Typologie de la parole spontanée à des fins d'analyse linguistique et de développement de systèmes de reconnaissance automatique de la parole

Solène Evain, Solange Rossato, François Portet, Benjamin Lecouteux

Typologie de la parole spontanée à des fins d'analyse linguistique et de développement de systèmes de reconnaissance automatique de la parole

Solène Evain Solange Rossato Benjamin Lecouteux François Portet
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, 38000 Grenoble, France
prenom.nom@univ-grenoble-alpes.fr

RÉSUMÉ

Les systèmes de Reconnaissance Automatique de la Parole (RAP) ont montré ces dernières années des performances toujours plus impressionnantes. Néanmoins, la RAP spontanée reste un problème ouvert : la littérature a pu montrer qu'il est difficile de la définir (pas de consensus) et combien elle est difficile à modéliser. Dans notre travail, nous revenons sur la notion même de parole spontanée avant de nous appuyer sur les dénominations et définitions trouvées pour proposer une typologie de la parole spontanée sur quatre axes (contexte situationnel, type et canal de communication, degré d'intimité entre les locuteurs). Cette catégorisation offre la possibilité de rassembler des situations propices à une parole plus ou moins spontanée. Les objectifs sont multiples : créer une nouvelle typologie des corpus oraux de parole spontanée pour faciliter son analyse linguistique et améliorer les performances de systèmes de RAP sur ce type de parole.

ABSTRACT

Typology of spontaneous speech for its linguistic analysis and the development of automatic speech recognition systems

The performance of Automatic Speech Recognition (ASR) systems increased significantly in recent years. However spontaneous speech reccognition is still an open problem : scientific publications have shown that it is difficult to define it and how hard it is to model it. In our paper, we come back to the concept of spontaneous speech, before using its denominations and definitions to create a typology of spontaneous speech with four axis (situational context, type and mode of communication, degree of intimacy between speakers). This categorization offers the possibility of gathering situations conductive to speaking more or less spontaneously. There are multiple objectives : create a new typology of spontaneous speech allowing a new categorization of spontaneous corpora to facilitate the analysis of this type of speech and improve the performance of ASR systems.

MOTS-CLÉS : corpus oraux, reconnaissance automatique de la parole, parole spontanée.

KEYWORDS: Speech corpora, automatic speech recognition, spontaneous speech.

1 Introduction

Les avancées récentes en reconnaissance automatique de la parole (RAP) ont permis d'obtenir des systèmes de plus en plus performants (Malik *et al.*, 2021; Szymański *et al.*, 2020) sur la parole lue (Mekki, 2020) ou préparée (Desnous *et al.*, 2018). Cependant, lorsque la spontanéité de la parole

10.21437/JEP.2022-23

augmente, les performances ont tendance à baisser (voir tableau 1).

Auteur, date	Corpus de test	ML	Syst.	Détail	Résultats
(Jousse *et al.*, 2008)	11 fichiers de journaux d'information de radios	oui	HMM-GMM*	Nivx 1-3 : parole prep., 4-6 : spont. faible, 7-10 : spont. fort	1-3 : 23 à 26% / 4-6 : 24 à 30% / 7+ : 28 à 44%
(Dufour *et al.*, 2010)	11 fichiers de journaux d'information de radios	oui	HMM-GMM*	Nv 1 : parole prep., 2-4 : spont. faible, 5-8 : spont. fort	1 : 10,1% / 2-4 : 18,4% / 5-8 : 28,5%
(Garnerin, 2018)	Emissions de Tv et radio (REPERE, ETAPE, ESTER1,2)	oui	HMM-DNN	regroupement des émissions spontanée	62.11%
(Evain *et al.*, 2021)	ETAPE test	oui	HMM-DNN	utilisation d'un modèle pré-appris + fine-tuning	25.22%
(Evain *et al.*, 2021)	ETAPE test	non	E2E	utilisation d'un modèle pré-appris + fine-tuning	26.14%

*Même système, voir (Deléglise *et al.*, 2005)

TABLE 1 – Quelques résultats de reconnaissance de parole spontanée en français

La reconnaissance de la parole spontanée (RAPS) reste donc l'un des défis actuels du domaine. De nombreuses pistes ont été explorées pour tenter de comprendre au mieux les points de blocage de la RAPS. Dufour (2008) a montré le besoin en corpus écrits de parole spontanée. En effet, dans le cas de systèmes HMM-GMM ou HMM-DNN, un poids important est accordé au modèle de langue (ML). Il est donc important d'y intégrer des données textuelles représentant au mieux ce type de parole, la parole spontanée étant différente de la parole préparée avec un ordre de mots souvent bousculé et un nombre plus important de disfluences (hésitations, reprises, amorces de mots ou répétitions) (Adda-Decker *et al.*, 2004), ce qui complique la transcription automatique. Comme le montre Dufour (2008), il est également nécessaire d'intégrer des données qui ressemblent aux données à décoder par la suite. Les auteurs indiquent également que l'intégration de variantes de prononciation au lexique (pour modéliser les métaplasmes et élisions plus nombreux en parole spontanée) semble aller en faveur d'une meilleure reconnaissance de la parole spontanée. Enfin, ils ont pu montrer une corrélation entre des degrés de spontanéité et le *WER* : plus la parole est spontanée plus les performances se dégradent (voir tableau 1). Les trois corpus de référence pour la RAPS en français étant (voir tableau 2) ETAPE (Gravier *et al.*, 2012), REPERE (Giraudel *et al.*, 2012), et EPAC (Estève *et al.*, 2010), la quantité de données de parole spontanée manuellement annotée est assez limitée (160 heures au total).

Corpus	Durée	Transcrit	Licence	Type de parole
BRAF100	30 h	oui	indisponible	lecture
BREF80 (ELRA-S0006)	≈80 h	oui	ELRA, €	lecture (dictation)
BREF120 (ELRA-S0067)	100 h	oui	ELRA, €	lecture (dictation)
ESTER1 (ELRA-E0021)	1 700 h	≈100 h	ELRA, €	préparée
EPAC (ELRA-S0305)	ESTER1	100 h man., 1 700 h auto.	ELRA, €	spontanée
ESTER2 (ELRA-S0338)	ESTER1 transcrit + EPAC transcrit + 150 h	100 h trans. riche, 50 h trans. rapide	ELRA, €	préparée + spontanée
ETAPE (ELRA-E0046)	30 h	oui	ELRA, €**	spontanée + préparée
REPERE (ELRA-E0046)	30 h	oui	ELRA, €**	préparée + spontanée

€ : corpus payant ; **gratuit pour recherche académique sans usage commercial

TABLE 2 – Corpus usuels en français, créés spécifiquement pour une tâche d'ASR

La RAPS en français s'est concentrée jusqu'alors sur la parole spontanée dans un contexte spécifique (radio et TV) où la parole est maniée principalement par des professionnels. Il y a donc un réel besoin de corpus de test plus diversifiés afin de permettre une évaluation plus précise des performances des

systèmes sur ce type de parole. Nos objectifs de recherche sont donc multiples et étroitement liés : (1) Qu'est-ce que la parole spontanée ? (2) Comment étiqueter les corpus de parole spontanée afin de pouvoir créer des sous-ensembles cohérents ? (3) Comment sélectionner et utiliser les sous-ensembles créés pour l'entraînement et l'évaluation de modèles de RAPS ? Afin de pouvoir y répondre, nous commençons par revenir sur la notion de parole spontanée, sur ses dénominations et sur ce qui la caractérise (section 2). Nous proposons ensuite une typologie de la parole spontanée (section 3), basée sur les recherches précédentes et sur les corpus de parole spontanée disponibles en linguistique. Enfin, nous discutons des nouvelles perspectives qui s'ouvrent pour la RAPS (section 4).

2 Les multiples dénominations de la parole spontanée

Dans la littérature, les travaux s'intéressant à la parole spontanée, que cela soit d'un point de vue linguistique ou RAP, ont utilisé plusieurs dénominations qui témoignent de la complexité de ce type de parole, parfois vue comme simple élément d'une dichotomie parole préparée - parole non préparée. Cette diversité est représentée dans la figure 1, qui regroupe différentes dénominations en quatre ensembles distincts. Nous analysons chacun des termes avant de discuter de l'existence d'un continuum parole préparée - parole spontanée et de degrés de spontanéité.

2.1 Dénominations et perceptions de la parole spontanée

Une parole non préparée ?

Certaines dénominations semblent indiquer qu'il existe une frontière nette entre parole spontanée et parole préparée. La parole préparée serait toute parole qui aurait été préalablement écrite sur support écrit, ou qui aurait été répétée, tandis que la parole spontanée serait celle qui serait énoncée sans aucune préparation du locuteur. Néanmoins, pour (Dufour *et al.*, 2010), la limite entre ces deux types de parole ne semble pas être aussi franche que ce que le terme utilisé laisse entendre. Nous y reviendrons en section 3.

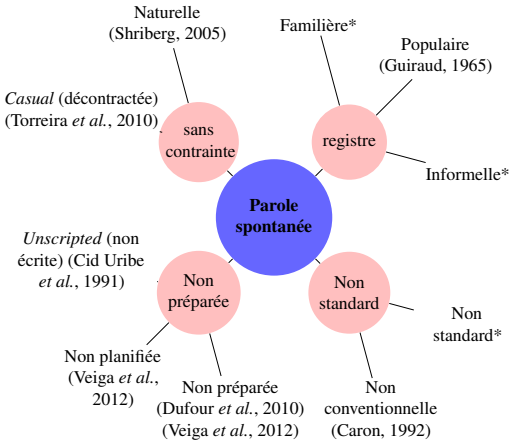

FIGURE 1 – Différentes dénominations de la parole spontanée
*Dénominations rapportées dans (Blanche-Benveniste *et al.*, 1999), sans qu'il soit fait référence aux auteurs

Une parole non standard ?

Les qualificatifs “non standard” ou “non conventionnelle” de la parole spontanée ramènent à une vision de la linguistique à forte tradition d’étude de l’écrit, qui considérait la parole comme “quelque chose de ‘naturel’, qui s’opposerait aux aspects ‘culturels’ des règles de la langue écrite apprises à l’école” (Blanche-Benveniste *et al.*, 1999). Certains chercheurs utilisaient le terme de “parlé parlé”, afin de distinguer cette parole considérée comme agrammaticale, du “parlé écrit”, contenant des constructions plus proches de celles de l’écrit et donc conformes aux règles. Ces dénominations ne sont plus utilisées aujourd’hui.

Un registre ?

Certains auteurs comme (Guiraud, 1965) semblent considérer la parole spontanée comme un registre. *Familier* se définit par ce "dont on use dans l’intimité, dans la conversation courante". La parole spontanée serait donc pour cet auteur une parole utilisée dans l’intimité. Quant aux qualificatifs *populaire* ("qui est propre aux couches les plus modestes de la société, au peuple et qui est inusité par les gens cultivés et la bourgeoisie") et *informel* ("dégagé de tout formalisme, de toute structuration ou institution" [1]), ils se rapportent à une vision de la parole spontanée réfutée, elle aussi, depuis plus de vingt ans (Blanche-Benveniste *et al.*, 1987), (Blanche-Benveniste *et al.*, 1999).

Une parole non contrainte ?

Enfin, pour d’autres auteurs (Torreira *et al.*, 2010), (Shriberg, 2005), la parole spontanée est une parole détendue ou relâchée, qui survient de façon naturelle. Bien que rangées dans la même catégorie, ces deux notions désignent différents aspects de la parole. Le relâchement peut se référer à l’aspect articulatoire. Par exemple, Torreira *et al.* (2010) ont construit le corpus NCCFr à des fins d’analyse phonétique. La notion de parole ‘naturelle’ est reliée quant à elle à l’aspect linguistique/syntaxique de cette parole, dans le cadre de son traitement automatique. Il y est question d’une parole où la phrase n’existe plus, où les disfluences sont omniprésentes et où les chevauchements sont nombreux.

Notre travail de récolte de différentes dénominations de la parole spontanée montre à quel point celle-ci est difficile à définir. On peut remarquer également que les dénominations sont dépendantes du sous-domaine de la linguistique dans lequel elles sont utilisées (*‘casual’* pour la phonétique, *’naturelle’* ou *’non préparée’* pour le TAL), mais aussi parfois en fonction du regard que l’on porte sur la parole spontanée (‘non conventionnelle’ dans une opposition écrit/oral). Si l’on devait définir la parole spontanée à partir de ces termes, ce serait donc une parole considérée à la fois comme non planifiée (avec une limite plus ou moins franche avec la parole préparée), apparaissant particulièrement dans l’intimité (et donc dans des contextes spécifiques), dérogeant aux règles strictes de l’écrit (pas de phrases bien définies, des disfluences pour corriger ou modifier le discours...), exprimée de façon relâchée, et où les chevauchements peuvent être nombreux.

2.2 Continuum et degrés de spontanéité

D’après les travaux de la littérature, il ne semble pas exister de frontière nette entre parole préparée et non-préparée. Ainsi Dufour *et al.* (2010) mesure trois degrés de spontanéité (parole préparée, spontané faible, spontané fort). Ceci rejoint les travaux de Labov (1972), Beckman (1997) et Fujisaki (1997), qui proposaient l’existence d’un continuum entre différents types de parole, et mesuraient

1. Les définitions proviennent du CNRTL https ://www.cnrtl.fr/

les degrés de spontanéité de *'careful'*, à *'casual'* [2] pour l'un, selon le contexte dans lequel la parole est produite pour l'autre et enfin selon le niveau de préparation de la parole pour le dernier. Les travaux de Torreira *et al.* (2010) et Veiga *et al.* (2012) nuancent cette idée de continuum en séparant parole préparée et parole spontanée, tout en conservant l'idée de degrés de spontanéité, définis selon la relation plus ou moins familière entre les locuteurs ou encore le nombre d'hésitations (pauses remplies et allongement de la voyelle).

Qu'il y ait consensus ou non sur l'existence d'un continuum entre parole préparée et parole spontanée, on peut remarquer que l'ensemble de ces travaux mentionnent l'existence de divers degrés de spontanéité, définis selon le contexte (Labov, 1972; Beckman, 1997), la relation plus ou moins intime entre les locuteurs (Torreira *et al.*, 2010), le nombre d'hésitations (Veiga *et al.*, 2012), ou encore différents paramètres prosodiques (durée des voyelles ou des fins de mots, vitesse d'élocution, hauteur de la voix) et linguistiques (pauses remplies, répétitions, faux départs) (Dufour *et al.*, 2010). Les multiples dénominations et les degrés de spontanéité rapportés dans cette section nous servirons de base de réflexion pour une typologie de la parole spontanée développée ci-dessous.

3 Proposition d'une typologie de la parole spontanée

3.1 Contexte situationnel et relation entre les locuteurs

Il existe plusieurs corpus de parole spontanée en français (hors RAP) : le PFC (Phonologie du Français Contemporain) (Laks *et al.*, 2009), CLAPI (Corpus de LAngue Parlée en Interaction) (Baldauf-Quilliatre *et al.*, 2016), CEFC (Corpus d'Etude pour le Français Contemporain) (Benzitoun *et al.*, 2016), Rhapsodie (corpus prosodique de référence en français parlé) (Lacheret *et al.*, 2014) et ESLO (Enquêtes Socio-Linguistiques à Orléans) (Eshkol-Taravella *et al.*, 2011). Il n'existe cependant pas à ce jour de moyen pour visualiser les rapprochements possibles entre ces différents corpus. Il ne serait pas intéressant d'essayer de les catégoriser en fonction de paramètres linguistiques ou prosodiques et ce pour plusieurs raisons : d'une part parce que ces paramètres ne sont pas réservés à la parole spontanée, d'autre part parce qu'ils ne permettent pas de justifier réellement de la diversité des données existantes. A l'inverse, leur catégorisation selon le contexte situationnel ou le niveau d'intimité entre les locuteurs est plus informatif. Le corpus "apéritif entre amis" issus de CLAPI en est un parfait exemple. Il y a dans ce simple titre à la fois le contexte - un apéritif - et la relation entre les locuteurs - des amis. De même pour un enregistrement du corpus ESLO2 (ESLO2_REPAS_1261) pour lequel les métadonnées nous apprennent que c'est un dîner (contexte) entre amis et voisins de palier (relation(s) entre les locuteurs). Ces simples informations permettent de se projeter sur l'éventualité de l'apparition d'une parole spontanée et sur le degré de spontanéité auquel on peut d'attendre, et permettent de rassembler des enregistrements issus de corpus différents.

3.2 Canal et type de communication

Le contexte situationnel et le degré d'intimité de la relation entre les locuteurs sont deux paramètres dont dépendent des degrés de spontanéité : la parole spontanée se définie donc de manière multidimensionnelle. Néanmoins, ces deux dimensions seules ont leur limite. La pandémie nous aura prouvé

2. Cet auteur fait une différence entre 'casual speech' et 'spontaneous speech'. Le première apparaîtrait en contextes informels alors que le second serait réservé aux contextes formels

qu'il est dorénavant possible de se retrouver dans un contexte d'apéritif entre amis via des outils de visioconférence. Nous pouvons donc nous interroger sur la comparaison entre cette situation et celles vues précédemment. En effet, Ruhleder & Jordan (2001), ten Bosch *et al.* (2004) et Galiano *et al.* (2008) ont pu montrer que la communication était différente lorsque ces outils sont utilisés. Les problèmes de réseau/son, les décalages temporels image/parole ou encore la parole inaudible semblent contraindre la spontanéité. Le canal de communication utilisé serait donc une dimension de plus de la parole spontanée. Nous avons jusqu'ici parlé d'une parole qui se restreint au contexte privé. Les corpus de parole spontanée créés pour la RAPS (corpus d'émission de radio et de TV) et les travaux de (Toledano *et al.*, 2005) ne sont pas sans rappeler que ce type de parole peut aussi être publique et qu'il existe bien une différence entre les deux. Ceci nous amène à considérer une quatrième dimension à la parole spontanée : le type de communication (public/privé). Au vue de l'analyse de la littérature et de l'exploration des corpus de parole spontanée en français, il semblerait donc qu'il y ait au moins quatre axes régissant des degrés de spontanéité.

3.3 Typologie de la parole spontanée selon quatre axes

Nous proposons de représenter les quatre dimensions présentées précédemment comme sur la figure 2. **Le contexte situationnel** (S) se référait jusqu'ici au lieu (radio, maison par exemple). Nous proposons

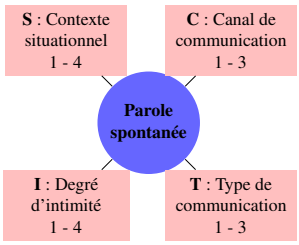

FIGURE 2 – Quatre axes impactant la spontanéité de la parole

d'y intégrer le rôle des locuteurs, étroitement lié au lieu dans lequel la parole est produite. L'échelon 1 correspond à un lieu neutre et une conversation libre où tout le monde est sur un même plan. L'échelon 2 correspond à la situation dans laquelle le lieu *ou* le rôle influence le discours. L'échelon 3 est la situation où le lieu *et* le rôle influencent le discours. Enfin, l'échelon 4 représente la situation où le lieu *et* le rôle influencent fortement la communication. Cet axe fait écho à la notion d'*'informalité'* désignant la parole spontanée dans la figure 1 et aux différents contextes dans lesquels elle peut apparaître tels que définis par (Labov, 1972) et (Beckman, 1997). Nous proposons ensuite de prendre en compte la relation entre les locuteurs en précisant le **degré d'intimité** (I) qu'il y a entre eux. La graduation 1 représente un degré d'intimité fort, la 2 la relation de personnes qui partagent un quotidien sans degré d'intimité fort, la 3 la relation de simples connaissances, et la 4 étant la relation entre des inconnus. Cette distinction s'inspire notamment du qualificatif *'familier'* pour désigner la parole spontanée et de l'influence de la relation plus ou moins familière entre les locuteurs sur la spontanéitée, mentionnée par (Torreira *et al.*, 2010). Pour ce qui est du **canal de communication** (C), nous choisissons de différentier les situations en (1) face-à-face, (2) via un outil permettant la visio, (3) via un outil ne permettant pas la visio. Enfin, le **type de communication** (T) est définit selon trois graduations : (1) communication interpersonnelle, (2) communication de groupe, (3) communication de masse. Les deux premières représentent une parole privée, tandis que la communication de masse est une parole publique, médiatique. Pour chacun des axes, les graduations les plus proches de 1 sont

celles qui annoncent un fort niveau de spontanéité. Le tableau 3 montre différentes catégorisation de corpus ou enregistrements de parole spontanée selon nos quatre axes.

Corpus	Description	S	I	C	T
NCCFr (Torreira *et al.*, 2010)	Conversations entre amis enregistrées dans un laboratoire	2	1	1	1
ESLO2 repas (Eshkol-Taravella *et al.*, 2011)	Discussion entre amis ou en famille pendant le repas, à la maison	1	1	1	1, 2
ESLO2 cinéma	Passants interviewés à leur sortie du cinéma, par des étudiants en linguistique	2	4	1	1, 2
ESLO2 entretiens jeunes	Jeunes interrogés (seuls ou en groupe) par un jeune chercheur	2	4	1	1, 2
ESLO2 itinéraires	Demande d'itinéraire et rapide interview de passants par des étudiants en linguistique.	2	4	1	1, 2
MPF (Gadet & Guerin, 2016)	Discussion entre personnes qui se connaissent. Entretiens traditionnels, entretiens de proximité ou événements écologiques (sans enquêteur)	2	3	1	1, 2
PFC (Laks *et al.*, 2009)	Deux enquêteurs, dont un (A) proche de l'enquêté et l'autre présenté comme un ami (B). Conversations guidées (menées par B) ou libres (menées par A). A domicile ?	2	1, 4	1	1, 2
CFPP2000* (Branca-Rosoff *et al.*, 2012)	Entretiens guidés sur le rapport des parisiens à leur quartier. Plusieurs situations : Enquêteurs et enquêtés ne se connaissent pas ou sont de la même famille. L'enregistrement a lieu à domicile ou à l'université.	2, 3	1, 4	1	1, 2
CLAPI* apéritif entre amis (Baldauf-Quilliatre *et al.*, 2016)	Interaction entre amis durant un apéritif, à domicile	1	1	1	1, 2
CLAPI* consultation médicale	Première consultation avec un psychiatre, présence de l'enquêteur	3	4	1	1, 2
CLAPI* conversations en ligne	Conversations entre deux locuteurs en visio, sur un thème imposé	1	?	2	1
CLAPI* conv. téléphoniques	Conversations entre deux locuteurs téléphone pour inviter à sortir	1	1	3	1
C-ORAL-ROM* (Cresti *et al.*, 2004)	Parole publique et privée (monologue, dialogue et multi-dialogue). Discours politique, cours, conférence, interviews, journaux télévisés, conversations téléphoniques...	1, 2, 3	1, 2, 3, 4	1, 3	1, 2, 3
CRFP* parole privée (Equipe Delic *et al.*, 2004)	Récits de vie (voyage, souvenir...) ou présentation d'un "savoir-faire" professionnel.	2	4	1	1, 2
CRFP* parole professionnelle	Locuteurs enregistrés dans l'exercice de leur fonction ou quand ils parlent de leur profession sur leur lieu de travail	3	4	1	1, 2
CRFP* parole publique	Présence de public. Entretiens sollicités, émissions de radio (interview, table ronde), cours, conférences, réunions politiques/associatives...	2, 3, 4	3, 4	1, 3	2, 3
FLEURON* (André, 2017)	Actions et interactions dans des situations universitaires variées (inscription auprès de l'administration, discuter avec des étudiants ou un enseignant...), témoignages	1, 2, 3	1, 4	1, 2	1, 2
OFROM* (Avanzi *et al.*, 2016)	Entretiens guidés, avec plus ou moins d'interactions. Thèmes : métiers, passe-temps, relations de voisinage, système politique... Enregistrements réalisés par des étudiants ou chercheurs.	1	4	1	1, 2
TCOF* (André & Canut, 2010)	Enregistrements recueillis par des chercheurs ou étudiants. Récits de vie, événements, savoir-faire professionnel, réunions publiques ou professionnelles...	2, 3, 4	3, 4	1	1, 2, 3

TABLE 3 – Catégorisation de corpus de parole spontanée selon quatre dimensions
∗Corpus accessibles via le corpus CEFC-ORFEO ; la virgule dans les colonnes **S**, **I**, **C** ou **T** représente la potentialité d'un score ou d'un autre.

4 Nouvelles perspectives pour la RAPS

Cette nouvelle typologie apporte de nouvelles perspectives pour la RAPS. Tout d'abord, elle permet de proposer une façon d'utiliser des corpus de parole spontanée qui n'étaient jusqu'alors pas utilisés pour

la reconnaissance automatique de cette parole. En plus des corpus d'émission de radio et de TV EPAC, REPERE et ETAPE, il devient possible d'utiliser des corpus d'autres sous-domaines de la linguistique comme CLAPI. Leur utilisation est facilitée puisque la nouvelle catégorisation permet de rassembler rapidement des contextes de production similaires, encourageant ainsi des travaux qui deviennent plus facilement comparables. A titre d'exemple, il est possible de rassembler d'un coup d'oeil les corpus ESLO2 'repas' et CLAPI 'apéritif entre amis' ou encore certains enregistrements du corpus PFC et du CRFP 'parole privée'. Les différents ensembles de test ainsi créés permettent de diversifier les données, en faisant en sorte que les données ne proviennent pas toujours du même corpus, tout en contrôlant les données et en ayant la possibilité d'étudier certains paramètres influençant la parole spontanée. Il devient possible par exemple d'étudier la différence de performance sur des enregistrements où des amis ou des inconnus se parlent, en choisissant de ne sélectionner que les corpus où le degré d'intimité entre les locuteurs est égal à 1 ou à 4.

Les corpus de parole spontanée sont généralement de petite taille par rapport aux volumes généralement rencontrés pour apprendre des systèmes de RAP. L'application de notre catégorisation de parole spontanée risque de partitionner ceux-ci en sous-corpus encore plus petits. Néanmoins, le manque de données n'est plus forcément une barrière aujourd'hui grâce aux opportunités amenées par les nouvelles architectures de RAP. Il est dorénavant possible d'utiliser des modèles pré-appris sur des milliers d'heures que l'on peut adapter avec seulement quelques heures de parole, tout en conservant de bonnes performances (Baevski *et al.*, 2020; Evain *et al.*, 2021).

5 Conclusion et perspectives

Nous détaillons dans cet article une proposition de typologie de la parole spontanée que nous illustrons avec une catégorisation de corpus de parole spontanée. Cette typologie repose sur une vision multidimensionnelle de la parole spontanée, dont les degrés de spontanéité dépendent du contexte situationnel (rôle des locuteurs, environnement), du canal de communication (face-à-face, distant), du type de communication (interpersonnel, de groupe, de masse) et du degré d'intimité entre les locuteurs. Ces quatre axes sont à la fois issus de recherches sur la parole spontanée et des corpus de parole spontanée en français. Ils permettent de mesurer la richesse de ces corpus. Une fois l'ensemble des corpus catégorisés selon ce principe, notre typologie permettrait de créer une base de données interrogeable des corpus de parole spontanée et d'étudier chacun des différents axes influençant la spontanéité de la parole. Leur bonne catégorisation *a posteriori* est toutefois contrainte par la qualité des fichiers de métadonnées associés aux enregistrements. Nous pensons que cette catégorisation des corpus pourrait être propice à la RAPS en permettant de mélanger les données de différents corpus, tout en offrant la possibilité de rassembler des enregistrements similaires et de créer des ensembles de données cohérents. Nous proposons en section 4 une façon d'utiliser efficacement les ensembles créés par l'utilisation et l'adaptation de modèles pré-appris. Nos travaux immédiats consisteront à étudier l'impact de l'adaptation de ce type de modèles sur la parole spontanée sachant que ces modèles sont majoritairement appris sur de la parole préparée ou lue avec seulement quelques heures de parole spontanée.

Remerciements

Ce travail est financé par MIAI@Grenoble Alpes (ANR-19-P3IA-0003). Nous remercions les relecteurs et relectrices pour leur relecture attentive et leurs commentaires très constructifs.

Références

ADDA-DECKER M. *et al.* (2004). Une étude des disfluences pour la transcription automatique de la parole spontanée et l'amélioration des modèles de langage. In *Actes des JEP 2004*, Fès, Maroc.

ANDRÉ V. (2017). Un corpus multimédia pour apprendre à interagir en situations universitaires en France. In *Actes de l'ATPF « Enseigner le français : s'engager et innover »*, Bangkok, Thaïlande.

ANDRÉ V. & CANUT E. (2010). Mise à disposition de corpus oraux interactifs : le projet TCOF (Traitement de Corpus Oraux en Français). *Pratiques : théorie, pratique, pédagogie*, (147-148).

AVANZI M. *et al.* (2016). De l'archive de parole au corpus de référence : la base de données orales du français de Suisse romande (OFROM). *Corpus*, (15).

BAEVSKI A. *et al.* (2020). Wav2vec 2.0 : A Framework for Self-Supervised Learning of Speech Representations. In *actes de NeurIPS 2020*, Vancouver, Canada.

BALDAUF-QUILLIATRE H. *et al.* (2016). CLAPI, une base de données multimodale pour la parole en interaction : apports et dilemmes. *Corpus*, (15).

BECKMAN M. E. (1997). A Typology of Spontaneous Speech. In *Computing Prosody : Computational Models for Processing Spontaneous Speech*. New York, sagisaka, campbell, higuchi edition.

BENZITOUN C. *et al.* (2016). Le projet ORFÉO : un corpus d'étude pour le français contemporain. *Corpus*, (15).

BLANCHE-BENVENISTE C. *et al.* (1987). *Le français parlé - transcription et édition*. Didier érudition, paris edition.

BLANCHE-BENVENISTE C. *et al.* (1999). "Français parlé - oral spontané". Quelques réflexions. *Revue française de linguistique appliquée*, **IV**(2), 21.

BRANCA-ROSOFF S. *et al.* (2012). Discours sur la ville. Présentation du Corpus de Français parlé Parisien des années 2000 (CFPP2000).

CARON P. (1992). L'écriture de la noblesse vers 1680. In *Grammaire des fautes et français non conventionnels*. Paris, France, presses de l'ecole normale supérieure, rue d'ulm edition.

CID URIBE M. *et al.* (1991). The construction of a corpus of spoken Spanish : Phonetic and phonological parameters. In *ESCA Workhop 'Phonetics and Phonology of Speaking Styles : Reduction and Elaboration in Speech communications'*, Barcelona, Spain.

CRESTI E. *et al.* (2004). The C-ORAL-ROM CORPUS. A Multilingual Resource of Spontaneous Speech for Romance Languages. In *Proceedings of LREC 2004*, Lisbon, Portugal.

DELÉGLISE P. *et al.* (2005). The LIUM speech transcription system : a CMU Sphinx III-based system for French broadcast news. In *Interspeech 2005*, p. 1653–1656 : ISCA.

DESNOUS F. *et al.* (2018). Impact de la détection de la parole pour différentes tâches de traitement automatique de la parole. In *XXXIIe Journées d'Études sur la Parole*, Aix-en-Provence, France.

DUFOUR R. (2008). From prepared speech to spontaneous speech recognition system : a comparative study applied to French language. In *Proceedings of CSTST 2008*, New York, NY, USA.

DUFOUR R. *et al.* (2010). Automatic indexing of speech segments with spontaneity levels on large audio database. In *Proceedings of SSCS 2010*, Firenze, Italy.

EQUIPE DELIC *et al.* (2004). Présentation du Corpus de référence du français parlé. *Recherches sur le français parlé*, **18**.

ESHKOL-TARAVELLA I. *et al.* (2011). Un grand corpus oral "disponible" : le corpus d'Orléans 1968-2012. *TAL*, **53**(2).

ESTÈVE Y. *et al.* (2010). The EPAC Corpus : Manual and Automatic Annotations of Conversational Speech in French Broadcast News. In *Proceedings of (LREC'10)*, Valletta, Malta.

EVAIN S. *et al.* (2021). Task Agnostic and Task Specific Self-Supervised Learning from Speech with LeBenchmark. In *NeurIPS 2021 Datasets and Benchmarks Track*, on-line.

FUJISAKI H. (1997). Prosody, Models, and Spontaneous Speech. In *Computing Prosody*. New York, NY, sagisaka, campbell and higuchi edition.

GADET F. & GUERIN E. (2016). Construire un corpus pour des façons de parler non standard : « Multicultural Paris French ». *Corpus*, (15).

GALIANO A. R. *et al.* (2008). Les interactions verbales chez le sujet aveugle : le rôle de la vision dans la communication.

GARNERIN M. (2018). *Répartition hommes/femmes dans les systèmes d'IA : une étude pilote sur les grands corpus pour la transcription automatique de la parole*. Mémoire, Université Grenoble-Alpes, Grenoble, France.

GIRAUDEL A. *et al.* (2012). The REPERE Corpus : a multimodal corpus for person recognition. *LREC*.

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

GUIRAUD P. (1965). *Le français populaire*. Number 1172 in "Que sais-je ?". Paris, France, presses universitaires de france edition.

JOUSSE V. *et al.* (2008). Caractérisation et détection de parole spontanée dans de larges collections de documents audio.

LABOV W. (1972). The Isolation of Contextual Styles. In *Sociolinguistic patterns*. University of Pennsylvania Press.

LACHERET A. *et al.* (2014). Rhapsodie : un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé. *SHS Web of Conferences*, **8**.

LAKS B. *et al.* (2009). Le projet PFC (Phonologie du Français Contemporain) : une source de données primaires structurées. In *Phonologie, variation et accents du français*. Hermes science publications edition.

MALIK M. *et al.* (2021). Automatic speech recognition : a survey. *Multimedia Tools and Applications*, **80**(6), 9411–9457.

MEKKI I. I. (2020). *Automatic Speech Recognition in the French Language*. PhD thesis.

RUHLEDER K. & JORDAN B. (2001). Co-Constructing Non-Mutual Realities : Delay-Generated Trouble in Distributed Interaction. *Computer Supported Cooperative Work (CSCW)*, **10**(1).

SHRIBERG E. (2005). Spontaneous Speech : How People Really Talk and Why Engineers Should Care. In *Proceedings of Interspeech*, Lisbon, Portugal.

SZYMAŃSKI P. *et al.* (2020). WER we are and WER we think we are. In *ACL : EMNLP 2020*, Online.

TEN BOSCH L. *et al.* (2004). Durational Aspects of Turn-Taking in Spontaneous Face-to-Face and Telephone Dialogues. In *Text, Speech and Dialogue*, p. 563–570, Berlin, Germany.

TOLEDANO D. T. *et al.* (2005). Acoustic-phonetic decoding of different types of spontaneous speech in Spanish. In *Proceedings of DiSS'05*, France.

TORREIRA F. *et al.* (2010). The Nijmegen Corpus of Casual French. *Speech Communication*, **52**(3).

VEIGA A. *et al.* (2012). Towards Automatic Classification of Speech Styles. In *conference on Computational Processing of the Portuguese Language*, Coimbra, Portugal.