



HAL
open science

Contributions expérimentales à l'élaboration d'un protocole robuste d'évaluation perceptive des troubles de la voix et de la parole

Alain Ghio, Sophie Dufour, Gilles Pouchoulin, Joana Revis, Danièle Robert,
Antoine Giovanni

► To cite this version:

Alain Ghio, Sophie Dufour, Gilles Pouchoulin, Joana Revis, Danièle Robert, et al.. Contributions expérimentales à l'élaboration d'un protocole robuste d'évaluation perceptive des troubles de la voix et de la parole . Revue PAROLE, 2014, pp.85-101. hal-01294774

HAL Id: hal-01294774

<https://hal.science/hal-01294774v1>

Submitted on 31 Mar 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Contributions expérimentales à l'élaboration d'un protocole robuste d'évaluation perceptive des troubles de la voix et de la parole

Alain GHIO¹, Sophie DUFOUR¹, Gilles POUCHOULIN¹,
Joana REVIS², Danièle ROBERT², Antoine GIOVANNI²

¹Laboratoire Parole et Langage, Université
d'Aix-Marseille, CNRS, Aix-en-Provence

²Centre Hospitalier Universitaire de la
Timone, Marseille, France

1. Introduction

La perception d'une voix ou d'une parole perturbée comme celle produite par des locuteurs dysphoniques ou dysarthriques reste peu explorée. Nous ne savons que peu de choses sur les mécanismes perceptifs et cognitifs mis en jeu chez l'auditeur lors de l'écoute d'un locuteur atteint de troubles phonatoires, articulatoires ou prosodiques. Malgré cela, le recours à la perception auditive est extrêmement répandu en pratique clinique comme procédé d'évaluation des marques de dysfonctionnement lors de la production de parole. Ainsi, dans la prise en charge de patients dysphoniques, la perception de la qualité vocale reste essentielle car ces patients se décident généralement à consulter à partir du moment où ils entendent des changements dans leur résultat vocal. De même, la plupart du temps, ils jugent le succès d'un traitement chirurgical ou orthophonique par l'amélioration de l'impression auditive laissée par leur voix (Gerrat et al., 2000). Pour le clinicien, l'évaluation perceptive est le moyen le plus couramment utilisé pour caractériser la voix du patient (soufflée, rauque...) ou pour rendre compte de la sévérité du dysfonctionnement (Dejonckere et al., 2001). Elle présente en effet de nombreux avantages : facile à mettre en œuvre, peu coûteuse, directement accessible à tout clinicien. Classiquement, l'évaluation perceptive est menée à l'aide d'une échelle standardisée comme par exemple, l'échelle GRBAS (Hirano,

1981) contenant plusieurs paramètres que l'auditeur doit juger à l'oreille.

Pourtant, si l'analyse perceptive reste la référence en la matière, cette méthode soulève encore une question centrale: celle de la fiabilité. En effet, la littérature rapporte une importante variabilité dans les jugements perceptifs de la voix (Kreiman et al., 1992, 1993 ; Bele, 2005 ; De Bodt et al., 1997 ; Wuyts et al., 1999 ; Révis et al., 1999 ; Shrivastav, 2006). Cette variabilité se constate dans la possible inconstance d'un auditeur qui peut fournir des jugements différents entre plusieurs sessions d'écoute de la même voix (variabilité intra-auditeur). Elle est aussi observable dans le manque de cohérence des jugements à l'intérieur d'un groupe d'auditeurs (variabilité inter-auditeurs). Ces variabilités ont été largement étudiées dans le but de réduire ces phénomènes considérés comme indésirables en pratique clinique. Les auteurs ont notamment cherché différents points d'entrée pour en réduire son ampleur : recrutement d'auditeurs experts vs. naïfs (De Bodt et al., 1997), choix d'échelles analogiques vs. catégorielles (Wuyts et al., 1999), type d'énoncés comme voyelle tenue vs. phrases (Révis et al., 1999), dimensions perçues comme la qualité globale vs. dimension soufflée vs. raucité (Shrivastav, 2006)...

Ce manque de fiabilité peut être expliqué par les fondements mêmes du procédé. En effet, l'évaluation perceptive d'un trouble de la voix ou de la parole repose sur la capacité à percevoir un écart à une norme à la fois en terme de direction (dimensions rauque, soufflée, hypo/hyper...) mais aussi de distance (légèrement, moyennement, sévèrement dysphonique) et ce indépendamment du contexte d'écoute. Dans ce cadre, deux expériences ont été dans un premier temps menées dans le but de mesurer les aspects liés à la sensibilité au contexte (Expérience 1) et à l'universalité ou non de la norme (Expérience 2). Enfin, dans une 3^{ième} expérience (Expérience 3) nous avons tenté de réduire ces sources de variabilité génératrice de manque de fiabilité, par le biais d'un protocole d'apprentissage.

2. *Expérience 1 : la variabilité liée au contexte de perception*

Juger la qualité vocale d'un locuteur est au premier abord un processus de perception ascendant (bottom-up), c'est-à-dire qu'à partir de l'échantillon vocal, l'auditeur va catégoriser la voix par interprétation des indices acoustiques détectés perceptivement. C'est l'un des fondements de l'évaluation perceptive des troubles de la voix et de la parole.

Mais, comme dans tout autre processus de perception de la parole, nous émettons l'hypothèse que ce processus ne se réduit pas à ce simple trajet ascendant de l'acoustique vers le cognitif. Des processus descendants (top-down) interviennent et influencent la perception. En effet, lorsque nous entendons un énoncé dégradé, bruité ou phonétiquement appauvri, ces processus top-down entrent en jeu pour restaurer ce qui est dégradé et optimiser l'intelligibilité du message. L'attention portée au signal viendra maximiser ou minimiser les effets de ces processus de restauration (Warren et al., 1970).

Cette attention pouvant varier considérablement d'un auditeur à l'autre ou d'un contexte à l'autre, elle introduira des changements importants dans l'interprétation des indices acoustiques détectés perceptivement et modifiera l'évaluation, la rendant peu fiable.

2.1. *Corpus et méthode*

Notre protocole expérimental consistait en l'évaluation, par un jury expérimenté, de voix dysphoniques, présentées en aveugle dans une première séance, puis accompagnées d'informations sur le parcours médical du patient dans une deuxième séance (Ghio, Merienne et al., 2011). Les différences entre les résultats des deux phases d'expérimentation pourraient être attribuées à l'apport d'information contextuelle et donc aux processus descendant. Nous avons sélectionné 50 locuteurs dysphoniques enregistrés plusieurs fois sur de la lecture de texte dans le service ORL du CHU Timone à Marseille (Ghio, Pouchoulin et al., 2012). Les voix ont été présentées par paire, chaque paire étant constituée de deux enregistrements d'un même locuteur. Ces enregistrements ayant été effectués à des dates

différentes, la qualité vocale des deux éléments de la paire était la plupart du temps différente. Les auditeurs devaient évaluer ces paires de voix par comparaison: après écoute de chacune des deux voix, l'auditeur devait juger leur degré de dysphonie en utilisant une échelle comparative à 7 points : la voix A est (1) nettement moins dysphonique, (2) moins dysphonique, (3) légèrement moins dysphonique, (4) de même qualité, (5) légèrement plus dysphonique, (6) plus dysphonique, (7) nettement plus dysphonique que la voix B. Nous avons opté pour cette échelle de façon à placer l'auditeur dans des conditions proches des usages en pratique clinique où l'intérêt principal réside souvent dans la perception de la quantité de changement (amélioration ou dégradation) lors de la prise en charge thérapeutique. D'autre part, ce choix a été fait pour avoir une sensibilité de mesure suffisante. Les 7 participants à l'expérimentation étaient des auditeurs régulièrement confrontés à l'écoute de voix dysphoniques. Bien évidemment, afin de ne pas biaiser les résultats de l'expérience, ils ignoraient l'objectif réel de l'étude qui était présentée comme une mise au point d'un protocole informatisé de jugement de la dysphonie en conditions hospitalières.

L'expérience se déroulait en quatre phases : deux sessions d'écoute en aveugle (test-retest) suivies de deux sessions d'écoute influencée (test-retest). Pour chacune des sessions, les paires étaient présentées dans un ordre aléatoire dans le but de minimiser les effets de liste. A l'issue de l'écoute en aveugle, chaque paire de voix a été jugée 14 fois (7 auditeurs x 2 sessions). Chaque jugement a été converti en note : [nettement moins dysphonique] \Leftrightarrow 3, [moins dysphonique] \Leftrightarrow 2, [légèrement moins dysphonique] \Leftrightarrow 1, [équivalent] \Leftrightarrow 0, [légèrement plus dysphonique] \Leftrightarrow -1, [plus dysphonique] \Leftrightarrow -2, [nettement plus dysphonique] \Leftrightarrow -3

La moyenne des 14 notes permettait d'établir un classement décroissant des paires. Plus la moyenne obtenue était proche de 3 en valeur absolue, plus la quantité de changement entre la voix A et la voix B était importante (+3 renseignant une nette préférence pour A, -3 une nette préférence pour B). Plus la moyenne était proche de 0, plus les auditeurs avaient considéré que les stimuli A et B étaient

équivalents en terme de qualité vocale. A partir des résultats de cette première expérience, nous avons scindé le corpus en deux parties équivalentes en terme de distribution de notes (jeu de données α et β) sur un principe simple. Le jeu de données α était constitué des paires positionnées en 1,3,5,7,...45, 47,49 dans le classement ordonné des notes. Le jeu de données β était constitué des paires 2,4,6,8 ...46, 48,50 dans le même classement. Sur le jeu de données α , l'information fournie à l'auditeur pour la deuxième expérience allait être cohérente au sens clinique du terme dans la mesure où les voix jugées moins dysphoniques en aveugle allaient être déclarées comme le résultat post thérapeutique. Sur le jeu de données β , l'information fournie à l'auditeur pour la deuxième expérience allait être incohérente au sens clinique du terme dans la mesure où les voix jugées moins dysphoniques en aveugle allaient être déclarées comme l'état en pré traitement, impliquant ainsi un résultat thérapeutique défavorable. Pour éviter que les informations incohérentes du jeu de données β apparaissent trop invraisemblables, nous n'avons sélectionné que 32 paires dont la note moyenne en aveugle se situait entre + 1.5 et -1.5, la même restriction étant appliquée aux deux jeux de données α et β . Bien évidemment, lors de la passation du test en condition contextuelle, les stimuli de la cohorte α ou β étaient mélangés et présentés de façon aléatoire. Le recueil informatisé des réponses étaient effectués par le logiciel libre PERCEVAL-LANCELOT (www.lpl-aix.fr/~lpldev).

2.2. Résultats

Nous avons utilisé un modèle de régression linéaire dans lequel la note obtenue en écoute influencée est comparée à la note obtenue en écoute aveugle en fonction des deux conditions (Figure 1). La droite d'ordonnée à l'origine 0 et de pente +1 traduit l'absence d'effet de contexte. Dans la condition α ou l'apport d'information est cohérent, l'analyse statistique montre que la pente de la droite de régression est de 1.60 +/- 0.19. Cette pente importante de coefficient directeur 1.6 rend compte de l'effet amplificateur du contexte cohérent : la note est accentuée de 60% par rapport à celle

obtenue en aveugle. Dans la condition β où l'apport d'information est incohérent, l'analyse statistique indique que la pente de la régression linéaire est de 0.17 ± 0.14 , valeur de pente faible traduisant l'effet inhibiteur de la condition. Rappelons qu'une pente nulle indiquerait la totale indépendance de la note fournie en contexte par rapport à celle obtenue en aveugle. La valeur faible de 0.17 indique que les auditeurs confrontés à une information en contradiction avec leur perception ont eu tendance à se fier à l'information contextuelle et à atténuer fortement les effets perçus en aveugle. On peut même observer des inversions de préférences visibles dans les points situés dans les quadrants supérieurs gauches ou inférieurs droits de la Figure 2b. Ces inversions de préférence représentent 50 % des cas de situation incohérente. 100% des cas d'inversion de préférence sont contraints par l'information contextuelle.

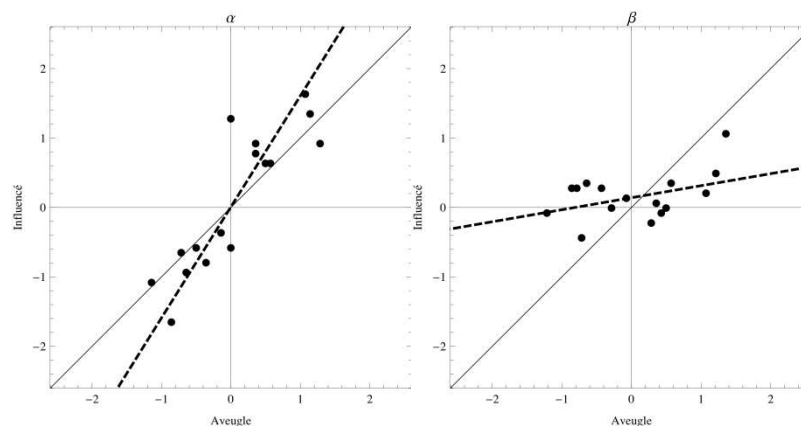


Figure 1 – Résultats de l'écoute contextuelle influencée (verticalement) en fonction de l'écoute aveugle (horizontalement). A gauche : apport d'information cohérente (cohorte alpha). A droite : apport d'information incohérente (cohorte beta). La droite bissectrice en trait plein est celle de l'absence d'effet (notes identiques entre la condition aveugle et la condition contextuelle). La droite en pointillé est la régression linéaire obtenue à partir des observations.

2.3. Discussion

Dans la condition cohérente, l'apport d'information est amplificateur : nous retrouvons les effets observés par Herz (2003) dans la perception des odeurs où les préférences sont généralement amplifiées par l'association d'information verbale au stimulus olfactif. Dans notre expérience, les auditeurs sont confortés dans leur jugement par la cohérence des informations fournies en

contexte : la qualité vocale est meilleure après qu'avant traitement. Dans la condition incohérente, nous observons là aussi des analogies avec les résultats de Herz. Les auteurs constatent que pour certaines odeurs, et sous l'unique effet du contexte verbal, jusqu'à 88 % des sujets ont une interprétation perceptive complètement différente entre deux sessions avec connotation positive vs négative. Dans notre expérience, nous observons une inversion de polarité du jugement dans 50 % des cas incohérents. Dans le cadre de notre étude, les stimuli auditifs étaient identiques entre les deux conditions d'écoute. Seule variait l'information fournie aux auditeurs sur la nature du locuteur et cette donnée a fait varier de façon importante le résultat. La perception du thérapeute est ainsi influencée par des processus cognitifs top-down (« une voix post thérapeutique est meilleure qu'une voix pré thérapeutique ») et que cette information le rend « sourd » à des phénomènes qu'il a perçus lors de la phase aveugle. Nous pouvons aussi interpréter ce phénomène comme une capture attentionnelle (« on m'indique que c'est post thérapeutique, je n'entends que ce que je m'attends à entendre : l'amélioration »). Nos auditeurs étaient des professionnels de la prise en charge de la voix. Ils étaient donc, de façon légitime, en situation de forte implication par rapport aux aspects liés à la réussite thérapeutique.

Les résultats de cette étude semblent converger vers l'extrême nécessité d'utiliser uniquement des évaluations perceptives en aveugle pour réaliser un bilan perceptif d'une dysphonie.

3. Expérience 2 : une norme universelle de perception ?

Le but de cette deuxième expérience était de comparer la perception de la qualité vocale à travers deux langues (le français et l'italien) dans un contexte clinique de la dysphonie. L'objectif était de vérifier si les relations entre la dysphonie et ses caractéristiques perceptives (sévérité globale, raucité, souffle) sont universelles ou variables.

En effet, si en phonétique, de nombreux résultats montrent une dépendance à la langue des aspects liés à la qualité vocale (Vaissière, 1997 ; Ní Chasaide et al., 1999 ; Ladefoged et al., 1996 ;

Kreiman et al., 2010), ces résultats ne se retrouvent pas en contexte clinique où la plupart des études ne laissent apparaître aucun effet de la langue dans la perception de la qualité vocale (Anders et al, 1988 ; Yamaguchi et al., 2003 ; Yiu et al., 2008), laissant acceptée l'hypothèse d'universalité de la norme. Notre objectif était donc de vérifier cela notamment en testant, de façon croisée, des locuteurs dysphoniques francophones et italophones par des phoniâtres français et italiens afin d'évaluer si la langue des locuteurs et des auditeurs peut influencer le résultat de l'évaluation de la dysphonie (Ghio, Weisz et al., 2011). Si un écart conséquent était observé, une telle mise en évidence pourrait s'avérer finalement inquiétante dans d'autres formes de variations plus visibles en pratique clinique telles que des variations régionales (parler méridional vs francilien) ou sociolinguistiques (parler des banlieues/quartiers...)

3.1. *Corpus et méthode*

67 locuteurs dysphoniques ont été enregistrés sur de la lecture de texte (« Chèvre de M Seguin » pour le français, « Il deserto » pour l'italien) : 6 hommes et 34 femmes chez les francophones, 7 hommes et 20 femmes chez les italophones. La sévérité de la dysphonie était variable : de voix normale à sévèrement dysphonique. Les pathologies étaient diverses : nodules, polypes, oedèmes de Reinke, paralysies laryngées, kystes, dysphonies dysfonctionnelles à larynx normal.

Nous avons soumis nos voix françaises et italiennes à 4 centres hospitaliers d'écoute : deux centres français (Marseille, Toulouse) et deux centres italiens (Milan et Ravenne). 3 personnes expertes (ORL, phoniâtre ou orthophoniste) par centre ont participé individuellement à l'expérience qui consistait au test d'évaluation GRB de l'échelle d'Hirano (1981). Les stimuli ont été présentés par bloc (français puis italien), l'ordre étant aléatoire dans un bloc. La présentation automatique des stimuli et le recueil informatisé des réponses ont été effectués par le logiciel libre PERCEVAL (www.lpl-aix.fr/~lpldev).

3.2. Résultats

Les résultats portent sur 2412 jugements perceptifs individuels. Chaque dimension G, R et B a été analysée séparément. Nous avons regroupé les résultats en prenant la valeur modale (majoritaire) parmi les jurys d'un pays et nous avons comparé les résultats du jury FRA vs ITA. L'analyse statistique des données montre une bonne cohérence sur le grade G entre auditeurs FRA et ITA (Kappa de Cohen = 0.78) et une pente de corrélation proche de 1, ce qui montre que le degré de sévérité est perçue de façon analogue à travers les 2 langues observées (Figure 2, à gauche). Il n'existe pas non plus de tendance différente sur la dimension soufflée B (Figure 2, à droite), bien que la cohérence soit moins nette ($K = 0.55$). Le résultat le plus remarquable porte sur la raucité R avec une très nette sous-évaluation du degré R par les juges italiens ($R_{ita} = 0.63 * R_{fra}$). En effet, si nous éliminons les voix R0 qui peuvent pas être sous-estimées, nous observons que 33% des données restantes sont sous évaluées par les auditeurs transalpins. Ces résultats sont à rapprocher des analyses acoustiques effectuées par Wagner & Braun (2003) sur des locuteurs normophoniques à travers plusieurs langues et où les auteurs observaient une tendance « rauque » des locuteurs italiens. Cela reste à confirmer par de plus amples données mais nous pensons mettre en évidence un cas d'adaptation de la perception (tolérance à la raucité) à la production (plus grande instabilité laryngée normale). Dans tous les cas, nos résultats semblent converger vers un aspect non universel de la norme.

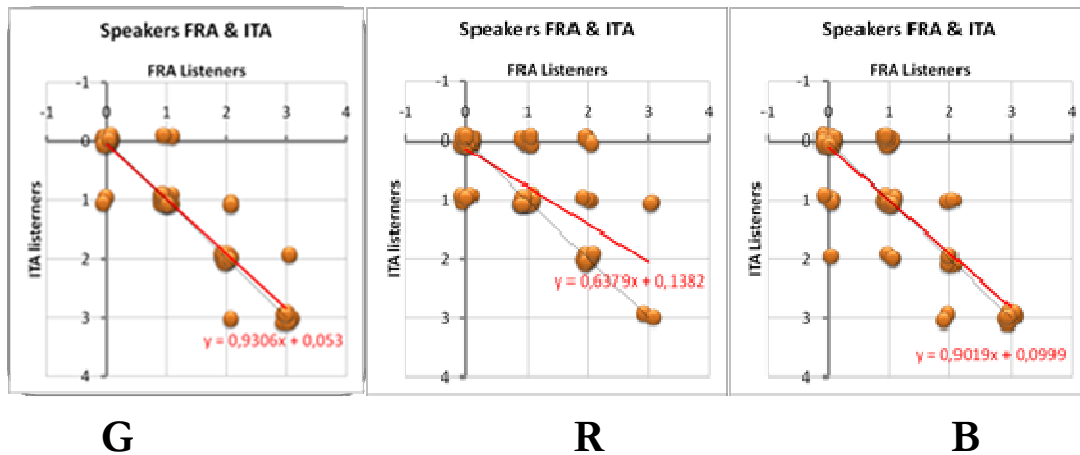


FIGURE 2 – Perception des voix par des auditeurs italiens (axe vertical) par rapport à des auditeurs français (axe horizontal) pour le grade globale de dysphonie (G), la raucité (R) et le souffle (B). La droite rouge est la courbe de tendance linéaire.

4. Expérience 3 : un protocole d'apprentissage perceptif

Le but de cette troisième expérience était d'examiner s'il était possible d'améliorer la fiabilité de l'évaluation perceptive par le biais d'un apprentissage (Ghio, Dufour et al., 2011). En effet, l'évaluation perceptive de la qualité vocale s'apparente à un processus de catégorisation (associer un grade à une production vocale) qui nécessite d'être capable d'associer dans une même catégorie des productions vocales proches mais aussi d'être capable de distinguer des productions vocales de deux catégories différentes. L'intérêt majeur d'un tel apprentissage est qu'il pourrait permettre de réduire la subjectivité dans les jugements des auditeurs. Un panel d'auditeurs naïfs a donc été soumis à un apprentissage intensif de catégorisation de la sévérité de patients dysphoniques et nous avons examiné l'impact de cet apprentissage sur la convergence des jugements entre auditeurs ainsi que sa persistance dans le temps. Cette démarche s'apparente à celles de Martin et al. (1996), Chan et al. (2002), Wipf (2003) mais à la différence de ces études, nous avons choisi de faire porter l'apprentissage sur des voix dysphoniques réelles et non synthétiques ou simulées. De plus, nous avons opté pour un apprentissage à travers la notion de réponses correctes/incorrectes, notion absente chez Wipf.

4.1. Corpus et méthode

La sélection des échantillons vocaux utilisés pour le processus d'apprentissage a fait l'objet d'un développement complexe et rigoureux. En effet, il n'existe aucun consensus ni collection de données susceptible de servir d'exemplaires à nos catégories de voix légèrement, moyennement ou sévèrement dysphoniques. Il a donc fallu construire cet échantillon. Pour cela, nous avons sélectionné un nombre important d'enregistrements extrait de la base de données de locuteurs dysphoniques du CHU de la Timone à Marseille (Ghio, Pouchoulin et al., 2012). N'ont été finalement retenus comme exemplaires de chaque grade que les échantillons vocaux qui recevaient un consensus entre 3 experts qui était lui-même confronté à l'évaluation par un dispositif de caractérisation automatique de la dysphonie (Fredouille et al., 2009). Ce procédé complexe de sélection de voix croisant de multiples évaluations perceptives et mesures instrumentales était la garantie d'obtenir au final des exemplaires que nous pouvions considérer comme représentatifs et suffisamment variés pour refléter de façon naturelle les expressions de la dysphonie. Nous disposons ainsi d'un corpus complet de 142 voix réparties comme suit: 33 voix normales (G0), 32 voix légèrement dysphoniques (G1), 35 voix moyennement dysphoniques (G2), 42 voix sévèrement dysphoniques (G3).

L'expérience a ensuite porté sur 38 auditeurs naïfs et s'est déroulée en 3 phases : une phase de pré-test permettant d'évaluer la performance initiale de nos auditeurs dans la catégorisation des dysphonies, une phase d'apprentissage et une phase de post-test permettant de tester les changements dans la catégorisation des dysphonies liés à notre protocole d'apprentissage. Pour la passation informatisée du protocole, nous avons utilisé le logiciel PERCEVAL avec son extension LANCELOT (www.lpl-aix.fr/~lpldev).

Durant le pré-test, les auditeurs avaient à évaluer 20 voix, 5 par grades, présentées 3 fois aléatoirement. Durant la phase d'apprentissage, 4 blocs de voix étaient présentés aux participants. Chaque bloc contenait 6 voix par grade et chaque voix était

présentée deux fois. Durant, les trois premiers blocs, les participants avaient une indication quant à la réponse correcte alors que dans le dernier bloc aucune indication n'était donnée aux participants. Durant le post-test, la même procédure que celle du pré-test a été utilisée, ce qui nous permettait de tester les changements liés à l'apprentissage dans la capacité des auditeurs à catégoriser la dysphonie. Dans le but d'évaluer la persistance de l'apprentissage, le post-test a été administré à deux reprises, immédiatement après et une semaine après l'apprentissage.

4.2. *Résultats*

Les résultats ont révélé une augmentation de la performance entre le pré-test et le post-test seulement pour les grades G2 et G3 (détails dans Ghio, Dufour et al., 2011). Aucune amélioration n'a été observée pour le grade G0 pour qui la performance de départ était déjà très bonne, ni pour grade G1 qui semble donc être le plus résistant à un apprentissage. L'amélioration observée pour les grades G2 et G3 était toujours présente à une semaine. Des analyses additionnelles ont par ailleurs montré une amélioration dans la cohérence des jugements sur l'ensemble des grades. Enfin, il faut préciser que même si des améliorations ont été observées sur G2, la performance sur ce grade restait moins bonne que celle observée sur les grades extrêmes G0 et G3.

4.3. *Discussion*

Le protocole d'apprentissage s'avère efficace. Il a l'avantage d'être construit sur de la parole naturelle, ce qui rend les acquis utilisables pour le jugement de la voix de patients en situation clinique. Les exemplaires choisis pour chaque grade de dysphonie ont fait l'objet d'une sélection sévère fondée sur de multiples analyses pour lesquelles la concordance était exigée. Seul un tel échantillonnage non ambigu permet à l'auditeur de se forger une représentation assez solide de la sévérité de la dysphonie dans ses multiples formes. La difficulté observée sur l'apprentissage du grade G1 et plus généralement sur les grades intermédiaires G1/G2 pourrait laisser penser que la catégorisation de la sévérité de la

dysphonie sur 4 niveaux n'est pas optimale. On peut ainsi partir de l'hypothèse qu'une réduction de la densité de l'espace perceptif sur 3 catégories (voix normale, moyennement dysphonique, sévèrement dysphonique) serait plus adéquate et correspondrait mieux aux capacités perceptives des auditeurs. A l'opposé, on pourrait remettre en cause l'aspect catégoriel de cette tâche de perception de la dysphonie et adopter un point de vue analogique s'apparentant à une mesure de quantité, c'est-à-dire utiliser des échelles visuelles analogiques. Wuyts et al. (1999) ont pointé les difficultés liées à de telles échelles avec notamment une faiblesse de reproductibilité et de concordance inter-auditeurs mais dans leurs études, les auteurs ont opéré à une discrétisation des grandeurs continues en catégories discrètes, ce qui repose à nouveau le problème de la densité de catégories et des frontières inter-catégorielles. Une perspective pourrait être d'envisager la perception de la dysphonie sous une forme quantitative (pour le jugement de la sévérité) et de corrélérer cela à des mesures instrumentales en analysant, non pas de la concordance de catégories, mais plutôt de la corrélation de grandeurs.

5. Conclusion

Ces expériences ont permis de mettre en évidence la complexité du processus d'évaluation perceptive de la dysphonie et la nécessité de mieux comprendre comment un expert ou un auditeur « naïf » perçoit de la parole perturbée. Les effets de contexte sont très importants et il convient donc de les contrôler parfaitement. L'approche aveugle apparaît comme la seule garante de résultats non entachés de considérations extra-acoustiques. La non universalité de la norme est évidemment une raison essentielle aux différences observées entre les auditeurs. Des études sont en cours pour explorer ce paradigme sur des nuances plus étroites comme des variations régionales. Une des solutions pour obtenir un procédé fiable d'évaluation perceptive réside probablement dans un apprentissage « dirigé » offrant un référentiel consensuel.

Remerciements

Ce projet a été soutenu par l'Agence Nationale de la Recherche (ANR) dans le cadre du projet DESPHO-APADY ANR-08-BLAN-0125 ainsi que par le projet COST Action 2103 "Advanced Voice Function Assessment" de l'Union Européenne.

Nous remercions S Mérienne, M Rouaze, V Bokanowski, F Weisz, G Baracca, G Cantarella, V Woisard, F Fussi, MD Guarella, C Spezza, M Puech, S Cretani, T Fuschini, V Geminiani, A Zambarbieri pour leur participation aux expériences.

Références bibliographiques

- ANDERS, HOLLIEN H, HURME, SONNINEN A, WENDLER J, "Perception of Hoarseness by Several Classes of Listeners", *Folia Phoniatr Logop*, 40:2, 91-100, 1988
- BELE I, Reliability in Perceptual Analysis of Voice Quality. *Journal of Voice*. 2005; 19(4), 555-573.
- CHAN KMK, YIU EM, The Effect of Anchors and Training on the Reliability of Perceptual Voice Evaluation. *Journal of Speech, Language, and Hearing Research*. 2002; 45(1), 111-126.
- DE BODT MS, WUYTS FL, VAN DE HEYNING PH, CROUX C, Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *J Voice*. 1997; 11(1), 74-80.
- DEJONCKERE PH, BRADLEY P, CLEMENTE P, CORNUT G, CREVIER-BUCHMAN L, FRIEDRICH G, VAN DE HEYNING P, REMACLE M, WOISARD V, A basic protocol for functional assessment of voice pathology, especially for investigating the efficacy of (phonosurgical) treatments and evaluating new assessment techniques. Guideline elaborated by the Committee on Phoniatics of the European Laryngological Society (ELS). *Eur Arch Otorhinolaryngol*. 2001; 258(2), 77-82.
- FREDOUILLE C, POUCHOULIN G, GHIO A, REVIS J, BONASTRE J, GIOVANNI A, Back-and-Forth Methodology for Objective Voice Quality Assessment: From/to Expert Knowledge to/from Automatic Classification of Dysphonia. *EURASIP Journal on Advances in Signal Processing*. 2009; 2009, 1-14.
- GERRATT B, KREIMAN J, Theoretical and methodological development in the study of pathological voice quality. *Journal of Phonetics*. 2000; 28(3), 335-342.
- GHIO A; DUFOUR S; ROUAZE, M; BOKANOWSKI, V; POUCHOULIN, G; RÉVIS, J; GIOVANNI, A (2011). Mise au point et évaluation d'un protocole d'apprentissage de jugement perceptif de la sévérité de dysphonies sur de la parole naturelle. *Revue de laryngologie, otologie, rhinologie*, vol. 132, no. 1. 2011, p. 19-27

- GHIO, A; MERIENNE, S; GIOVANNI, A (2011). Influence du contexte d'écoute dans la perception des dysphonies : effets de la connaissance de l'état thérapeutique du patient sur le jugement de sa qualité vocale. *Revue de laryngologie, otologie, rhinologie*, vol. 132, no. 1. 2011, p. 9-17.
- GHIO, A; WEISZ, F; BARACCA, G; CANTARELLA, G; ROBERT, D; WOISARD, V; FUSSI, F; GIOVANNI, A (2011). Is the perception of voice quality language-dependant? A comparison of French and Italian listeners and dysphonic speakers. *Proceedings of Interspeech, Firenze: ISCA*. 2011, p. 525-528.
- GHIO, A; POUCHOULIN, G; TESTON, B; PINTO, S; FREDOUILLE, C; DE LOOZE, C; ROBERT, D; VIALLET, F; GIOVANNI, A (2012). How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers? *Speech Communication*, vol. 54. 2012, p. 664-679.
- HERZ, R., *The effect of verbal context on olfactory perception*. (2003) *J Exp Psychol Gen*. 132(4), 595-606.
- HIRANO M, *Clinical Examination of Voice*. 1981; Springer Verlag, Wien
- KREIMAN J, GERRATT BR, PRECODA K, BERKE GS, Individual differences in voice quality perception. *J Speech Hear Res*. 1992; 35(3), 512-520.
- KREIMAN J, GERRATT BR, KEMPSTER GB, ERMAN A, BERKE GS, Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Hear Res*. 1993; 36(1), 21-40.
- KREIMAN J, GERRAT B, KHAN S, "Effects of native language on perception of voice quality", *Journal of Phonetics*, 38: 4, 588-593, 2010.
- LADEFOGED, MADDIESON, *The sounds of the world's languages*. Blackwell Publishers, 1996.
- MARTIN DP, WOLFE VI, Effects of perceptual training based upon synthesized voice signals. *Percept Mot Skills*. 1996; 83(3 Pt 2), 1291-1298.
- NI CHASAIDE A, GOBL A, "Voice Source Variation", in *The Handbook of Phonetic Sciences*, W. J. Hardcastle, J. LAVER, Ed. Oxford, UK: Blackwell Pub. Ltd, 1999.
- REVIS J, GIOVANNI A, WUYTS F, TRIGLIA J, Comparison of different voice samples for perceptual analysis. *Folia Phoniatr Logop*. 1999; 51(3), 108-116.
- SHRIVASTAV R, Multidimensional Scaling of Breathly Voice Quality: Individual Differences in Perception. *Journal of Voice*. 2006; 20(2), 211-222.
- VAISSIERE J, "Phonological Use of the Larynx", *Proc. Larynx 1997*, Marseille, France, 1997, 115-126.
- WAGNER, BRAUN, "is voice quality language dependant ?", *Proc. ICPhS, Barcelona*, 2003, 651-654
- WARREN RM., WARREN RP. (1970), *Auditory illusions and confusions*. *Sci. Am.*; 223, 30-36.
- WIPF A.-L., (2003) *Elaboration d'un protocole d'entraînement à l'analyse perceptive des dysphonies pour un jury inexpérimenté. Mémoire pour le certificat de capacité d'orthophonie*, Université Aix-Marseille II.
- WUYTS FL, DE BODT MS, VAN DE HEYNING PH, Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of dysphonia. *J Voice*. 1999; 13(4), 508-517.

- YAMAGUCHI H, SHRIVASTAV R, ANDREWS M, NIJMI S, "A Comparison of Voice Quality Ratings Made by Japanese and American Listeners Using the GRBAS Scale", *Folia Phoniatr Logop*, 55: 3, 147-157, 2003.
- YIU E, MURDOCH B., HIRD K, LAU P, HO E, "Cultural and language differences in voice quality perception: a preliminary investigation using synthesized signals", *Folia Phoniatr Logop*, 60: 3, 107-119, 2008.