus in social sciences. We propose, to study this notion to automatically detect interruptions and gender in speech.

Overlapped speech is a phenomenon that regularly appears in conversations but remains limited in duration. Therefore there is a need for large speech corpora to get enough overlap samples for these studies to be significant. To facilitate and accelerate the analysis carried by social sciences, we propose to implement automatic tools able to segment speaker turns and overlaps, and to extract high-level clues such as gender, in order to provide large amounts of pre-processed data. Most of existing segmentation and characterisation tools are trained in a supervised manner on data related to the task. There are actually many corpora which contain hundreds hours of speech, and numerous speakers. However, while speaker annotation protocols are fairly homogeneous across these corpora, no standard exist for annotating overlapped speech segments. In order to automatically characterise overlaps on the basis of gender and roles labels, one need consistent annotations. Unfortunately, the characterisation of speaker gender and role is not always included in annotation guidelines. To conclude, the speech data usable to train automatic overlaps detection and characterisation models is scarce and inconsistent.

Section 2 provides a review of existing corpora and practices for overlaps detection, and gender annotations. Section 3 presents an analysis of different annotation protocols of various corpora used for this task, a special attention is paid for gender annotations which are critical for GEM project. Finally, section 4 proposes a chronological visualisation of interactions based on speaker turns, using overlap information.

2. Overlap and Gender : Corpora and Practices

To the best of our knowledge, no speech database has been designed and collected for the specific study of speech interruptions between speakers according to their gender or role. Consequently, the data used to train an automatic overlap speech detection system has been originally collected for other tasks. In fact, it is possible to retrieve speech segments where more than one speaker is speaking, directly from a speech turn segmentation.

2.1. Practices in speech segmentation

In tasks using speaker turn segmentation, one can find corpora usable for overlap detection. From then on, we will focus on two tasks: diarisation (speaker segmentation and clustering which answers the question "who speaks when") and speech transcription. Among the speech databases used for one of these two tasks, we can cite AMI (Mccowan et al., 2005), which is a multi-modal corpus of meeting recordings. Corpora collected for the DIHARD I, II and III campains (Ryant et al., 2021) which goal was to evaluate diarisation systems under challenging conditions, i.e. data with a lot of overlaps, background noise and different recording qualities and environments. The DIHARD corpora regroup other existing corpora and are rich and diverse for the study of overlaps.

Focusing on French data, speech turn segmentation mainly comes from corpora designed for automatic speech recognition and speaker identification tasks. Among others, one can mention the radio and television broadcasting corpora REPERE (Giraudel et al., 2012), ETAPE (Gravier et al., 2012), ESTER (Galliano et al., 2006) and EPAC (Estève et al., 2010). If speaker identities (name, gender, role) are annotated, these corpora can also be used in diarisation (Broux et al., 2018). In such case, the transcription is ignored and only the speech turn segmentation is considered. Specific corpora to the study of conversational speech are deemed to contain a lot of overlapped speech. The French NCCFr (Torreira et al., 2010) corpus which contains dyadic conversations between relatives, has been collected to study casual speech, and also contains speech segmentation and speaker identity.

The general pipeline for diarisation usually consists of three steps in which overlapping speech is not treated. The voice activity detection aims at identifying signal segments where at least one participant is speaking, i.e. speech segments, discarding areas of silence, noise or music. Speech segments are then divided into homogeneous segments containing the speech of a single speaker. Finally, the segments are grouped by unique speaker. Conventional systems ignore the presence of overlapped speech or consider it negligible in terms of error. In this case, segments of overlapped speech are simply removed from the evaluation despite their negative impact on the training of speaker models(Huijbregts and Wooters, 2007). Detecting overlapped speech is thus needed to reach a realistic use of diarisation systems but also to improve speaker representation training and improve the overall system performance (Bullock et al., 2020).

In transcription (Automatic Speech Recognition - ASR), most approaches consider single-speaker speech segments in input, consequently neglecting overlaps. It has been shown that transcription errors generated in overlap areas and contiguous areas (Çetin and Shriberg, 2006) are more important than in single-speaker areas, confirming the relevance of detecting overlaps.

As aforementioned, speech portions containing overlaps are interesting to study for some aspects of conversation, such as the study of interruptions (Adda et al., 2007), or the study of reparation mechanisms and disfluencies caused by an interruption (Sacks et al., 1974). Different categories of interruptions have been defined in (Adda-Decker et al., 2008): backchannels (short acknowledgement), addition of complementary information, interruption to take the floor by anticipating the end of another speaker's sentence. The joint analysis of these categories and speaker roles made it possible to identify unexpected links between journalists and their interviewees (Adda-Decker et al., 2008). These works are close to the topics addressed in the GEM project.

2.2. Overlap detection

Overlap detection systems generally output a per-frame segmentation, typically with a 10 ms step, corresponding to a sequence indicating whether the frame contains one or more speakers. Such systems take as input an acoustic representation of the signal frame by frame: MFCCs (Mel Frequency Cepstral Coefficients), mel-spectrograms or direct representations of the waveform with SincNet extraction (Ravanelli and Bengio, 2018).

Most of the current detection systems use sequence-to-sequence neural architecture: recurrent networks including LSTMs (Long Short-Term Memory) (Geiger et al., 2013), or the recent TCNs (Temporal Convolutional Network) which allow the network to have a large temporal context (Cornell et al., 2020). For example, the neural system PyAnnote (Bredin et al., 2020), designed for diarisation, provides a pre-trained overlapped speech detector which has been used in recent challenges such as DIHARD (Ryant et al., 2021) or VoxCeleb speaker identification campaign (Kim et al., 2021; Wang et al., 2021).

As mentioned in the introduction, overlaps appear regularly in a conversation, but remain brief in total duration. Consequently, speech corpora are completely unbalanced towards overlaps. Corpora containing animated debates and spontaneous speech, usually reach the highest overlaps duration, nevertheless contain little overlapped speech. To train an efficient overlap detection model, it is necessary to compensate for this imbalance and increase the number of overlap segments. The

Corpus	Total duration	Overlap proportion	Language	Recording period
ESTER1	99h	0.67%	fr	1998-2004
ESTER2	161h	0.67%	fr	1999-2008
EPAC	105h	5.29%	fr	2003-2004
ETAPE	34h	1.11%	fr	2010-2011
REPERE	58h	3.36%	fr	2011-2013
DIHARD	34h	11.6%	en	NA
AMI	96h	13.87%	en	NA

Table 1: Total duration and proportion of overlaps duration for different speech corpora.

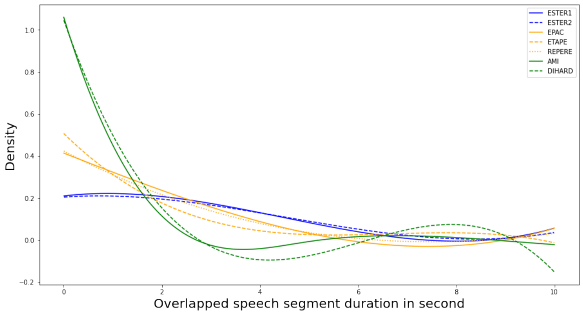

Figure 1: Normalized distribution of overlap segments durations in seconds for different corpora.

solution used so far is to add artificial overlaps (Bullock et al., 2020). This solution makes it possible to balance overlap segments distribution during the training of the system, yet most overlap segments remain artificial and therefore induce a new bias.

3. Annotation Protocols Analysis

This section presents an analysis of different annotation guidelines for two English corpora AMI and DIHARD, and five French corpora of broadcast news.

3.1. Overlapped Speech

Natural language processing community share common annotation rules described in annotation guidelines[1]. However, these rules differ according to the needs for which the corpora have been collected. In order to correctly process overlaps in a given corpus, one must not only be aware of the general guidelines used during the annotation process, but also of the different issues and their proposed solutions encountered at that time. This aspect highlights the difficulty of gathering annotations coming from different domains, each having its own historical context.

Diarisation annotations mainly use a speech turn annotation in MDTM tabular format. The MDTM format has been developed by NIST for the Rich Transcription evaluation campaign in the early 2000s. The segmentation consists of speech turns with start and end times, the name of the speaker and the name of the show from which the segment is extracted. Some other information associated with the speaker such as gender and role are often added. Overlapped speech annotation is not explicit. The overlap segments need to be generated by taking the intersection between speech turns. When speech turns overlap, we are in the presence of overlapped speech.

Corpora developed for transcription generally use the XML format of Transcriber (Barras et al., 1998) in the corpora selected for our studies (ESTER 1 and 2, EPAC, ETAPE, REPERE). Each turn is associated with one or several speaker names and information on gender, accent and channel quality. In the case of multi-speaker turn, the words spoken by each speaker within a turn are also tagged so the start and the end of the words from each speaker can be found. However it exists several major differences between corpora. In ESTER 1 and 2 corpora as well as EPAC corpus, when two speakers speak simultaneously in an intelligible way, the transcription corresponding to these speakers is often annotated although the annotation guide only explicitly requires the transcription of the main speaker. In ETAPE corpus, temporal indications of overlapped speech are not always indicated within the speaker turn making impossible to identify when the overlap segment starts and ends. On the other hand, overlap category is precisely annotated according to the definitions given in (Adda-Decker et al., 2008). In REPERE, when two speakers speak at the same time, only the transcription of the main speaker is annotated.

Table 1 gives statistics for the duration of studied corpora as well as the proportion (in terms of duration) of overlaps. The smallest proportion of overlaps can be found in ESTER 1 and 2. These corpora are essentially composed of radio broadcasts with newscasters, columnists, experts, and politics. These corpora contain little interviews and no debate, what explains the relatively low amount of overlaps.

In contrast, REPERE, ETAPE and EPAC corpora reach a higher proportion of overlap, mainly because they contain debate shows where overlapping speech segments are more frequent and longer. This proportion can be as high as 10% on a show such as *Ça vous regarde* from LCP French channel, present in REPERE and ETAPE corpora.

AMI and DIHARD mainly contain conversational speech corpora including spontaneous speech. The overlap proportion is higher than in the five other corpora. AMI is a corpus recorded in meetings, and DI-

[1] Annotation guideline from Transcriber `http://trans.sourceforge.net/en/transguidFR.php`.

	Annotated Proportion		
Corpus	**Female**	**Male**	**Non specified proportion**
ESTER1	30.3%	69.7%	0.1%
ESTER2	25.1%	74.9%	29.8%
ETAPE	18.5%	81.5%	35.2%
EPAC	18.21%	81.8%	1.1%
REPERE	20.0%	80.0%	0.1%

Table 2: Speech time proportion for women, men and non specified speech time for different french corpora.

Show	Total Duration	Overlapped speech	Female/Male Proportion
Ça vous regarde	15h	10.44%	16.4% / 81.6%
Entre les lignes	16h	8.08%	7.9% / 91.4%
Pile et face	17h	10.99%	16.7% / 83.2%

Table 3: Total duration, overlaps duration proportion, and proportion of speech for each gender for the three chosen debate shows from ALLIES corpus.

HARD contains a lot of highly interactive data with few communication specialists, for example data collected in a restaurant. The GEM project focuses on the study of interactions between speakers according to their gender and role in broadcast media. Although they are useful to learn a detection system, the characteristics of AMI and DIHARD corpora do not fit well with GEM goals.

Figure 1 shows the distribution of the duration of overlap segments for the seven aforementioned corpora. We can see that the more spontaneous speech the corpus contains, the shorter the speech segments and overlaps. While AMI and DIHARD contain the shortest segments (< 1 second), REPERE, ETAPE and EPAC corpora contain mostly segments between 1 and 2 seconds. ESTER 1 and 2 corpora, on the other hand, have relatively uniform duration. Few short segments have been annotated due to the complexity of the task. Additionally, overlap segments are relatively rare due to the nature of the corpora.

3.2. Gender Annotations

GEM project focuses on the interaction between speakers, according to the gender and role of the speakers. Contrary to the speaker role, speaker gender is available in most of the studied corpora. Therefore we decided to focus on gender only, letting speaker role for future works. Amongst the different databases, we find three categories of gender : Male, Female, and unknown.

The Table 2 shows that the proportion of Females and Males in the five French corpora is not balanced. This phenomenon is well known (Garnerin et al., 2019), and is all the more important as the data studied predate the French regulations on equality between women and men[2]. This imbalance proportion obviously leads to a bias in the data for any system trained with these corpora (Garnerin et al., 2019).

We can note that a third of the data from ESTER2 and ETAPE do not specify a gender. Gender is not annotated at all in DIHARD and AMI corpora. Only ESTER1 campaign proposed a gender identification task. The other corpora were not created with the objective of working on gender, which partly explains the absence of annotation.

3.3. Interactions and Gender

In this section we intend to drive some statistical analysis in order to establish some correlations between gender and interruptions. More precisely, we would like to assess if women are more likely to be interrupted by a man rather than by another woman (*mansterrupting)*, and if this interruption is realised with overlapping speech or not. To do so, we use data from the soon-to-be-available ALLIES corpus (Larcher et al., 2021), which is a diarisation corpus that follows on from the ESTER1 and 2, REPERE and ETAPE campaigns, and contains data taken from French television and radio shows. Within this corpus, we focused on three shows from the LCP channel recorded between 2010 and 2014 (*Ça vous regarde, Entre les lignes, Pile et face*). These shows present a proportion of more than 8% of overlaps, as indicated in the Table 3.

In the followings, we focus on interactions between two speakers (S_1 speaks first, then S_2 takes the floor) which are characterised according three different ensembles. The universe Ω contains all the interactions, O: contains interactions with an overlap segment, and SC: contains interactions with a speaker change, *i.e.* the first speaker is S_1 and the last speaker is S_2.

Figure 2 details the three categories used to count the number of interactions:

- O^c: the speech turn of the first speaker S_1 is directly followed by a new turn corresponding to a second speaker S_2 without overlap,

[2] French law of 4 August 2014 for the real equality between women and men, and law of 27 January 2011 on the balanced representation of women and men on directors and supervisory boards

Gender interaction	**Interactions number**		
$S_1 - S_2$	Ω	**O**	**O ∩ SC**
M-M	12008	4420 (36.81%)	2208 (49.95%)
M-F	1359	640 (47.09%)	371 (57.97%)
F-M	1357	716 (52.76%)	311 (43.44%)
F-F	467	78 (16.70%)	31 (39.74%)

Table 4: Number of interactions, number of interactions with overlaps and number of interactions with speaker change depending on the gender of the first (S_1) and second (S_2) speaker for three debate shows from ALLIES. O is the relative proportion of interactions with overlap while $O \cap SC$ is the relative proportion of speaker changes with overlap for a given gender interaction .

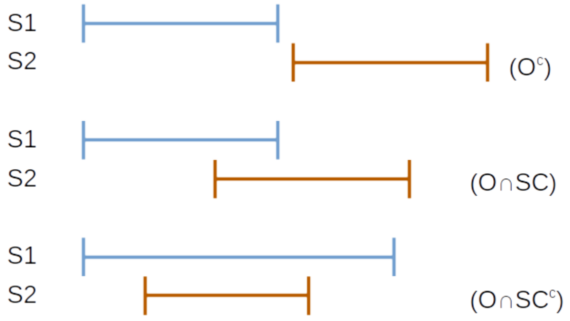

Figure 2: Types of interactions considered.

- O ∩ SC: the speech turn of S_1 ends after the following speech turn of S_2, indicating the presence of an overlap between S_1 and S_2,

- O ∩ SC^c: the speech turn of S_2 starts and ends before S_1 turn is finished, indicating an overlap between S_1 and S_2.

The number of interactions between two speakers is extracted directly from the reference segmentation. Table 4, summarises the total number of interactions (Ω), interactions with overlaps (O), and interactions with overlaps and speaker change corresponding to the interruption of S_1 by S_2 (O∩SC). These numbers are given for all gender interactions (first and second speakers are males, first is male, second is female, etc.) calculated on the corpus of the three LCP shows. The frequencies calculated on separated shows are consistent with those obtained on all the shows.

Because the number of females is clearly lower than the number of males, it is not possible to compare the total number of interactions for the different $S_1 - S_2$ combinations in absolute. That is the reason why, we have included the overlap (respectively overlap with change) proportion relatively to the total number of interactions (resp. the interactions with overlap).

From these results, we can draw several comments. First, the proportion of interactions with overlaps is almost balanced between F-M and M-F interactions. These proportions are higher than those measured between two women (16.70%) or two men (36.81%), thus indicating a potential gender effect in the interactions. Secondly, we can note that the number of interactions with overlaps between two women is quite low (16.70%) and these overlaps rather correspond to backchannels or complementary speech ($O \cap SC^c$) than to interruptions ($O \cap SC$) (39.74%) of cases. This reveals that interactions between women involve less overlapping speech, and if so, the interruption rarely appears to take the floor. It can also be noted that it is more common for a woman to interrupt a man with an overlap (57.97%) than the opposite (43.44%).

These first results are based on 93 recordings of 3 shows collected between 2010 and 2014 and a study on a wider range of shows and years is needed to confirm it. We hypothesize a gender effect, however, further investigations are needed on how gender and role impact these results.

4. Visualisation of Highly Interactive Speech Zones

So far, no consensual categories of speaker interaction have emerged, mainly due to the difficulty and the subjectivity of the phenomenon. As no automatic tool exists, the study of interactions often requires human listening in order to characterise interactions or carry out a detailed analysis. In this part, we present a visualisation method which assists a human annotator with the visualisation of high interactivity speech zones from audio data.

For a given audio file previously segmented speech turns, we extract the cumulative duration of overlap segments, called d_{acc} as a function of their starting time, called t_{deb}. We obtain N tuples (d_{acc}, t_{deb}) that follow the show timeline. We then calculate the derivative $\Delta[n]$ as described in eq. 1 where N is the total number of overlap segments, for each point n corresponding to the order of the segments contained t_{deb} and d_{acc}.

$$\Delta[n] = \frac{d_{acc}[n+h] - d_{acc}[n-h]}{t_{deb}[n+h] - t_{deb}[n-h]} \quad \forall n \in [0; N] \tag{1}$$

The set empirically $h = 9$ to smooth Δ on a large temporal window. The cumulative duration being a strictly increasing function, $\Delta > 0$. A value of 1 would imply that the duration of an overlap segment d_{acc} is equal to the time between two consecutive segments. This situation is impossible unless there is an error in the annotation, therefore $\Delta < 1$. More precisely, a high value of

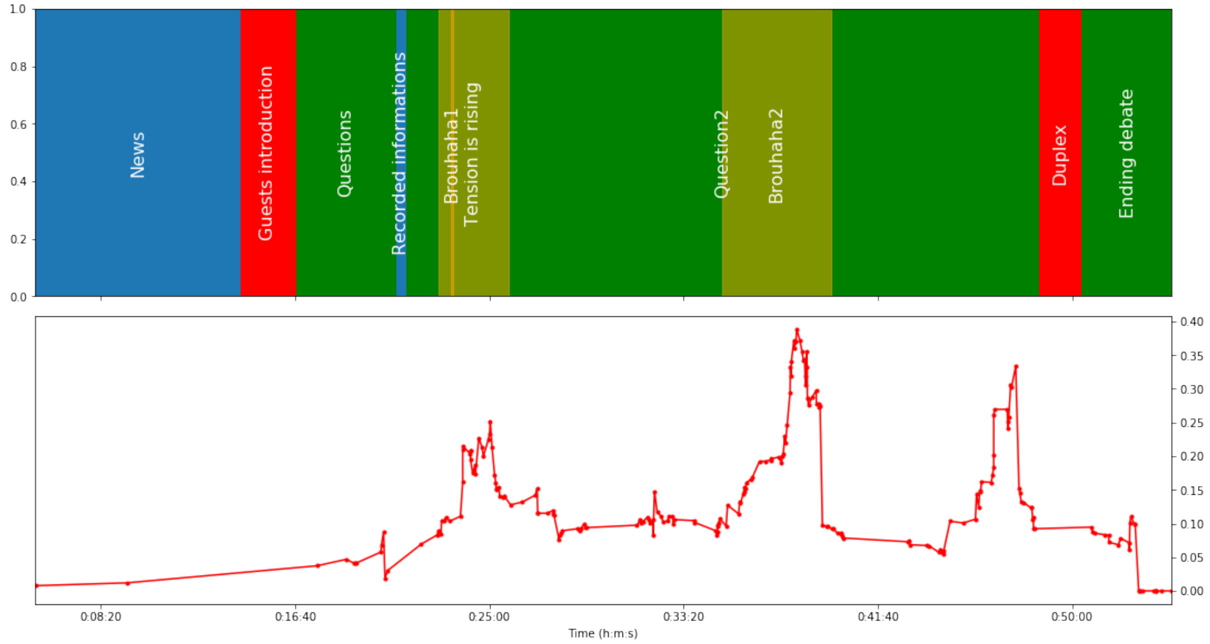

Figure 3: Chronological evolution of the derivative Δ (bottom) and manual annotations of the show *Ça vous regarde* recorded the 29/04/2014 (up).

Δ indicates the presence of either a fast succession of consecutive short overlaps, or long overlap segments. Consequently, high Δ values are characteristic of areas of high interactivity.

We then draw the curve corresponding to Δ as a function of t_{deb} as represented at the bottom of Figure 3 for the show *Ca vous regarde* recorded on the 29/04/2104. To measure the usefulness of our representation, we manually structured the show into different phases, represented in the Figure 3 by the timeline (up). The curve of Δ, associated with the phases of the show, clearly shows that high interactivity areas (light green) are correlated with a peak in the derivative.

This preliminary study shows the high potential of such a representation. Further works on additional shows (including recent recordings prior to the gender equality laws), with an automatic overlap segmentation (instead of the reference) are needed to confirm it. Also, a validation by GEM humanities partners should confirm that the targeted high interactivity speech areas are relevant to study interruptions in broadcast media.

5. Conclusion

In the framework of GEM project, we intend to study the interactions between speakers according to their gender and role in French broadcast media. This paper proposes an extensive analysis of gender and overlaps annotations in various speech corpora, thus pointing out the heterogeneity of guidelines and treatments.

This statistical study shows that the speech content (casual speech, debates, interviews, etc.) highly impacts the distribution of overlaps duration. We also highlight the fact that overlaps can be considered as a rare phenomenon in broadcast news.

We conducted a complete analysis of 93 recordings from three *LCP* shows regarding the interaction categories according to the gender of the speakers. We conclude that interactions which involve two different genders are more likely to include overlap, thus indicating a potential gender effect. We also found that it is more common for a woman to interrupt a man with overlap than the opposite. Given our findings on the link between overlaps and gender, it would be interesting to continue similar studies on a larger scale using automatic overlap detection and to include role information to confirm or refute our observations. To do so, we face the problem of the scarcity of overlapping speech in existing databases and the difficulty of merging heterogeneous annotations. In order to overcome the lack of overlapped speech and to allow easiest merging of corpora, it is necessary to pursue the agreement between the major actors of the fields. Although the official conventions are similar, the implicit rules currently vary too much. Initiatives such as the ALLIES corpus will certainly move the community forward by providing a large source of harmonised data. We also proposed a visualisation which could be used to speed-up the annotation and characterisation process, by indicating interactive speech areas of interest for human analysis. Additional information such as the proportion of women on a sliding window will be included to this representation in a future work.

6. Acknowledgement

Our research is funded by the French national agency project GEM, Gender Equality Monitoring (ANR-19-CE38-0012).

7. Bibliographical References

Adda, G., Adda-Decker, M., Barras, C., Boula de Mareüil, P., Habert, B., and Paroubek, P. (2007). Speech overlap and interplay with disfluencies in political interviews. *International Workshop on Paralinguistic Speech-between models and data, ParaLing*, pages 41–46.

Adda-Decker, M., Barras, C., Adda, G., Paroubek, P., de Mareüil, P. B., and Habert, B. (2008). Annotation and analysis of overlapping speech in political interviews. In *LREC*, pages 3105–3111, Marrakech, Morocco.

Beattie, G. (1982). Turn-taking and interruption in political interviews: Margaret Thatcher and Jim Callaghan compared and contrasted. *Semiotica*, 39:93–114.

Bennett, J. (2015). How not to be 'manterrupted' in meetings. *Time*, January, 14.

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*, pages 7124–7128, Barcelona, Spain.

Broux, P.-A., Doukhan, D., Petitrenaud, S., Meignier, S., and Carrive, J. (2018). Computer-assisted Speaker Diarization: How to Evaluate Human Corrections. In *LREC*, Miyazaki, Japan, May.

Bullock, L., Bredin, H., and Garcia-Perera, L. P. (2020). Overlap-Aware Diarization: Resegmentation Using Neural End-to-End Overlapped Speech Detection. In *ICASSP*, pages 7114–7118, Barcelona, Spain.

Cornell, S., Omologo, M., Squartini, S., and Vincent, E. (2020). Detecting and Counting Overlapping Speakers in Distant Speech Scenarios. In *Interspeech*, pages 3107–3111, Shanghai, China.

Garnerin, M., Rossato, S., and Besacier, L. (2019). Gender Representation in French Broadcast Corpora and Its Impact on ASR Performance. In *AI for Smart TV Content Production, Access and Delivery*, AI4TV '19, pages 3–9, New York, NY, USA.

Geiger, J., Eyben, F., Schuller, B., and Rigoll, G. (2013). Detecting overlapping speech with long short-term memory recurrent neural networks. In *Interspeech*, pages 1668–1672, Lyon, France.

Huijbregts, M. and Wooters, C. (2007). The blame game: performance analysis of speaker diarization system components. In *Interspeech*.

Kim, M., Ki, T., Anshu, A., and Apsingekar, V. R. (2021). North America Bixby Speaker Diarization System for the VoxCeleb Speaker Recognition Challenge 2021. *arXiv:2109.13518 [eess]*.

Ravanelli, M. and Bengio, Y. (2018). Speaker Recognition from Raw Waveform with SincNet. *SLT*, pages 1021–1028.

Wang, K., Mao, X., Wu, H., Ding, C., Shang, C., Xia, R., and Wang, Y. (2021). The ByteDance Speaker Diarization System for the VoxCeleb Speaker Recognition Challenge 2021. *arXiv:2109.02047 [cs, eess]*.

Çetin, O. and Shriberg, E. (2006). Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: insights for automatic speech recognition. In *Interspeech*, pages paper 1915–Mon2A2O.6, Pittsburgh, USA.

8. Language Resource References

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (1998). Transcriber: a free tool for segmenting, labeling and transcribing speech. In *LREC*, pages 1373–1376, Granada, Spain.

Estève, Y., Bazillon, T., Antoine, J.-Y., Béchet, F., and Farinas, J. (2010). The EPAC corpus: manual and automatic annotations of conversational speech in French broadcast news. In *LREC*, pages 1686–1689, Valetta, Malta.

Galliano, S., Geoffrois, E., Gravier, G., Bonastre, J.-F., Mostefa, D., and Choukri, K. (2006). Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news. In *LREC*, pages 139–142, Genoa, Italy, May.

Giraudel, A., Carré, M., Mapelli, V., Kahn, J., Galibert, O., and Quintard, L. (2012). The REPERE Corpus : a multimodal corpus for person recognition. In *LREC*, pages 1102–1107, Istanbul, Turkey.

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 *LREC*, pages 114–118, Istanbul, Turkey.

Larcher, A., Mehrish, A., Tahon, M., Meignier, S., Carrive, J., Doukhan, D., Galibert, O., and Evans, N. (2021). Speaker Embedding For Diarization Of Broadcast Data In The ALLIES Challenge. In *ICASSP*, pages 5799–5803, Toronto, Canada.

Mccowan, I., Lathoud, G., Lincoln, M., Lisowska, A., Post, W., Reidsma, D., and Wellner, P. (2005). The ami meeting corpus. In *Proceedings on the Conference on Methods and Techniques in Behavioral Research*, page 4, Wageningen, Netherlands.

Ryant, N., Singh, P., Krishnamohan, V., Varma, R., Church, K., Cieri, C., Du, J., Ganapathy, S., and Liberman, M. (2021). The third dihard diarization challenge.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Torreira, F., Adda-Decker, M., and Ernestus, M. (2010). The Nijmegen Corpus of Casual French. *Speech Communication*, 52:201–212.