Caractérisation du locuteur par CNN à l'aide des contours d'intensité et d'intonation : comparaison avec le spectrogramme

Gabriele Chignoli, Cédric Gendrot, Emmanuel Ferragne

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Caractérisation du locuteur par CNN à l'aide des contours d'intensité et d'intonation : comparaison avec le spectrogramme

Gabriele Chignoli Cédric Gendrot Emmanuel Ferragne
Laboratoire de Phonétique et Phonologie, 19 rue des Bernardins, 75005 Paris, France
gabriele.chignoli@sorbonne-nouvelle.fr,
cedric.gendrot@sorbonne-nouvelle.fr, emmanuel.ferragne@u-paris.fr

RÉSUMÉ

Dans ce travail nous avons recours aux variations de f0 et d'intensité de 44 locuteurs francophones à partir de séquences de 4 secondes de parole spontanée pour comprendre comment ces paramètres prosodiques peuvent être utilisés pour caractériser des locuteurs. Une classification automatique est effectuée avec un réseau de neurones convolutifs, fournissant comme réponse des scores de probabilité pour chacun des 44 locuteurs modélisés. Une représentation par spectrogrammes a été utilisée comme référence pour le même système de classification. Nous avons pu mettre en avant la pertinence de l'intensité, et lorsque les deux paramètres prosodiques sont combinés pour représenter les locuteurs nous observons un score qui atteint en moyenne 59 % de bonnes classifications.

ABSTRACT

CNN speaker characterisation through prosody : spectrogram comparison

In this study we focused on f0 and intensity variation in four-second spontaneous speech sequences by 44 French speakers in order to evaluate the strength of prosodic parameters for speaker characterisation. We used a deep Convolutional Neural Network (CNN) with f0 and/or intensity values as input in a classification task, where the system had to classify 44 speakers in a closed dataset. Spectrograms were also used as input with the same CNN architecture as a benchmark for maximum possible performance. Results show that f0 and intensity are complementary as together they yield 59 % classification precision.

MOTS-CLÉS : Caractérisation du locuteur, intensité, f_0, prosodie, réseaux de neurones convolutifs.

KEYWORDS: Speaker characterisation, intensity, fundamental frequency, prosody, convolutional neural network.

1 Introduction

Les recherches sur la caractérisation phonétique du locuteur ont fréquemment pour objectif d'identifier des facteurs cohérents expliquant la variabilité inter-individuelle (Kahn, 2011). Ces facteurs comprennent notamment des mesures acoustiques issues de variables d'ordre biologique telles que l'âge ou des troubles de la parole. Par exemple, dans Schötz (2007), parmi plusieurs mesures acoustiques étudiées en relation avec l'âge, l'étendue de la pression acoustique se montre stable. Le style de parole (Arvaniti & Rodriquez, 2013) ou le sexe sont également des critères de variation importants. Dans Keating & Kuo (2012), des locuteurs hommes et femmes natifs de l'anglais ou du mandarin sont comparés à partir de mesures de fréquence fondamentale dans des tâches de lecture de mots isolés,

Volume 1 : Journées d'Études sur la Parole, pages 91–99.
HAL : hal-02798521.

de texte et de vocalisation dans différents contextes. Plusieurs mesures de ce descripteur, valeurs moyennes, maximales, minimales, moyenne de l'étendue ou écart type sont comparées dans deux analyses de variance prenant en compte la langue et le sexe comme facteurs de variabilité inter-locuteurs. Les variations de f_0 d'ordre physiologique sont jugées similaires dans les deux langues par les auteurs.

Étant donné la robustesse de la f_0 face à plusieurs facteurs perturbateurs comme les canaux de diffusion, l'effort vocal ou l'état d'esprit du locuteur (Lindh & Eriksson, 2007), elle s'est affirmée comme un des descripteurs les plus analysés non seulement pour essayer de différencier le sexe des locuteurs mais aussi pour distinguer les locuteurs de même sexe. Hudson *et al.* (2007) montrent comment la distribution des valeurs de f_0 pour 100 locuteurs hommes reste cohérente indépendamment de la tâche exécutée, avec la majorité des locuteurs (65 %) qui présentent une variation intra-individuelle ne dépassant pas les 20 Hz. Niebuhr & Skarnitzl (2019) analysent plusieurs mesures de f_0 sur 51 locuteurs, hommes et femmes, lors d'une tâche de conversation semi-spontanée pour décrire comment celles-ci peuvent représenter la variabilité des locuteurs. Les mesures prises en compte, dont la moyenne, l'étendue (pour 80% des locuteurs) et le kurtosis de la distribution des valeurs de f_0 pris tous les 5 ms, s'avèrent utiles pour expliquer la variabilité inter-locuteurs. Les résultats de cette étude montrent également que la f0 est un paramètre très dépendant du sexe. Dans Adami *et al.* (2003), c'est la complémentarité des indices qui est analysée avec 40 locuteurs, 20 hommes et 20 femmes, dont le contour intonatif et l'énergie de 15 mots sont utilisés pour modéliser les locuteurs et successivement tester un système de reconnaissance automatique du locuteur pour une des tâches du protocole NIST 2001. Les deux paramètres sont analysés d'abord séparément puis ensemble et montrent une forte complémentarité puisque l'erreur totale passe de 13 % à 3 %.

Contrairement à la f_0, l'intensité est un facteur prosodique peu utilisé dans la description de la variabilité inter-locuteurs. Sorin (1981) montre la cohérence de ce paramètre et la sensibilité de l'oreille humaine à sa variation ; cependant le rôle linguistique de l'intensité reste à approfondir. Pardo (2006) cite l'intensité comme l'un des facteurs pouvant potentiellement déterminer la similarité entre locuteurs. Dans la lignée de ces considérations sur l'intensité, Tweedy & Culling (2014) montrent comment ce facteur reste peu influencé chez des locuteurs soumis à des perturbations lors de tâches de conversation. Plus récemment nous retrouvons dans He *et al.* (2015) l'utilisation de mesures d'intensité parallèlement à des corrélats rythmiques pour la classification de 16 locuteurs à travers un système de réseaux de neurones. Dans cette étude, l'intensité permet d'atteindre un taux de bonnes classifications de 30 % qui augmente à 36 % lorsque les corrélats rythmiques et d'intensité sont observés ensemble. Les mesures utilisées se rapprochent des corrélats rythmiques déjà utilisés par les auteurs, nous renvoyons à cette étude et à d'autres (Dellwo *et al.*, 2015) pour plus de précision à ce sujet. La complémentarité des mesures prosodiques est un des points fondamentaux de l'étude que nous allons présenter puisque la production de la parole, et dans notre cas, la représentation de la variabilité inter-individuelle ne peut être expliquée que par l'interaction de plusieurs facteurs.

Dans ce travail nous essayons d'apporter un élément d'analyse supplémentaire à la caractérisation du locuteur à partir d'indices prosodiques, pour cela nous allons prendre en compte deux dimensions prosodiques classiques : la f_0 et l'intensité. Choisir ces deux mesures nous permettra de comparer leurs résultats respectifs et leur éventuelle complémentarité dans la caractérisation du locuteur. Dans la section suivante nous décrirons de manière plus précise les mesures acoustiques qui ont été sélectionnées, leurs méthodes d'extraction et de classification. La complémentarité entre les résultats des différentes mesures sera présentée dans la section 3 ainsi qu'une analyse plus spécifique de certains locuteurs et cas particuliers.

2 Protocole expérimental

Le corpus sur lequel nous avons travaillé est composé d'extraits de 4 secondes provenant de chacun des 44 locuteurs du corpus NCCFr (Torreira *et al.*, 2010) (pour les 45ème et 46ème les transcriptions n'étant pas intégralement présentes, nous les avons écartés). Nous utiliserons les chiffres de 1 à 44 pour identifier les locuteurs à la place des codes présents dans la base. Le corpus NCCFr se compose de conversations sur des thèmes divers entre binômes ou trinômes d'amis, dont la durée totale est d'une heure en moyenne, et enregistrées dans une chambre insonorisée à l'aide d'un micro casque, ce qui limite les variations d'intensité dûes aux mouvements de la tête. Les transcriptions ont été effectuées de manière semi-automatique au laboratoire LIMSI (Gauvain *et al.*, 2002). Nous avons considéré uniquement les extraits comportant un minimum de 20 phonèmes, en dessous de ce seuil, après écoute des extraits, nous avons souvent remarqué la présence de la voix des autres locuteurs de l'enregistrement. À partir de ces extraits nous avons obtenus 4 types de données que nous avons utilisées dans un système de classification à l'aide de réseaux de neurones : le spectrogramme complet de chaque séquence ; le contour de f_0 ; les valeurs d'intensité ; la représentation conjointe de f_0 et d'intensité. Le spectrogramme servira ici de référence puisqu'il comprend l'information la plus complète sur la séquence analysée.

Ces représentations ont été extraites avec les valeurs par défaut de Praat (Boersma, 2001) et converties en images codées sur 8 bits en niveaux de gris afin de préserver la mémoire du GPU. Pour l'extraction des spectrogrammes nous avons utilisé une fenêtre avec des trames de 5 ms avec un chevauchement de 90 % et une fréquence d'échantillonage de 16 kHz. Sur chaque trame de signal a été appliquée une fenêtre de Hamming pour obtenir 512 échantillons sur lesquels une FFT a été effectuée. Nous n'avons pas appliqué de pré-emphase et la dynamique a été fixée à 70 dB. Dans le but de ne pas saturer la mémoire du GPU, les spectrogrammes ont été redimensionnés à 800×257 pixels ($L \times H$), où chaque pixel correspond à 5 ms sur l'axe du temps (après compression de 8000 à 800 par interpolation bicubique), et à 31.13 Hz sur celui de la fréquence. Le même redimensionnement sur l'axe horizontal est utilisé pour les répresentations d'intensité et de f_0, sur l'axe vertical nous avons 1 seul pixel dont les niveaux de gris correspondent aux valeurs d'intensité ou de f_0. Pour la répresentation conjointe d'intensité et f_0 le format des images passe à 800×2 pixels.

Le modèle que nous avons utilisé pour la classification est une version légèrement modifiée d'un réseau de neurones convolutif avec architecture ResNet-18, à la base employé pour de la reconnaissance d'images. Les données de chacun des 44 locuteurs ont été aléatoirement divisées selon des ensembles d'entraînement (70 %), validation (10 %) et test (20 %). Le même découpage est utilisé les quatre expériences (f_0, intensité, f_0 et intensité, spectrogrammes). L'entraînement du réseau a été effectué avec l'optimiseur Adam et des mini-batches de taille 32 sur un GPU GTX 1080. Le nombre maximum d'itérations défini étant de 30, l'entraînement du modèle s'arrête avant cette limite si la valeur de la perte atteint un chiffre inférieur ou égal à sa valeur minimale dans les 10 dernières itérations. Tous les modèles ont été entraînés et testés à l'aide de la MATLAB Deep Learning Toolbox (Mathworks, 2019).

Pour pouvoir comparer et interpréter les résultats obtenus par la classification automatique nous avons aussi extrait des valeurs acoustiques à l'aide du logiciel Praat (Boersma, 2001) à partir des mêmes séquences de 4s. L'analyse de celles-ci nous a permis de comprendre dans quelle mesure les locuteurs se différencient ou se rapprochent du point de vue acoustique et ainsi d'interpréter les confusions faites par la classification automatique avec une approche phonétique. L'ensemble des mesures effectuées pour la f_0 et l'intensité comprend : les valeurs minimales, maximales, la moyenne,

l'écart type, le premier décile, le dernier, le décile du milieu et la différence entre les valeurs de f_0 et d'intensité (voir paragraphe 3.2). Nous avons aussi extrait d'autres valeurs comme la pente de f_0, le nombre de pics de f_0 et d'intensité mais nous les avons écartées de notre analyse finale car elles n'apportent pas d'éléments intéressants à celle-ci.

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	
Intonation	5	15	17	8	44	14	10	30	24	72	17	
Intensité	37	15	37	25	50	31	22	40	35	45	41	
Intens-into	45	41	77	50	88	62	62	88	80	82	40	
Spectrogramme	88	97	92	97	97	88	98	92	97	98	92	
	12	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	
Intonation	5	27	27	21	62	34	25	27	20	37	25	
Intensité	45	10	12	27	67	88	47	47	5	15	37	
Intens-into	72	48	51	51	87	100	71	65	51	42	70	
Spectrogramme	97	81	90	87	97	98	98	98	92	95	95	
	23	**24**	**25**	**26**	**27**	**28**	**29**	**30**	**31**	**32**	**33**	
Intonation	15	12	11	41	20	55	37	34	15	25	8	
Intensité	10	10	4	55	44	34	44	17	42	8	18	
Intens-into	47	68	28	78	57	95	87	54	44	22	44	
Spectrogramme	94	97	94	92	94	91	95	95	92	85	98	
	34	**35**	**36**	**37**	**38**	**39**	**40**	**41**	**42**	**43**	**44**	**Tot.**
Intonation	4	20	44	68	25	37	42	25	25	62	20	28
Intensité	17	8	22	62	24	62	12	31	21	61	48	32
Intens-into	27	37	62	62	50	44	60	31	28	80	48	59
Spectrogramme	88	75	91	97	95	94	77	92	100	97	92	93

TABLE 1 – Résultats obtenus par le réseau de neurones lors de la classification des différentes représentations, pourcentage de réussite pour l'ensemble des 44 locuteurs

3 Résultats

La tâche de classification consiste, pour chacun des 44 locuteurs, à donner au réseau de neurones les 70 extraits pris aléatoirement dans l'ensemble de départ et ne faisant pas partie des extraits d'entraînement. Cette même expérience a été répétée 4 fois pour chacune des représentations. La première considération concerne le score obtenu par les spectrogrammes, qui atteint un taux de bonnes classifications de 93 %, résultat attendu en considérant que cette représentation décline plusieurs éléments de la production de la parole alors que les 3 autres ne prennent en compte qu'une dimension voire deux à la fois. L'utilisation du contour intonatif, la répresentation à la fois des valeurs moyennes et des modulations de f_0, atteint le résultat moyen le moins élevé avec 28 %. Les valeurs d'intensité apportent quelques informations supplémentaires en totalisant un taux total moyen de 32 %, mais c'est lorsque les deux représentations sont combinées que le score est presque doublé et atteint 59 %.

3.1 Complémentarité des paramètres prosodiques

En considérant le spectrogramme comme la valeur de référence, nous remarquons deux locuteurs qui sont malgré tout mieux reconnus grâce à la prosodie seule : les locutrices 17 et 28. Elles présentent des profils similaires même si l'une se caractérise de manière plus marquée à travers l'intensité : la locutrice 17 obtient 88 % de bonnes classifications à partir des valeurs d'intensité et 34 % à partir de l'intonation, contre respectivement 24 % et 39 % pour la locutrice 28. Lorsque les deux dimensions sont combinées la locutrice 17 enregistre 100 % de bonnes classifications alors que la locutrice 28 obtient 95 %.

La locutrice 17 est caractérisée par des valeurs d'intensité toujours maximalement basses. La Figure 1(a) met en évidence cette tendance avec les valeurs minimales d'intensité pour les 44 locuteurs, nous prenons en exemple cette mesure puisque, avec les valeurs maximales et le dernier décile, elles permettent de distinguer de manière nette quelques groupes de locuteurs. Nous comprenons aussi pourquoi lors de la classification de la locutrice 17 la seule confusion existante se fait avec la locutrice 16, son binôme d'enregistrement, qui apparaît également dans la partie gauche de ce schéma avec des valeurs bien inférieures à la moyenne. Ceci nous renvoie aux considérations faites plus haut sur la convergence phonétique de l'intensité. Nous constatons la même tendance à avoir des valeurs d'intensité proches pour d'autres binômes : 25 et 26, 32 et 33, 34 et 35, 41 et 42. Cependant l'observation des valeurs de f_0 pour la locutrice 17 nous fait remarquer qu'elles ne s'écartent pas de la moyenne et forment un groupe compact avec les locutrices 24, 28, 35 et parfois 25. Nous observons cette fois une certaine distance avec la locutrice 16, qui quant à elle fait partie d'un groupe plus large. Cette différence évidente nous montre un exemple saillant de la complémentarité entre les mesures de f_0 et d'intensité, puisque grâce à l'une nous pouvons éliminer les confusions obtenues avec l'autre. Pour la locutrice 28, en utilisant la prosodie, nous n'observons pas une élimination totale

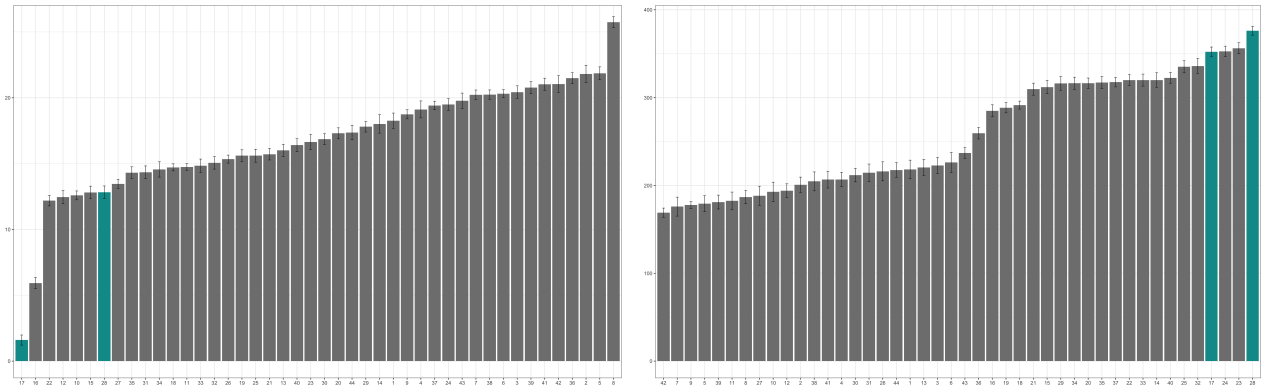

FIGURE 1 – Comparaisons des valeurs minimales d'intensité, en dB (Figure 1a, à gauche) et des valeurs maximales de f_0, en Hz (Figure 1b, à droite) pour les 44 locuteurs. Les locutrices 17 et 28 mises en évidence

des confusions mais plutôt une réduction du nombre de confusions effectuées par la classification. D'un côté, pour ses valeurs d'intensité, nous observons un petit groupe de locuteurs ayant des valeurs très basses mais rapprochés entre eux (Figure 1a) : il s'agit des locuteurs 10, 12 (les deux étant des hommes), 15, 22 et 28. Nous retrouvons tous ces locuteurs dans l'ensemble des confusions faites à partir des valeurs d'intensité pour la locutrice 28, mais aucun d'entre eux n'apparaît dans les résultats

obtenus à partir de la f_0. Pour cet autre paramètre nous retrouvons une confusion importante avec les locutrices 17 et 24, avec qui 28 est confondue respectivement 12 et 8 fois. Ces deux dernières locutrices, comme observé précédemment, font partie d'un groupe assez compact dont la locutrice 28 s'écarte grâce à ses valeurs maximales de f_0, que nous présentons en Figure 1(b). Le décile moyen et la valeur moyenne de f_0 sont aussi des valeurs permettant de distinguer certains groupes de locuteurs.

À partir des valeurs de f_0 une distinction nette entre hommes et femmes peut être faite. Effectivement, nous voyons au milieu de la Figure 1(b) un écart entre le dernier locuteur, 43, et la première locutrice, 36. Par conséquent, nous comprenons que les confusions entre locuteurs de sexes différents faites à partir des valeurs d'intensité peuvent disparaître avec cette autre dimension prosodique. La seule similarité qui n'est pas complètement effacée par la prosodie dans le cas de la locutrice 28, est avec la locutrice 24. Ces deux locutrices sont très écartées selon les valeurs d'intensité mais confondues dans tous les autres cas, même à partir des spectrogrammes. À partir des mesures que nous avons à notre disposition lors de cette étude nous pouvons affirmer que ces deux locutrices se distinguent de l'ensemble des locuteurs par leur intonation mais nous n'arrivons pas à déterminer quel élément prosodique peut nous permettre de les différencier nettement entre elles.

3.2 Cas de non-complémentarité

Les deux cas que nous venons d'évoquer montrent les deux tendances de la majorité de nos locuteurs : une caractérisation accentuée par l'intensité suivie d'un apport mineur mais complémentaire de la f_0 pour réduire les confusions ; ou des taux de classification légèrement différents entre les deux mesures. Dans les deux cas nous observons une augmentation importante du taux de bonnes réponses lorsque la représentation conjointe de f_0 et d'intensité est utilisée.

Cependant sur les 44 locuteurs, nous avons remarqué 6 cas dans lesquels la prosodie totalise un score égal ou inférieur à celui des deux mesures prises singulièrement : les locuteurs 11, 39, 41, 44 et les locutrices 32, 37. Parmi ces locuteurs, un premier groupe comprend les locuteurs 41 et 44 dont l'intensité totalise le même score des deux dimensions conjointes, avec des taux de bonnes classifications respectivement de 31 % et de 48 %. D'un autre côté, nous avons les locuteurs 11 et 39 qui ont aussi une meilleure classification à partir de l'intensité mais des scores inférieurs dans tous les autres cas. Enfin les locuteurs 32 et 37 atteignent un score plus important à partir de leur f_0.

Pour comprendre quels éléments influencent la non complémentarité nous avons décidé d'observer le locuteur 11 et ses confusions avec le locuteur 27, puisque ce dernier présente quant à lui des indices prosodiques complémentaires. Selon ses valeurs de f_0, le locuteur 11 est plus souvent identifié comme le 27 et inversement. En reprenant la Figure 1(a) comme exemple, nous observons que ces deux locuteurs ont des valeurs très proches et légèrement moins élevées que la moyenne. La f_0 et l'intensité du locuteur 11 ne montrant pas de complémentarité, il continue à être confondu avec le locuteur 27 à partir de la représentation conjointe de ces indices alors que cet autre locuteur atteint un taux de reconnaissance beaucoup plus important et ne montre pas de confusion avec le locuteur 11.

Nous avons essayé d'expliquer cette non complémentarité en analysant la relation entre les variations d'intensité et de f_0 à l'intérieur des séquences analysées. Nous avons observé cette relation sous forme de ratio et de différence des valeurs de f_0 et d'intensité. Nos observations montrent que le ratio n'apporte aucune information supplémentaire, pas de significativité statistique observée contrairement à la valeurs de la différence, et nous l'avons exclu de l'analyse statistique présentée dans le paragraphe suivant. La différence nous permet d'écarter les locuteurs 11 et 27 et d'autres cas similaires. Cependant,

ceci ne nous permet pas d'expliquer complètement pourquoi ces locuteurs sont moins caractérisés par la prosodie. Des indices contenus dans le spectrogramme – mais qui ne font pas partie de nos indices prosodiques – permettent ainsi d'améliorer les taux de classification de ces locuteurs. Effectivement, les locuteurs 11 et 27 atteignent respectivement 40 % et 57 % de bonnes classifications avec les deux indices conjoints alors qu'à partir du spectrogramme, ils atteignent respectivement 92 % et 94 %. Ces locuteurs, mais aussi les autres cités dans ces cas de non-complémentarité, atteignent à travers le spectrogramme plus de 90 % de bonnes classifications, nous pouvons supposer par conséquent qu'ils font partie d'un groupe de locuteurs dont la prosodie seule ne peut pas expliquer la variation et pour lesquels une représentation globale est nécessaire.

3.3 Analyse en Composantes Principales

Nous avons effectué une Analyse en Composantes Principales pour l'ensemble des mesures acoustiques afin de mieux comprendre leurs interactions. La Figure 2(a) montre les valeurs moyennées pour chacun des 44 locuteurs dans l'espace obtenus à partir des deux premières composantes. Nous n'avons retenu que les deux premières pour la représentation puisqu'elles expliquent ensemble 87.5 % de la variance, respectivement 55.3 % et 32.2 %. La Figure 2(b) montre la contribution des mesures acoustiques à la définition de chacune des composantes. Nous observons que les mesures de f_0 et la différence entre f_0 et intensité, dernière ligne du tableau, contribuent de manière importante à la première dimension. L'intensité contribue plutôt à la définition de la deuxième composante. Cette division nous confirme la non corrélation et par conséquent la complémentarité des deux paramètres. Seul l'écart type de l'intensité contribue de manière importante à définir la troisième composante, restant presque absent des deux premières. Tous les locuteurs sont représentés par les points dans la

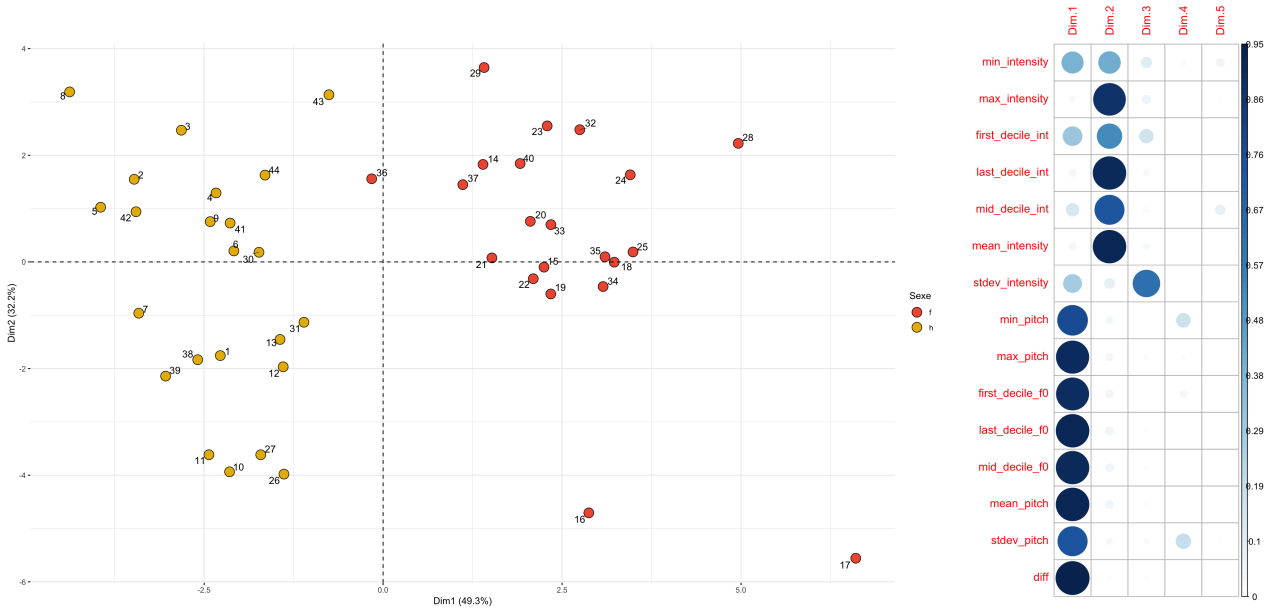

FIGURE 2 – Résultats de l'Analyse en Composantes Principales. À gauche (Figure 2a) la distribution des individus selon les dimensions 1 et 2 est représentée. La Figure de droite (2b) montre la contribution de nos mesures acoustiques à la définition de chaque composante

Figure 2(a) où l'espace défini par les deux premières composantes permet d'isoler 10 locuteurs : 3, 7, 8, 16, 17, 24, 28, 29, 36 et 43. Nous remarquons la correspondance entre certains de ces locuteurs et les cas de complémentarité des indices mis en évidence par la classification automatique.

La contribution des indices acoustiques n'étant pas la même pour les deux premières composantes, nous observons l'axe horizontal de l'ACP qui respecte la division par sexe grâce aux mesures de f_0. Sur le côté gauche, en jaune, sont disposés tous les locuteurs et sur la droite en rouge toutes les locutrices. Comme observé précédemment aussi (Figure 1b), le locuteur 43 et la locutrice 36 marquent la limite entre les deux sexes. À travers cette analyse nous pouvons également observer comment les paramètres prosodiques permettent de différencier 10 groupes de locuteurs plus ou moins compacts. Certains locuteurs sont plus proches que d'autres dans la représentation de l'ACP. Nous retrouvons parmi les groupes de locuteurs les mêmes confusions présentes lors de la classification à travers les deux paramètres conjoints. Certaines confusions, 38-39 et 11-10 par exemple, montrent une convergence phonétique pour les locuteurs d'un même binôme mais ne représentent qu'une minorité. Ceci renforce les observations faites dans le paragraphe précédent, concernant certains locuteurs qui sont moins caractérisés à partir de leur prosodie mais pouvant être ainsi associés entre eux et ensuite différenciés selon d'autres paramètres présents dans la production de la parole.

4 Discussion et conclusion

Dans ce travail nous avons présenté une analyse acoustique pour caractériser des locuteurs à partir de f_0 et d'intensité. L'utilisation d'une classification par réseaux de neurones et l'analyse des confusions opérées par le réseau nous ont aidé à comprendre dans quelle mesure le système utilise la prosodie comme facteur déterminant pour classer les locuteurs. Si la f_0 se montre très efficace pour trancher entre des locuteurs de sexes différents l'intensité atteint un score global de classification plus élevé. Nous avons ici analysé la contribution de ces dimensions prosodiques face au spectrogramme qui représente la référence puisqu'il englobe d'autres dimensions en plus de la prosodie et permet ainsi d'expliquer la majorité de la variation. Pour le sous-groupe de données que le réseau n'arrive pas à classer correctement en se basant sur le spectrogramme (7 % des données totale) l'utilisation conjointe du contour de f_0 et d'intensité arrive à donner 33 % de bonnes classifications.

La f_0 et l'intensité sont des paramètres classiques utilisés pour décrire deux aspects différents de la prosodie. Nous avons pu observer, à travers notre étude, comment l'utilisation d'une technique de classification non classique en phonétique peut faire apparaître des nouvelles caractéristiques liées à la description prosodique des locuteurs. L'Analyse en Composantes Principales, présentée en Figure 2, montre comment avec une approche statistique les mesures de f_0 contribuent à la description d'une partie conséquente de notre ensemble. L'intensité obtient un meilleur score que la f_0 lors de la classification par réseaux de neurones mais elle est moins présente dans l'ACP.

Pour diminuer cette disparité, il est envisageable d'aller regarder de manière plus approfondie d'autres mesures d'intensité plus robustes que celles que nous avons pu présenter. Nous avons observé pour ce paramètre que les valeurs minimales et le dernier décile sont les mesures plus précises pour la caractérisation des locuteurs. En ce qui concerne la f_0, la valeur moyenne et le décile du milieu sont les mesures qui se montrent plus adaptées pour la caractérisation.

Remerciements

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

Références

ADAMI A. G., MIHAESCU R., REYNOLDS D. A. & GODFREY J. J. (2003). Modeling prosodic dynamics for speaker recognition. *Proceedings of ICASSP*, p. 788–791.

ARVANITI A. & RODRIQUEZ T. (2013). The role of rhythm class, speaking rate, and f0 in language discrimination. *Laboratory Phonology*, **4**.

BOERSMA P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, **5 :9/10**, 341–345.

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.

GAUVAIN J., LAMEL L. & ADDA G. (2002). The limsi broadcast news transcription system. *Speech Communication*, **37**, 89–108.

HE L., GLAVITSCH U. & DELLWO V. (2015). Comparisons of speaker recognition strengths using suprasegmental duration and intensity variability : an artificial neural networks approach. *Proceedings of ICPhS*.

HUDSON T., DE JONG G., MCDOUGALL K., HARRISON P. & NOLAN F. (2007). F0 statistics for 100 young male speakers of standard southern british english. *Proceedings of ICPhS*, p. 1809–1812.

KAHN J. (2011). *Parole de locuteur : performance et confiance en identification biométrique vocale*. Thèse de doctorat, ED 536.

KEATING P. & KUO G. (2012). Comparison of speaking fundamental frequency in english and mandarin. *Journal of Acoustical Society of America*, **132**, 1050–1060.

LINDH J. & ERIKSSON A. (2007). Robustness of long time measures of fundamental frequency. *Proceedings of Interspeech*, p. 2025–2028.

MATHWORKS (2019). *MATLAB Deep Learning Toolbox R2019a*. Mathworks, Natick, MA, USA.

NIEBUHR O. & SKARNITZL R. (2019). Measuring a speaker's acoustic correlates of pitch - but which ? a constrastive analysis based on perceived speaker charisma. *Proceedings of ICPhS*, p. 1774–1778.

PARDO J. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, **119**.

SCHÖTZ S. (2007). Analysis and synthesis of speaker age. *Proceedings of ICPhS*.

SORIN C. (1981). Functions, roles and treatments of intensity in speech. *Journal of Phonetics*, **9**, 359–374.

TORREIRA F., ADDA-DECKER M. & ERNESTUS M. (2010). The nijmegen corpus of casual french. *Speech Communication*, **52**, 201–212.

TWEEDY R. S. & CULLING J. F. (2014). Does the signal-to-noise ratio of an interlocutor influence a speaker's vocal intensity ? *Computer Speech and Language*, **28**, 572—-579.