Perceptual Evaluation of Dysphonic Voices: Can a Training Protocol Lead to the Development of Perceptual Categories?

*Alain Ghio, *Sophie Dufour, †Aude Wengler, *Gilles Pouchoulin, *,†Joana Revis, and *,†Antoine Giovanni,
*Aix-en-Provence and †Marseille, France

Summary: The purpose of this study was to develop and test a training protocol for the perceptual evaluation of dysphonia. A group of 38 inexperienced listeners participated in a three-phase experiment: a pretest to evaluate their initial performance on categorization of dysphonic voices, a training phase, and a posttest to detect training-related changes in performance. In parallel, a different group of 14 listeners who were experts in voice assessment took a test that was identical to the posttest taken by the inexperienced subjects. The corpus used for the tests was made up of recordings of 142 voices of women reading aloud, with a sampling of voice qualities ranging from normal to severely degraded. The learners' performance on judgments of moderate and severe dysphonia improved between the pretest and the posttest. No improvement was observed for normal voices, whose initial detection was already good, nor for slight dysphonias, which appear to be the most difficult to learn. The improvements were still present on a delayed posttest taken a week later. Unexpectedly, the inexperienced listeners' initial performance was similar to that of the experts. After the training phase, their scores for severely deteriorated voices were even better than the experts'. In conclusion, our training protocol seems to be effective and could therefore be proposed to voice therapists. However, judging intermediate degrees of dysphonia remains fragile and therefore needs to be reinforced by repeated training.
Key Words: Perceptual evaluation of dysphonia–Training–Voice–Expertise.

INTRODUCTION

Perceptual evaluation of dysphonia

To treat patients with dysphonia, it is essential to assess the quality of their voices. Such patients usually decide to consult a voice specialist when they begin to hear changes in their own vocal output. Similarly, after having undergone surgical treatment or speech therapy, they generally judge the treatment's success in terms of the auditory impression they have of their voice.[1] Perceptual evaluation is the most widespread method used by clinicians to describe a patient's voice (breathy, hoarse, rough, etc.) or measure the severity of the dysfunction. This method was recommended by Dejonckere et al[2] in their basic protocol for the functional evaluation of voice pathologies. It has many advantages: it is easy to implement, inexpensive, and directly accessible to any clinician. However, although perceptual analysis remains the standard in this field, it nevertheless raises a key question: Just how reliable is it? Reports in the literature indicate substantial variability in perceptual judgments of the voice.[3–9] Variability shows up as inconsistencies between ratings of the same voice made by different listeners (between-listener variability) and between ratings made by the same listener at different times (within-listener variability).

Assessment variability has been widely studied in view of alleviating these phenomena, deemed undesirable from the clinical standpoint. Various authors have taken different approaches in an attempt to reduce its magnitude, including recruitment of experts vs naive listeners,[6] the use of analog vs categorical scales,[7] judgments of different types of utterances (sustained vowels vs sentences),[8] and ratings along various dimensions (overall quality vs breathiness[9] vs roughness).

Variability vs reliability

As a general rule, studies in this area have focused on observing variability-related phenomena, considered indicative of the assessment method's lack of reliability.[10] We think it is preferable to test method reliability directly by measuring the accuracy of the listeners' responses. Although it seems legitimate to assume that reliable responses exhibit little variability—because, by definition, a correct response is invariable—the reverse is not trivial. For example, if dysphonia severity is being rated on a four-level scale like Hirano's,[11] where G0 is a normal voice, G1 a slightly dysphonic voice, G2 a moderately dysphonic voice, and G3 a severely dysphonic voice, then listeners who systematically give a rating of G0 for globally normal voices and G2 for globally abnormal voices would obtain a low degree of variability because of their simplified response strategy. The responses might also be highly valid, if these listeners are correctly rating normal voices as G0 and dysphonic voices as G2. But their responses could not be considered as sensitive enough because they did not use all four levels of the rating scale.

According to Bele,[5] the reliability of an evaluation is related to the degree to which the results are void of measurement errors. This author makes the distinction between random errors (listener distraction, poor use of the response choices, etc.) and systematic errors. Although the former type of error can be minimized by repeating the tests and increasing the number of participants, the second affects the listener's score and

Accepted for publication July 14, 2014.
From the *Laboratoire Parole et Langage, Aix-Marseille University, CNRS, Aix-en-Provence, France; and the †Department of Otolaryngology — Head and Neck Surgery, Timone University Hospital, Marseille, France.
Address correspondence and reprint requests to Alain Ghio, Laboratoire Parole et Langage, Aix-Marseille University, CNRS, 5 Avenue Pasteur, BP 80975, 13604 Aix-en-Provence Cedex 1, France. E-mail: alain.ghio@lpl-aix.fr
Journal of Voice, Vol. 29, No. 3, pp. 304-311
0892-1997/$36.00

http://dx.doi.org/10.1016/j.jvoice.2014.07.006

reveals an idiosyncrasy of that particular participant. If a systematic error is made by all listeners, then this points either to a specific feature of the stimulus or to a limitation of the evaluation method.

The key issue raised in the present study, then, concerns how to categorize dysphonic voices. What is a normal voice? What is a slightly, moderately, or severely dysphonic voice? This question is all the more complex because of the multiple manifestations of dysphonia—a dysphonic voice can be breathy, hoarse, or rough, diplophonic, hyperfunctional, or hypofunctional, and so on—making it difficult to define clear-cut categories, whether in terms of quality or quantity. This brings us to the topic of category exemplars and prototypes.

Categorization, exemplars, and prototypes

Dysphonia evaluation involves a categorization process (assigning a grade to a vocal production), which requires:

- being able to put similar vocal productions into the same category and
- being able to distinguish between vocal productions that belong to two different categories.

According to the so-called "classical" theory, a category is defined by the properties its members share, and any entity that possesses those properties belongs to the said category: if an exemplar possesses a given property then it belongs to category X; if it does not possess that property, then it does not belong to category X. This approach was questioned by Rosch,[12] whose experiments showed that in most cases, it is difficult to define a set of necessary and sufficient properties to characterize a category. The most common example is the ostrich, which has many of the defining features of the bird category (two feet, two wings, feathers, a beak, lays eggs, etc.), but does not fly. Having noted also that certain exemplars are more representative of a category than others, Rosch introduced the notion of prototype: the prototype is the best representative of the category. However, what is prototypical for some may not be for others, so prototypicality differs across individuals, depending on their particular experience. In this view, membership in a category is no longer a yes/no question but depends on the object's degree of similarity to the prototype, with certain exemplars being central to the category, whereas others are more peripheral. As we shall see later, the notion of prototype proposed in cognitive psychology is similar to the notion of internal standard described by Kreiman et al.[3,4,13]

TRAINING-BASED REINFORCEMENT OF INTERNAL STANDARDS

Unstable and listener-specific prototypes

The lack of reliability in perceptual evaluations of dysphonia is largely dependent on what strategies and mechanisms are used by listeners to classify voices, particularly the auditory standards used by the judges. Kreiman et al[3] introduced the notion of internal standard, which is equivalent to the notion of prototype used in cognitive psychology: each listener judges the quality of a voice by comparing it to his or her internal auditory standard or prototype, which is based on what the person thinks a normal or dysphonic voice sounds like. It is the perceptually estimated distance between one's internal standard and the voice heard that determines the degree of severity assigned to the voice. As a whole, however, such internal standards are listener specific and are more or less precisely defined in accordance with the listener's perceptual experience with dysphonic voices.

External anchoring based on comparison

An alternative assessment method was used by Gerratt et al,[13] who proposed replacing internal standards by a set of external anchors or perceptual references. By supplying a voice scale that is constant and the same for everyone, this method allows listeners to categorize voice samples by comparing them to the reference set.[14] In one study, these authors demonstrated the effectiveness of using an externally anchored scale to evaluate voice roughness.[13] Their use of synthesized voices[15–18] seemed justified because it could provide a range of voice samples representing the diverse manifestations of dysphonia, both in terms of quality (breathiness and roughness) and severity. But a paradigm with synthesized and/or pseudonatural voice anchors (To obtain the full range of dysphonic voice samples, Chan and Yiu[17] had a healthy speaker simulate various degrees of roughness and breathiness, hence our use of the term "pseudonatural.") leaves much to be desired, not only because synthesized voices are too artificial to be compared with natural voices but also because difficulty calibrating stimuli derived from natural voices makes it hard to obtain a representative set of voice samples. Last, a method based on systematic comparisons with external references is far removed from everyday speech perception and therefore does not leave listeners in a position to judge for themselves after training.

Our proposal: internal anchoring via training on natural voices

The use of external anchors can generate "unnatural" situations for speech perception. In an attempt to avoid this problem, we designed a training method similar to the ones developed for learning new words or phonemes,[19,20] which do not supply external references to subjects. This approach is similar to the approach of Martin and Wolfe[15] and Chan and Yiu[18] but differs from theirs by the fact that our training protocol makes use of real dysphonic voices rather than synthesized or simulated ones.

For the inexperienced listeners, the experiment took place in three phases: (1) a pretest for measuring the listener's initial performance on dysphonia categorization, (2) a training phase, and (3) a posttest to detect any changes resulting from the training. To assess the learners' final performance, we compared them to a group of experts who were given a test identical to the posttest taken by the inexperienced participants. Two questions were raised in this study. The first was aimed at finding out whether inexperienced listeners can learn to categorize dysphonia severity. The second was aimed at