Estimation automatique de caractéristiques acoustiques pour l'étude diachronique du français oral dans les médias

Simon Devauchelle, David Doukhan, Lucas Ondel Yang, Benjamin Élie, Albert Rilliard

Estimation automatique de caractéristiques acoustiques pour l'étude diachronique du français oral dans les médias

Simon Devauchelle [*,**], David Doukhan[**],
Lucas Ondel-Yang[*], Benjamin Élie[***], Albert Rilliard [*]

[*] Université Paris Saclay, CNRS, LISN, France.
[**] Institut national de l'audiovisuel (INA), France.
[***] The University of Edinburgh, UK.

1 Introduction

La description des évolutions diachroniques des langues orales nécessite de relever un certain nombre de défis méthodologiques : obtenir des données suffisamment représentatives des populations en contrôlant des facteurs explicatifs tels que l'âge, le genre, les années et les contextes d'élocution. Les rares corpus collectés à des fins de recherche contiennent plusieurs limitations : périodes d'enregistrement et nombre de locuteur·rice·s limités, absence de certaines catégories de personnes (Pemberton et al., 1998; Berg et al., 2016). Les analyses de ces corpus ont conclu que la hauteur de la voix des femmes avait baissé au cours du XX^e siècle : un résultat relayé par un grand nombre d'études académiques ainsi que par la presse grand public et souvent présenté comme une évidence (Ellis et al., 2023; Robson, 2018). En l'absence de comparaisons avec des mesures réalisées à partir de voix d'hommes, il devient compliqué de comprendre si ces tendances sont exclusives à un genre en particulier ou non.

Si les archives TV et radio peuvent permettre de réduire les biais d'échantillonnage liés aux faibles quantités de données des études précédentes, leur exploitation nécessite de mettre en place des méthodologies d'analyses adaptées aux particularités de ce matériau, pour lequel le contenu lexical n'est pas contrôlé, les prises de parole et l'identité des personnes ne sont pas connues a priori, le signal de parole peut être mélangé à d'autres sources tels que les bruits d'environnement et la musique de fond, susceptibles de perturber les méthodes de description automatique de la voix. Cette communication vise à détailler les méthodes proposées pour analyser la parole dans les fonds TV et radio dans le but de décrire l'évolution des voix des années 1950 à nos jours : qu'il s'agisse de la hauteur de la voix estimée à l'aide de la fréquence fondamentale (f_0) et de la longueur du conduit vocal (VTL), ou encore des paramètres articulatoires (protrusion des lèvres, hauteur du larynx), et tenter de confirmer ou d'infirmer l'hypothèse selon laquelle la voix des femmes a pu baisser au cours du XX^e siècle.

2 Estimation automatique des caratéristiques acoustiques

Le corpus se compose de 111 heures de parole, obtenues par une sélection de 1028 locuteur·rice·s équilibrée en genre et en termes de catégories d'âge (20 - 35, 36 - 50, 51 - 65, < 65

ans), identifiés dans les archives de l'Institut national de l'audiovisuel à l'aide d'une chaîne de traitements semi-automatiques développée par Uro et al. (2022), et d'une aide documentaliste indispensable pour la présélection des documents.

Les segments de paroles ont été transcrits automatiquement en utilisant Whisper (Radford et al. (2022)) puis alignés phonétiquement en utilisant MFA (McAuliffe et al. (2017)). L'évaluation de cette approche sur nos données est associée à un score WER de 12.8% et une précision de 96.9% sur les voyelles (Elie et al. (2024)), des résultats jugés suffisants pour réaliser des études automatisées. Les analyses portent sur les trames centrales de plus d'un million de voyelles en prenant soin d'exclure les phones non voisés grâce à un processus de filtrage décrit par Rilliard et al. (2023). Les quatre premiers formants (F_1, F_2, F_3, F_4) ont été estimés avec l'algorithme de Burg implémenté dans Pratt (par Boersma et Weenink (2024)), en appliquant la méthode d'optimisation proposée par Escudero et al. (2009). À partir de ces formants et des équations proposées par Lammert et Narayanan (2015), des estimations de la longueur du conduit vocal ont également été analysées. Des régressions linéaires mixtes ont ensuite été apprises sur les différentes mesures acoustiques. Cette approche automatique a également donné lieu à une autre étude sur l'évolution des configurations articulatoires en utilisant une technique d'inversion détaillée par Elie et al. (2024), s'intéressant quant à elle à l'évolution de la hauteur du larynx et de la protrusion des lèvres.

3 Résultats et discussions

Un allongement du conduit vocal en fonction du temps est capturé par le modèle – c.-à-d. une baisse de la hauteur de la voix – mais dans le cadre de notre étude, cette observation se constate autant pour les hommes que pour les femmes. On observe également des valeurs de F_1 plus importantes pour les voyelles ouvertes dans les années 1950 comparées aux périodes plus récentes, toujours indépendamment du genre. Au niveau de la f_0, on remarque une baisse des valeurs chez les femmes en fonction de l'âge et le phénomène contraire chez les hommes. Au niveau articulatoire, on observe au cours du temps une baisse du larynx et une faible tendance vers plus de protrusion.

Les résultats de notre analyse acoustique réalisée sur de la parole issue des archives audiovisuelles ne soutiennent pas la thèse d'un changement de la hauteur de la voix au cours du temps qui serait seulement spécifique au genre féminin – comme soutenu par Pemberton et al. (1998) –, mais avancent plutôt l'hypothèse d'une baisse générale de la hauteur de la voix. Néanmoins, notre corpus est principalement constitué d'interviews, provenant de *talk shows* – avec un style de parole propre qui peut donner lieu à des variations acoustiques de la parole (voir Hollien et al. (1997)) – limitant intrinsèquement nos conclusions à ce type de matériel. Par exemple, la baisse observée des valeurs des F_1 au cours du temps peut typiquement s'expliquer par un effort vocal moins important. La cause de cette tendance pourrait en partie résider dans l'évolution des pratiques technologiques médiatiques, comme celle de la distance des microphones aux locuteur·rice·s, plus importante dans les archives plus anciennes.

Pour mieux saisir la subtilité des trajectoires phonétiques et améliorer l'interprétation des changements diachroniques, nous cherchons à nous affranchir du paradigme d'analyse fréquentiste (Candea et al. (2013)). La conclusion de la présentation portera sur les travaux en cours qui privilégient désormais une modélisation acoustique probabiliste de la parole, permettant de mieux isoler les différentes sources potentielles de variation phonétique.

Références

Berg, M., M. Fuchs, K. Wirkner, M. Loeffler, C. Engel, et T. Berger (2016). The speaking voice in the general population : Normative data and associations to sociodemographic and lifestyle factors. *Journal of Voice 31*.

Boersma, P. et D. Weenink (2024). Praat : doing phonetics by computer [computer program]. version 6.4.23. Technical report.

Candea, M., M. Adda-Decker, et L. Lamel (2013). Recent evolution of non-standard consonantal variants in french broadcast news. pp. 412–416.

Elie, B., D. Doukhan, R. Uro, L. Ondel-Yang, A. Rilliard, et S. Devauchelle (2024). Articulatory configurations across genders and periods in french radio and tv archives.

Ellis, S., S. Goetze, et H. Christensen (2023). Moving towards non-binary gender identification via analysis of system errors in binary gender classification.

Escudero, P., P. Boersma, A. S. Rauber, et R. A. H. Bion (2009). A cross-dialect acoustic description of vowels : Brazilian and european portuguese. *The Journal of the Acoustical Society of America 126*(3), 1379–1393.

Hollien, H., P. A. Hollien, et G. de Jong (1997). Effects of three parameters on speaking fundamental frequency. *Journal of the Acoustical Society of America 102*(5), 2984–2992.

Lammert, A. C. et S. S. Narayanan (2015). On short-time estimation of vocal tract length from formant frequencies. *PLOS ONE 10*, 1–23.

McAuliffe, M., M. Socolof, S. Mihuc, M. Wagner, et M. Sonderegger (2017). Montreal forced aligner : Trainable text-speech alignment using kaldi. In *Interspeech 2017*, pp. 498–502.

Pemberton, C., P. McCormack, et A. Russell (1998). Have women's voices lowered across time ? a cross sectional study of australian women's voices. *Journal of Voice 12*(2), 208–213.

Radford, A., J. W. Kim, T. Xu, G. Brockman, C. McLeavey, et I. Sutskever (2022). Robust speech recognition via large-scale weak supervision.

Rilliard, A., D. Doukhan, R. Uro, et S. Devauchelle (2023). Evolution of voices in french audiovisual media across genders and age in a diachronic perspective. In *20th International Congress of Phonetic Sciences (ICPhS)*.

Robson, D. (2018). The reasons why women's voices are deeper today. bbc. *BBC*.

Uro, R., D. Doukhan, A. Rilliard, L. Larcher, A.-C. Adgharouamane, M. Tahon, et A. Laurent (2022). A semi-automatic approach to create large gender- and age-balanced speaker corpora : Usefulness of speaker diarization & identification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France, pp. 3271–3280.

Summary

This communication details the implementation of an automatic pipeline to extract acoustic cues from audiovisual speech excerpts in order to describe the diachronic evolution of spoken French. Acoustic results and difficulties inherent to archive data will be discussed.