Speech Overlap and Interplay with Disfluencies in Political Interviews

Gilles Adda[1], *Martine Adda-Decker*[1], *Claude Barras*[1,2],
Philippe Boula de Mareüil[1], *Benoît Habert*[1,3], *Patrick Paroubek*[1]

[1] LIMSI-CNRS, Orsay, France
[2] Univ Paris-Sud Orsay France
[3] Univ Paris-X Nanterre France

{Gilles.Adda,Martine.Adda,Claude.Barras,Philippe.Boula.de.Mareuil,
Benoit.Habert,Patrick.Paroubek}@limsi.fr

Abstract

The reported study focuses on overlapping speech, transcription, annotation and disfluency analysis in an 8-hour audio corpus of French political interviews. Overlaps are frequent (on average 3-4 overlaps per minute) and of short duration (5% of data), non-intrusive overlaps being significantly shorter than intrusive ones. Disfluencies include repetitions, revisions and filled pauses. Manual annotation achieved a higher inter-annotator agreement when based on the four overlap types: back-channel, turn request, anticipated turn taking and complementary. Discourse markers are also considered in this study. The disfluency rate in overlaps is almost double of the one in non-overlapping speech. Repetitions are the most involved disfluency type, especially for intrusive overlaps (turn requests and complementary). The study highlights interesting differences between active (incoming) and passive (floor holding) overlap speakers, as well as between journalists and interviewees.

Index Terms: disfluencies, speech annotation, overlapping speech

1. Introduction

Overlapping speech tends to become a hot topic in automatic speech recognition (ASR) research. In earlier years, transcription tasks have focused on situations where speakers can be considered as separate audio streams, without making unwarranted assumptions and overlapping speech have been neglected. Speech overlaps and turn-taking however are a major focus in more traditional research areas such as discourse analysis [7, 8].

Viewing oral communication between several actors as a sequence of single speaker turns is a too strong assumption, because overlapping speech, i.e. speech portions simultaneously involving more than one speaker, is very common in natural communication contexts [6]. Overlaps may entail disfluencies (hesitations, repetitions, restarts) and are likely to contribute to speaker turn regulation. They definitely cause problems for automatic processing [9].

In the present study, real-world interactive speech data is annotated with a primary objective of elaborating valuable information for ASR language modeling of interactive speech. An additional aim is to provide a description useful for linguistic studies, such as interaction and discourse analysis. Multi-party speaker related productions, as well as speech disfluencies in overlapping speech are then examined with respect to speaker roles and attitudes, reflected by the proposed active/passive (incoming/floor holding) and intrusive/non-intrusive (speech flow disrupting/preserving) overlap types. Our contribution focuses on overlapping speech phenomena in TV political interviews, where overlaps do occur, even though their overall ratio remains relatively low as compared to ratios reported for conversational or meeting speech [9]. As roles in these interviews are asymmetrical, it may be enlightening to analyze overlapping speech and disfluency measures with respect to the speaker's role in the communication context.

The questions addressed in this study are the following ones:

- How to annotate overlapping speech for both automatic processing and linguistic studies?
- Are there different types of overlapping speech and if so, can they be qualified as more or less intrusive?

A large amount of speech overlaps can be seen as anticipated speaker turns and may be understood as a particular case of speaker synchronization as opposed to inter-speaker silences. A major point of interest concerns the link between overlapping speech and disfluency production.

- Do overlap types impact disfluency rates and types?
- Do disfluency rates significantly differ in active versus passive roles in the overlap situation?

In section 2 we present both the material and the methodology for segmentation and annotation of overlapping speech. Section 3 presents results on speech overlaps and disfluencies with analyses along different axis for overlaps: intrusive versus non-intrusive, passive versus active. Finally section 4 recapitulates our findings.

2. Material and methodology

2.1. Terminology

Prior to further developments, let us explain the meaning of some common terms we use hereafter.

Segmentation is the result of dividing the speech flow into segments according to particular criteria (e.g. speaker iden-

tity), segment boundaries being specified by time codes. Hereafter manual segmentation produces either single speaker segments or may include overlapping (multi-speaker) speech parts. Speaker turns may be composed of one or more segments.

Transcription provides a normative orthographic (verbatim) word flow of the segmented audio data.

Annotation is carried out on overlapping speech segments. Part of the work aims at proposing labels, both relevant for ASR and discourse analysis studies. Disfluency labels used in this study stem from earlier work [1].

The proposed overlap labels are further grouped as **intrusive/non-intrusive** to characterize their more or less disrupting effect on the speech flow of the current speaker and to study their relationship with disfluencies.

Active/passive overlap speakers are distinguished according to their roles in the overlap situation. The incoming speaker generates the overlap situation and hence corresponds to the active overlap speaker, no matter whether he/she keeps the floor at the end of the overlap segment.

2.2. Corpus outline

The reported study is based upon French broadcast interviews recorded in the early nineties. The corpus is composed of 8 one-hour TV shows during which a major figure from either political or civil society is interviewed by 3 journalists and a chairman. The chairman watches over the schedule and may interrupt interviewees or interviewers to have them stick to previously determined topics and timing. This configuration favours speech overlaps and disfluencies among interlocutors.

The audio corpus benefits from exact orthographic transcripts including disfluent speech events and specific annotations concerning discourse markers (DM) and disfluencies, namely filled pauses (FP), repetitions (RP) and revisions (RV) [1] in line with the LDC annotation guidelines [1] and the French GARS conventions [3]. The Transcriber software[2] has been customized to facilitate and speed up the manual annotation process through contextual menus and a coloured display of the various disfluency and overlap types. The new overlap and disfluency annotation tags are embedded into XML transcription files.

2.3. Overlap segmentation, transcription

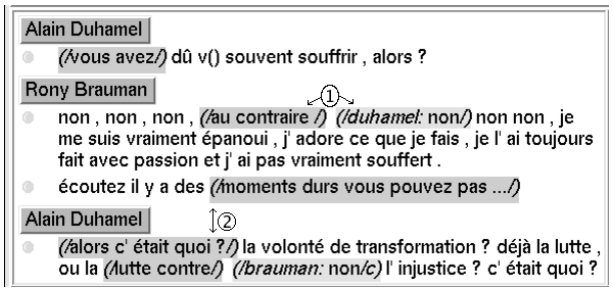
In telephone conversations or meetings, overlaps are very frequent (with more than 10% of overlapping words [9]). It may hence be convenient to transcribe each speaker as a separate synchronized stream. On the opposite, broadcast news is a very controlled form of communication, which includes a high proportion of monologues and thus very few overlaps. For automatic speech transcription, it is usual to partition broadcast news data as a sequence of individual speaker turns, setting aside overlapping segments with a precise temporal anchoring [2], as their processing is still beyond the scope of state-of-the-art systems. Our corpus of political interviews is less controlled than broadcast news, and a crude segmentation of overlapping segments has the drawback of breaking the logical interaction stream. We thus chose to preserve the interaction structure by relaxing temporal synchronization constraints at turn boundaries in the case of overlaps.

An overlap occurs when a first speaker (primary) keeps talking while a second speaker comes in. The more complex situation of more than two people speaking at the same time appeared to be negligible in our data. For overlap segmentation and transcription, we distinguished two situations:

1. the overlap does not entail a speaker change: the primary speaker remains the same at the end of the overlap.
2. the overlap results in a speaker change: the primary speaker stops and the second speaker becomes the primary speaker of the new turn.

Fig. 1 shows examples of segmentation and transcription in both cases. Overlapping words are highlighted in the transcription. For case ① the first portion of words corresponds to the default (primary) speaker, followed by the portion from the overlapping speaker, explicitly named here (*"duhamel"*). Case ② features overlapping words in sequential speaker-dependent streams, the highlighted section marking lose overlap boundaries. Overlapped speech is hence marked on the transcription level, with approximate time stamps in the audio stream but remaining highly legible in the transcription.

Figure 1: *Examples of overlapped speech transcription and its display in the customised Transcriber annotation editor. ①: no speaker change ; ②: primary speaker stops.*



2.4. Overlap tagset

Speech overlap types may be viewed as a continuum, the extremes of which are back-channels and turn requests. Back-channels (***bkch***) indicate that we follow our interlocutor, understand him/her, agree with him/her [5]. They barely disturb the main speaker, in opposition to turn taking overlaps, where the entering speaker tries to get the floor. Two types of turn taking overlaps have been distinguished: turn requests (***trq***) are clear attempts to interrupt the main speaker, although these may fail (as may any other "speech acts"). A second type corresponds to anticipated turn taking (***att***). An *att* may occur at a (potential) turn end of the primary speaker. The incoming speaker seems to perceive specific cues for *att* (clause or phrase boundaries, falling pitch...). Finally a complementary label (***cmpl***) has been introduced for overlaps which aim at complementing the main speaker's topic. This somewhat heterogeneous category covers: a possibly paraphrased repetition of the primary speaker's statement; an explicit agreement or disagreement; a

[1] http://www.ldc.upenn.edu/Projects/MDE/
[2] http://sf.net/projects/trans/

short anticipated answer; a precision forwarded or required, not only on the content but also on the form of the exchange (time limit, approached topic); a witty remark or the continuation of the utterance. Contrary to turn request, the complementary label is assigned to self-sufficient utterances or comments: the intervening speaker does not take the floor to develop an argument. This type of overlap may be favored by the communication situation: actors of the show may wish to provide additional information to the audience, beyond the actively involved speakers. Fig. 2 shows an example for each overlap tag.

Figure 2: *Examples of the different overlap types.*

bkch: backchannel
A: it is simply /the fact/ /B: hmm/ that...

cmpl: complementary
A: I have a last question /about/ /B: very short/ about your...

trq: turn request
A: and in /this case.../
B: /I want to/ come back...

att: anticipated turn taking
A: and this leads to humanitary /action?/
B: /well I/ think

2.5. Overlap annotation

Beyond the problem of precisely locating overlap events, differences between overlap tags happen to be subtle and may give rise to diverging interpretations. A unique label assignment is not always straightforward. For example some *att* events can be seen as *trq*. Even *"hmm"*s may have additional communicative functions of complementing or of signaling that one is eager to jump in. In fact, progressive transitions from back-channeling *"hmm"*s to complementary or turn requesting items are very common during a long lasting turn. The distinction between intrusive or non-intrusive overlaps can rely on prosody but also on possible segmentation points (sentence, clause or phrase ends for instance).

Several options may be taken for labeling speech overlaps. Multiple annotators were felt necessary for this time-consuming task. First, 2 shows were annotated by 5 annotators and the reference annotation resulted from harmonizing the different annotations through first negotiation, then adjudication, for the disputed labels. Table 1 presents the label distribution for the different annotators. It confirms the intermediate nature of the complementary label, and shows a rather high confusion value of 24% between *att* and *trq*. Compared to the reference, the 5 annotations show an inter-annotation agreement Kappa measure [4] between 0.7 and 0.8, which decreases to a 0.6–0.7 interval when only a *trq* vs. *att* binary choice is considered. Each of the remaining six shows was processed by one single annotator and passed over to a colleague for verification. Corrections involved between 3% and 6% of the labels. This can be taken as an estimate of the residual disagreement rate, but it also reflects the problem of assigning a unique label when two categories appear to be relevant.

Table 1: *Overlap label distribution from 5 annotators relative to the final annotation of one show (1 hour).*

		annotator labels (%)			
label	count	*bkch*	*cmpl*	*trq*	*att*
bkch	63	91.1	8.0	1.0	0.0
cmpl	50	9.2	75.8	15.0	0.0
trq	107	0.4	3.6	89.2	6.8
att	26	0.0	0.0	**24.0**	76.0

3. Experimental results

We first report results concerning speech overlaps, before investigating their link with disfluencies.

3.1. Overlapping speech

Although speech overlaps occur frequently (on average 3-4 overlaps per minute), their cumulative duration remains short (below 5% of the data). Overlaps, averaging 2.5 words per segment, are very brief compared to single speaker turns (30 words on average).

3.1.1. Intrusive/non-intrusive overlaps

The *bkch* label typically corresponds to a very short non-intrusive overlap, meant to encourage a fluid interaction. *cmpl* overlaps do not aim at a speaker change, and may be felt as non-intrusive by their author. However, both their length and informational content are likely to disturb the primary speaker and thus to generate disfluencies in their speech flow. They are hence considered as intrusive. *trq* is clearly intrusive and either implicitly or explicitly asks for a speaker change. *att* is a non-intrusive form of overlap, occuring slightly in advance of a commonly agreed speaker change. The message of the primary speaker, even though not yet completely uttered, has already been received by the audience.

3.1.2. Active/passive overlaps

Overlapping speech can be analyzed by comparing productions from active overlap speakers to those of primary speakers who are considered as passive with respect to the overlap situation. Table 2 shows overlapping segment counts, their frequency, word counts and mean length for intrusive (*trq*, *cmpl*) and non-intrusive overlaps (*bkch*, *att*). Overall, non-intrusive overlaps are shorter than intrusive overlaps. Active and passive figures are quite comparable. The highest production is measured for active turn requests, if not in number of occurrences, at least in number of words. In this challenging situation, active speakers tend to speak faster than passive competitors.

Table 2: *Overlap segment counts (# of segments of a given overlap type), frequency, word counts (# of words included in segments) and mean segment length (in words) for passive (P) and active (A) roles and for* bkch*,* cmpl*,* trq *and* att*.*

category		segment #	freq. /min.	words #	words %	mean length
bkch	P	461	1.2	719	0.8	1.6
	A			550	0.6	1.2
att	P	168	0.4	345	0.4	2.1
	A			391	0.5	2.3
cmpl	P	278	0.7	955	1.1	3.4
	A			974	1.1	3.5
trq	P	438	1.1	1447	1.7	3.3
	A			1658	1.9	3.8

3.2. Disfluencies

Overlapping speech is assumed to increase disfluency rates as compared to overall rates measured in single speaker regions. The first three lines of Table 3 show discourse marker and disfluency rates in non-overlapping and overlapping speech. Measures exhibit an important increase of disfluencies in overlap regions. More disfluencies are produced by the active (incoming) overlap speakers than by the passive (primary) overlap speakers. Higher rates of repetitions and discourse markers are measured on active overlap speaker segments in comparison to their passive counterparts. These high rates cannot be explained alone by the overlapping nature of speech. It is worth mentionning here an exploratory study of local disfluency rates in non-overlapping regions. This analysis suggests and confirms previous studies [9] that disfluency rates globally follow a *"declension line"* over time: high figures for repetitions and discourse markers were observed at the very beginning of speech turns, followed by quickly dropping rates for more turn-internal positions. The following lines of Table 3 give separate figures

Table 3: *DM and disfluency rates in non-overlapping speech (*non-over*), for passive (P) and active (A) roles for intrusive and non-intrusive overlap segments. (*over*: overlap;* non-intr*: non-intrusive;* intr*: intrusive).* DM *means discourse markers,* FP*: filled pauses,* RV*: revisions,* RP*: repetitions.*

category		% DM	% disfluencies FP	RV	RP	All
non-over		2.4	2.0	2.5	2.5	6.9
over	P	2.1	1.6	2.3	7.2	11.1
	A	5.9	0.5	3.0	11.0	14.5
non-intr	P	2.4	1.6	2.0	1.3	4.9
	A	7.2	0.6	0.9	5.2	6.7
intr	P	2.0	1.6	2.5	9.5	13.6
	A	5.4	0.4	3.8	13.0	17.2

for both intrusive vs non-intrusive and passive vs active conditions. Very few disfluencies are found for the non-intrusive condition. Detailed figures by overlap type (see Table 4) reveal that *bkch* does not raise the average disfluency rates of passive overlap speakers.

Table 4: *DM and disfluencies rates in non-overlapping speech (non-over) and for passive (P) and active (A) roles for each of the overlap types.*

category		% DM	% disfluencies FP	RV	RP	All
non-over		2.4	2.0	2.5	2.5	6.9
over	P	2.1	1.6	2.3	7.2	11.1
	A	5.9	0.5	3.0	11.0	14.5
non intrusive overlaps						
bkch	P	3.1	2.2	2.5	1.9	6.7
	A	2.0	0.2	1.3	1.1	2.5
att	P	1.2	0.3	0.9	-	1.2
	A	14.6	1.3	0.3	11.0	12.5
intrusive overlaps						
cmpl	P	1.9	1.5	3.7	16.3	21.5
	A	4.1	0.2	0.7	7.6	8.5
trq	P	2.1	1.2	1.8	5.7	8.6
	A	6.2	0.6	5.7	16.4	22.7

By contrast, active overlap speakers happen to be very disfluent during *att* overlaps. Beyond the overlap situation, this can also be related to turn-start positions, where high disfluency rates (in particular repetitions, which may have other interpretations than disfluencies in this position) are observed even in non-overlapping speech. Concerning the overall high overlap rates in the intrusive condition, a detailed analysis by overlap type also highlights interesting differences between active and passive roles. Whereas *trq* favors the production of disfluencies by active speakers, passive (primary) speakers become dramatically disfluent on *cmpl* segments which correspond to overlapping comments from the entering speaker (see subsection 3.1.1).

To get a more statistically informed view of the presented disfluency results, Fig. 3 makes use of a box-and-whiskers[3] representation. To do so, we consider the whole population of speakers, while the points for the different speaker categories (Interviewees, Chairman, Journalists) are the mean values produced by cumulating the occurrences of disfluences for each speaker category. As previously, disfluency rates are given for the different types of regions: non-overlapping (*non-over*) and overlapping (*over*). The latter are then analyzed with respect to both passive and active roles in overlaps, corresponding respectively to the primary and overlapping (secondary) speakers. Overlap types are examined in intrusive (*intr*) and non-intrusive (*non-intr*) conditions. Disfluency rates are lower for non-intrusive (*bkch*, *att*) segments as compared to intrusive

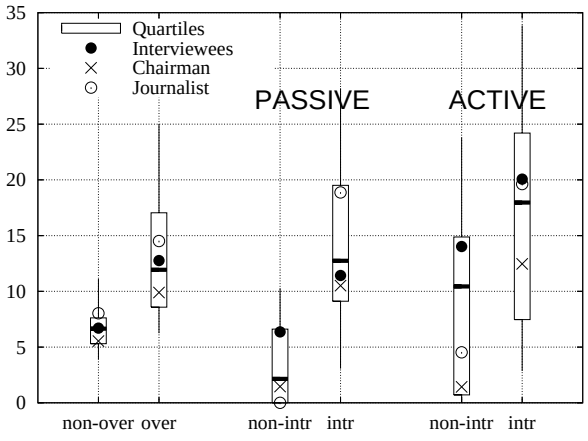
[3]The horizontal bar within each box represents the median; the box includes 50% of the population (the 2nd and the 3rd quartiles). The remaining 1st and 4th quartiles (without the outliers) are represented by the vertical bars above and below the box. A synthetic description of box-and-whiskers can be found at
`http://en.wikipedia.org/wiki/Box_plot`

overlaps. In passive conditions, they are even lower than the average disfluency rate in non-overlapping speech: backchannels are known to preserve the speech flow and the primary speaker of an *att* overlap is by definition reaching the end of his/her turn.

The higher rate achieved for active (secondary speaker) anticipated turn-taking concerns non-overlapping sections at the turn beginning which contain more disfluencies than the middle or the end of the segment.

In intrusive segments we may see that the situation is dissymetric for active and passive complementary segments, with a very high disfluency rate for the passive speakers. Concerning the relation with the speaker's role, we can see in Fig. 3 that although overall disfluency rates are almost the same for Journalists, Interviewees, and the Chairman, condition-dependent rates in overlapping speech are quite different. In non-intrusive segments, Interviewees have higher disfluency rates; for intrusive segments the situation is dissymmetric for passive and active conditions: in the passive case, Journalists have higher rates, while for the active condition, rates are comparable. Possible explanations include an exchange of standard roles (active overlap for Journalists and passive overlap for Interviewees) and a greater resistance of journalists towards overlapping speech. The Chairman achieves lower disfluency rates in all conditions.

Figure 3: *Disfluency rates for the different segment types (non-overlapping, overlapping, passive and active type), and the different speaker role.* intr *(resp.* non-intr*) means intrusive (resp. non-intrusive) segments.*



4. Summary

Broadcast TV interviews have been annoted for overlapping speech. The TRANSCRIBER tool has been customized to deal with four overlap types: back-channel, anticipated turn taking (non-intrusive), turn request and complementary (intrusive). High inter-annotator agreement were achieved with 5 annotators. Overlaps are frequent (3-4 overlaps per minute) and of short duration (5% of data), non-intrusive overlaps being shorter than intrusive ones. Disfluency rates on overlaps almost double as compared to non-overlapping speech. Repetitions are the most involved disfluency type, especially for intrusive overlaps (turn requests and complementary). The study highlights interesting differences between active and passive overlap speakers, as well as between journalists and interviewees. More accurate models for both speech recognition and speech interaction may be envisioned in future works.

Our belief is that drawing up a descriptive overlap inventory may contribute to the design of pragmatics models. Gained insights may be of help to improve language modeling for conversational speech and to contribute to automatic conversational speech transcription whose performance is still poor as compared to recognition of read and prepared speech.

5. Acknowledgements

We are indebted to the INA Research & Experimentation Directorate (http://www.ina.fr/) for the *L'Heure de Vérité* corpus. Part of this work was funded by the French INFOM@GIC project.

6. References

[1] Boula de Mareüil, Ph. et al, A quantitative study of disfluencies in French broadcast interviews. *Proceedings of Disfluency In Spontaneous Speech (DiSS) Workshop* Aix-en-Provence. 27–32.September 2005.

[2] Barras, C., et al. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication* 33(1-2), Jan 2001. 5–22.

[3] Blanche-Benveniste, C. Le français parlé, études grammaticales. *Éditions du CNRS, Paris*. 1990.

[4] Carletta, J. Assessing agreement on classification tasks: the Kappa statistic. *Computational Linguistics* 22(2), 249–254. 1996.

[5] Cerrato, L. and D'Imperio, M. Duration and tonal characteristics of short expressions in Italian. *Proceedings of the 15th International Congress of Phonetic Sciences (ICPhS)* Barcelona. 1213–1216. August 2003.

[6] Delmonte, R. Modeling conversational styles in Italian by means of overlaps. *Proceedings of Disfluency In Spontaneous Speech (DiSS) Workshop* Aix-en-Provence. 65–70. September 2005.

[7] Sachs, H., Schegloff, E., Jefferson, G. A simplest systematics for the organization of turn-taking for conversation. *Language* 50, 696–735. 1974.

[8] Schegloff, E., Jefferson, G., Sachs, H. The preference for self-correction in the organization of repair in conversation. *Language* 53, 361–382. 1977.

[9] Shriberg, E., Stolcke, A., Baron, D. Observations on Overlap: Findings and Implications for Automatic Processing of Multi-Party Conversation. *Proceedings of the 7th European Conference of EUROSPEECH* Aalborg. 1359–1362. September 2001.