# Variabilité inter et intra locuteurs de mesures spectrales et prosodiques en parole lue

Cédric Gendrot, Gabriele Chignoli, Nicolas Audibert, Cécile Fougeron
Laboratoire de Phonétique et Phonologie
19 rue des bernardins 75005 Paris
cedric.gendrot@univ-paris3.fr, gabriele.chignoli@gmail.com,
nicolas.audibert@univ-paris3.fr, cecile.fougeron@univ-paris3.fr

## RESUME

Cette étude préliminaire tente de déterminer des paramètres spectraux et prosodiques qui délimitent la variabilité inter et intra locuteurs pour un corpus de parole lue par 9 locuteurs d'une liste de mots et du texte « la bise et le soleil », enregistré tous les ans depuis 2012, incluant 3 répétitions à chaque enregistrement. Parmi l'ensemble des paramètres testés, nous montrons que s'avèrent pertinents le Centre de Gravité spectral (CoG), la f0 moyenne - pour les paramètres identifiés auparavant dans la littérature-, mais également le ratio harmoniques sur bruit (HNR), l'indice de variabilité temporelle entre les segments (Cnpvi/Vnpvi), ainsi que les variations syntagmatiques de f0.

## ABSTRACT

This preliminary study aims at distinguishing spectral and prosodic parameters that delimitate inter and intra speakers' variability for a corpus of words and a small text ("la bise et le soleil") read by 9 speakers every year since 2012, with 3 repeats for each recording. Amongst all tested parameters, we show that spectral Center of Gravity (CoG), mean f0 – for parameters mentioned previously in the literature-, but also the Harmonics to Noise Ratio (HNR), the index of temporal variability between segments (Cnpvi/Vnpvi), as well as syntagmatic variations of f0.

MOTS-CLES : variabilité inter locuteurs, variabilité intra locuteurs, moments spectraux, rythme, prosodie.

KEYWORDS: inter speakers' variability, intra speakers' variability, spectral moments, rhythm, prosody.

# 1 Etat de l'art

Les travaux en phonétique se concentrent majoritairement sur la recherche d'invariants au sein du signal acoustique pour un phénomène linguistique tel que par exemple l'accentuation et l'organisation


10.21437/JEP.2018-6

prosodique, et visant - au mieux - à la mise en évidence de quelques stratégies bien identifiées. Le signal acoustique de parole est soumis à de fortes variations, parmi lesquelles le contexte segmental, le contexte prosodique, les méthodes d'enregistrement du signal, le type de parole, etc. Dans le cadre de ce travail, nous considérons le locuteur comme un facteur de variation supplémentaire et notre but principal est de déterminer ses limites de variations, se rapprochant ainsi des travaux portant sur la reconnaissance du locuteur. Puisque le choix des extraits de parole peut impliquer une forte variation, nous avons choisi une analyse contrôlée avec un protocole d'enregistrement et un corpus identiques pour tous les locuteurs.

Pour les systèmes automatiques de reconnaissance du locuteur, l'objectif est de chercher les caractéristiques propres au locuteur, il est généralement admis que ces systèmes procèdent en trois phases : une phase de paramétrisation du signal acoustique, avant de déterminer un modèle du locuteur qui permettra au final d'aboutir à la phase de décision. Le taux d'erreurs ('Equal Error Rate' qui délimite le seuil entre le taux de fausses acceptations et de faux rejets) pour des extraits de 2,5 minutes avoisine les 5,3% pour les meilleures séries. Nous ne procéderons à aucune de ces phases, en nous concentrant sur les mesures acoustiques utilisées classiquement en phonétique expérimentale, et qui donnent des indices concrets sur la performance articulatoire ou le timbre du locuteur. Nous nous positionnons dans la continuité du travail de Kahn (2011) qui cherchait les indices acoustiques pertinents pour distinguer plusieurs locuteurs au moyen de la taille d'effet (éta carré ou $\eta^2$, Levine & Hullet, 2002) du facteur locuteur considéré comme variable dépendante lors d'une ANOVA. Les valeurs de F et de p ne sont pas présentées car elles s'avèrent systématiquement significatives et nous cherchons à mettre en valeur l'importance de l'effet, plus que sa significativité.

Dans les systèmes automatiques de reconnaissance du locuteur, l'information temporelle a été majoritairement abandonnée (Haton et al., 2006) ou alors considérée uniquement d'après les variations d'une trame d'analyse à une autre (en moyenne 15 à 20 ms), et nous choisirons ici d'inclure également des mesures prenant en compte la variation temporelle, depuis le niveau phonémique jusqu'à la totalité de la production. Si certains systèmes ont malgré tout tenté d'incorporer des informations temporelles comme la durée des mots, des phonèmes et des pauses (Shriberg et Stolcke, 2008), voire également en modélisant la courbe de fréquence fondamentale (Kockmann et al., 2010), ceux-ci étaient dans tous les cas combinés à des systèmes « classiques », i.e. intégrant des coefficients cepstraux par ailleurs.

Ce travail permettra une meilleure connaissance du facteur locuteur et de ses limites de variation dans un cadre expérimental contrôlé. L'objectif de ce travail pourrait avoir pour but d'améliorer les connaissances en matière de reconnaissance du locuteur, mais également de mieux identifier les invariants dans le signal acoustique. Nous n'aborderons pas ici la capacité humaine à identifier un locuteur par sa voix. Kahn (2011) a montré que les performances des auditeurs dans ce cadre sont excessivement dépendantes des conditions et de la tâche demandée (longueur des extraits présentés, connaissance préalable des locuteurs, état émotionnel des locuteurs, etc.), ainsi que de leur capacités.

Nous cherchons les indices idiosyncratiques des locuteurs contenus dans le signal acoustique. Comme décrit par Kahn (2011), ces indices ne sont pas uniformément répartis dans le signal de parole, ce qui implique que la pertinence des indices acoustiques varie en fonction des extraits de parole. De par le choix d'un corpus lu et identique pour tous les locuteurs, les résultats présentés ici sont considérés comme 'text dependent', mais nous aborderons dans la discussion l'apport que ces résultats peuvent fournir sur des systèmes de reconnaissance du locuteur dits 'text-independent'.

# 2 Protocole expérimental

## 2.1 Corpus

Les données acoustiques analysées ici correspondent à des extraits du corpus PATATRA (Parole AdulTe A TRavers les Ages, élaboré et recueilli au Laboratoire de Phonétique et Phonologie) (Fougeron et al., 2017), qui consiste en un enregistrement annuel de 9 locuteurs (4 hommes et 5 femmes) depuis 2012, soit actuellement 5 années, incluant une liste de mots, un texte lu (« la bise et le soleil »), des exercices de phonation, et une courte séquence de parole spontanée. Seuls la liste de mots et le texte sont utilisés pour l'analyse des données dans le présent travail. Ces items étant systématiquement répétés trois fois pour chaque enregistrement, nous avons donc au total 5 années * 3 répétitions * 9 locuteurs pour une production de 58 mots et un texte. Les locuteurs sont enregistrés au moyen d'un microphone casque AKG C 520, après calibration avec un sonomètre. Ce corpus a été conçu pour évaluer le vieillissement de la voix, mais nous l'exploitons ici pour au contraire valider la cohérence des mesures acoustiques d'une année à la suivante, tout en multipliant les répétitions.

## 2.2 Mesures acoustiques

Dans les travaux de Kahn (2011), les mesures acoustiques effectuées comprennent la f0, le jitter et le shimmer mesurés sur les voyelles, les formants 1 à 4 mesurés sur les voyelles orales, et pour l'ensemble des phonèmes le centre de gravité spectral, la durée, ainsi que les MFCC (Mel Frequency Cepstral Coefficients) et les LFCC (Linear Frequency Cepstral Coefficients).

Dans ses travaux, le centre de gravité spectral a été mesuré comme un paramètre fiable de variation chez le locuteur, notamment pour les fricatives et les nasales (en comparaison des occlusives). Pour les voyelles orales, plus la voyelle est ouverte et antérieure, plus l'effet du locuteur est élevé sur les valeurs de centre de gravité. Ce dernier résultat se confirme sur les valeurs de formants (et notamment F3 et F4). Les transitions formantiques (Mc Dougall, 2006) s'avèrent moins pertinentes. Il apparaît en résumé que les voyelles ne discriminent pas de façon égale les différents locuteurs, les voyelles ouvertes et les voyelles nasales étant les plus informatives. L'hypothèse avancée pour les voyelles nasales est que l'intégration de la cavité nasale par son ouverture fournit une information supplémentaire par la qualité des résonances qui sont propres aux locuteurs. Il en serait de même pour les voyelles ouvertes pour lesquelles le conduit vocal est plus large.

Nous nous proposons de prolonger ces travaux en partant des mesures de CoG afin de comparer nos résultats à ceux de Kahn (2011) dans un premier temps, puis d'étendre nos mesures aux autres moments spectraux ('skewness', 'standard deviation', 'kurtosis'), mais également des mesures de rapport Harmoniques sur bruit (HNR) qui déterminent la proportion de bruit et de voisement dans le signal acoustique, les variations mélodiques (minimum, maximum, étendue et pente de f0) sur les mots et l'énoncé. Toutes les mesures ont été effectuées à l'aide de PRAAT en utilisant les paramètres par défaut, la segmentation en phonèmes a été effectuée par un aligneur (EasyAlign), puis corrigée manuellement par les 2 premiers auteurs. La combinaison de l'effet de ces différents paramètres sera également évaluée puisqu'ils pourraient s'avérer complémentaires dans la distinction entre différents locuteurs. Les statistiques présentées dans les sections suivantes ont été effectuées avec R (version 3.4.2. 2017).

# 3 Analyses

## 3.1 Analyses sur les mots

Une ANOVA est effectuée pour chaque phonème avec comme variable dépendante chaque valeur acoustique (voir 2.2.), et comme variable indépendante (facteur fixe) les locuteurs et l'année d'enregistrement, afin de mesurer l'influence des locuteurs sur les mesures acoustiques. Nous ne présentons pas ici les résultats pour les 3 répétitions par année car leur taille d'effet est systématiquement proche de 0. Les résultats présentés dans les tables 1 et 2 ci-dessous confirment les résultats observés par Kahn (2011), à savoir que les fricatives, les nasales, mais également les sonantes présentent une taille d'effet plus importante que les occlusives, de même les voyelles antérieures, ouvertes et nasales, comparativement aux voyelles fermées et postérieures. Les paramètres de CoG, auquel nous ajoutons le HNR et le maximum de f0 mesurée sur le mot sont donc des facteurs pertinents pour distinguer les locuteurs, notamment pour certains phonèmes comme /m/, /l/ ou /ɑ̃/ Le facteur « année » (d'enregistrement) présente des valeurs de taille d'effet entre 0 et 0.06, ce qui montre que ce facteur n'est pas sensible à nos mesures, il sera donc éludé pour la suite de cet article.

Les autres mesures acoustiques effectuées, à savoir les trois autres moments spectraux ('kurtosis', 'standard deviation' et 'skewness' révèlent des valeurs semblables mais inférieures à celles présentées pour le CoG et ne sont donc pas présentées ici ; de même pour les autres mesures de f0 (moyenne, étendue, minimum et pente mesurés sur chaque mot).

| Taille d'effet | | /b/ | /d/ | /p/ | /t/ | /k/ | /f/ | /v/ | /ʃ/ | /ʒ/ | /l/ | /m/ | /ʁ/ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| COG | Année | 0,00 | 0,00 | 0,11 | 0,00 | 0,02 | 0,02 | 0,06 | 0,00 | 0,00 | 0,01 | 0,01 | 0,02 |
| | Locuteur | 0,15 | 0,27 | 0,13 | 0,04 | 0,07 | **0,43** | 0,29 | **0,42** | 0,40 | **0,45** | **0,47** | 0,32 |
| HNR | Année | 0,01 | 0,01 | 0,01 | 0,02 | 0,01 | 0,01 | 0,02 | 0,02 | 0,01 | 0,00 | 0,02 | 0,03 |
| | Locuteur | 0,33 | 0,31 | 0,07 | 0,12 | 0,03 | **0,13** | **0,28** | **0,33** | **0,26** | **0,33** | **0,49** | 0,12 |
| maximum f0 | Année | 0,00 | 0,00 | | | | | 0,03 | | 0,00 | 0,01 | 0,00 | 0,00 |
| | Locuteur | 0,83 | 0,80 | | | | | **0,62** | | **0,56** | **0,60** | **0,92** | 0,29 |

TABLE 1 : Tailles d'effet du facteur locuteur mesuré sur chaque consonne pour les mesures de CoG, HNR et f0

| Taille d'effet | | i | u | a | ɑ̃ |
|---|---|---|---|---|---|
| COG | Année | 0,00 | 0,00 | 0,03 | 0,01 |
| | Locuteur | **0,47** | **0,40** | **0,55** | **0,62** |
| HNR | Année | 0,01 | 0,01 | 0,01 | 0,02 |
| | Locuteur | **0,33** | **0,31** | 0,07 | 0,12 |
| maximum f0 | Année | 0,00 | 0,00 | 0,00 | 0,04 |
| | Locuteur | **0,83** | **0,80** | **0,64** | **0,96** |

TABLE 2 : Tailles d'effet du facteur locuteur mesuré sur chaque voyelle pour les mesures de CoG, HNR et f0

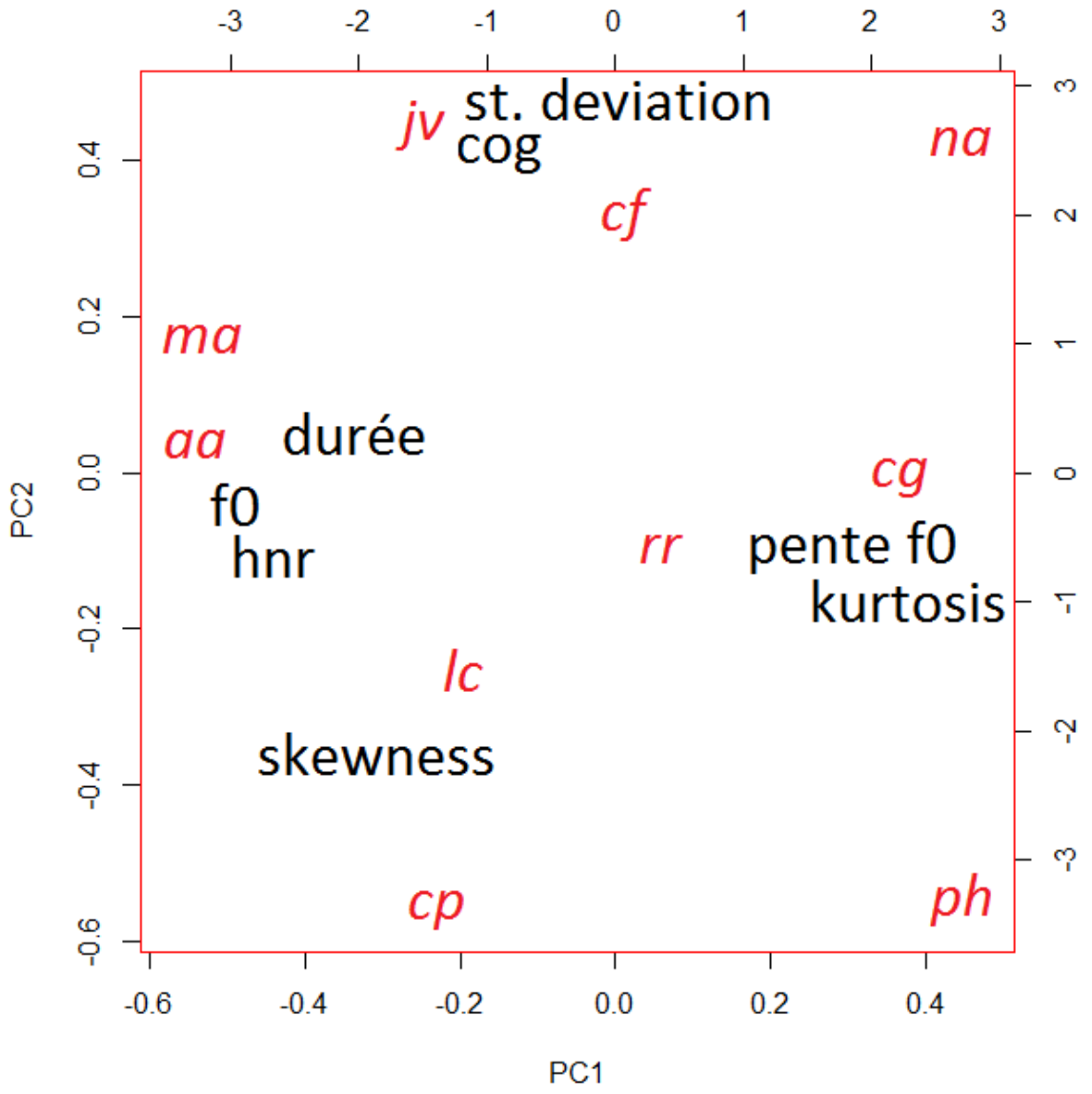


FIGURE 1: Analyse en Composantes Principales des mesures acoustiques effectuées sur /ɑ̃/

Nous présentons maintenant les analyses qui combinent les différents facteurs mesurés afin de tester leur complémentarité. Les analyses ont été effectuées sur / ɑ̃ / car c'est le phonème ayant montré la taille d'effet la plus élevée sur les tables 1 et 2 pour le CoG et le maximum de f0 (seul le HNR est plus pertinent pour les consonnes.

Une Analyse en Composantes Principales (figure 1), avec les valeurs de chaque locuteur moyennées par année et répétition, montre que sont orthogonaux les mesures de CoG et de HNR, alors que les autres paramètres leurs sont corrélés. La proportion de variance de la 3ème composante étant de seulement 8.5% (comparativement à 60% et 21% pour les 2 premières), nous ne présentons ici que les 2 premières composantes principales. Par la suite, nous avons effectué un arbre de classification pour comprendre dans quelle mesure ces paramètres retenus interagissent entre eux pour classifier les

locuteurs en fonction de leur variance. La figure 2 montre que les paramètres de f0, de CoG et de HNR interagissent pour délimiter les différents locuteurs et doivent donc être pris en compte simultanément et non un par un.

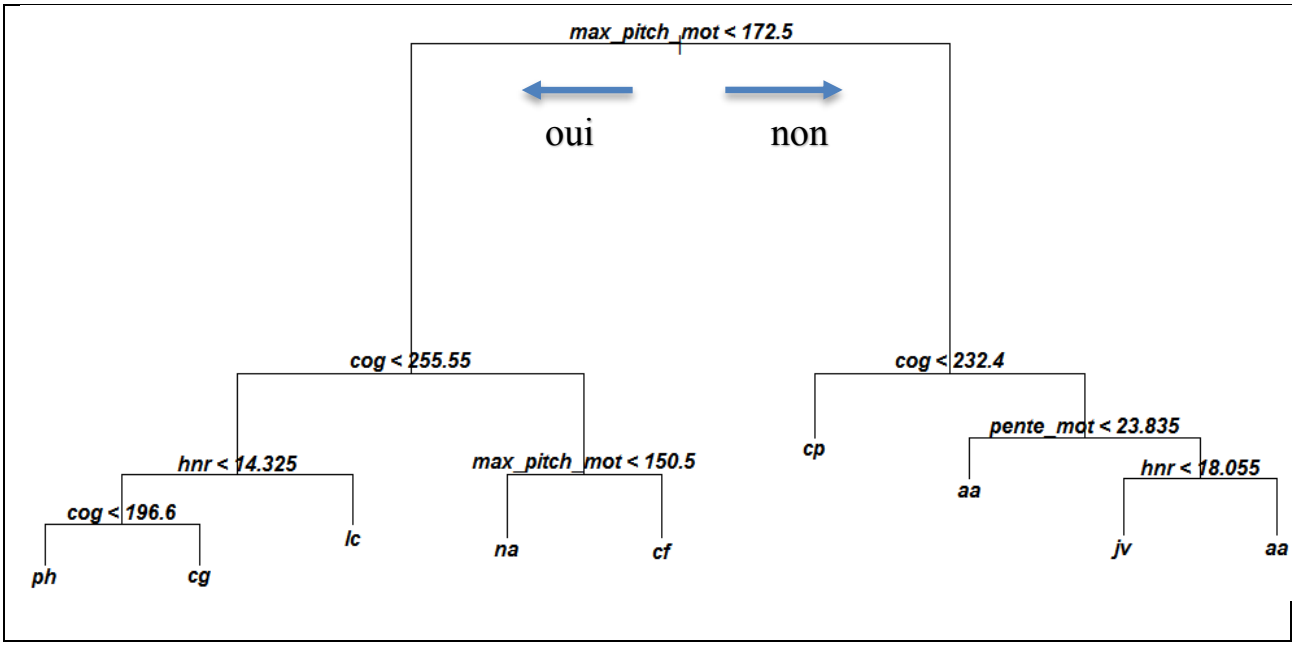


FIGURE 2: Arbre de classification : interaction entre les différentes mesures de CoG et les mesures de HNR effectuées sur la voyelle /ɑ̃/

## 3.2 Analyses sur le texte « La bise et le soleil »

Dans cette section, nous présentons les résultats des analyses effectuées sur le texte « la bise et le soleil ». Seules les 5 dernières séquences « alors le soleil a commencé à briller », « et au bout d'un moment », « le voyageur réchauffé a ôté son manteau », « ainsi la bise a dû reconnaître » et « que le soleil était le plus fort des deux » ont été analysées, les autres étant actuellement en cours de correction manuelle. Les mesures de rythme que nous proposons ici ont été calculées à l'aide de Correlatore 2.2 (Mairano & Romano, 2010) qui permet de calculer %V, ΔV, ΔC, VarcoV, VarcoC, rPVI, nPVI et CCI définis ci-dessous.

- %V : Pourcentage de la durée de la phrase composée de ses intervalles vocaliques
- ΔV/C : Ecart-type des intervalles de durée vocaliques / consonantiques
- VarcoV/C : Ecart-type des intervalles de durée vocaliques / consonantiques divisé par la durée moyenne des intervalles vocaliques / consonantiques (et multiplié par 100)
- CnPVI / VnPVI : Consonant / Vocalic Normalized pairwise variability index; moyenne des différences entre consonnes / voyelles successives, divisées par la moyenne des deux durées)
- CCI : Control/Compensation Index. Normalisation du rPVI; moyenne des différences entre intervalles successifs, où chaque intervalle est divisé par le nombre de segments dans l'intervalle.

Après application de l'ensemble de ces mesures, les variations intra-locuteurs et les différences inter-locuteurs les plus importantes ont été observées pour les mesures de CnPVI et VnPVI que nous présentons dans la figure 3.

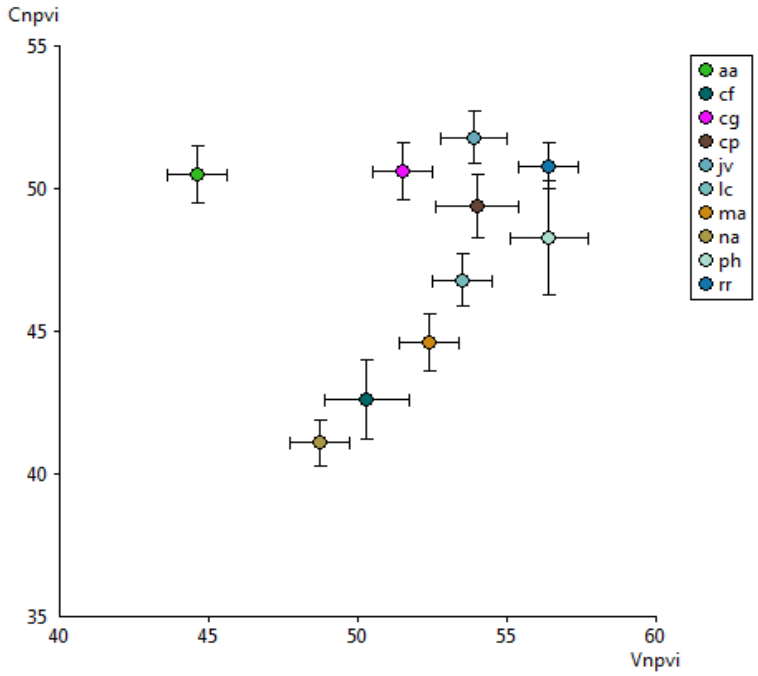


FIGURE 3 : VnPVI sur CnPVI pour le texte « la bise et le soleil »

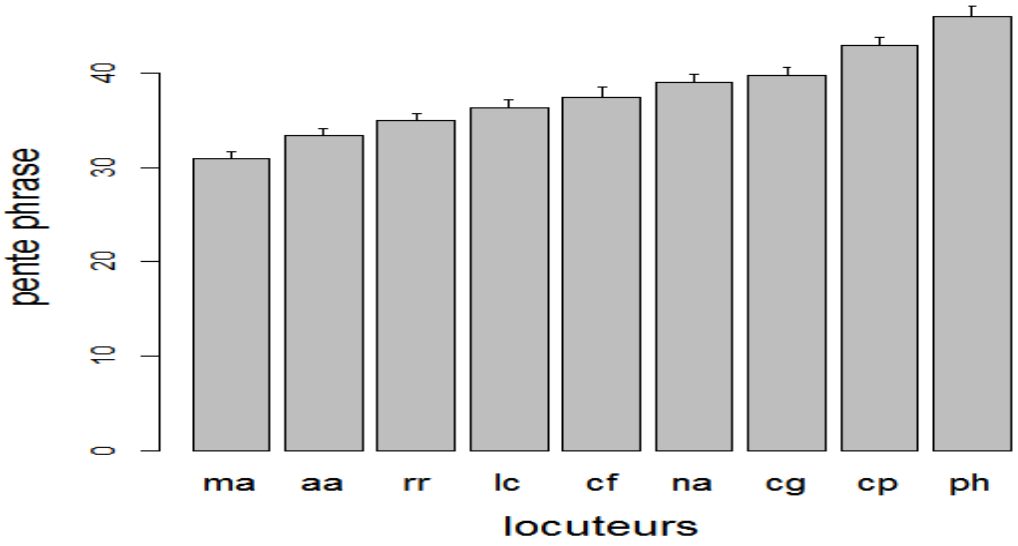


FIGURE 4 : Mesure de pente de f0 (en Hz/s) sur les phrases du texte « la bise et le soleil »

La pente de f0 mesurée comme la différence en demi-tons sur toute la longueur de la phrase (valeur absolue moyenne). La taille d'effet du locuteur sur la pente de f0 mesurée pour une ANOVA avec

comme variables dépendantes la pente de f0 et comme facteurs fixes le locuteur et l'année d'enregistrement est de 0.56. Nous voyons sur la figure 4 comment les locuteurs se répartissent sur les différentes valeurs de pente.

# 4 Discussion et conclusion

Nous avons montré dans cette étude préliminaire que d'autres paramètres s'avéraient pertinents et complémentaires de ceux déjà connus dans la littérature (valeurs moyenne de f0, CoG, formants), à savoir le HNR (ratio harmoniques sur bruit), la pente de f0 mesurée sur le mot ou la phrase et le rythme (VnPVI/CnPVI) dans l'énoncé, ainsi que la combinaison de ces paramètres. Il sera bien sûr nécessaire de tester ces paramètres sur un nombre plus important de locuteurs. Nous avons également montré que les autres moments spectraux ('skewness', 'kurtosis' et 'standard deviation') n'apportent qu'une information redondante par rapport au centre de gravité spectral. Comme observé par Kahn (2011), les segments qui montrent une taille d'effet plus conséquente dans notre corpus sont les nasales (consonnes et voyelles), les sonantes, et dans une moindre mesure les fricatives. Il semble que les segments porteurs d'informations idiosyncratiques relèvent souvent des différences morphologiques, comme par exemple la forme des fosses nasales pour les voyelles nasales, la forme des dents et du palais pour les fricatives dentales et palatales, etc. En effet, les caractéristiques propres au locuteur consistent en deux éléments principaux : les variations physiques statiques dans le signal qui correspondent aux caractéristiques physiques des articulateurs mises en évidence ici par le CoG, le HNR et la valeur moyenne de f0, alors que les variations dynamiques reflètent les différences comportementales qui seraient plus vraisemblablement révélées par le rythme et la pente de f0.

Kahn (2011) soulignait en conclusion de sa thèse la nécessité de distinguer le locuteur et l'extrait de parole dans une tâche d'identification des indices idiosyncratiques du locuteur. Elle mentionnait également que plus la parole est contrôlée, plus le locuteur est contraint dans son énonciation, plus le risque de ne pas trouver d'indices discriminants pour le locuteur est grand. Le corpus PATATRA comprend également des séquences de parole spontanée pour chaque enregistrement que nous analyserons dans la suite de ce travail, le but premier ayant été ici de déterminer les paramètres acoustiques les plus pertinents et leur complémentarité dans des contextes phonémiques contrôlés. Les mesures présentées ici sont pour la plupart des mesures dites 'text-dependent' puisqu'elles impliquent un étiquetage et un alignement en phonèmes préalable, mais les mesures liées à la f0 (maximum et pente) pourraient être effectuées sans connaissances linguistiques à priori. Dans la suite de ce travail, nous viserons également à tester si les mesurées classiquement effectuées par les systèmes automatiques de reconnaissance du locuteur, à savoir les MFCC/LFCC (MEL Frequency /Linear Predictive Cepstral Coefficients), montrent un effet du locuteur plus important que les paramètres mesurés ici.

# Remerciements

Nous remercions l'ANR VOXCRIM (ANR-17-CE39-0016) ainsi que le LaBeX Empirical Foundations of Linguistics (EFL).

# Références


FOUGERON, C., DELVAUX, V., GENDROT, C., LAGANARO, M., MENARD, L. (2016). Introducing two databases of spoken French throughout adulthood. in Workshop on Speech Perception and Production across the Lifespan, London, England, 2017

HATON, J., CERISARA, C., FOHR, D., LAPRIE, Y. ET SMAÏLI, K. (2006). Reconnaissance automatique de la parole, Du signal à son interprétation. Paris : Dunod.

KAHN, J. (2011). Parole de locuteur : performance et confiance en identification biométrique vocale, Thèse de Doctorat, Avignon.

KOCKMANN, M., BURGET, L. ET CERNOCKY, J. (2010). Investigations into prosodic syllabe contour features for speaker recognition. in International Conference in Acoustics, Speech and Signal Processing (ICASSP), Dallas, 4418–4421.

LEVINE, T. ET HULLET, C. (2002). Eta squared, partial eta squared, and misreporting of effect size in communication research. Human Communication Research 28(4), 612–625.

MAIRANO, P. ET ROMANO, A. (2010). Un confronto tra diverse metriche ritmiche usando Correlatore. In: Schmid, S., Schwarzenbach, M. & Studer, D. (eds.) La dimensione temporale del parlato, Proc. of the V Natioanl AISV Congress (Associazione Italiana di Scienze della Voce) (University of Zurich, Collegiengebaude, 4th-6th February 2009), Torriana (RN): EDK, 79-100.

MCDOUGALL, K. (2006). Dynamic features of speech and characterization of speakers : towards a new approach using formant frequencies. Speech, Language and the Law 13, 89–126.

R CORE TEAM (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

SHRIBERG, E. ET STOLCKE, A. (2008) The case for automatic Higher-Level features in forensic speaker recognition. Dans les actes de International Conference on Speech Communication and Technology (Interspeech), Brisbane, 1509–1512.