Quand la connaissance de l'état du locuteur nous fait entendre sa voix autrement

Alain Ghio[1], Sabine Merienne[2], Antoine Giovanni[1,2]
(1) LPL, CNRS, Université d'Aix-Marseille, Aix-en-Provence, France
(2) CHU Timone, Service ORL, Marseille, France
alain.ghio@lpl-aix.fr

RESUME

L'objectif de l'étude était d'évaluer dans quelle mesure les connaissances a priori qu'un thérapeute possède sur un locuteur dysphonique peut influencer le jugement de la voix du patient. 53 patients dysphoniques ont été enregistrés deux fois dans des circonstances différentes. Sept auditeurs cliniciens ont été soumis en aveugle à l'écoute de ces paires de voix et devaient fournir un jugement comparatif. Quelques semaines plus tard, les mêmes auditeurs ont subi le test à l'identique sauf que dans cette deuxième session, une information sur le statut des voix du locuteur était donnée : pré ou post traitement. Nous avons équilibré cette information de façon soit à renforcer le jugement porté au préalable, soit à contrarier le jugement. Dans l'écoute influencée renforcée, la préférence est amplifiée de façon significative. Dans l'écoute influencée contrariée, nous observons des inversions de préférences et la note avec information contradictoire est presque indépendante de la note obtenue en aveugle.

ABSTRACT

When the knowledge of the speaker's state can modify the perception of voice quality

Two experiments were conducted to examine how the knowledge of the patient's clinical state affects the results of perception of voice quality. This study involved 53 dysphonic speakers recorded twice in different circumstances. These pairs of voices were presented to seven listeners. The task was to perceptually compare the severity of the dysphonia between the 2 recordings of the pair. Stimuli were presented first in a blind test, then several weeks later with accompanying information about the patient (pre- or post-treatment). We balanced this artificial contextual information in order to reinforce the blind judgment or be inconsistent in a clinical point of view compared to the blind test. Results revealed that in the clinical-consistent context, the preference was amplified in a significant way. In clinical-inconsistent condition, we observed an inhibition effect or a change of decision. In this condition, the judgment was more dependent on the contextual information than on the auditory sensation obtained in blind condition.

MOTS-CLES : perception, qualité vocale, dysphonie, processus descendant montant
KEYWORDS : perception, voice quality, dysphonia, bottom-up top-down processes

1 La perception de la qualité vocale

Dans le cadre de la prise en charge de la dysphonie, l'évaluation perceptive de la qualité de la voix permet de faire un bilan clinique de la forme et de la sévérité du dysfonctionnement. Elle s'effectue généralement à l'aide d'une échelle standardisée

Actes de la conférence conjointe JEP-TALN-RECITAL 2012, volume 1: JEP, pages 843–850,
Grenoble, 4 au 8 juin 2012.

contenant plusieurs paramètres que l'auditeur doit juger à l'oreille. L'échelle GRBAS (Hirano, 1981) est l'échelle la plus couramment utilisée. Dans la pratique, l'évaluation perceptive est considérée comme le « gold standard » par les spécialistes de la voix. Cependant, même si elle est très largement utilisée et si une écoute attentive contribue de façon indéniable à dresser le tableau clinique complet du patient, son importance réelle dans un procédé fiable d'évaluation est régulièrement discutée. En effet, de nombreuses études ont mis en évidence une variabilité certaine, observée dans le jugement porté par différents auditeurs sur une même voix ou pour un même auditeur entre diverses séances d'écoute (Kreiman et al., 1993). Ce manque de fiabilité peut être expliqué par la sensibilité au contexte des mécanismes de perception de la parole. (Gerratt et al., 1993) ont mis en évidence que le contexte de présentation des voix peut entraîner une modification des jugements. Dans (Martens et al., 2007), les auteurs montrent que la lecture du spectrogramme de la voix simultanément à son écoute augmente la fiabilité inter-auditeurs pour le jugement perceptif de la qualité vocale. Ces résultats illustrent le fait que la décision issue de la perception est en fait une décision complexe faisant appel non seulement au système perceptif « pur » mais aussi, pour ce dernier travail, aux informations visuelles du spectrogramme et aux connaissances possédées par les auditeurs sur l'interprétation d'une telle représentation spectro-temporelle. Ces phénomènes de variabilité de jugement, reflet d'un manque de fiabilité du procédé, sont ainsi déjà observés en conditions expérimentales contrôlées. Qu'en est-il en pratique clinique quotidienne ? Comment peut-on expliquer ces phénomènes ?

2 L'importance des processus descendants dans la perception

Pour (Gaillard et al., 2007), *« la perception de la réalité sonore n'est pas un enregistrement direct de la réalité. C'est une construction mentale opérée à la suite d'un traitement de l'information disponible, contrainte par nos sens ainsi que nos habitudes sélectives »*. Ainsi, évaluer une voix à l'oreille consiste à interpréter à un moment précis le signal sonore qui nous est donné à entendre, avec le risque que le résultat de cette interprétation diffère avec celle d'un autre auditeur sous l'influence d'habitudes sélectives différentes ou diffère lors d'une évaluation ultérieure sous l'influence d'un changement de l'information disponible. 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. Mais, comme dans tout autre processus de perception de la parole, il 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 message viendra maximiser ou minimiser les effets de ces processus de restauration (Warren et al., 1970). Dans le domaine de la perception visuelle, (Simons et al., 1999) ont montré que nous pouvons être aveugles à certains éléments saillants et inattendus d'une scène visuelle lorsque notre attention est focalisée sur une autre tâche ou un autre objet de cette scène. Il définit cela par le terme de cécité inattentionnelle (inattentional blindness). Dans l'expérience de (Vitevitch, 2003), les participants avaient pour consigne principale de répéter des mots dont la complexité lexicale variait. Au milieu de la liste, la voix utilisée pour produire les mots à répéter

pouvait changer. Au moins 40% des participants ne détectaient pas ce changement de locuteur. L'odorat n'est pas épargné par les effets de distorsion ou d'illusion perceptive avec notamment, une interférence du langage dans cette modalité de perception. Les études de (Herz, 2003) ont mis en évidence l'influence du contexte verbal dans la perception des odeurs et le simple fait d'associer un label à une odeur pouvait provoquer une illusion olfactive. Les auteurs ont constaté en effet qu'une même odeur pouvait être jugée différemment selon le nom qu'on lui donne.

3 Corpus et méthode

Notre protocole expérimental consistait en l'évaluation, par un jury expérimenté, de voix dysphoniques, présentées en aveugle dans une première expérience, puis accompagnées d'informations sur le parcours médical du patient dans une deuxième séance. Les différences entre les résultats des deux expérimentations pourraient être attribuées à l'apport d'information sous réserve de rigueur méthodologique. 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, et ce, plusieurs fois s'il le souhaitait, 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 (1) 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 (2) pour avoir une sensibilité de mesure suffisante. Les participants à l'expérimentation étaient des auditeurs régulièrement confrontés à l'écoute de voix dysphoniques: 3 chirurgiens ORL, 3 orthophonistes et 1 phoniatre. 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. Les enregistrements proposés aux auditeurs ont été sélectionnés dans la base de données MTO de locuteurs dysphoniques enregistrés dans le service ORL du CHU de la Timone à Marseille (Ghio et al., 2011). Les 53 patients retenus, pour lesquels nous disposions d'au moins deux enregistrements réalisés à des dates différentes, étaient des adultes, porteurs de nodules ou de polypes (44 femmes et 9 hommes). La restriction à ces deux pathologies a été retenue afin de limiter l'hétérogénéité des formes d'expression de la dysphonie mais aussi car elles peuvent être prises en charge à la fois par des traitements chirurgicaux et orthophoniques, conditions nécessaires à la deuxième partie de l'expérience. Le style de parole choisi était de la lecture de texte effectuée sur le premier chapitre de « La chèvre de Monsieur Seguin » d'Alphonse Daudet. La durée moyenne des énoncés était de 20 secondes.

4 Déroulement des expériences et précautions méthodologiques

Pour le déroulement des expériences, nous avons utilisé le logiciel PERCEVAL avec son extension LANCELOT, développé par le Laboratoire Parole et Langage d'Aix en Provence

(www.lpl-aix.fr/~lpldev). Les sessions d'écoute ont été effectuées dans un local fermé, sur le même ordinateur, avec la même carte son et le même casque audiophonique. L'expérience se déroulait en quatre phases : deux sessions d'écoute en aveugle (test-retest) suivies de deux sessions d'écoute influencée (test-retest). Dans la condition aveugle, l'auditeur n'avait aucune information sur les locuteurs qu'il écoutait. Dans les sessions d'écoute influencée, une information était affichée à l'écran indiquant la nature du traitement suivi par le patient (chirurgie ou rééducation) et pour chacune des voix de la paire le statut pré ou post traitement. Pour chacune des sessions, la consigne était présentée par écrit sur l'écran de l'ordinateur. Avant de démarrer le test proprement dit, 3 items d'entraînement étaient proposés à l'auditeur, lui permettant de s'approprier la tâche et l'échelle. Les paires étaient présentées dans un ordre aléatoire dans le but de minimiser les effets de liste. Enfin, pour chaque auditeur, les sessions d'écoute étaient séparées les unes des autres d'au moins une semaine pour s'affranchir d'éventuels phénomènes de mémorisation. Chaque modalité de présentation des stimuli (aveugle ou avec information) était constituée d'un test et d'un retest pour lesquels l'ordre de présentation des stimuli variait d'une part entre les auditeurs, d'autre part entre le test et le retest pour le même auditeur. La répétition du test en retest a permis, pour les deux types d'écoute, de moyenner les résultats et donc de diminuer une part des effets d'erreurs aléatoires. 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] ⇔ 3, [moins dysphonique] ⇔ 2, [légèrement moins dysphonique] ⇔ 1, [équivalent] ⇔ 0, [légèrement plus dysphonique] ⇔ -1, [plus dysphonique] ⇔ -2, [nettement plus dysphonique] ⇔ -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. Nous précisons que les informations qui allaient accompagner les voix étaient construites artificiellement pour équilibrer parfaitement les conditions de test et rendre le design expérimental symétrique. Elles ne tenaient pas compte de la réelle situation pré-post traitement. Pour éviter que les informations incohérentes du jeu de données β apparaissent trop invraisemblables et sèment le doute dans l'esprit des auditeurs, nous avons exclus les paires de voix dont la note moyenne en aveugle était proche des extrêmes, c'est à dire avec une différence de qualité importante entre la voix A

et la voix B. En effet, dans ces cas là, la voix évaluée comme nettement plus dysphonique aurait été déclarée dans le jeu de données β comme post traitement, hypothèse peu vraisemblable et difficilement acceptable par des thérapeutes de la voix. Finalement, nous avons sélectionné les paires dont la note moyenne en aveugle se situait entre + 1,5 et -1,5, ce qui représentait un échantillon de 32 paires, la même restriction étant appliquée aux deux jeux de données α et β. Signalons enfin une dernière précaution méthodologique. Afin d'éviter un effet d'ordre d'écoute dans la paire, la voix déclarée comme pré-thérapeutique correspondait parfois à la voix « A » écoutée en premier et parfois à la voix « B », ceci pour limiter l'effet de récence (meilleure trace en mémoire de la dernière voix écoutée, laquelle aurait été favorisée). Par la suite, nous appellerons la condition α comme « information cohérente » ou « jugement renforcé ». De même, la condition β sera dénommée « information incohérente », « jugement contraire » ou « jugement contrarié ».

5 Résultats

L'analyse statistique a été effectuée avec le logiciel 'R' version 2.12.0. Que ce soit en situation aveugle ou en écoute contextuelle, la note retenue par paire de voix est la note moyenne obtenue sur les 14 jugements (7 auditeurs * 2 écoutes pour chaque condition). Au total, les expériences ont porté sur 700 écoutes de paires de voix (14 jugements * 50 paires) pour l'expérience en aveugle et 448 (14 * 32) pour l'expérience en contexte, soit 2296 écoutes de voix. 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. La répartition des notes obtenues sur les 32 paires de voix est fournie en Fig 1.

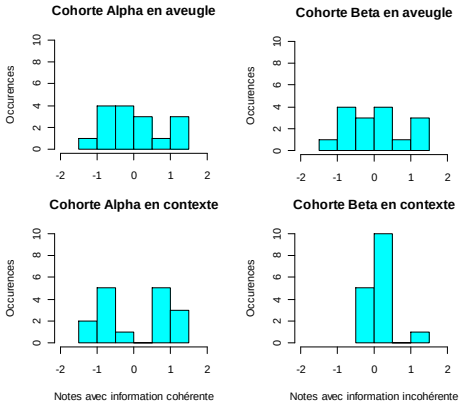

FIGURE 1 –Résultats de l'écoute aveugle (en haut) et de l'écoute contextuelle (en bas). La cohorte alpha (resp. beta) a été utilisée en fournissant une information cohérente (resp . incohérente) en situation contextuelle.

Nous observons clairement sur les distributions de la Figure 1 un effet net de l'apport d'information lors de l'écoute des voix. Dans la condition ou l'information est cohérente, nous observons une distribution quasi bimodale (Figure 1, en bas à gauche) correspondant à une décision plus tranchée. Inversement, dans la condition ou

l'information est incohérente, nous observons une distribution où la majorité des appréciations est centrée autour de zéro (équivalence de qualité vocale des 2 voix de la paire) correspondant à une indécision (Figure 1, en bas à droite). Nous émettons l'hypothèse que dans le cas de la condition cohérente, l'effet de l'apport d'information est amplificateur : les voix jugées comme légèrement moins dysphoniques en écoute aveugle sont jugées clairement moins dysphoniques car elles sont annoncées comme post thérapeutiques et les voix jugées comme légèrement plus dysphoniques en écoute aveugle sont jugées clairement plus dysphoniques car elles sont annoncées comme pré thérapeutiques. Cette hypothèse rendrait compte de la bimodalité de la distribution de la Figure 1. Inversement, nous émettons l'hypothèse que dans le cas de la condition incohérente, l'effet de l'apport d'information est inhibiteur : les voix jugées comme légèrement moins dysphoniques en écoute aveugle sont jugées de qualité vocale équivalente à l'autre voix car elles sont annoncées comme pré thérapeutiques et les voix jugées comme légèrement plus dysphoniques en écoute aveugle sont jugées de qualité vocale équivalente à l'autre voix car elles sont annoncées comme post thérapeutiques. Pour vérifier cette hypothèse, nous avons réalisé une régression linéaire entre les notes obtenues en contexte en fonction des notes fournies en aveugle pour les catégories α ou β. Les observations sont fournies en Figure 2.

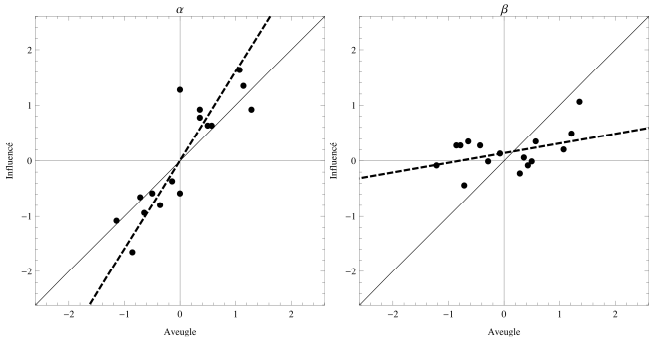

FIGURE 2 – Résultats de l'écoute contextuelle influencée (verticalement) en fonction de l'écoute aveugle (horizontalement). A gauche : apport d'information cohérente (cohorte alpha). A droite : apport d'information incohérente (cohorte beta). La droite bissectrice en trait plein est celle de l'absence 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.

La droite d'ordonnée à l'origine 0 et de pente +1 traduit l'absence d'effet de contexte. En effet, un point sur cette droite correspond à deux notes égales en écoute contextuelle et en aveugle. La répartition des notes observées par rapport à cette droite est donc importante. 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 a été obtenue par la méthode des moindres carrés pondérés en tenant compte simultanément des incertitudes sur les données aveugle et contextuelle, où chaque note (aveugle ou contextuelle) est affectée d'un poids qui est inversement proportionnel à la variabilité (erreur standard) inter-juges. Cette pente importante de coefficient directeur 1.6 valide

l'hypothèse 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 validant l'hypothèse inhibitrice 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.

6 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. Nos résultats confirment que la perception « *est une construction mentale opérée à la suite d'un traitement de l'information disponible* » (Gaillard et al., 2007). 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 perçu 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. Il serait intéressant d'effectuer ces expériences sur des auditeurs détachés de cette problématique et de vérifier si les résultats restent présents ou disparaissent. De plus, les auditeurs étaient composés de chirurgiens ORL et d'orthophonistes. Ce choix avaient été guidé par le fait que lors de l'écoute contextuelle, nous manipulions non seulement la situation pré/post traitement mais aussi la nature du traitement : chirurgie ou rééducation. Nous souhaitions mesurer si les phénomènes d'influence contextuelle pouvaient varier selon l'origine professionnelle : par exemple, nous émettions l'hypothèse que les orthophonistes seraient plus sensibles dans le cas de rééducation que de chirurgie. Le faible effectif de chaque groupe (3 auditeurs par groupe) ne permet pas de mesurer de tels éventuels effets de groupe. Cela nécessite une

augmentation du nombre d'auditeurs.

7 Conclusion

L'évaluation d'un résultat thérapeutique lié à une dysphonie est une préoccupation essentielle du phoniatre et de l'orthophoniste. La chirurgie ou la rééducation a-t-elle eu un effet positif, négatif ou négligeable ? L'écoute attentive de la voix avant et après peut être un moyen d'obtenir cette réponse. Mais peut-on estimer qu'il s'agit réellement d'une évaluation dans la mesure où le thérapeute est parfois lui-même juge de son travail thérapeutique et qu'il dispose de nombreuses informations sur le parcours médical de son patient ? 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.

Remerciements

Nous remercions l'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 été utilisé dans cette étude.

Références

GAILLARD, P., BILLIERES, M., MAGNEN, C. (2007) *La surdité phonologique illustrée par une étude de catégorisation des voyelles françaises perçues par les hispanophones*. In: Proc. Percepción y Realidad., Valladolid, Spain, 2007; pp. 187-196.

GHIO, A., POUCHOULIN, G., TESTON, B., PINTO, S., FREDOUILLE, C., DE LOOZE, C., ROBERT, D., VIALLET, F., GIOVANNI, A. (2011), *How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers?* Speech Communication. 2011; Special Issue "Advanced Voice Assessment".

GERRAT, B., KREIMAN, J., ANTONANZAS-BARROSO, N., BERKE, G. (1993) *Comparing internal and external standards in voice quality judgments*. J Speech Hear Res.; 36(1), 14-20.

Herz, R., *The effect of verbal context on olfactory perception*. (2003)J Exp Psychol Gen. 132(4), 595-606.

HIRANO, M. (1981). Clinical Examination of Voice. Springer Verlag, Wien

KREIMAN, J., GERRAT, B., KEMPSTER, G., ERMAN, A., BERKE, G. (1993). *Perceptual evaluation of voice quality: review, tutorial, and a framework for future research*. J Speech Hear Res. 36(1), 21-40.

MARTENS, J., VERSNEL, H., DEJONCKERE, P. (2007) *The effect of visible speech in the perceptual rating of pathological voices*. Arch. Otolar. Head Neck Surg. 133(2), 178-185.

SIMONS, D., CHABRIS, C. (1999) *Gorillas in our midst: sustained inattentional blindness for dynamic events*. Perception.; 28(9), 1059-1074.

VITEVITCH, M. (2003) *Change deafness: the inability to detect changes between two voices*. J Exp Psychol Hum Percept Perform.; 29(2), 333-342.

WARREN RM., WARREN RP. (1970), *Auditory illusions and confusions*. Sci. Am.; 223, 30-36.