Etude des voyelles et de la force de voix par analyse discriminante

Jean-Sylvain Liénard[1], Claude Barras[1,2]
(1) LIMSI-CNRS
(2) Université Paris Sud
{jean-sylvain.lienard, claude.barras}@limsi.fr

RESUME

L'effort vocal, représenté ici par une mesure d'intensité objective appelée force de voix, est à la fois un facteur de variabilité de la parole et une grandeur acoustique utilisée par les interlocuteurs pour échanger diverses informations dans une situation donnée. La présente étude s'intéresse aux indices acoustiques codant ces informations dans le spectre des voyelles. L'Analyse Discriminante est mise en œuvre d'une part pour identifier les voyelles et d'autre part pour estimer la force de voix en dépit de leurs variabilités mutuelles. Les résultats, établis sur deux bases de données différentes, montrent que la force de voix peut être estimée avec précision à partir du spectre des voyelles et que la connaissance préalable de la force de voix permet d'améliorer la classification des voyelles.

ABSTRACT

Vocal Effort, represented here by an objective intensity measurement called Voice Strength, is both a speech variability factor, and a physical quantity used by interlocutors in order to exchange some types of information in a given situation. The present study deals with the acoustical features coding this information in the vowel spectrum. Discriminant Analysis is used to identify the vowels as well as to estimate the voice strength despite their mutual variabilities. The results established on two different databases, show that voice strength may be precisely estimated from the vowel spectra, and that knowing the voice strength improves the classification of the vowels.

MOTS-CLES : Voix, parole, communication orale, voyelles françaises, effort vocal, analyse acoustique, analyse discriminante, interactions voix-parole.
KEYWORDS: Voice, speech, oral communication, French vowels, vocal effort, acoustic analysis, discriminant analysis, voice-speech interactions.

1 Introduction

La communication orale est ***située*** et ***interactive***. Lorsqu'on parle, on s'adresse toujours à un interlocuteur, réel ou imaginaire, proche ou lointain, familier ou non. Le signal vocal encode divers types d'information, de nature linguistique (parole) ou non (voix), intimement mélangés. L'effort vocal (EV) est ajusté par le locuteur, de manière consciente ou non, afin d'assurer la transmission de ces informations à l'interlocuteur, en fonction des caractéristiques de ce dernier, de leur situation mutuelle, du cours de l'échange oral et des conditions de la transmission acoustique (distance, bruit ambiant).

L'EV est une notion qualitative, issue du domaine de la phoniatrie. Nous faisons l'hypothèse qu'elle peut être indexée par l'intensité acoustique mesurée à une distance donnée de la bouche du locuteur, selon des modalités particulières (pics d'intensité mesurés sur les noyaux vocaliques selon une fenêtre temporelle de l'ordre de 50 ms). Dans la suite nous appelons cette grandeur physique "force de voix" (FDV). L'expérience quotidienne montre à l'évidence que la force de voix est distincte de l'intensité du signal reçu par l'auditeur: on ne change pas une voix faible en une voix forte simplement en l'amplifiant de quelques dB ou en s'approchant du locuteur.

L'effort vocal relève de plusieurs problématiques différentes, concernant l'intelligibilité de la voix criée (ROSTOLLAND, 1982), le rôle de la distance entre les interlocuteurs (TRAUNMULLER and ERIKSSON, 2000), l'écoute dans le bruit et l'effet Lombard (JUNQUA, 1992), (GARNIER, 2007), le fonctionnement de la source glottique (FANT et al, 1985), (DOVAL et al 2006), (HANSON,1997), l'incidence de l'EV sur les formants (LIENARD and DI BENEDETTO, 1999), (HUBER et al, 1999), sur la qualité de la voix chantée (HENRICH et al, 2005), ainsi que sur la prosodie (D'ALESSANDRO et al, 2006). Les travaux menés en reconnaissance de la parole (ZELINKA et al, 2012) ou du locuteur (BRUNGART et al, 2001) (HANSEN and VARADARAJAN, 2009) montrent une grande dégradation des résultats avec l'EV, catégorisé en un petit nombre de types (chuchoté, faible, normal, fort, crié ou Lombard), mais jamais envisagé comme repérable par une grandeur mesurable telle que la FDV. Les références données ici à titre indicatif n'ont rien d'exhaustif mais permettent de comprendre la diversité des points de vue sur l'EV.

Les différentes études convergent sur les observations suivantes. Le fait de parler plus fort entraîne le déplacement vers l'aigu du "formant glottique" variant avec f0 et avec le quotient ouvert (Oq) des cordes vocales. Le renforcement peut se manifester dans toute l'étendue du spectre. La pente spectrale à long terme ("spectral tilt") représente globalement le rehaussement des composantes spectrales élevées. Le fondamental f0 augmente avec l'EV, d'environ une octave entre voix très faible et voix très forte. La fréquence du 1er formant F1 augmente de manière significative avec l'EV. Enfin la différence d'amplitude entre le 1er harmonique (fondamental) et le second semble en relation avec le formant glottique, mais ceci dépend de f0 et de la voyelle.

Dans ce qui suit nous décrivons les deux bases de données utilisées ainsi que les outils d'analyse des voyelles françaises pour diverses valeurs de la FDV. Nous présentons l'application de l'Analyse Discriminante à l'identification des voyelles, puis à l'estimation de la FDV, et nous examinons en quoi l'une peut être utilisée pour améliorer l'autre.

2 Données

Pour étudier la force de voix de manière quantitative il faut des données étalonnées en niveau sonore, dont on puisse être sûr du niveau sonore émis (en dB SPL), et surtout de la constance des conditions de prise de son (distance bouche-microphone, chaîne de traitement électroacoustique). Il n'existe pas, à ce jour, de base de données de grande taille offrant toutes les garanties d'étalonnage et de constance de la prise de son. Nous avons donc dû nous contenter de deux bases de données de taille réduite disponibles dans notre laboratoire.

2.1 Données CRC

La base de données CRC, utilisée dans l'article (LIENARD et DI BENEDETTO, 1999) comporte 12 voyelles françaises (9 orales, 3 nasales) prononcées isolément par 6 locuteurs et 7 locutrices, selon 3 degrés d'effort vocal suscités par des variations de distance entre locuteur et opérateur. L'ensemble a été enregistré en 3 sessions, dont la 3e six mois après la première. Le locuteur était assis, sa bouche à 30 cm du microphone omnidirectionnel, et devait répéter une courte phrase puis des voyelles émises par l'opérateur, lui-même debout et placé à une distance choisie parmi trois: 0.40 m (condition "p" pour proximité), 1.50 m (condition "n" pour normal) ou 6 m (condition "l" pour loin). Les locuteurs n'ont pas tous participé aux 3 sessions; le corpus se compose de 720 sons ou "tokens", soit 20 séries de 36 tokens, chaque série comportant les 12 voyelles émises par un même locuteur selon les 3 conditions p, n et l. Les voix restent dans la dynamique usuelle de la conversation (30 à 35 dB entre les sons les plus faibles et les plus forts) et les conditions d'enregistrement sont maintenues constantes.

2.2 Données JAE

Ce corpus (AUGUSTE-ETIENNE, 1999) comporte des syllabes de type CV et des voyelles isolées /a, i, u/ produites par 7 hommes, 7 femmes et 3 enfants. Chaque syllabe est répétée ad libitum en partant d'une voix que le locuteur considère comme neutre. La consigne est de produire un effort vocal croissant par paliers jusqu'à un maximum proche de la voix criée, puis décroissant, et de recommencer cette fois en baissant la voix jusqu'à un minimum proche de la voix chuchotée, puis en retournant à la valeur neutre. La prise de son est effectuée avec deux microphones placés à distance fixe (20 cm) de la bouche du locuteur. L'un des deux est un microphone de mesure, étalonné de temps en temps au moyen d'une source étalon (1 kHz, 94 dB). La différence de niveau sonore entre les sons les plus faibles et les plus forts excède 60 dB. La présente étude ne prend en compte que les voyelles isolées, produites en nombre variant de 39 à 232 selon les locuteurs, pour un total de 1452 tokens.

3 Outils et méthodes

3.1 Sélection et prétraitement

Tous les calculs sont effectués en Praat (BOERSMA and WEENINK, 2012). L'intensité ***ax*** (en dB par rapport à une référence arbitraire) et le fondamental ***f0*** (en Hz) sont calculés sur une fenêtre gaussienne de 50 ms. Le signal est ensuite pondéré selon la courbe A des sonomètres (atténuation des fréquences inférieures à 1 kHz) et son intensité ***ap*** est calculée (en dBA). L'analyse utilise la méthode BarkFilter en 18 bandes de 1 Bark. Le spectre Bark est prélevé à l'instant d'intensité maximale. Le fondamental ***f0*** et les amplitudes du spectre Bark normalisé par rapport à son maximum (arbitrairement fixé à 50), notées ***b1*** à ***b18***, sont retenus comme indices acoustiques. Les intensités utilisées dans la suite ne correspondent pas directement à des niveaux sonores absolus (SPL) mesurables dans les conditions de référence (à 1m, fenêtre de 125 ms, environnement anéchoïque). Mais elles sont comparables entre elles pour les deux bases de données, centrées sur une valeur moyenne arbitraire de 50 dBA.

3.2 Analyse discriminante

L'Analyse Discriminante (AD) considère des observations définies chacune par un ensemble de valeurs numériques et une catégorie d'appartenance. L'AD calcule un nouvel espace orthogonal de représentation des données, qui simultanément minimise la variance des observations appartenant à une même catégorie et maximise la variance entre catégories (pour une introduction à l'AD on pourra se reporter à l'aide en ligne de Praat). Le résultat est un classifieur (discriminant), dont la qualité peut être mesurée par le coefficient lambda de Wilks. L'utilisation du classifieur est évaluée par le taux d'observations mal classées, les données étant soit celles qui ont servi pour apprendre le classifieur (mode classification, ou autocohérence), soit de nouvelles données (mode prédiction). Dans la présente étude nous nous limiterons au mode classification.

4 Discrimination des voyelles

Les données soumises à l'AD sont, pour chaque token, la catégorie phonétique (l'une des voyelles françaises représentées dans le corpus) et les 19 indices acoustiques f0+18Bk.

4.1 Données CRC

La figure 1 présente les ellipses de dispersion à 1 écart-type dans le plan factoriel principal, le profil des deux premiers vecteurs propres, les valeurs propres attachées à chaque axe (en % d'explication de la variance) et en cumul, le coefficient lambda de Wilks et les taux d'erreur obtenus en classification. On ne peut pas juger de la classification à partir du seul plan principal lorsque les vecteurs propres d'ordre supérieur à 2 contribuent notablement à l'explication de la variance, ce qui est le cas ici.

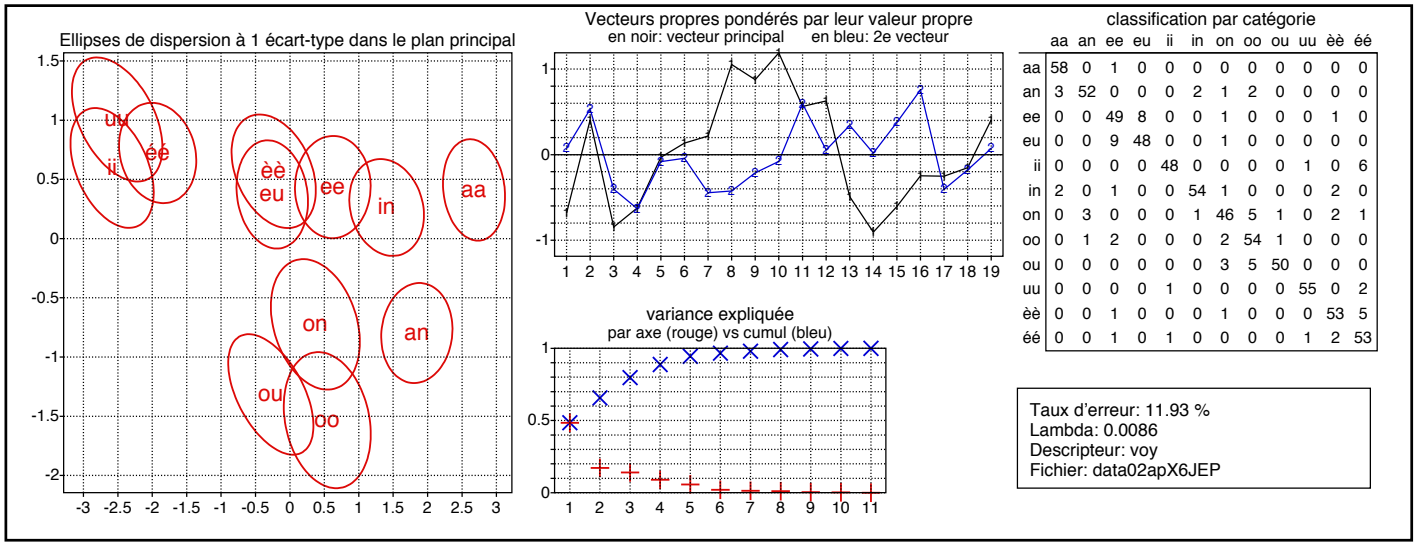

FIGURE 1 – Analyse Discriminante appliquée à la classification des 12 voyelles (9 orales et 3 nasales) du corpus CRC

Dans chaque groupe vocalique la variabilité est due à plusieurs causes: force de voix, locuteur, petites différences d'émission au cours de la session ou entre sessions. Malgré ces variations l'AD classifie les 12 voyelles de manière satisfaisante, avec un taux d'erreur de l'ordre de 12%. Les confusions constatées sont pour l'essentiel présentes dans les données elles-mêmes: un test perceptif (identification directe des voyelles par un groupe

d'auditeurs francophones) rapporté dans l'article (LIENARD and DI BENEDETTO, 1999) montrait un taux d'erreur de 9,3% sur les seules voyelles orales. Le test de classification ci-dessus donne sur celles-ci un taux d'erreur comparable (10,0%). Parmi les indices acoustiques dégagés par l'AD on observe que ***f0*** occupe une place secondaire.

4.2 Données JAE

Ces données ne comportent que 3 catégories vocaliques. Les résultats de l'AD sont présentés dans la figure 2.

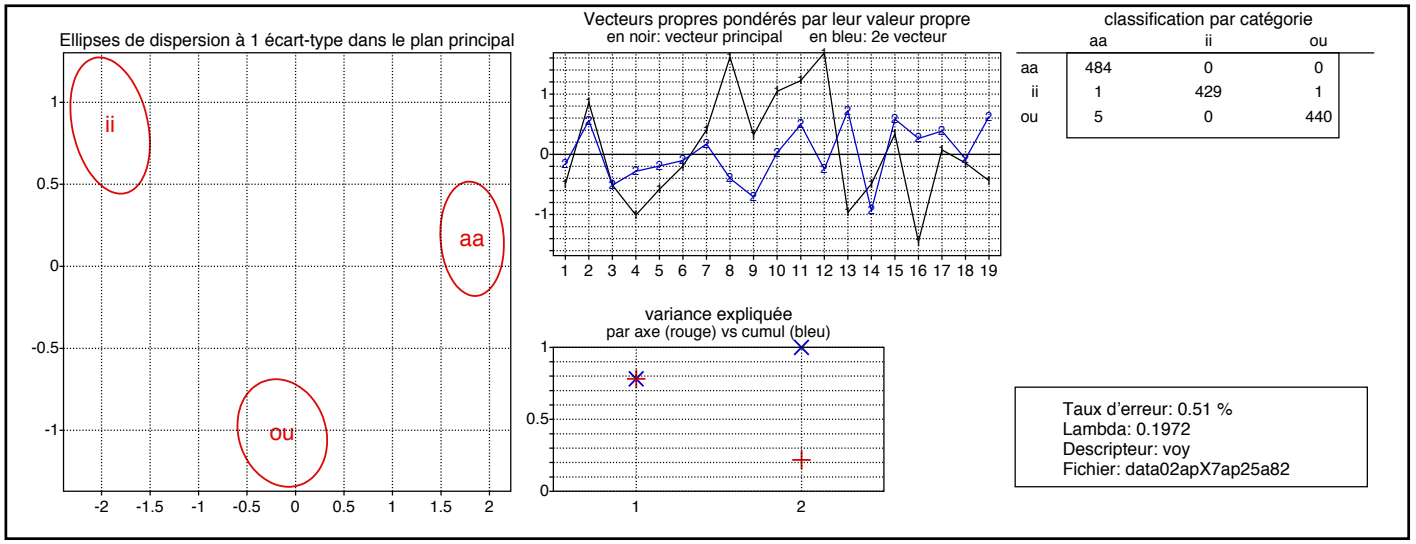

FIGURE 2 – Analyse Discriminante appliquée à la classification des 3 voyelles [a, i, u] du corpus JAE

Le taux d'erreur en classification est très faible, et identique à celui des données CRC réduites à ces mêmes voyelles cardinales. Lambda est nettement plus élevé, ce qui reflète une plus grande variance intragroupe par rapport à la variance intergroupe, due à la plus grande variabilité des données JAE selon la FDV et le locuteur.

4.3 Commentaire des résultats

Les essais menés sur les deux bases de données montrent que les voyelles sont correctement discriminées malgré les variations dues à la force de voix et au locuteur. Le nombre de dimensions nécessaires pour rendre compte d'au moins 95% de la variance est de 5 (sur 11) quand on considère 12 voyelles et de 2 (sur 2) quand on considère seulement les 3 voyelles cardinales.

5 Estimation de la force de voix

Nous utilisons le même outil et les mêmes données pour étudier la variabilité due à la force de voix. Celle-ci est une grandeur intrinsèquement continue, exprimée en dB, mais rien ne permet d'affirmer a priori qu'elle soit en relation linéaire avec une quelconque combinaison linéaire d'indices acoustiques. C'est pourquoi nous distinguons de manière quelque peu artificielle des paliers de FDV, espacés de 6 dB, qui seront considérés comme les centroïdes des catégories par rapport auxquelles l'AD regroupera les observations de manière optimale. Comme il y a continuité de la grandeur sous-jacente d'un palier à

l'autre, la valeur prise par lambda sera moins pertinente que dans le cas où les catégories sont par nature disjointes (comme dans le cas des voyelles). En revanche, lors de la phase de classification on pourra effectuer une interpolation entre centroïdes voisins pour retrouver une valeur continue. Ainsi, pour chaque token, l'écart entre la valeur initiale ***ap*** de la FDV et la valeur ***apc*** calculée par le classifieur et interpolée servira de base pour calculer une marge d'erreur statistique, qui constituera le véritable critère de réussite de l'analyse.

5.1 Données CRC

La figure 3 montre, de gauche à droite: i) le plan principal de l'analyse factorielle et les ellipses de dispersion à 1 écart-type, ii) la représentation des deux premiers vecteurs propres en fonction des 19 indices acoustiques retenus, et iii) la correspondance entre la valeur mesurée ***ap*** de la FDV (en abscisse) et la valeur ***apc*** (en ordonnée) estimée au moyen du classifieur établi à partir des 19 indices acoustiques (f0+18 Bk).

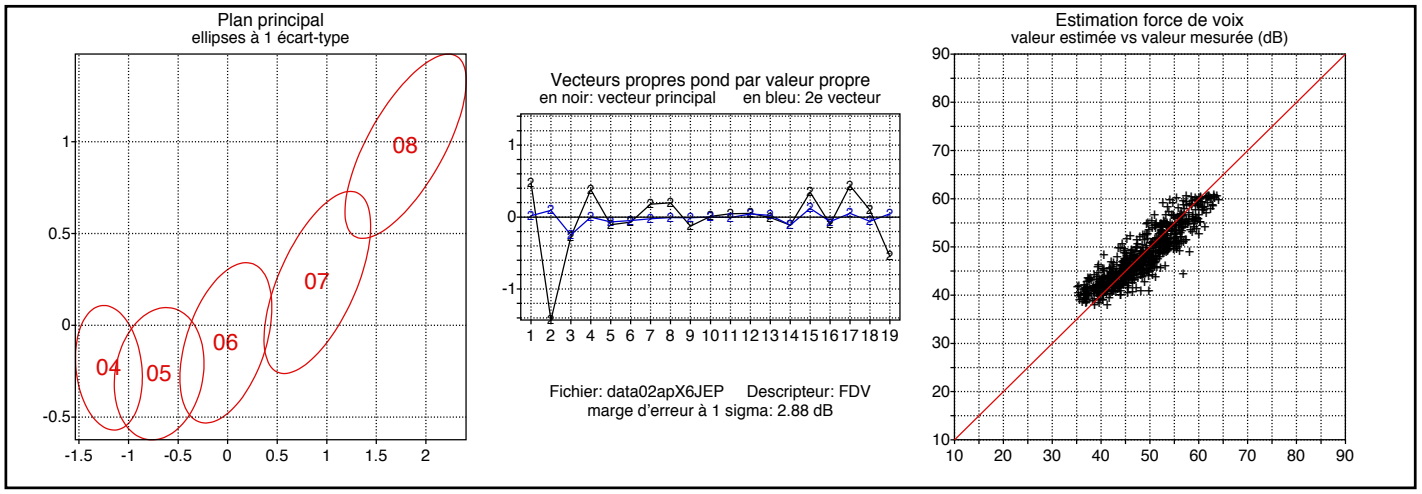

FIGURE 3 – Analyse Discriminante appliquée à l'estimation de la force de voix pour un sous-ensemble du corpus CRC (les classes incomplètes 03 et 09 ont été omises)

On constate une performance remarquable: la marge d'erreur (à ±1 écart-type de la valeur correcte) est de moins de 3 dB en classification, ce qui indique que l'information permettant de discerner 5 à 10 nuances de FDV dans la dynamique considérée (30 dB) est présente dans le spectre Bark du signal. L'axe 1, où intervient majoritairement l'indice 2 (***b1***) suffit presque à lui seul à estimer la FDV (explique 92.7% de la variance). On observe une forte non-linéarité dans le plan principal pour les faibles FDV, manifestée surtout sur l'axe 2 au niveau de ***b2***, ce qui justifie a posteriori le choix d'une technique de catégorisation par paliers. Le fondamental ***f0*** intervient peu dans les vecteurs propres, mais une partie de l'information qu'il porte se retrouve dans les bandes ***b1*** et ***b2***.

5.2 Données JAE

La moindre performance illustrée dans la figure 4 (marge d'erreur 6.2 dB), s'explique par la plus grande dynamique de ces données. Elle permet de discerner statistiquement 8 à 10 nuances de FDV. Les axes principaux sont semblables à ceux obtenus avec le corpus CRC, ainsi que la non-linéarité qui concerne surtout les voix faibles. Ces résultats confirment la pertinence des indices ***b1*** et ***b2*** pour l'estimation de la FDV.

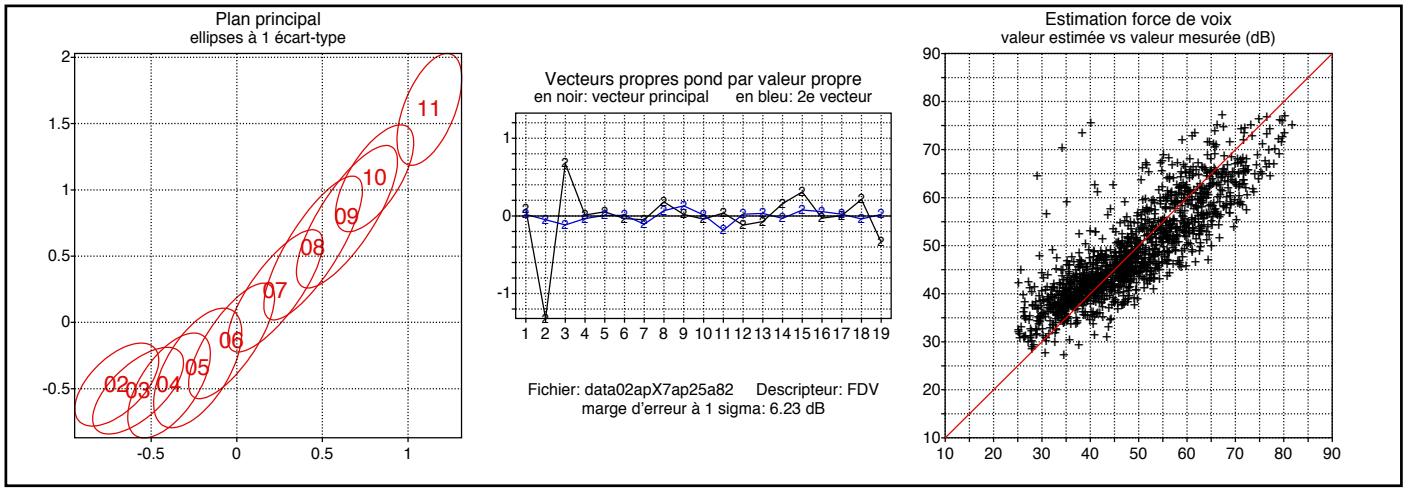

FIGURE 4 – Analyse Discriminante appliquée à l'estimation de la force de voix pour un sous-ensemble du corpus JAE (les classes incomplètes 01 et 12 ont été omises)

6 Interactions entre voyelle et FDV

Nous avons vu dans les sections précédentes que tant les voyelles que la force de voix pouvaient être estimées avec un faible taux d'erreur à partir des spectres Bark normalisés. Nous allons maintenant chercher s'il existe des interactions entre les deux descripteurs. Pour cela nous nous posons la double question: le fait de connaître la voyelle (par une première analyse ou par une information descendante) permet-il d'améliorer l'estimation de la force de voix ? Et, réciproquement, la connaissance de la FDV peut-elle être utilisée pour mieux discriminer les voyelles ?

6.1 Connaître la voyelle permet-il de mieux estimer la FDV ?

Pour tester cette proposition un classifieur de la FDV est calculé pour chaque catégorie vocalique du corpus, cad les 12 voyelles de CRC et les 3 de JAE, et dans un second temps ce classifieur est appliqué à ces mêmes tokens. Les résultats sont consignés dans la table 1, pour les deux bases de données. La colonne "voyelle inconnue" indique les valeurs sur l'ensemble des données, la classification étant effectuée sur toutes les voyelles. La colonne "voyelle connue" indique la moyenne des marges d'erreur trouvées pour chaque voyelle.

Marge d'erreur sur la FDV (dB)	voy inconnue	voy connue
CRC (12 voy, 720 tok, dyn 30 dB)	2,97	2,12
JAE (3 voy, 1395 tok, dyn 60 dB)	6,51	4,87

TABLE 1 – Amélioration de l'estimation de la FDV lorsque la voyelle est connue

Les chiffres précis diffèrent légèrement de ceux rapportés plus haut car les conditions de calcul ne sont pas exactement identiques. Mais le résultat est sans ambiguïté: dans les deux cas la connaissance préalable de la voyelle entraîne une amélioration notable de l'estimation de la FDV.

6.2 Connaître la FDV permet-il de mieux discriminer les voyelles ?

Pour tester cette proposition on ne peut guère prendre comme seul critère le taux d'erreur de classification, intrinsèque aux données dans le cas de CRC et très faible dans le cas de JAE. Il est donc accompagné entre parenthèses de la valeur de lambda (table 2). Dans le cas de CRC on a considéré les tokens produits selon la consigne p, n ou l, tout en sachant que les FDV moyennes dans ces trois tranches sont proches (resp. 62,7 pour p, 66,6 pour n et 74,1 dB pour l) et se recouvrent partiellement, ce qui explique les valeurs faibles de lambda. Dans le cas de JAE on a considéré 4 tranches de FDV espacées de 12 dB; le taux d'erreur affiché est la moyenne des taux obtenus pour chaque tranche.

Erreurs de classification voy % (lambda)	FDV inconnue	FDV connue
CRC (12 voy, moy consignes FDV: p,n,l)	12,5 (0,0089)	9,0 (0,0048)
JAE (3 voy, moy 4 tranches de FDV de 12 dB)	0,5 (0,2018)	0,3 (0,1311)

TABLE 2 – Amélioration de la discrimination des voyelles lorsque la FDV est connue

Tout en gardant à l'esprit les réserves formulées plus haut, on constate une amélioration sur les deux critères, pour les deux bases de données: la connaissance, même approchée, de la FDV permet d'améliorer la classification des voyelles.

7 Conclusion

Cette étude a mis en évidence le caractère quantifiable des variations spectrales dues à l'effort vocal, qui se distinguent ainsi d'un simple changement de timbre. Ce caractère systématique permet de retrouver la force de voix à partir du spectre dont a été retirée l'information directe de niveau sonore, et ceci avec une précision comprise entre 3 et 6 dB. On peut donc penser que les degrés d'effort vocal perceptibles d'emblée dans la voix conversationnelle sont plus nombreux que les 3 nuances habituellement mentionnées dans la littérature. Les résultats laissent entrevoir, à terme, la possibilité de retrouver dans tout enregistrement de parole l'intensité véritablement émise par le locuteur, information habituellement perdue dès la prise de son.

Sur un plan plus général, la démarche suivie s'inscrit dans la problématique des interactions entre voix et parole. Souvent considérés comme étrangers l'un à l'autre, ces deux aspects du signal oral apparaissent comme partiellement liés et complémentaires, au point que la connaissance même approximative de l'un peut faciliter l'estimation de l'autre.

Remerciements

Les auteurs remercient les instances scientifiques du LIMSI-CNRS, ainsi que leurs collègues Nicolas Audibert, Philippe Boula de Mareüil, Christophe d'Alessandro et Albert Rilliard qui ont apporté divers éclairages sur le problème de l'effort vocal.

Références

AUGUSTE-ETIENNE, J. (1999). "Etude d'un protocole d'enregistrement pour l'analyse du timbre de la voix", mémoire de recherche, ENS Louis Lumière, Noisy le grand.

BOERSMA, P. and WEENINK, D. (2012). "Praat: doing phonetics by computer" version 5.3.32, retrieved 17 October 2012 from http://www.praat.org/.

BRUNGART, D., SCOTT, K. and SIMPSON, B. (2001) "The influence of vocal effort on human speaker identification", ***Eurospeech***, Scandinavia.

D'ALESSANDRO, C. (2006). "Voice source parameters and prosodic analysis", in S. Sudhoff et al [Eds] ***Methods in Empirical Prosody Research***, 63-87, Walter de Gruyter.

DOVAL, B., D'ALESSANDRO, C. HENRICH, N. (2006). "The spectrum of glottal flow models", ***Acustica united with Acta Acustica***, 92:1026-1046, 2006.

FANT, G., LILJENCRANTS, J. and LIN, Q. (1985). "A four parameter model of glottal flow", STL-QPRS, 26(4):1-13, 1985.

GARNIER, M. (2007). "Communiquer en environnement bruyant: de l'adaptation jusqu'au forçage vocal", thèse de doctorat, université Paris VI.

HANSEN, J. and VARADARAJAN, V. (2009). "Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition", IEEE tr. on ASLP, 17 (2), 366-378.

HANSON, H. (1997). "Glottal characteristics of female speakers: acoustic correlates", ***J. Acoust. Soc. Am.*** 101 (1), 466-481, 1997.

HENRICH, N., D'ALESSANDRO, C., DOVAL, B. and CASTELLENGO, M. (2005). "Glottal open quotient in singing: measurement and correlation with laryngeal mechanisms, vocal intensity, and fundamental frequency", ***J. Acoust. Soc. Am.*** 117 (3), 1417-1430.

HUBER, J.E., STATHOPOULOS, E.T., CURIONE, G.M., ASH T.A. and JOHNSON, K. (1999). "Formants of children, women, and men: the effects of vocal intensity variation", ***J. Acoust. Soc. Am.*** 106 (3), 1532-1542.

JUNQUA, J.-C. (1992). "The Lombard reflex and its role on human listeners and automatic speech recognizers", ***J. Acoust. Soc. Am.*** 93, 510-524.

LIENARD, J.S. and DI BENEDETTO, M.G. (1999). "Effect of vocal effort on spectral properties of vowels", ***J. Acoust. Soc. Am.*** 106 (1), 411-422.

LIENARD J.S. and BARRAS C. (2013). "Fine-grain voice strength estimation from vowel spectral cues", ***InterSpeech***, Lyon.

ROSTOLLAND, D. (1982). "Acoustic features of shouted voice", ***Acustica***, vol 50, 118-125.

TRAUNMULLER, H. and ERIKSSON, A. (2000). "Acoustic effects of variation in vocal effort by men, women and children", ***J. Acoust. Soc. Am.*** 107 (6), 3438-3451.

ZELINKA, P., SIGMUND, M. and SCHIMMEL, J. (2012). "Impact of vocal effort variability on automatic speech recognition", ***Speech Communication*** (54), 732-742.