Mise au point et évaluation d'un protocole d'apprentissage de jugement perceptif de la sévérité de dysphonies sur de la parole naturelle

Alain Ghio, Sophie Dufour, Maud Rouaze, Valérie Bokanowski, Gilles Pouchoulin, Joana Révis, Antoine Giovanni

Rev Laryngol Otol Rhinol.
2011;132,1:1-9.

PHONIATRIE

ARTICLE ORIGINAL

Mise au point et évaluation d'un protocole d'apprentissage de jugement perceptif de la sévérité de dysphonies sur de la parole naturelle [1]

Perceptual assessment of dysphonia: A training protocol with natural speech

Ghio A. [2]
Dufour S. [2]
Rouaze M. [3]
Bokanowski V. [3]
Pouchoulin G. [2]
Révis J. [2/3]
Giovanni A. [2/3]
(Marseille)

Résumé

Introduction *: l'objectif de l'étude était de mettre au point et d'évaluer un protocole d'apprentissage d'évaluation perceptive des dysphonies. En particulier, des auditeurs naïfs apprenaient à juger de la sévérité du trouble vocal.* ***Matériels et méthodes*** *: le corpus utilisé était constitué de 142 voix de femmes en lecture avec un échantillonnage en qualité vocale variée allant de voix normales à des voix sévèrement dégradées. L'expérience de perception a porté sur 38 auditeurs naïfs et s'est déroulée en 3 phases : (1) un pré-test permettant de mesurer la performance de base des auditeurs dans la catégorisation des dysphonies, (2) un apprentissage proprement dit et (3) un post-test permettant de tester les changements liés à l'apprentissage. Dans le but d'évaluer la persistance de l'apprentissage, le post-test a été administré planifié à deux reprises, immédiatement après et une semaine après l'apprentissage.* ***Résultats*** *: les résultats ont révélé une augmentation de la performance entre le pré-test et le post-test pour le jugement des dysphonies moyennes et sévères. Aucune amélioration n'a été observée pour le jugement des voix normales pour qui la performance de départ était déjà bonne, ni pour les dysphonies légères qui semblent donc être les plus résistantes à un apprentissage. Les améliorations observées sur les dysphonies moyennes et sévères étaient toujours présentes à une semaine.* ***Discussion*** *: le protocole d'apprentissage semble efficace et pourrait être proposé, par exemple, dans la formation des thérapeutes de la voix même si le jugement des grades intermédiaires de dysphonies reste, malgré tout, fragile et demande à être renforcé.*

Mots-clés : Évaluation perceptive des dysphonies, apprentissage, voix.

Summary

Introduction*: In this study, we proposed and tested the effectiveness of a training procedure on perceptual evaluation of dysphonia. Participants were naive listeners who learned to judge the severity of vocal disorders.* ***Materials and methods****: The corpus consisted of 142 female voices ranging from normal to severely dysphonic voices. A total of 38 naive listeners were tested, and the experiment was divided in 3 phases: (1) A pretest to assess the level of our listeners in the categorization of dysphonia, (2) the training phase and (3) a post-test to assess the improvement associated with the training. To examine the persistence of the training, the post-test was administrated twice: Immediately after the training session and one week later.* ***Results****: Improvement in the performance between the pretest and the post-test was found for moderate and severe dysphonia. No improvement was observed in the categorization of slightly dysphonic voices, which seem the more resistant to learning. Normal voices also showed no improvement, which is likely due to the high performance on these voices during the pre-test. The improvement observed in the categorization of moderate and severe dysphonia was still present even one week after the training.* ***Discussion****: The protocol tested in this study appears to be effective and could constitute an element of training courses for speech and voice therapists. The perception of the intermediate levels of dysphonia, however, remains difficult to quantify and needs to be improved.*

Key-words: Perceptual voice assessment, training, voice.

INTRODUCTION

L'évaluation perceptive de la dysphonie

Dans la prise en charge de patients dysphoniques, l'évaluation perceptive de la qualité vocale est essentielle. En effet, 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 audi-

1. Communication présentée au LXVI ème Congrès, Société Française de Phoniatrie et Pathologies de la Communication, Paris, 18/10/10.
2. Laboratoire Parole et Langage, UMR 6057 CNRS, Université Aix-Marseille, France.
E-mails: alain.ghio@lpl-aix.fr, sophie.dufour@lpl-aix.fr
3. CHU La Timone, Fédération d'ORL et de chirurgie Cervico-faciale, Université Aix-Marseille, France.

Article reçu : 04/04/11 accepté : 04/05/11

tive laissée par leur voix [1]. 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 [2]. Cette méthode est ainsi recommandée dans le protocole minimal d'évaluation fonctionnelle des pathologies de la voix proposé par De Jonckere et al [3], en particulier pour estimer l'efficacité de traitements en phonochirurgie. Elle présente en effet de nombreux avantages : facile à mettre en œuvre, peu coûteuse, directement accessible à tout clinicien. 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 [4-10]. 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 [7], choix d'échelles analogiques vs. catégorielles [8], type d'énoncés (voyelle tenue vs. phrases) [9], dimensions perçues (qualité globale vs. dimension soufflée [10] vs. raucité.)…

Variabilité vs. fiabilité

De manière générale, les études en la matière portent leur attention sur l'observation des phénomènes de variabilité vue comme un marqueur du manque de fiabilité de la méthode. Il nous apparait préférable de tester directement la fiabilité de la méthode par une mesure de la justesse des réponses. En effet, s'il semble légitime de postuler qu'une bonne fiabilité des réponses aura pour conséquence une faible variabilité, une réponse correcte étant par essence non variable, la réciproque n'est pas triviale. En effet, si on se place dans une tâche de jugement sur 4 niveaux de la sévérité d'une dysphonie comme dans le cas de l'échelle d'Hirano [11] où G0 est une voix normale, G1 une voix légèrement dysphonique, G2 une voix moyennement dysphonique, et G3 une voix sévèrement dysphonique, un auditeur qui répondrait systématiquement G0 pour des voix globalement «normales» et G2 pour des voix globalement «anormales» obtiendrait une faible variabilité du fait de sa stratégie simplifiée de réponses mais ne pourrait pas être considéré comme fiable dans la mesure où il ne répond pas à la tâche de catégorisation sur 4 niveaux.

Selon Bele [6], la fiabilité d'une évaluation est liée au degré avec lequel les résultats sont exempts d'erreurs de mesure. L'auteur distingue notamment (1) les erreurs dues au hasard (distraction de l'auditeur, mauvaise manipulation dans le choix de la réponse…) (2) des erreurs systématiques. Si la première source d'erreurs peut se minimiser par la répétition des tests et par le nombre important de participants, la deuxième affecte le score de l'auditeur et révèle une caractéristique particulière de ce participant. Si l'erreur est générale à l'ensemble des auditeurs, ces erreurs systématiques dévoilent la nature particulière du stimulus ou les limites de la méthode d'évaluation. Notre approche dans ce travail va dans ce sens là.

Se pose alors la question centrale de la catégorisation des voix dysphoniques. Qu'est-ce qu'une voix «normale» ? Qu'est-ce qu'une voix moyennement, légèrement ou sévèrement dysphonique ? Cette question est d'autant plus complexe que les manifestations multiples de la dysphonie sous la forme de voix soufflée, rauque, diplophonique, hyper ou hypo fonctionnelle, etc. empêchent la constitution de catégories claires et précises que ce soit en termes de qualité ou de quantité. Nous abordons ici la notion d'exemplaires ou de prototypes catégoriels.

Catégorisation, exemplaires et prototypes

L'évaluation des dysphonies 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»

• d'être capable de distinguer des productions vocales de deux catégories différentes

Selon la théorie dite «classique», les catégories se forment sur la base de propriétés communes et appartient à la dite catégorie toute entité possédant au moins ces propriétés. Si un élément possède telle propriété alors il appartient à une catégorie X ; par contre s'il ne possède pas telle propriété, il n'appartient pas à cette catégorie. Cette approche a été remise en question par Rosch [12] qui au travers d'expériences a montré que pour la plupart des catégories, il était difficile de définir un ensemble de propriétés nécessaires et suffisantes les caractérisant. L'exemple le plus souvent cité est celui de l'autruche qui tout en possédant beaucoup de traits définissant la catégorie oiseau (avoir deux pattes, deux ailes, des plumes, un bec, pondre des œufs, etc.) ne vole pas. Rosch constate aussi que certains membres sont plus représentatifs que d'autres d'une catégorie et introduit ainsi la notion de prototypes. Le prototype est le meilleur représentant d'une catégorie et peut ainsi varier d'un individu à l'autre en fonction de leur expérience particulière. L'appartenance à une catégorie n'est ainsi plus en termes de vrai ou faux mais est fonction du degré de similarité avec le prototype, certains éléments étant centraux dans la catégorie alors que d'autres étant plus périphériques. Comme nous le verrons dans ce qui suit, la notion de prototypes émanant principalement des recherches en psychologie cognitive s'apparente à la notion de référents internes décrits par Kreiman et Gerrat [4, 5, 13].

LE RENFORCEMENT DES STANDARDS INTERNES PAR APPRENTISSAGE

Des prototypes peu stables et peu partagés

Le manque de fiabilité dans l'évaluation perceptive des dysphonies avec ses variabilités intra et inter-individuelles qui en découlent, sont largement dépendantes des stratégies et des mécanismes mis en jeu par les auditeurs lors de l'évaluation d'une voix, et en particulier des références auditives de ce jury [4]. Kreiman et al ont ainsi introduit la notion de «standard interne» qui n'est autre que la notion de prototype utilisée en psychologie cognitive : chaque auditeur juge la qualité d'une voix en la confrontant à un standard interne auditif ou prototype, le standard interne auditif renvoyant au modèle sonore que l'on se fait d'une voix normale ou dysphonique. C'est la distance estimée perceptivement entre ce standard interne et la voix entendue qui permet d'attribuer à cette dernière un degré de sévérité. Or, ces standards internes sont globalement propres à chaque individu et plus ou moins bien définis en fonction de l'expérience perceptive de l'auditeur face à ce type de voix.

Ancrage externe par comparaison

Une des alternatives, proposées notamment par Gerrat et al [13], a été de substituer les standards internes par des ancrages externes pouvant constituer un ensemble de références perceptives. En fournissant une palette constante et partagée de voix, la méthode permet à tous les auditeurs de catégoriser les échantillons par comparaison à de mêmes références [14]. Les auteurs ont effectivement démontré l'efficacité de cet apport d'ancrages externes dans l'évaluation de la raucité de la voix [13]. Le recours à des voix synthétiques [15, 17, 18] est apparu de façon légitime afin de pouvoir disposer d'une palette de voix représentatives des diverses expressions de la dysphonie en termes de qualité (soufflée, rauque) et de sévérité. Mais ce paradigme proposant des voix d'ancrage synthétiques et/ou pseudo-naturelles* reste peu satisfaisant car d'une part, les voix synthétiques sont trop artificielles pour être comparées aux voix naturelles et d'autre part, les voix naturelles ne sont pas suffisamment calibrées pour constituer un exemplaire représentatif. Enfin, ce procédé ayant recours à une comparaison systématique à des références externes s'écarte du processus régulier de la perception de la parole et ne place pas l'auditeur en situation «d'autonomie» de jugement par la suite.

Ancrage interne par apprentissage

En effet, dans le jugement perceptif de la dysphonie tel qu'il est pratiqué actuellement, nous sommes assez éloignés d'un processus de mesure d'intelligibilité comme celui de paires minimales pratiqué, par exemple, dans l'évaluation de troubles arthriques. Les tests d'intelligibilité font appel aux représentations perceptivo-cognitives stables et partagées entre les auditeurs d'une même langue. Or, un tel procédé, bien que perceptif, ne laisse part qu'à une faible subjectivité du fait de la solidité des représentations acquises par des années d'apprentissage.

Notre objectif était donc de mettre en place un processus d'apprentissage s'inspirant de protocoles déjà mis en place dans le cadre, par exemple, d'un apprentissage de mots nouveaux ou de phonèmes [19, 20]. Cette démarche s'apparente à celles de Martin [15], Chan [18], Wipf [21] mais à la différence de ces études [15, 18], 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 dans [21].

MATÉRIEL ET MÉTHODE

Méthode d'évaluation et de catégorisation

Nous avons opté pour l'échelle GRBAS d'Hirano [11] qui est la méthode d'évaluation perceptive de la dysphonie la plus couramment utilisée. Afin de simplifier le processus de catégorisation, nous avons choisi de focaliser l'attention des participants sur la seule dimension G, représentative de la sévérité globale de la dysphonie. Ainsi, le processus de catégorisation se réduisait à quatre classes : G0 = voix normale, G1 = voix légèrement dysphonique, G2 = voix moyennement dysphonique, G3 = voix sévèrement dysphonique. Ce choix peut être discuté dans la mesure où la dimension G ne rend pas compte des aspects multidimensionnels du dysfonctionnement vocal. En effet, une voix moyennement dysphonique (G2) peut être à la fois soufflée (ex : G2 R0 B2) ou rauque (ex : G2 R2 B0) ou les deux (ex : G2 R2 B2). D'autres dimensions peuvent aussi intervenir de façon variable comme la bitonalité, le caractère tendu/lâche, l'instabilité temporelle. Par conséquent, les exemplaires sonores d'une même classe sont plutôt hétérogènes d'un point de vue acoustique. En revanche, nous avons postulé que la dimension G était la dimension la plus fiable dans cette échelle [22], celle-ci demandant aux auditeurs de catégoriser la sévérité de la dysphonie en intégrant ses expressions multiples.

Enoncés

Notre objectif était de nous placer dans un contexte le plus écologique possible. Nous avons donc écarté les voyelles tenues qui s'éloignent des énoncés produits en parole naturelle et qui introduisent une sous-estimation du niveau de dysphonie [9]. Nous avons aussi exclu la parole spontanée qui ne permet pas de standardiser le protocole. Notre choix s'est donc porté sur de la lecture de texte (« La chèvre de Monsieur Seguin » d'Alphonse Daudet), support utilisé depuis une vingtaine d'année dans le protocole d'enregistrement des patients dysphoniques au service ORL du CHU de la Timone à Marseille. La dysphonie pouvant se manifester temporellement de

* Afin d'obtenir une palette complète de voix dysphoniques, Chan et al [17] ont eu recours à un locuteur sain simulant certains degrés de raucité ou de souffle, d'où notre terminologie de «pseudo-naturelle».

façon non uniforme, nous avons extrait une phrase (« il les perdait toutes de la même façon ») de cette lecture. Cette phrase a été choisie car elle comprend une succession de transitions voisé/non voisé, des continuums vocaliques, et sa structure prosodique induit une accentuation naturelle sur le mot « toutes », structure particulièrement intéressante car la montée nécessaire de F0, pas toujours bien réalisée, couplée à une structure syllabique [occlusive sourde + voyelle + occlusive sourde + voyelle**] peut être révélatrice dans le cas de difficultés laryngées.

Locuteurs et corpus

Un choix méthodologique pour des exemplaires naturels

La difficulté majeure dans ce type d'expérimentation est la sélection des exemplaires représentatifs des catégories qui vont constituer les modèles d'apprentissage. Or, dans l'étude de la dysphonie, l'absence de modèle théorique représente une difficulté majeure pour le choix de ces exemplaires. En effet, rien dans la littérature, et tel est le problème majeur actuellement, ne permet de définir ce qu'est une voix normale, légèrement, moyennement ou sévèrement dysphonique. Martin et Wolfe [15] ont simplifié et contourné le problème en synthétisant des voix artificielles en faisant varier le jitter (instabilité de la fréquence fondamentale instantanée) et le rapport signal/bruit des stimuli de synthèse. Chan et Yiu [18] ont suivi le même principe en manipulant «l'amplitude d'aspiration» et «l'index de diplophonie» du synthétiseur de Klatt HLSyn. Dans notre étude, nous avons procédé différemment afin de manipuler uniquement des voix naturelles.

Patients et locuteurs

Nous avons utilisé l'importante quantité de locuteurs dysphoniques et contrôles enregistrés au Service ORL du CHU de la Timone à Marseille ou dans le Service de Neurologie du CH du Pays d'Aix. Cette base de données inclut environ 2500 locuteurs dysphoniques ou dysarthriques [23]. Afin de réduire la tâche de catégorisation, nous avons choisi de centrer notre corpus sur des voix de femmes, présentant des dysphonies dysfonctionnelles diverses telles que nodules, polypes, kystes ou oedèmes de Reinke. 400 voix de femmes ont ainsi été extraites. Outre la date d'enregistrement, le nom de la patiente, sa pathologie et la précision sur le contexte pré ou post-opératoire, nous disposions de l'évaluation perceptive du grade G d'Hirano réalisée par un ou deux expert(s) orthophoniste(s) ou phoniatre(s) présent(s) au moment de l'enregistrement.

Le processus de sélection en entonnoir

Chaque extrait vocal a été ensuite analysé par un dispositif de caractérisation automatique de la dysphonie (CAD). Cet analyseur, issu de la reconnaissance automatique du locuteur, a été adapté à l'analyse de la qualité vocale et a prouvé ses capacités à caractériser la sévérité de la dysphonie dans environ 80 % des cas [24]. Le principe d'un tel système est le suivant. Dans une première phase dite d'apprentissage, le dispositif construit des modèles statistiques de catégories (ex : les 4 grades de dysphonie) à partir d'exemples sonores qui lui sont fournis. Cela s'apparente finalement à l'établissement de prototypes à partir d'exemplaires. Une fois ces modèles de grade appris, lorsqu'on soumet au dispositif un échantillon de voix à tester, il fournit en sortie des mesures de vraisemblance estimées sur chacune des 4 catégories, la plus forte valeur indiquant ainsi la classe d'appartenance. La phase d'apprentissage du système avait déjà été opérée auparavant [24] et nous avons utilisé la CAD uniquement en phase de test pour évaluer nos voix du corpus. Nous avons finalement gardé les échantillons du corpus pour lesquels nous obtenions une concordance exacte entre le ou les 2 grades perceptifs fournis au moment de l'enregistrement et le dispositif de CAD.

Dans un deuxième temps, nous avons extrait la phrase cible retenue pour notre protocole (voir paragraphe précédent) et avons à nouveau procédé à une analyse par CAD sur cet extrait afin de vérifier si la catégorisation était identique sur le texte entier et sur la phrase seule. Les enregistrements ne remplissant pas ce critère ont été éliminés. Au terme de cette étape, nous n'avions pas la taille de corpus nécessaire. Il nous a donc fallu revoir nos critères de choix. Nous avons réintégré dans notre corpus certaines voix pour lesquelles la catégorisation par la CAD et les évaluations perceptives faites au moment de l'enregistrement n'était pas parfaitement concordante entre elles, considérant que l'évaluation non aveugle par un unique auditeur était discutable compte-tenu des conditions dans lesquelles elle avait été réalisée. Nous avons donc procédé sur toutes les voix présélectionnées à une évaluation perceptive en aveugle réalisée par consensus par un jury d'experts composé d'une phoniatre et deux orthophonistes sur la phrase «il les perdait toutes de la même façon». Le jury pouvait écouter plusieurs fois le même stimulus.

Le corpus final

Au final, seules les voix faisant consensus entre les 3 experts et le dispositif de mesure instrumentale étaient retenues. Parmi ces dernières, nous avons opéré une ultime sélection : nous avons retenu les voix qui, d'après le système de CAD, présentaient la plus grande probabilité d'appartenance au grade retenu. Il fallait également que la probabilité d'appartenir à un autre grade soit la plus faible possible. Pour chaque grade, nous avons donc classé les voix par ordre décroissant de probabilité d'appartenance. Si deux voix avaient une probabilité d'appartenance identique pour le grade en question, nous choisissions celle qui présentait les plus faibles probabilités d'appartenance aux autres grades.

Nous disposions ainsi d'un corpus complet de 142 voix réparties comme suit : 33 voix en grade G0, 32 voix en grade G1, 35 voix en grade G2, 42 voix en grade G3.

** Nos locuteurs, la plupart d'origine méridionale, prononcent le «e» final (schwa) dans le mot «toutes».

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. Ce procédé a été rendu possible du fait de l'importante masse de données structurées et accessibles dont nous disposons à présent [23]. L'ensemble des voix utilisées pour notre expérience a été extrait de ce corpus final.

Auditeurs

38 étudiants de première année d'école d'orthophonie ont participé à l'expérience. Ils étaient tous de langue maternelle française et n'ont rapporté aucun trouble de l'audition. Pour les besoins de l'expérience, les participants ont été divisés en quatre groupes (trois groupes de 10 et un groupe de 8).

Protocole

Pour le déroulement de l'expérience, nous avons utilisé le logiciel PERCEVAL avec son extension LANCELOT [25]. L'expérience 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.

Pré-test

Durant le pré-test, les auditeurs entendaient une série de voix qu'ils devaient associer à l'un des quatre grades testés. Un essai se déroulait de la façon suivante : Une voix était tout d'abord entendue puis apparaissait à l'écran les intitulés correspondant aux quatre grades (grade 0, grade 1, grade 2, grade 3). L'auditeur devait alors cliquer sur le grade correspondant à la voix entendue. Aucune indication sur la réponse correcte n'était donnée aux participants. Le pré-test était constitué d'un bloc de 20 voix, 5 par grade, présentées dans un ordre aléatoire. Afin de s'assurer que le pré-test ne permettait pas à lui seul une augmentation de la performance résultant d'une certaine habituation aux différentes sévérités, le bloc a été présenté 3 fois et nous avons comparé la performance au travers la répétition du bloc.

Apprentissage

Durant l'apprentissage, les auditeurs devaient apprendre à catégoriser des nouvelles voix en les associant à l'un des quatre grades. Dans la mesure où les grades G1 et G2 sont ceux pour lesquels la variabilité est la plus grande, ils ont été appris séparément ce qui permettait de renforcer les standards internes correspondant à chacun de ces deux grades. Cette démarche a été inspirée d'une technique de rééducation orthophonique consistant dans le cas de confusions auditives à renforcer un phonème puis l'autre avant de les opposer. L'apprentissage était divisé en 4 blocs. Chaque bloc contenait 6 voix par grade et l'ordre de présentation des voix à l'intérieur des blocs était aléatoire. Les 2 premiers blocs étaient constitués de 3 grades : G0, G1 et G3 pour le premier puis G0, G2 et G3 pour le second bloc. L'ordre de présentation de ces deux blocs a été contrebalancé au travers les participants de sorte à ce que G1 soit consolidé en premier pour la moitié des participants, et que G2 soit consolidé en premier pour l'autre moitié des participants. Le troisième bloc était constitué des 4 grades (G0, G1, G2 et G3) de façon à opposer G1 et G2 au cours même de l'apprentissage. Les 3 premiers blocs étaient présentés deux fois de façon successive et un essai se constituait de la façon suivante : une voix était présentée aux participants, puis apparaissait à l'écran l'intitulé des grades. Les participants devaient alors cliquer sur le grade correspondant à la voix entendue. Une fois qu'ils avaient donné leur réponse, le grade correct s'affichait et la voix était répétée. A la fin de chaque bloc, les participants étaient informés de leur performance sur l'ensemble du bloc. Enfin dans le dernier bloc (bloc 4), les 4 grades étaient à nouveau présentés mais les participants n'avaient aucune indication quant à la réponse correcte et quand à leur performance globale sur ce bloc. Une illustration de la procédure d'apprentissage est fournie dans le tableau I.

TABLEAU I : Illustration du protocole d'apprentissage (* G1 ou G2 suivant le groupe d'apprentissage).

Bloc	Grade et nombre de voix	Répétition du bloc	Feed-back
1	6G0, 6**G1/G2***, 6G3	oui	oui
2	6G0, 6**G2/G1***, 6G3	oui	oui
3	6G0, 6G1, 6G2, 6G3	oui	oui
4	6G0, 6G1, 6G2, 6G3	non	non

Post-test

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 de nos auditeurs à catégoriser des dysphonies. Dans le but d'évaluer la persistance de l'apprentissage, le post-test a été administré à deux reprises, immédiatement après l'apprentissage (J0), et une semaine après (J+1). Comme pour le pré-test, le post-test était composé d'un bloc de 20 voix, 5 par grade répété 3 fois. Notons que de façon à éviter qu'une éventuelle amélioration de la performance entre le pré-test et le post-test soit simplement liée aux particularités des voix utilisées dans chacun des tests, un contrebalancement a été effectué à l'intérieur des groupes d'apprentissage (G1 ou G2 consolidé en premier). Les 20 voix utilisées en pré-test pour la moitié des sujets ont été utilisées en post-test pour l'autre moitié et inversement, les 20 voix utilisées en post-test pour la moitié des sujets ont été utilisées en pré-test pour l'autre moitié des sujets.

RÉSULTATS

Pré-test

Les performances au pré-test sont indiquées dans le tableau II. Dans un souci de clarté, seuls les seuils de significativité pour les résultats significatifs sont précisés. Aucune mention aux statistiques n'est faite pour les comparaisons non significatives (valeurs du $p > .20$)

TABLEAU II : Pourcentage moyen de réponses correctes en fonction du type de grade lors du pré-test.

	G0	G1	G2	G3
1ere présentation	85	49	38	68
2ème présentation	85	49	38	69
3ème présentation	88	49	37	67
Sur l'ensemble des répétitions	86	49	38	68

Comme nous pouvons le constater la performance sur chaque grade n'a pas évoluée au fur et à mesure des répétitions, ce qui laisse suggérer qu'un simple jugement de la qualité des voix sans aucune information sur les réponses correctes ne suffit pas à lui seul à augmenter les performances dans la catégorisation de dysphonies. Dans la suite, nous présenterons donc uniquement les résultats obtenus sur l'ensemble des répétitions pour chaque grade.

Nous pouvons constater que les performances les plus basses ont été obtenues pour les grades intermédiaires G1 et G2. Les voix de grade G1 se sont toutefois révélées être plus faciles à catégoriser que les voix de grade G2 [$F (1,37) = 9.36$; $p < .01$]. Les meilleures performances ont été obtenues pour le grade G0 avec en moyenne 86 % de bonnes réponses [G0/G1 : $F (1, 37) = 65.55$; $p < .0001$, G0/G2 : $F (1, 37) = 100.44$; $p < .0001$, G0/G3 : $F (1, 37) = 15.81$; $p < .001$]. Le score obtenu pour le grade G3 est moins bon que celui obtenu pour le grade G0 mais il reste toutefois supérieur aux scores des grades G1 [$F(1,37) = 17.47$; $p < .001$] et G2 [$F (1,37) = 117.37$; $p < .0001$].

Apprentissage

Les résultats obtenus lors de l'apprentissage sont présentés dans les tableaux III et IV. Il va de soi que la meilleure façon d'évaluer l'impact d'un tel apprentissage est de comparer les performances entre le pré-test et le post-test. Nous pointons néanmoins ici les résultats de l'apprentissage qui nous semblait le plus pertinent. Par souci de clarté et étant données les nombreuses comparaisons effectuées, nous nous sommes abstenus de donner les significativités. Notons toutefois que tous les résultats discutés ci-dessous sont significatifs ($p < .05$).

Que ce soit pour le groupe qui a consolidé G1 en premier ou pour le groupe qui a consolidé G2 en premier, on constate que les performances sur G1, G2 et G3 se sont dès les 2 premiers blocs considérablement améliorées en comparaison avec les performances obtenues lors du pré-test. Par contre, dès que les grades G1 et G2 sont présentés au sein du même bloc, les performances obtenues notamment sur ces 2 grades intermédiaires rechutent même si celles-ci s'avèrent tout de même plus élevées qu'en pré-test. Enfin, les performances obtenues au bloc 4 sont assez élevées pour les grades G0 et G3. Soulignons d'ailleurs que contrairement au pré-test, elles sont au même niveau de performance. En revanche, pour les grades intermédiaires G1 et G2 les performances sont inférieures ou égales à 60 % mais tout de même plus élevées que celles obtenues lors du pré-test.

Post-test

Les résultats obtenus lors des post-tests sont présentés dans le tableau V.

Aucun effet du type d'apprentissage (G1 ou G2 consolidé en premier) n'a été mis en évidence. Par conséquent, nous discuterons les performances globales indépendamment du type d'apprentissage. Dans un souci de clarté, seuls les seuils de significativité pour les résultats significatifs sont précisés. Aucune mention aux statistiques n'est faite pour les comparaisons non significatives (valeurs du $p > .20$)

Immédiatement après l'apprentissage (J0) : de façon similaire au pré-test, les performances étaient significativement meilleures pour les grades G0 et G3 [G0/G1 : $F (1,37) = 78.03$; $p < .0001$, G0/G2 : $F (1,37) = 48.54$; $p < .0001$, G2/G3 : $F (1,37) = 252.75$; $p < .0001$, G1/G3 : $F (1,37) = 219.97$; $p < .0001$]. Contrairement au pré-test, la performance pour les voix de grade G3 était significati-

TABLEAU III : Pourcentage moyen de réponses correctes en fonction du type de grade pour le groupe ayant consolidé G1 en premier.

	Grade	G0	G1	G2	G3
Bloc 1	1ère présentation	74	65		89
	2ème présentation	81	77		93
Bloc 2	1ère présentation	98		88	83
	2ème présentation	98		88	83
Bloc 3	1ère présentation	87	63	63	82
	2ème présentation	90	58	61	93
Bloc 4	(sans feed-back)	80	54	58	88

TABLEAU IV : Pourcentage moyen de réponses correctes en fonction du type de grade pour le groupe ayant consolidé G2 en premier.

	Grade	G0	G1	G2	G3
Bloc 1	1ère présentation	81		75	93
	2ème présentation	81		75	93
Bloc 2	1ère présentation	91	81		78
	2ème présentation	96	89		85
Bloc 3	1ère présentation	83	65	59	87
	2ème présentation	92	69	60	93
Bloc 4	(sans feed-back)	92	59	60	89

TABLEAU V : Pourcentage moyen de réponses correctes en fonction du type de grade lors des post-tests (J0 = immédiatement après l'apprentissage ; J+1 = une semaine après l'apprentissage).

	G0	G1	G2	G3
G1 consolidé en premier (J0)	84	53	59	97
G2 consolidé en premier (J0)	84	53	59	96
Total (J0)	**84**	**53**	**59**	**96**
G1 consolidé en premier (J+1)	88	47	52	88
G2 consolidé en premier (J+1)	88	49	55	94
Total (J+1)	**88**	**48**	**53**	**90**

vement plus élevée que celle obtenue pour les voix de grade G0 [G0/G3 : F (1,37) = 28.16, p < .0001] et les performances obtenues pour les grades G1 et G2 ne différaient pas de façon significative.

Si l'on compare les performances obtenues au pré-test et au post-test immédiat, nous constatons un maintien de la performance pour les grades G0 (86 % versus 84 % pour le pré- et le post-test respectivement) et G1 (49 versus 53 % pour le pré- et le post-test respectivement). Par contre nous pouvons constater une nette amélioration de la performance pour les grades G2 (38 % versus 59 % pour le pré- et le post-test respectivement ; [F (1,37) = 19.5 ; p < .0001]) et G3 (68 % versus 96 % pour le pré- et le post-test respectivement ; [F (1,37) = 47.7 ; p < .0001]).

Une semaine après l'apprentissage (J+1) : un participant ne s'étant pas présenté au post-test à une semaine, les performances ont été calculées que sur 37 auditeurs. Ce participant a été également supprimé des analyses lors de la comparaison entre le pré-test et le post-test.

A nouveau, les performances étaient significativement meilleures pour les grades G0 et G3 [G0/G1 : F (1,36) = 85.97 ; p < .0001, G0/G2 : F (1,36) = 82.32 ; p < .0001, G1/G3 : F (1,36) = 115.56 ; p < .0001, G2/G3 : F (1,36) = 209.59 ; p < .0001]. Les performances obtenues pour les grades G0 et G3 étaient ici identiques. Les voix de grades G1 et G2 restaient les plus difficiles à catégoriser et aucune différence significative entre ces deux grades intermédiaires n'a été obtenue.

Les comparaisons entre les performances obtenues au pré-test au post-test à une semaine montrent que les améliorations observées pour les grades G2 et G3 se sont maintenues pendant la semaine. On constate en effet une amélioration de 38 à 53 % [F (1,36) = 9.9 ; p < .01] pour le grade G2 et de 68 à 90 % [F (1,36) = 34.2 ; p < .0001] pour le grade G3. Un maintien des performances pour le grade G0 et G1 a de nouveau été observé.

Les comparaisons entre les performances obtenues au pos-test immédiat et celles obtenues au post-test à une semaine montrent une légère diminution significative des performances pour les voix de grades G1 [F (1,36) = 4.14 ; p < .05] et G3 [F (1,36) = 10.5 ; p < .05]. Pour les voix de grade G2, une diminution bien que marginalement significative a été également observée [F (1,36) = 2.98 ; p < .09]. L'augmentation de la performance pour les voix de grade 0 était quant à elle non significative.

DISCUSSION

Les résultats obtenus au pré-test ont montré de bonnes performances chez des auditeurs naïfs dans la catégorisation des grades G0 et G3. La bonne performance pour le grade G0 s'explique sans aucun doute par le fait que des auditeurs non spécialistes ont un standard interne stable des voix normales, ce qui n'est pas étonnant dans la mesure où tous les auditeurs sont confrontés depuis leur naissance à ce type de voix. La bonne performance obtenue pour G3 peut quant à elle s'expliquer par le fait que ce grade correspond à l'autre extrémité, ce qui implique une catégorisation assez facile des voix appartenant à ce grade. La performance sur G3 était malgré tout moins bonne que la performance sur G0, ceci pouvant s'expliquer par le fait que le standard interne pour ce grade est peu robuste, les auditeurs non expérimentés étant en effet rarement confrontés à ce type de dysphonie. Les grades G1 et G2 se sont quant à eux révélés les plus difficiles à catégoriser. Ceci n'est a priori pas surprenant puisque des auditeurs non expérimentés étant peu exposés à ce type de voix, n'ont probablement pas de standards internes pour ces grades intermédiaires. D'autre part, les résultats obtenus au pré-test ont montré que des auditeurs naïfs ont plus de facilité à catégoriser des voix de grade G1 que des voix de grade G2. Une explication possible est que le standard interne pour les voix non dysphoniques (G0) étant robuste, les voix de grade de G1 peuvent être différenciées des voix de grade 0. Par contre, les standards internes pour les voix sévèrement dysphoniques (G3) étant moins robustes, les voix de grade 2 sont plus sujettes à des erreurs, car elles sont moins bien consolidées par le standard interne correspondant à un grade 3.

Pour ce qui est du protocole d'apprentissage, les résultats ont montré d'assez bonnes performances et ceci dès le début de l'apprentissage. Même si les analyses réalisées au niveau du pré-test n'ont révélé aucune amélioration des performances avec la répétition des voix, les bonnes performances de départ lors de l'apprentissage pourraient s'expliquer par le fait que les auditeurs ont commencé à se former les bonnes catégories de voix lors de leur exposition durant le pré-test aux différents grades. Une autre explication a priori plus plausible, est liée au fait que les deux premiers blocs de notre protocole d'apprentissage ne comportaient que 3 grades, ce qui permettait aux auditeurs de s'affranchir de la difficulté inhérente aux grades intermédiaires G1 et G2. Une telle explication semble être soutenue par le fait que les performances diminuent lors de l'introduction du quatrième grade. Les résultats obtenus à la fin de l'apprentissage ont montré une nette amélioration des performances

sur les grades intermédiaires G1 et G2, même si ces derniers restaient les plus difficiles à catégoriser.

De façon cruciale, la comparaison des performances entre le pré- et le post-test a montré que notre protocole d'apprentissage s'est avéré efficace puisqu'il a permis une amélioration des performances au moins dans la catégorisation des grades 2 et 3. Ainsi, il est apparu que les progrès réalisés sont d'autant plus importants que la dysphonie est sévère. La catégorisation des voix de grade 3 ayant été améliorée, l'apprentissage a visiblement permis de renforcer le standard interne des voix de grade 3. Suivant le raisonnement adopté préalablement, il en résulte alors une amélioration dans la catégorisation des voix de grade 2, celle-ci bénéficiant du renforcement du standard interne voisin. Un tel raisonnement pourrait également expliquer l'absence d'amélioration après l'apprentissage dans la catégorisation des voix de grade G1. Le grade G0 ne s'étant pas lui-même renforcé, la catégorisation du grade G1 n'est pas plus facile au pré-test qu'au post-test. Les améliorations obtenues sur les grades G2 et G3 se sont maintenues à une semaine. Toutefois, une diminution des performances entre les deux post-tests a été observée dans la catégorisation des voix dysphoniques, ce qui laisse suggérer qu'un renforcement au niveau de l'apprentissage est nécessaire pour que celui-ci se maintienne.

CONCLUSION ET PERSPECTIVES

Le protocole d'apprentissage que nous proposons pour le jugement perceptif de la sévérité de la dysphonie 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. Cet aspect polymorphique de la dimension G de l'échelle d'Hirano pourrait être réduit en travaillant à présent sur les autres axes tels que la raucité R ou le souffle B. Cependant, dans l'état actuel, transposer le protocole aux dimensions R et/ou B nécessite la sélection de voix prototypiques clairement identifiées, opérations délicates du fait de la moins bonne fiabilité observée pour ces grandeurs [22].

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. Nous pouvons faire ici une analogie avec la densité des espaces vocaliques des langues du monde qui ont émergé en adaptant leur système phonologique aux capacités perceptives de l'être humain. On peut remarquer ainsi que l'ensemble des langues du monde ont adopté un système où la position avant/arrière des voyelles ne s'est mis en place que sur 3 niveaux (avant, médian, arrière) car une multiplication de catégories intermédiaires entrainerait trop de confusions perceptives chez l'auditeur. Cette analogie a ses limites dans la mesure où le jugement de la dysphonie concerne, non pas la catégorisation de phonèmes mais plutôt la perception fine de caractéristiques liées à la voix. En revanche, cette digression pointe les limites des capacités perceptives de l'être humain, notamment dans les processus de catégorisation. Par conséquent, une perspective de ce travail pourrait être d'appliquer à nouveau le procédé non pas sur 4 mais 3 catégories.

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 [8] 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 corréler cela à des mesures instrumentales en analysant, non pas de la concordance de catégories, mais plutôt de la corrélation de grandeurs.

Remerciements

Nous remercions l'Agence Nationale de la Recherche (ANR) pour le financement qu'elle a apporté dans le cadre du projet DESPHO-APADY ANR-08-BLAN-0125 ayant permis la structuration et l'exploitation du corpus de parole pathologique qui a été utilisé dans cette étude.

Références

1. GERRATT B, KREIMAN J. Theoretical and methodological development in the study of pathological voice quality. JOURNAL OF PHONETICS. 2000;28(3):335-342.
2. HARTL DM, HANS S, CREVIER BUCHMAN L, LACCOURREYE O, VAISSIERE J, BRASNU D. Méthodes actuelles d'évaluation des dysphonies. ANN OTOLARYNGOL CHIR CERVICOFAC. 2005;122(4):163-172.
3. 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 Phoniatrics of the European Laryngological Society (ELS). EUR ARCH OTORHINOLARYNGOL. 2001;258(2):77-82.
4. KREIMAN J, GERRATT BR, PRECODA K, BERKE GS. Individual differences in voice quality perception. J SPEECH HEAR RES. 1992;35(3):512-520.
5. 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.

6. BELE I. Reliability in Perceptual Analysis of Voice Quality. JOURNAL OF VOICE. 2005;19(4):555-573.
7. 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.
8. 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.
9. REVIS J, GIOVANNI A, WUYTS F, TRIGLIA J. Comparison of different voice samples for perceptual analysis. FOLIA PHONIATR LOGOP. 1999;51(3):108-116.
10. SHRIVASTAV R. Multidimensional Scaling of Breathy Voice Quality: Individual Differences in Perception. JOURNAL OF VOICE. 2006;20(2):211-222.
11. HIRANO M. Clinical Examination of Voice. *Springer Verlag, Wien*. 1981.
12. ROSCH E. Natural categories. COGNITIVE PSYCHOLOGY. 1973;4: 328- 350.
13. GERRATT BR, KREIMAN J, ANTONANZAS-BARROSO N, BERKE GS. Comparing internal and external standards in voice quality judgments. J SPEECH HEAR RES. 1993;36(1):14-20.
14. FEX S. Perceptual evaluation. JOURNAL OF VOICE. 1992;6(2):155-158.
15. MARTIN DP, WOLFE VI. Effects of perceptual training based upon synthesized voice signals. PERCEPT MOT SKILLS. 1996;83(3Pt 2):1291-1298.
16. GERRATT BR, KREIMAN J. Measuring vocal quality with speech synthesis. J. ACOUST. SOC. AM. 2001;110(5 Pt 1):2560-2566.
17. 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.
18. CHAN KMK, YIU EM. A comparison of two perceptual voice evaluation training programs for naive listeners. J VOICE. 2006; 20(2):229-241.
19. MAGNUSON JS, TANENHAUS MK, ASLIN RN, DAHAN D. The time course of spoken word recognition and learning: Studies with artificial lexicons. JOURNAL OF EXPERIMENTAL PSYCHOLOGY: GENERAL. 2003;132:202-227.
20. DUFOUR S, NGUYEN N, FRAUENFELDER UH. Does training on a phonemic contrast absent in the listener's dialect influence word recognition? Journal of the Acoustical Society of America. 2010;128:EL43-EL48.
21. WIPF AL. 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*.
22. DEJONCKERE PH, OBBENS C, DE MOOR GM, WIENEKE GH. Perceptual evaluation of dysphonia: reliability and relevance. FOLIA PHONIATR (BASEL). 1993;45(2):76-83.
23. GHIO A, POUCHOULIN G, TESTON B, PINTO S, FREDOUILLE C, DE LOOZE C, ROBERT D, VIALLET F, GIOVANNI A. How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers? *Speech Communication. Accepted in Special Issue "Advanced Voice Assessment"*. 2011.
24. 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. 2009:1-14.
25. GHIO A, ANDRÉ C, TESTON B, CAVÉ C. PERCEVAL : une station automatisée de tests de PERCeption et d'EVALuation auditive et visuelle. *Travaux interdisciplinaires du Laboratoire parole et langage d'Aix-en-Provence (TIPA)*. 2003;22:115-133.