Parole superposée et genre, étude des annotations pour les médias audiovisuels

Martin Lebourdais, Marie Tahon, Antoine Laurent, Anthony Larcher, Sylvain Meignier

Parole superposée et genre, étude des annotations pour les médias audiovisuels.

Martin Lebourdais Marie Tahon Antoine Laurent Anthony Larcher
Sylvain Meignier
LIUM, Avenue Olivier Messiaen, 72000 Le Mans, France
[Prenom].[Nom]@univ-lemans.fr

RÉSUMÉ

Notre objectif consiste à caractériser les représentations du genre dans les médias français au travers des interactions entre locuteurs en fonction de leur rôle. Cet article propose une étude des annotations en genre et en parole superposée dans plusieurs corpus dédiés à la segmentation en locuteurs et à la transcription, et met en évidence le problème d'hétérogénéité des protocoles d'annotations pour ces deux catégories. De plus, nous analysons l'influence du format des émissions sur la distribution des segments de parole superposée. Sur un sous-corpus de 93 enregistrements de la chaîne LCP, nous cherchons à caractériser les interactions entre locuteurs en fonction de leurs genres. Enfin, nous proposons une méthode qui vise à mettre en valeur les zones de parole fortement interactives. Une telle visualisation améliorerait l'efficacité des études qualitatives menées par les chercheurs en sciences humaines.

ABSTRACT

Overlaps and Gender Annotation Analysis in the Context of Broadcast Media

Our main goal is to characterise gender representations in French broadcast media through speaker interactions and according to their role. This paper proposes a study of gender and overlap annotations in various speech corpora mainly dedicated to diarisation or transcription tasks, and points out the issue of the heterogeneity of the annotation guidelines for both overlapping speech and gender categories. On top of that, we analyse the influence of the speech content (casual speech, meetings, debate, interviews, etc.) on 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 highly interactive speech areas. Such a visualisation tool would improve the efficiency of qualitative studies conducted by researchers in human sciences.

MOTS-CLÉS : Parole, Parole superposée, Corpus, Médias, Genre .

KEYWORDS: Speech, Overlap, Gender, Corpus, Medias.

1 Introduction

L'analyse de la parole conversationnelle permet de caractériser les phénomènes de communication et de préciser les relations entre les participants. De telles études sociologiques sont utiles dans des domaines comme la parole politique, les débats télévisés ou encore la parole journalistique (Beattie, 1982). Lors d'une interview ou d'un débat, les interactions entre les différents participants sont organisées principalement en fonction de leur rôle, par exemple journaliste ou invité. On peut se

demander si la structure des interactions est également influencée par le genre des participants. L'étude des interruptions en fonction du genre est une thématique importante en sciences humaines avec des travaux comme ceux de Zimmerman & West (1975) qui présentent une différence dans la proportion d'interruption des femmes par les hommes. Dans un cadre similaire, le projet Gender Equality Monitoring (GEM) vise à explorer les représentations du genre dans les médias français en utilisant des méthodes automatiques combinées à des études sociologiques plus fines. Plus précisément, nous cherchons à caractériser les interruptions relatives au genre dans les corpus de média audiovisuels en français.

La communication orale entre plusieurs personnes est habituellement caractérisée par une succession de tours de parole. Lorsqu'un participant en interrompt un autre, ceci favorise la parole superposée, définie par la présence d'au moins deux locuteurs parlant simultanément. La parole superposée est un des mécanismes qui structure les changements de tour de parole en fonction des rôles des locuteurs. C'est pourquoi la caractérisation des interruptions, et plus précisément des segments de parole superposée, permet d'analyser les relations entre les locuteurs en fonction de leur genre et rôle. La parole superposée est un phénomène qui apparaît régulièrement dans les conversations, mais reste rare par rapport à sa durée totale en raison d'une durée très courte des segments. Il est donc nécessaire d'avoir accès à de large corpus pour avoir assez de parole superposée afin que ces études soient significatives.

L'utilisation d'outils automatiques capables de segmenter en tour de parole incluant la parole superposée et d'extraire des indicateurs haut-niveau de ces segmentations facilitera et accélérera l'analyse par des chercheurs en sciences humaines, d'un grand nombre de données audio. La plupart des outils de segmentation et de caractérisation sont entraînés de manière supervisée sur des données annotées. Pour cela, il existe de nombreux corpus contenants des centaines d'heures de parole et beaucoup de locuteurs. Néanmoins, bien que les protocoles d'annotations en locuteur sont plutôt homogènes entre les corpus, il n'existe pas de règle commune pour l'annotation de la parole superposée. De plus, pour caractériser automatiquement la parole superposée en fonction du genre, nous avons besoin d'annotations consistantes. Malheureusement, la caractérisation du genre et du rôle n'est pas toujours incluse dans les règles d'annotations.

La section 2 propose une analyse des corpus existants et des pratiques pour la détection de genre, de rôles et de parole superposée. La section 3 présente une analyse de différents protocoles d'annotation de plusieurs corpus utilisables pour la détection de parole superposée, avec une attention particulièrement sur les annotations en genre qui sont cruciales pour le projet GEM. Enfin, la section 4 propose une visualisation chronologique des interactions basée sur les tours de parole qui inclue l'information de présence de parole superposée.

2 Parole superposée et genre : corpus et usages

En informatique, la détection du genre est une tâche de classification qui consiste à évaluer si les caractéristiques acoustiques d'une voix sont plutôt proches de celles d'un homme ou d'une femme, ou éventuellement d'une troisième catégorie non identifiable (enfants, locuteurs non nommés). L'obtention des temps de parole en fonction des genres permet d'automatiser les analyses manuelles des sociologues et ainsi les étendre à des données massives (Doukhan *et al.*, 2018). La détection de genre est également utile en prétraitement des systèmes de traitement automatique de la parole, pour permettre d'affiner les modèles au genre du locuteur et ainsi améliorer les performances. Les premières approches de détection du genre à partir de la voix reposent sur la détection de

caractéristiques acoustiques, telle que la F_0, pour différencier les voix, parfois associées à un modèle probabiliste (Parris & Carey, 1996). Actuellement, les systèmes à base de réseaux de neurones sont les plus performants pour la détection de genre à partir de la voix (Majkowski *et al.*, 2019).

L'étude du genre d'un locuteur est généralement associé à celle du rôle de cette personne. La détection du rôle peut simplifier la navigation dans les enregistrements audio grâce à une indexation automatique (Bigot, 2011). En ne considérant que la modalité audio, des systèmes récents s'appuient sur des caractéristiques acoustiques mais également sur une transcription automatique du contenu linguistique pour détecter automatiquement les journalistes, expert.e.s, chroniqueur.se.s dans les médias audiovisuels (Laurent *et al.*, 2014). Contrairement au rôle, le genre est disponible dans la plupart des corpus étudiés, nous avons donc choisi de nous concentrer sur le genre et de garder le rôle pour de futurs travaux.

À notre connaissance, aucun corpus pour l'étude spécifique de la parole superposée en fonction du genre et du rôle, n'a été collecté et annoté. Les données utilisées dans un système de détection automatique proviennent donc de corpus segmentés en tours de parole et collectés à l'origine pour d'autres tâches comme la segmentation et regroupement en locuteur (SRL) (qui répond à la question "Qui parle quand"), et la transcription. AMI (Mccowan *et al.*, 2005) est un corpus multimodal pour le SRL enregistré lors de réunions. Les campagnes DIHARD (Ryant *et al.*, 2021) portent sur la tâche de SRL avec des données difficiles contenant de la parole superposée et du bruit de fond dans différents qualités et environnements d'enregistrement. Concernant les données en français, les segmentations en tour de parole proviennent de corpus créés pour les tâches de transcription et de SRL, par exemple les corpus radiophoniques et télévisuels REPERE (Giraudel *et al.*, 2012), ETAPE (Gravier *et al.*, 2012), EPAC (Estève *et al.*, 2010), ESTER (Galliano *et al.*, 2006). Le corpus ALLIES (Larcher *et al.*, 2021) regroupe les corpus audiovisuels précédents, tous extraits des archives de l'Institut National de l'Audiovisuel (INA), en incluant des nouvelles données plus récentes.

Les systèmes classiques de SRL ignorent la présence de la parole superposée ou la considère comme négligeable en terme d'erreur, malgré l'impact négatif qu'elle peut avoir sur l'apprentissage des modèles de locuteurs (Huijbregts & Wooters, 2007). Cet impact est réduit en traitant cette parole superposée (Bullock *et al.*, 2020). De même, en transcription automatique de la parole, la majorité des approches considèrent que les segments d'entrée sont mono-locuteur et négligent aussi la parole superposée, malgré les erreurs de transcription générées dans ces zones et sur les segments contigus (Çetin & Shriberg, 2006).

Les segments de parole contenant de l'overlap sont intéressants pour caractériser les interruptions dans la conversation, avec des catégories telles que le backchannel (courte interjection montrant l'écoute), l'ajout d'informations complémentaires, une interruption pour prendre la parole ainsi que des départs qui anticipent la fin du tour de parole du locuteur précédent (Adda-Decker *et al.*, 2008). Cela peut également servir à l'étude des mécanismes de correction et des disfluences causées par de telles zones (Sacks *et al.*, 1974).

La plupart des systèmes de détection actuels utilisent des architectures neuronales de type *sequence-to-sequence* comme des réseaux récurrents incluant des LSTMs (Long Short-Term Memory) (Bredin *et al.*, 2020), ou les récents TCNs (Temporal Convolutional Network) qui permettent d'avoir un contexte temporel large (Cornell *et al.*, 2020). Comme mentionné dans l'introduction, les zones contenant de la parole superposée apparaissent régulièrement dans les conversations mais sont très courtes, et donc peu représentatives en terme de durée, ce qui induit un déséquilibre des corpus. La solution utilisée pour compenser ce déséquilibre est de rajouter de la parole superposée artificielle (Bullock *et al.*, 2020).

Corpus	Genre annoté		Genre non renseigné	Durée Totale
	Femmes	Hommes		
ESTER1	30.3%	69.7%	0.1%	99h
ESTER2	25.1%	74.9%	29.8%	161h
EPAC	18.21%	81.8%	1.1%	105h
ETAPE	18.5%	81.5%	35.2%	34h
REPERE	20.0%	80.0%	0.1%	58h
ALLIES-LCP-DEBATE	13.67%	85.43%	0.9%	48h

TABLE 1 – Proportion de temps de parole femme/homme, ainsi que la proportion de parole non annotée, pour les différents corpus en français.

3 Analyse des protocoles d'annotation

3.1 Annotations en genre

Le Tableau 1 présente la proportion du temps de parole en fonction du genre dans les corpus de médias en français. ALLIES-LCP-DEBATE est un sous-ensemble du corpus ALLIES contenant 3 émissions de débat de la chaîne de télévision LCP enregistrées entre 2010 et 2014 (*Ça vous regarde, Entre les lignes, Pile et face*). Ce sous-corpus représente 93 émissions au total. Ces résultats montrent que la proportion de femmes et d'hommes dans les corpus étudiés n'est pas équilibrée. Cette observation est commune à plusieurs études (Doukhan *et al.*, 2018) et induit un biais de donnée pour tout système entraîné sur celles-ci (Garnerin *et al.*, 2019). Ce phénomène est amplifié par le fait que ces corpus précèdent les réglementations françaises sur l'égalité entre les femmes et les hommes [1].

On notera qu'un tiers des données des corpus ESTER2 et ETAPE ne disposent pas d'information sur le genre. Dans le cas des corpus DIHARD et AMI, le genre des locuteurs n'est pas indiqué. Seule la campagne ESTER 1 a proposé une tâche d'identification du genre. Les autres corpus n'ont pas été créés avec l'objectif de travailler sur le genre, expliquant en partie l'absence d'annotation.

3.2 Parole superposée

La communauté du traitement automatique de la parole partage des règles communes d'annotation, décrites dans les guides d'annotation [2]. Cependant, ces règles diffèrent en fonction des tâches pour lesquelles les corpus ont été construits. Il est donc difficile d'utiliser les annotations pour des tâches différentes de celles d'origine sans en avoir une connaissance approfondie.

L'annotation des corpus de SRL DIHARD, AMI et du sous-corpus ALLIES-LCP-DEBATE consiste en une segmentation comportant pour chaque segment de tour de parole le temps de début et de fin, le nom du locuteur, le nom de l'émission d'où il extrait. Parfois, d'autres informations associées au locuteur comme son genre et son rôle viennent compléter les segments. L'annotation de la parole superposée n'est pas explicite. Les segments de parole superposée sont générés en calculant l'intersection entre les tours de parole.

1. Loi du 4 août 2014 pour l'égalité réelle entre les femmes et les hommes, loi du 27 janvier 2011 relative à la représentation équilibrée des femmes et des hommes au sein des conseils d'administration et de surveillance.
2. Guide d'annotation de Transcriber : `http://trans.sourceforge.net/en/transguidFR.php`.

Corpus	Total	Parole superposée	Langue	Période d'enregistrement
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
ALLIES-LCP-DEBATE	48h	9.85%	fr	2011-2014
AMI	96h	13.87%	en	NA
DIHARD	34h	11.6%	en	NA

TABLE 2 – Durée totale des corpus et proportion de parole superposée par rapport à la durée totale.

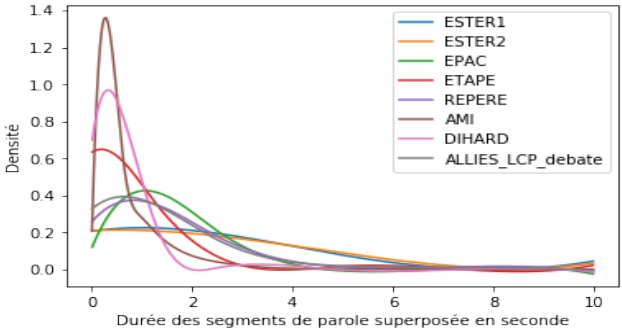

FIGURE 1 – Distribution normalisée des durées des segments de parole en superposée en secondes.

Dans les corpus français de transcription ESTER 1 et 2, EPAC, ETAPE et REPERE, chaque tour de parole est associé à un ou plusieurs noms de locuteurs ainsi que des informations sur le genre, l'accent et la qualité du canal. Dans le cas d'un tour de parole multi-locuteurs, les mots prononcés sont aussi balisés pour retrouver le début et la fin des propos de chaque locuteur. Dans les corpus ESTER 1 et 2, ainsi que le corpus EPAC, lorsque deux locuteurs parlent simultanément de manière intelligible, les transcriptions des deux locuteurs sont souvent présentes bien que le guide d'annotation ne demande que le locuteur principal. Dans le corpus ETAPE, les indications temporelles de la parole superposée ne sont pas indiquées, par contre le type de superposition défini par (Adda-Decker *et al.*, 2008) est annoté. Dans le corpus REPERE, lorsque deux locuteurs parlent en même temps, seuls les mots du locuteur principal sont transcrits ce qui ne permet pas de segmenter l'overlap.

D'après le Tableau 2, nous constatons que les corpus ESTER 1 et 2 sont les moins riches en parole superposée. Ils sont essentiellement composés de journaux radiophoniques avec des présentateurs, des chroniqueur.se.s, des expert.e.s et des hommes ou femmes politiques. Le corpus contient peu d'interviews et aucun débat. À l'opposé, les corpus REPERE, ETAPE, EPAC et ALLIES-LCP-DEBATE contiennent des émissions radiophoniques et télévisées de débat où la proportion de parole superposée est plus importante.

AMI et DIHARD sont des corpus de parole conversationnelle incluant de la parole spontanée. La proportion de parole superposée est supérieure aux autres corpus. AMI est un corpus d'enregistrement de réunion. DIHARD comporte beaucoup de données fortement interactives avec peu de professionnels de la communication, par exemple des données récoltées dans un restaurant. Bien qu'utiles pour

L_1-L_2	Total (a+b+c)	Interaction sans superposition (a)	Interaction avec superposition	
			sans changement de locuteur (b)	avec changement de locuteur (c)
M-M	12008	7588(63.19%)	2212(18.42%)	2208(18.39%)
M-F	1359	719(52.69%)	269(19.79%)	371(27.30%)
F-M	1357	641(47.24%)	405(29.85%)	311(22.92%)
F-F	467	389(83.30%)	47(10.06%)	31(6.64%)

TABLE 3 – Nombre d'interactions total, avec/sans parole superposée, avec/sans changement de locuteur en fonction du genre du premier (L1) et du second (L2) locuteur pour le sous corpus ALLIES-LCP-DEBATE .

FIGURE 2 – Type d'interactions prises en compte.

développer un système de détection de parole superposée, les caractéristiques des corpus AMI et DIHARD restent éloignées de notre sujet d'étude.

La Figure 1 montre la répartition des durées des segments de parole superposée pour les différents corpus cités précédemment. Plus un corpus contient de parole spontanée, plus les segments de parole superposée sont courts. Alors que AMI et DIHARD contiennent les segments les plus courts (< 1 seconde), les corpus REPERE, ETAPE et EPAC contiennent majoritairement des segments entre 1 à 2 secondes. Les corpus ESTER 1 et 2, quant à eux, ont des durées relativement uniformes. Peu de segments courts ont été annotés et ils sont peu présents de par la nature des corpus.

3.3 Interactions et genre

Dans les corpus ESTER 1 et 2, ETAPE, EPAC et REPERE, la nature des enregistrements explique en partie les statistiques mesurées sur la parole superposée. Mais le manque d'information sur le genre des locuteurs est plus difficilement explicable au regard des guides d'annotation qui prévoient cette information. Les différences d'annotation rendent difficiles les comparaisons entre corpus et une étude globale est impossible sans garder la provenance de l'émission. Dans la suite, nous proposons de nous concentrer sur le sous-corpus ALLIES-LCP-DEBATE présentant une proportion supérieure à 8% de parole superposée, comme indiqué dans le Tableau 2. À partir de la segmentation de référence d'une émission donnée, le nombre d'interactions entre deux locuteurs est compté comme indiqué dans la Figure 2. Une interaction correspond à un tour de parole d'un locuteur L_1 qui est suivi soit d'un nouveau tour de parole correspondant à un second locuteur L_2, soit d'une zone de parole superposée entre L_1 et L_2. Le Tableau 3, résume le nombre total d'interactions sans parole superposée (a), les interactions avec parole superposée et sans changement de locuteurs (b), et les interactions avec parole superposée et changement de locuteur correspondant à l'interruption de L_1 par L_2 (c). Les proportions calculées par émission sont consistantes avec celles obtenues sur les trois émissions (résultats non reportés dans cette article).

Premièrement, nous observons que la proportion de changements de locuteur avec parole superposée (c) est équilibrée, que ce soit entre une femme et un homme où inversement. Il y a tout de même une légère différence : les hommes ont tendance à employer de la parole superposée plus fréquemment

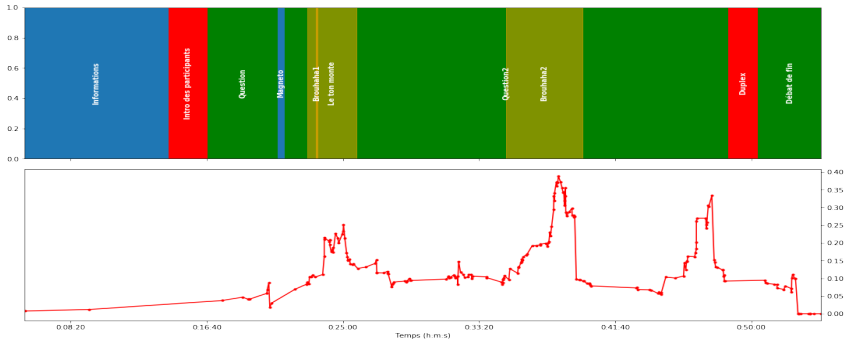

FIGURE 3 – Évolution chronologique de la dérivée de la durée cumulée des segments de parole superposée Δ (bas) et description manuelle des phases de l'émission *Ça vous regarde* du 29/04/2014 (haut).

(52.76 pour F-H contre 47.31 pour H-F). Ces proportions sont supérieures à celles mesurées avec L_1 et L_2 du même genre. Deuxièmement, le nombre d'interactions avec parole superposée entre deux femmes est assez faible (16.70%). Après écoute de ces interactions, nous constatons quelles correspondent plus à des *backchannels* ou des ajouts de compléments d'information (10.06%) qu'à des interruptions (6.64%). Par contre, on peut remarquer qu'il y a plus souvent une femme interrompant un homme avec superposition (27.30%) que l'inverse (22.92%). Ce premier résultat a été obtenu sur un sous-ensemble de ALLIES, une étude sur un plus large panel d'émissions et d'années est nécessaire pour le confirmer.

4 Visualisation des zones fortement interactives

L'étude des interactions nécessite souvent une écoute de celles-ci afin de les caractériser ou d'effectuer une analyse fine. Dans cette partie, nous présentons une méthode d'aide à l'analyse d'émissions radiophoniques qui permet de visualiser les zones de forte interactivité.

À partir des segments contenant de la parole superposée, nous calculons la durée cumulée des segments de parole superposée, notée d_{cum} en fonction de leur temps de début, noté t_{deb} pour chaque émission. Nous obtenons N couples (d_{cum}, t_{deb}) qui suivent l'ordre chronologique de l'émission. Nous calculons ensuite la dérivée Δ pour chaque segment n, grâce à l'équation 1 où N est le nombre total de segments de parole superposée.

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

La valeur de $h = 9$ a été choisie de manière empirique afin de réaliser un lissage de Δ sur une large fenêtre temporelle. Nous traçons ensuite la courbe correspondant à Δ en fonction de t_{deb}, comme le présente la Figure 3 pour l'émission *Ca vous regarde* en date du 29/04/2014.

La durée cumulée étant une fonction strictement croissante, on a donc $\Delta > 0$. $\Delta \geq 1$ impliquerait que la durée du segment de parole superposée est supérieure ou égale au temps entre ce segment et le suivant, ce qui est impossible sauf erreur d'annotation, donc $\Delta < 1$. Plus précisément, une forte valeur de Δ indique la présence d'une succession rapide de zones de parole superposée, ou bien de

longs segments de parole superposée. Ainsi, les valeurs élevées de Δ sont caractéristiques des zones de forte interactivité.

Pour évaluer l'utilité de notre courbe, nous avons analysé manuellement l'émission *Ca vous regarde* du 29/04/2014. Nous avons noté différentes phases de l'émission, représentées sur la Figure 3 par la frise chronologique. La courbe de Δ associée aux phases de l'émission, montre que les zones de brouhaha annotées après écoute (vert clair) sont corrélées avec un pic de la dérivée.

Cette étude préliminaire a montré son intérêt, elle devra être confirmée sur plusieurs points. Cette visualisation a été réalisée avec une segmentation issue de la référence manuelle utilisée, elle devra être remplacée par une segmentation en parole superposée automatiquement. Quel sera l'impact des erreurs de détection ? L'étude des différences entre ces courbes devra être confirmée sur d'autres émissions, avec différent.e.s invité.e.s, pour des périodes temporelles plus larges incluant des enregistrements antérieurs aux lois favorisant l'égalité entre les femmes et les hommes. Une validation par les partenaires de sciences humaines devra confirmer que les zones de parole superposées ciblées sont les zones pertinentes à l'étude des interruptions.

5 Conclusion

Dans le cadre du projet GEM, nous étudions les interactions entre les locuteurs au travers du spectre du rôle et du genre dans les médias français. Cet article propose une analyse complète des annotations en genre et en parole superposée dans plusieurs corpus, mettant ainsi en lumière l'hétérogénéité des protocoles et des traitements effectués. Grâce à cette étude statistique, nous avons pu vérifier que le contenu de discours (parole spontanée, débat, interview, etc.) a un impact important sur la distribution de la longueur des segments de parole superposée. Nous avons également mis en évidence le fait que la parole superposée peut-être considérée comme un phénomène rare. L'analyse complète des catégories d'interactions en fonction du genre dans plusieurs émissions de débats, permet de conclure que les interactions impliquant deux genres différents sont plus susceptibles d'inclure de la parole superposée, indiquant ainsi un effet potentiel de genre. Nous avons également trouvé qu'il est plus commun pour une femme d'interrompre un homme avec de la parole superposée que l'inverse. Compte tenu de nos observations sur le lien entre parole superposée et genre, il serait intéressant de poursuivre des études similaires sur une plus grande échelle pour confirmer ou infirmer nos observations. De telles études seront possibles grâce à la détection automatique de parole superposée. Le principal frein à cette méthode est la rareté de la parole superposée dans les corpus existants, ainsi que la difficulté de fusionner les corpus actuels, trop hétérogènes. Pour contourner ce problème, il est nécessaire de poursuivre les accords et collaborations entre les différents acteurs du domaine. Bien que les conventions d'annotations soient souvent très similaires, les pratiques et usages sont trop hétérogènes. Des initiatives comme le corpus ALLIES font avancer la communauté en fournissant une large source de données harmonisées. En seconde partie, nous avons proposé une méthode capable de visualiser les zones fortement interactives Ce type de représentation sera utilisé pour accélérer les annotations et la caractérisation de fichier audio, en indiquant les zones d'intérêt pour une analyse humaine. Des informations additionnelles telles que la proportion de femmes seront ajoutées sur cette courbe dans des travaux futurs.

Remerciements

Notre recherche est financée par le projet ANR GEM, Gender Equality Monitoring (ANR-19-CE38-0012).

Références

ADDA-DECKER M. *et al.* (2008). Annotation and analysis of overlapping speech in political interviews. In *LREC*, p. 3105–3111, Marrakech, Morocco.

BEATTIE G. (1982). Turn-taking and interruption in political interviews. *Semiotica*, **39**, 93–114.

BIGOT B. (2011). *Recherche du rôle des intervenants et de leurs interactions pour la structuration de documents audiovisuels*. Theses, Université Paul Sabatier - Toulouse III.

BREDIN H. *et al.* (2020). Pyannote.Audio. In *ICASSP*, p. 7124–7128, Barcelona, Spain.

BULLOCK L. *et al.* (2020). Overlap-Aware Diarization. In *ICASSP*, p. 7114–7118, Barcelona, Spain.

CORNELL S. *et al.* (2020). Detecting and Counting Overlapping Speakers in Distant Speech Scenarios. In *Interspeech*, p. 3107–3111, Shanghai, China.

DOUKHAN D. *et al.* (2018). Describing Gender Equality in French Audiovisual Streams with a Deep Learning Approach. *VIEW Journal of European Television History and Culture*, **7**(14), 103–122.

ESTÈVE Y. *et al.* (2010). The EPAC corpus. In *LREC*, p. 1686–1689, Valetta, Malta.

GALLIANO S. *et al.* (2006). Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news. In *LREC*, p. 139–142, Genoa, Italy.

GARNERIN M. *et al.* (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, p. 3–9, New York, NY, USA.

GIRAUDEL A. *et al.* (2012). The REPERE Corpus. In *LREC*, p. 1102–1107, Istanbul, Turkey.

GRAVIER G. *et al.* (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *LREC*, p. 114–118, Istanbul, Turkey.

HUIJBREGTS M. & WOOTERS C. (2007). The blame game. In *Interspeech*, p. 1857–1860, Antwerp,Belgium.

LARCHER A. *et al.* (2021). Speaker Embedding For Diarization Of Broadcast Data In The ALLIES Challenge. In *ICASSP*, p. 5799–5803, Toronto, Canada.

LAURENT A. *et al.* (2014). Boosting bonsai trees for efficient features combination : application to speaker role identification. In *Interspeech*, p. 76–80, Singapore.

MAJKOWSKI A. *et al.* (2019). Identification of Gender Based on Speech Signal. In *Conf. on Computational Problems of Electrical Engineering (CPEE)*, p. 1–4, Lviv-Slavske, Ukraine.

MCCOWAN I. *et al.* (2005). The AMI meeting corpus. In *Conf. on Methods and Techniques in Behavioral Research*, p.$\tilde{4}$, Wageningen, Netherlands.

PARRIS E. & CAREY M. (1996). Language independent gender identification. In *ICASSP*, volume 2, p. 685–688, Atlanta, GA, USA.

RYANT N. *et al.* (2021). The third dihard diarization challenge. In *Interspeech*, p. 3570–3574, Brno, Czechia.

SACKS H. *et al.* (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, **50**(4), 696–735.

ZIMMERMAN D. H. & WEST C. (1975). Sex roles, interruptions and silences in conversation. *Language and Sex*, p. 105–129.

ÇETIN O. & SHRIBERG E. (2006). Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site. In *Interspeech*, p. paper 1915–Mon2A2O.6, Pittsburgh, USA.