Suivre le rythme de tes paroles

Solange Rossato[1], Dan Zhang[1], Moez Ajili[2], Jean-François Bonastre[2]
(1) Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG, F-38000 Grenoble France
(2) Univ. Avignon et Pays de Vaucluse, LIA, F-84000 Avignon France
solange.rossato@univ-grenoble-alpes.fr

RÉSUMÉ

Différentes mesures temporelles, telles que la durée des voyelles et des consonnes, ont été proposées pour tenter de caractériser le rythme de la parole et classer ainsi les langues, les dialectes ou les idiolectes. C'est sur ce dernier rôle des paramètres temporels de la parole que cette étude se focalise en s'appuyant sur la base de données FABIOLE. Utilisée pour la comparaison de voix, elle est construite à partir d'émissions médiatiques (TV et radio). Elle nous permet ainsi d'étudier la variabilité de certains paramètres temporels, variabilité intra et inter locuteurs, à la recherche d'un idiolecte. Les résultats montrent que la part de variabilité que l'on peut attribuer au locuteur atteint 45% pour la variance de la durée des segments non voisés, 42% pour le pourcentage total de segments voisés. Ainsi, ces mesures temporelles dépendent du locuteur, de façon bien plus marquée que ne le sont les paramètres formantiques.

ABSTRACT

Following the rhythm of your speech
Various temporal measures, such as the duration of vowels and consonants, have been proposed to characterize the rhythm of speech and thus classify languages, dialects or idiotic expressions. It is on this last role of the temporal parameters of speech that this study focuses on, using the FABIOLE database. Used for voice comparison, it is constructed from media broadcasts (TV and radio). It allows us to study the variability of certain temporal parameters, within and between speakers, in search of idiosyncrasy. The results show that the percentage of variability that can be attributed to the speaker is 45% for the variance of the duration of un-voiced segments, 42% for the total percentage of voiced segments. Thus, these temporal measurements depend on the speaker, much more strongly than the formantic parameters.

MOTS-CLÉS : Mesures temporelles, rythme, voisement, locuteur, idiolect, idiosyncrasie
KEYWORDS: Timing, rythm, voicing, speaker, idiolect, idiosyncrasy

1 Parole et Identité : le choix de mesures rythmiques

Le rythme est une notion de structuration temporelle du flux de parole. Ainsi, une typologie ryhtmique classe les langues comme étant plutôt syllabiques, accentuelles ou moraiques. Les mesures ryhtmiques, quelles soient basées sur l'alternance Consonne Voyelle (Ramus, 1999) ou sur des alternances d'intervalles voisés et non voisées ou des intervalles entre syllabes accentuées (Dellwo & Fourcin, 2013a; Dellwo, Fourcin, & Abberton, 2007) montrent en effet un impact de la langue parlée important sans pour autant permettre une réelle classification typologique. En effet, d'autres facteurs entrent en jeu, tel que le style de parole ou phonostyle (de Mareüil, 2014; Fónagy,

10.21437/JEP.2018-5

1983; Simon, Auchlin, Avanzi, & Goldman, 2010) qui modifient l'organisation temporelle de la parole. Il a été ainsi montré que les pauses jouent un rôle spécifique dans la parole politique (Duez, 1999). (Eskénazi, 1993) écrit que le style de parole reflète une interaction entre un locuteur et son environnement : « *It is the perception of the various status levels of his listener and of the type of situation in which he finds himself.* » (p. 502). La situation de communication est donc un facteur important mais le rôle central du locuteur, ses projections et son histoire est également souligné par l'auteur et rejoint la définition de (Labov, 1972) pour qui le style de parole est un fait qui relève de l'individualité du locuteur. Le caractère idiosyncrasique des paramètres rythmiques a ainsi été mis en évidence par plusieurs études (Dellwo, Leemann, & Kolly, 2012, 2015; Leemann, Kolly, & Dellwo, 2014). Le rythme de la parole apparaît comme un élément dépendant du locuteur. Nous allons alors suivre le rythme des paroles de locuteurs pour tenter de quantifier la relation entre ryhtme et l'identité du locuteur. Pour cela, la base de données FABIOLE, contenant plus de 30h de parole médiatique, pour 30 locuteurs différents, soit 100 extraits de parole par locuteur, est un corpus contenant une grande variabilité intra- et inter- locuteurs. L'objectif de cette étude est de commencer par des paramètres rythmiques concernant le voisement afin de déterminer dans quelle mesure ces paramètres temporels nous renseignent sur la provenance de la voix. Après avoir détaillé la méthodologie mise en œuvre, les résultats seront présentés avant la partie discussion.

2 Méthodologie

Nous avons utilisé une approche basée sur un grand corpus de parole contenant environ 30h de parole, le corpus FABIOLE. Cette approche nous a conduit à envisager des mesures automatiques, ne nécéssitant pas ou un minimum d'interventions manuelles. Les mesures rythmiques qui en découlent sont donc d'un très bas niveau, et se limitent ici à une analyse grossière des alternances de parties voisées et non voisées dans les extraits de parole. Il ne s'agit en aucun cas d'une analyse fine, et nous sommes bien loin des mesures rythmiques s'appuyant sur les syllabes accentués. L'avantage de ce type d'analyse bas niveau est qu'elle permet une extraction de paramètres rythmiques qui peut se faire de façon automatique mais sur un très grand nombre d'extraits de parole.

Par ailleurs, nous n'avons pas de transcription orthographique pour l'intégralité de la base de données. Plutôt que de se baser sur une transcription obtenue par un système de transcription automatique (Ajili, 2017), nous avons pris le parti de nous appuyer sur des mesures ne nécessitant aucune transcription préalable, effectuées directement sur le signal acoustique sans faire appel à d'autres ressources. Les indices de voisement se prètent parfaitement à cela. Bien évidemment, ces indices dépendent de la phonotactique de la langue, mais ce facteur est neutralisé dans notre base de données contenant exclusivement des extraits de locuteurs francophones.

2.1 La base de données FABIOLE

La base de données utilisée est celle développée dans le cadre du projet ANR Fiabilité en biométrique vocale. Ce corpus FABIOLE (Ajili, Bonastre, Kahn, Rossato, & Bernard, 2016) a été extrait de programmes télévisuels ou radiophoniques francophones entre 2013 et 2014 avec l'objectif de capturer les variabilités de la parole intra- et inter- locuteurs dans un contexte de comparaison de voix. Le corpus s'est focalisé sur les voix d'hommes en raison de leur plus grande disponibilité dans les médias. Ainsi, un minimum de 100 extraits de minimum 30 sec chacun est obtenu pour 30 locuteurs différents. La base de données FABIOLE contient également 100 extraits de 30 sec produits par 100 locuteurs différents mais cette partie du corpus n'est pas pris en compte ici. Les variabilités dues canal de communication sont réduites. En effet, les extraits de parole sont généralement de très bonne qualité, étant enregistré avec du matériel audio professionnel. Ainsi, les

analyses ont pu être menées sur l'ensemble du corpus excepté pour un locuteur, chroniqueur, dont les enregistrements n'ont pas lieu en studio et qui sera par la suite enlevé de l'analyse. Certains signaux contiennent de la musique de fond.

Les locuteurs ont des métiers différents : journalistes, hommes politiques, chroniqueurs mais étant donné leur grande présence dans les médias ont tous une très grande pratique de la parole médiatique. Le diagramme de la Figure 1 présente cette répartition et l'on note clairement la grande proportion de journalistes, chroniqueurs et débatteurs. Cependant, plutôt que de considérer le métier des locuteurs, nous nous sommes intéressés à la situation de la communication, en prenant en compte les différentes émissions dont sont extraits les signaux de parole de la base FABIOLE comme autant de situations différentes.

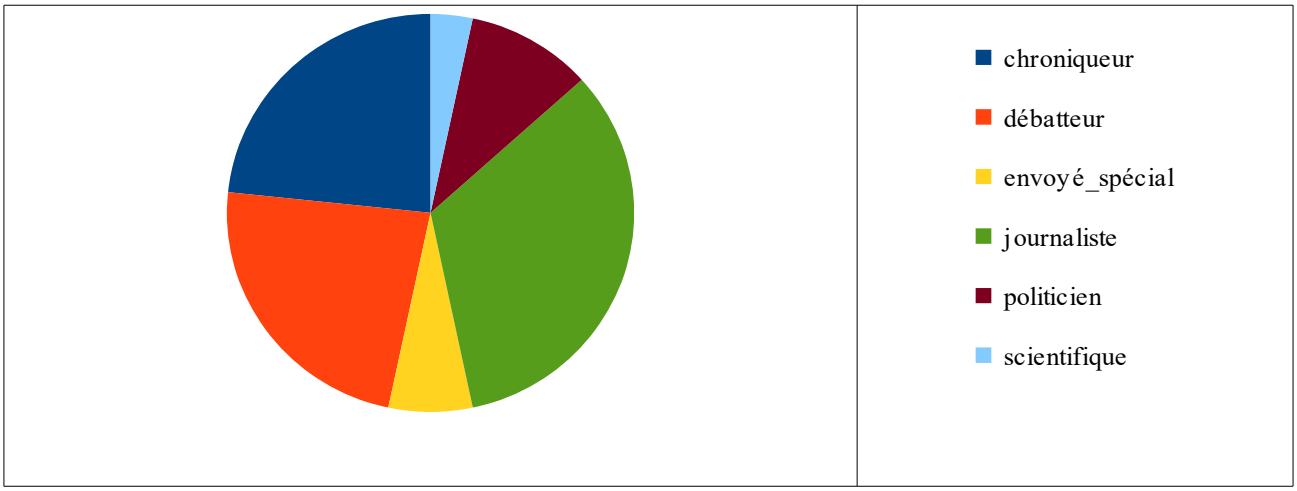

FIGURE 1: Diagramme de répartition des métiers des locuteurs

Ces extraits de parole proviennent de 9 émissions :

- des débats : "Ça vous regarde - Le débat" (cvgdde-bat), "Entre les Lignes" (entreligne), "Le masque et la plume" (MsqPlum),
- des chroniques : "Service public" (Spublic), "Comme on nous parle" (ComParle),
- des séances parlementaires : "Top Questions" (topquestions),
- des journeaux télévisés "BFM Story" (bfmstory), "LCP Info"(parlinfo), "Ca vous regarde-l'Info (cvgdinfo)".

La répartition entre les différentes émissions n'est pas uniforme, ainsi que le montre la Figure 2. Par rapport à (Ajili et al., 2016), il manque une émission qui correspond au locuteur que nous avons écarté pour des raisons de qualité acoustique des signaux. Il y a ainsi un fort lien entre locuteurs et émissions, certains locuteurs étant les animateurs des émissions. Ainsi, 19 locuteurs sur les 29 n'interviennent très majoritairement (plus de 90 % des extraits) que dans une seule émission. Les locuteurs interviennent dans plusieurs émissions mais avec des temps de parole très disparates. Il n'est pas possible, avec ces données, d'étudier conjointement l'influence du locuteur et de l'émission. Or ces deux facteurs influencent le ryhtme de la parole. Nous allons donc les étudier de façon indépendante tout en étant conscients que c'est une limite forte que de ne pas pouvoir étudier l'interaction entre ces deux facteurs.

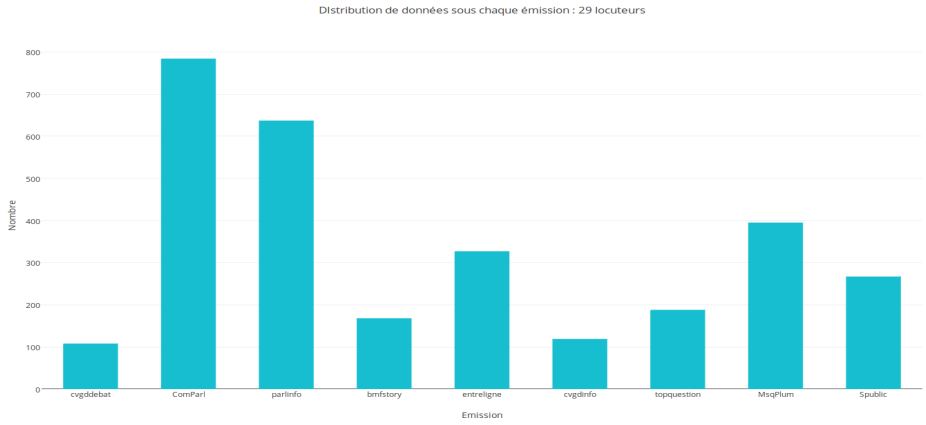

FIGURE 2: Répartition des extraits de parole en fonction des émissions

2.2 Les mesures temporelles du voisement

Ce premier travail s'est focalisé sur un premier aspect rythmique de la parole, à savoir l'alternance des intervalles voisés et de intervalles non voisés (incluant les pauses). Ce travail s'inspire largement des mesures effectuées des études de (Dellwo & Fourcin, 2013b; Dellwo et al., 2012, 2015; Leemann et al., 2014) Nous avons utilisé le script de ProsodyPro[1] développé par (Xu, 2013) pour automatiser la mesure du voisement. A partir des mesures de f0, nous avons obtenu les durées des intervalles voisés *dVO* et non voisés *dUV* pour chacun extrait. Pour chaque fichier, nous avons calculé les mesures suivantes :

- le pourcentage global de la durée cumulée des intervalles voisés dans l'extrait sonore, *%VO ;*
- la durée moyenne des intervalles voisés, *Moyenne(DuréeVO)*;
- le coefficient de variation de la durée des intervalles non voisés, *VarcoUV*, calculé de la façon suivante : 100*Var(*dUV*)/Moy(*dUV*)
- le coefficient de variation de la durée des intervalles voisés, *VarcoVO*.

Ainsi, un extrait de parole qui contient des pauses importantes aurait un *%VO* légèrement plus faible et un coefficient de variance *VarcoUV* plus grand que celui qui contient moins de longues pauses. A ces mesures classiques, nous avons également ajouté les moyenne et coefficient de variation de l'intervalle de temps entre le début de deux intervalles voisés successifs ou *paire*, le dernier intervalle voisé n'étant pas pris en compte. La *paire* est ainsi l'intervalle de durée dVO_i+dUV_{i+1}. Parmi ces *paires*, nous avons cherché la proportion de *paires* pour lesquelles l'intervalle voisé dVO_i est plus court que l'intervalle non voisé dUV_{i+1}. A la liste précédente, viennent donc s'ajouter :

- le proportion de *paires* pour lesquelles la durée d'un intervalle non voisé est supérieure à celle de l'intervalle voisé qui le précède, $\%(UV_{i+1} > VO_i)$;
- la durée moyenne de chaque *paire* d'intervalles voisés et non voisés, *Moyenne(paire)*;
- le coefficient de variation de la durée des *paires*, *VarcoPaire*.

La Figure 3 permet de visualiser les durées extraites du signal et permettant de calculer ces 7 paramètres temporels de voisement.

[1] http://www.homepages.ucl.ac.uk/~uclyyix/ProsodyPro/

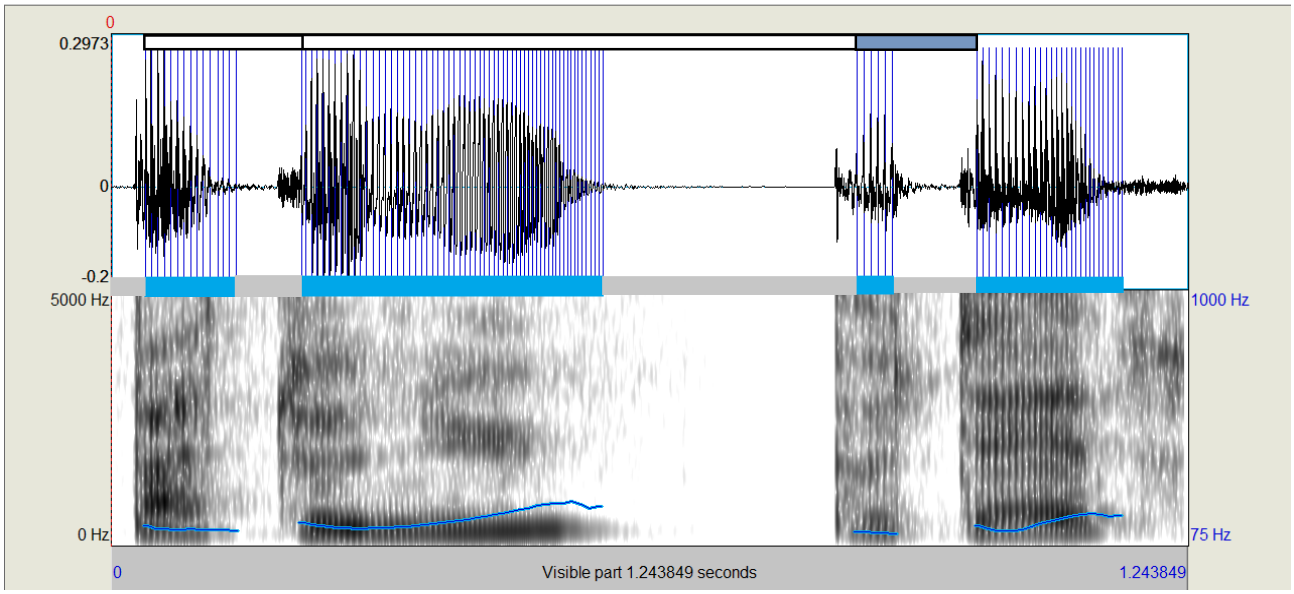

FIGURE 3: Signal de parole et sonagramme avec, au dessus du sonagramme, la visualisation des intervalles voisés (en bleu) et non voisés (en gris) et, au dessus du signal de parole, la visualisation des paires (encadrés noirs) avec en bleu, celle pour laquelle la portion voisée est plus courte que la portion non voisée qui la suit.

2.3 Analyses statistiques

Des ANOVA à un facteur, Locuteur ou Émission, sont appliquées sur chaque variable temporelle de façon indépendante. Les résultats sont significatifs étant donné le grand nombre de valeurs (avec 100 valeurs par locuteur). Pour quantifier l'influence d'un facteur sur chaque variable, nous avons calculé la taille de l'effet en utilisant l'êta-carré η^2. L'êta-carré décrit la force de la relation, et correspond au pourcentage de la variance totale expliquée par le facteur en question. Plus cette valeur est importante, plus le facteur en question est explicatif de la variance de la variable étudiée. Une interprétation fréquente (Cohen, 1988) indique une taille de l'effet faible lorsque la valeur est inférieure à 1 %, moyenne jusqu'à 6 %, et importante lorsque les valeurs sont supérieures à 14 %. La formule de calcul de l'éta-carré est la suivante, avec *SStotal* étant la variance totale et *SSbetween*, la variance des moyennes par locuteur :

$$\eta^2 = \frac{SSbetween}{SStotal} *100$$

Après avoir présenté les résultats sous forme de graphiques permettant de visualiser la variation intra- et inter- locuteurs, les résultats de l'analyse statistique sont présentés.

3 Résultats

3.1 Un rythme spécifique au locuteur ?

L'intégralité des mesures obtenues sont disponible dans le fichier .odt téléchargeable avec l'article. Les graphiques et analyses statistiques ont été réalisés avec Matlab®. La Figure 4 montre la

répartition du pourcentage de voisement *%VO* en fonction des locuteurs. Globalement, sur l'ensemble des extraits, 60,2 % du signal est voisé, cette proportion moyenne variant de 53,7 % (loc. 6) à 65,5 % (loc. 19) en fonction des locuteurs.

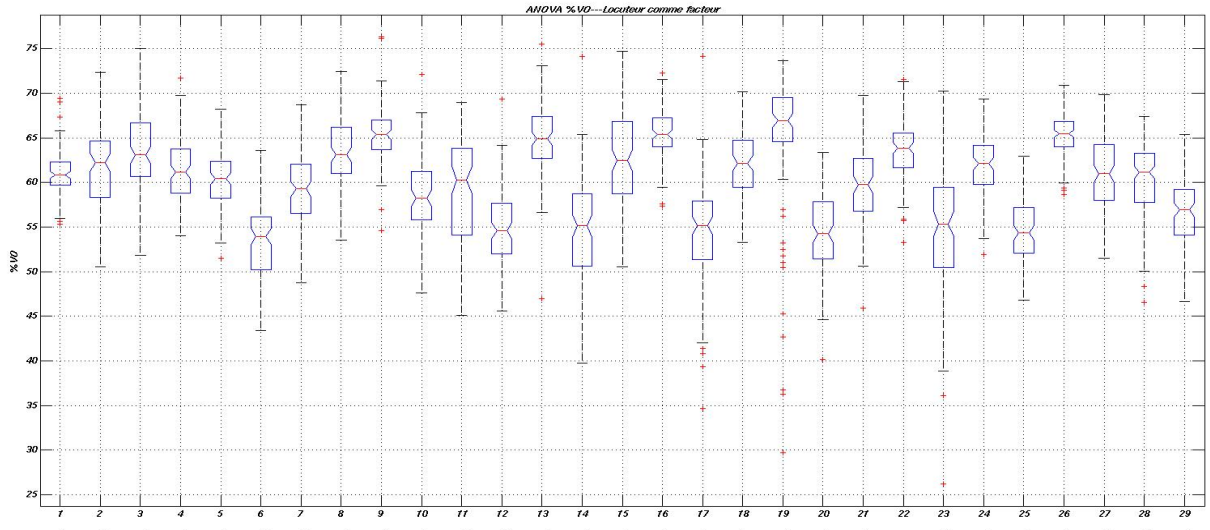

FIGURE 4: Répartition (médiane et quartile) du pourcentage de voisement %VO en fonction des locuteurs.

Une ANOVA à un facteur est effectuée pour chacune des 7 variables. Les résultats des ANOVA étudiant le facteur locuteur, tous significatifs, ainsi que les valeurs des êta-carré η^2 sont indiqués dans le tableau 5. Le facteur Locuteur a un effet important pour toutes les variables temporelles étudiées, avec des valeurs largement supérieures au seuil de 14 %. Pour certaines variables, la taille de l'effet est même très importante. Représenter les tailles sous forme d'un radar-chart permet de bien visualiser les variables la force de la relation entre le facteur Locuteur et chacune des variables (voir Figure 6). Il apparaît ainsi clairement que les variables qui montrent le plus de variation propre au locuteur sont la proportion de voisement (*%VO*), le coefficient de variance des intervalles non voisés (*VarcoUV*) ainsi que la durée moyenne des intervalles séparant le début de deux intervalles de voisement consécutifs (*Moyenne(paire)*).

Variables	*Résultats de l'ANOVA*	η^2
%VO	F(28,2964) = 76.83, p <.001	**42.056**
Moyenne(DuréeVO)	F(28,2964) = 64.13, p <.001	**37.727**
VarcoVO	F(28,2964) = 36.22, p <.001	**25.492**
VarcoUV	F(28,2964) = 88.32, p <.001	**45.485**
$\%(UV_{i+1} > VO_i)$	F(28,2964) = 60.02, p <.001	**37.686**
Moyenne(paire)	F(28,2964) = 73.99, p <.001	**41.142**
VarcoPaire	F(28,2964) = 25.47, p <.001	**19.391**

TABLE 5 : Résultats de l'ANOVA étudiant le facteur Locuteur

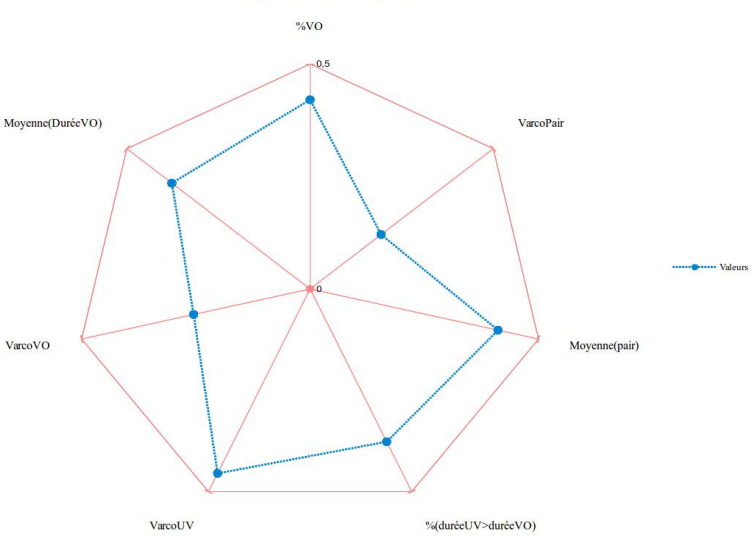

FIGURE 6: Radar-chart des tailles de l'effet Locuteur η^2 pour chaque variable.

3.2 Ou un rythme lié à l'émission médiatique ?

En suivant la même démarche pour les émissions que pour les locuteurs, nous obtenons également une influence significative du facteur Émission sur chaque variable étudiée. Dans l'ensemble, les tailles de l'effet sont plus faibles que celles obtenues pour le facteur Locuteur mais non négligeables ainsi que le montre le tableau 7. On observe ainsi que les variables qui mesurent la proportion de voisement *%VO* ainsi que le coefficient de variance des intervalles non voisés *VarcoUV* sont fortement dépendantes de l'émission.

Variables	*Résultats de l'anova*	η^2
%VO	F(8,891) = 36.56, p<.001	**24.71**
Moyenne(DuréeVO)	F(8,891) = 16.14, p<.001	*12.66*
VarcoVO	F(8,891) = 30.77, p<.001	**21.56**
VarcoUV	F(8,891) = 42.89, p<.001	**27.81**
$\%(UV_{i+1} > VO_i)$	F(8,891) = 27.99, p<.001	**20.08**
Moyenne(paire)	F(8,891) = 29.36, p<.001	**20.86**
VarcoPaire	F(8,891) = 4.07 , p<.001	3.52

TABLE 7 : Résultats de l'ANOVA étudiant le facteur Emission

4 Discussion et conclusion

L'organisation temporelle du voisement dépend clairement du locuteur, et ce lien de dépendance est très important. En effet, les mesures formantiques des voyelles orales, très largement étudiés comme

apportant des informations idiosyncratiques, ont été étudiées sur ces mêmes données FABIOLE, (Ajili, 2017) et les tailles de l'effet obtenues sont bien plus faibles pour les formants, y compris le F4 que pour les mesures temporelles étudiées ici, pour lesquelles la taille de l'effet varie entre 37,7 % et 45,5 % pour 5 d'entre elles, tandis qu'elle atteint difficilement 20 % pour les valeurs les plus importantes obtenues pour les formants des voyelles orales (voir le tableau 8).

Vowel	F1	F2	F3	F4
/i/	1.65	*6.16*	*8.09*	**14.08**
/y/	2.98	5.95	5.91	*11.68*
/u/	2.1	2.50	*6.51*	3.98
/e/	1.83	**17.46**	*9.72*	**20.56**
/ø/	5.79	*7.9*	4.13	**14.86**
/o/	*12.2*	*13.22*	*8.1*	8.10
/ɛ/	2.8	*11.30*	*10.75*	**18.48**
/œ/	*10.88*	*7.60*	*8.48*	**23.04**
/ɔ/	*12.51*	*10.56*	*7.34*	*13.18*
/a/	*12.85*	4.0	*13.21*	**19.82**

TABLE 8 : Tailles de l'effet η^2 pour les formants des voyelles orales sur la base de données FABIOLE (extrait de (Ajili, 2017) p 154).

Cependant, le rôle de l'émission n'est pas négligeable, loin de là, avec des tailles de l'effet allant jusqu'à 27.8 % pour le coefficient de variance des intervalles non voisés *VarcoUV*. Quelle part, dans l'organisation temporelle globale du voisement, peut-on attribuée au style de parole spécifique du locuteur (son idiolecte) et quelle part peut-on attribuer au style de parole correspondant à la situation de communication ? Ici, les situations de communication ont toutes en commun d'être d'une parole publique et médiatique et tous nos locuteurs sont des professionnels de la parole, mais cette parole médiatique n'a pas les mêmes objectifs lorsqu'il s'agit de journaux télévisés ou de sessions parlementaires, avec un aspect plus ou moins formel. La question de l'interaction entre idiolecte et situation de communication due à l'émission reste ainsi posée. Il faudrait, à l'instar des travaux de Dellwo (Dellwoa & Schmida, 2016) qui étudient l'interaction entre la langue et l'idiolecte en enregistrant des locuteurs bilingues, étudier cette problématique dans son ensemble en croisant locuteurs et situations de communication pour étudier leur influence respective et leur interaction sur les paramètres rythmiques. Peut-on, parmi les multiples mesures rythmiques possibles trouver celles qui relèvent plutôt de tel ou tel facteur ?

Une importante limitation à cette étude est sa focalisation sur des mesures d'organisation temporelle du voisement. Or le rythme ne peut se résumer à cette mesure bas niveau et nous devons compléter cette étude, sur cette même base de données avec une estimation des centres de pseudo-syllabes, une détection des pseudo-syllabes accentuées pour permettre de mesurer des intervalles entre syllabes accentuées.

Remerciements

Ce travail a pu se faire grâce au financement du projet ANR-12-BS03-0011 FABIOLE.

Références

Ajili, M. (2017, novembre 28). *Fiabilité de la comparaison des voix dans le cadre judiciaire*. Université d'Avignon et des Pays du Vaucluse, Avignon, France.

Ajili, M., Bonastre, J.-F., Kahn, J., Rossato, S., & Bernard, G. (2016). Fabiole, a speech database for forensic speaker comparison. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC* (p. 23–28).

Cohen, J. (1988). Statistical power analysis for the behavioral sciences second edition. *Lawrence Erlbaum Associates, Publishers*.

de Mareüil, P. B. (2014). Qu'est-ce qu'un (phono) style? *Cahiers de linguistique française*, (31), 9–19.

Dellwo, V., & Fourcin, A. (2013). Rhythmic characteristics of voice between and within languages. *Revue Tranel (Travaux neuchâtelois de linguistique)*, *59*, 87–107.

Dellwo, V., Fourcin, A., & Abberton, E. (2007). Rhythmical classification of languages based on voice parameters. *Proceedings of ICPhS XVI*, 1129–1132.

Dellwo, V., Leemann, A., & Kolly, M.-J. (2012). Speaker idiosyncratic rhythmic features in the speech signal. In *INTERSPEECH* (p. 1584–1587).

Dellwo, V., Leemann, A., & Kolly, M.-J. (2015). Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *The Journal of the Acoustical Society of America*, *137*(3), 1513-1528.

Dellwo, V., & Schmida, S. (2016). Speaker-individual rhythmic characteristics in read speech of German-Italian bilinguals. *TRENDS IN PHONETICS AND PHONOLOGY. STUDIES FROM GERMAN-SPEAKING EUROPE*, 349.

Duez, D. (1999). La fonction symbolique des pauses dans la parole de l'homme politique. *Faits de langues*, *7*(13), 91-97.

Eskénazi, M. (1993). Trends in Speaking Styles Research. In *Proceedings of the 3rd European Conference on Speech Communication and Technology* (p. 501-509). Berlin, Germany.

Fónagy, I. (1983). *La vive voix: essais de psycho-phonétique* (Vol. 20). Payot.

Labov, W. (1972). *Sociolinguistic patterns* (University of Pennsylvania Press). Philadelphia.

Leemann, A., Kolly, M.-J., & Dellwo, V. (2014). Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison. *Forensic Science International*, *238*, 59-67.

Ramus, F. (1999). La discrimination des langues par la prosodie: Modélisation linguistique et études comportementales. *De la caractérisation…… à l'identification des langues*, 131.

Simon, A.-C., Auchlin, A., Avanzi, M., & Goldman, J.-P. (2010). Les phonostyles: une description prosodique des styles de parole en français. *Les voix des Français: en parlant, en écrivant, Bern: Lang*, 71–88.

Xu, Y. (2013). ProsodyPro—A tool for large-scale systematic prosody analysis. In *TASP'2013*. Aix en Provence, France: Laboratoire Parole et Langage, France.