A Simplified Vocal Profile Analysis Protocol for the Assessment of Voice Quality and Speaker Similarity

*Eugenia San Segundo and †Jose A. Mompean, *York, UK, and †Murcia, Spain

Summary: Objectives. A simplified perceptual protocol for the assessment of voice quality (VQ) is attempted based on the Vocal Profile Analysis (VPA) scheme, with the aim of alleviating typical issues associated with the multidimensionality of VQ and enabling an easy quantification of speaker similarity.

Study Design. Twenty-four non-pathological male speakers (12 monozygotic twin pairs) of Standard Peninsular Spanish were perceptually evaluated by two trained phoneticians using the simplified VPA (SVPA). Based on their perceptual ratings, intra- and inter-rater agreement was measured, and an index of speaker similarity was calculated not only between twin pairs but also between non-twin pairs. For that purpose, one member of each twin pair was compared with a member of a different twin pair.

Methods. Intra- and inter-rater agreement measures were tested with unweighted and linear weighted kappa. Speaker similarity was measured with simple matching coefficients (SMC).

Results. The results show that analysts' internal consistency was very high, whereas inter-rater agreement was found to be strongly setting-dependent. SMCs between speakers indicate that twin pairs are, on average, more similar than non-twin pairs.

Conclusions. Agreement results suggest that the proposed SVPA is a reliable protocol for the perceptual characterization of VQ, and SMC results confirm that it can also be a useful tool for the assessment of speaker (dis)similarity. The extraction of a voice quality similarity index shows potential in fields like forensic phonetics, but could also be of interest in related areas of voice research and professional practice.

Key Words: Voice quality–Perceptual protocol–Rater agreement–Twins–Spanish.

INTRODUCTION

The perceptual assessment of voice quality

Voice quality (henceforth VQ) can be broadly defined as the combination of laryngeal and supralaryngeal features in someone's voice, producing a long-term effect in perception and making that voice recognizably different from others.[1] Methodologically, the assessment of VQ can be approached from an articulatory, acoustic, or perceptual point of view. In this investigation, we focus on the perceptual assessment of VQ. In this respect, it is well known that auditory protocols are sensitive to biases and errors[2] given analyst-related as well as speech-related factors. Both can call into question the reliability and validity of such perceptual methods.

As far as analyst-related factors are concerned, lack of agreement on definitions and terminology may lead to totally different assessments of the same speech material. Moreover, raters may have different internal standards to compare speakers' voices.[3,4] Regarding speech-related factors, VQ multidimensionality is often considered to be a problem. In this regard, some researchers opt for featural analyses, whereas others consider that VQ perception must involve a great component of holistic, gestalt-like pattern processing.[5–7] Anyhow, the perceptual assessment of voices has a quantifiable basis that can correlate with other forms of evaluation, such as laryngoscopic observations or acoustic analyses.[8] In fact, auditory assessment is still regarded as the "gold standard"[9] with which acoustic measures alone—or a combination of objective parameters—should be compared.

Perceptual evaluations are necessary in a variety of research areas. In clinical voice therapy, a considerable number of protocols have been proposed for the description and monitoring of a patient's VQ. These protocols typically require expert or trained listeners to rate several VQ features using scalar degrees, interval scales, or visual analog scales (see Wewers and Lowe[10] for a discussion). Forensic phoneticians have also benefited from the use of VQ perceptual assessment schemes in forensic speaker comparison (FSC) tasks, consisting in the analysis of the voice recording of an offender and its comparison with a voice sample of a suspect.[11] VQ is considered an extremely valuable voice feature by most authors.[12,13] In sociophonetic studies, the use of perceptual assessment protocols has resulted in thorough descriptions of several varieties of English,[14–17] often showing gender- and age-dependent differences in VQ.

The need for a simplified VPA protocol for research and professional practice

One of the best known perceptual assessment protocols among phoneticians is the Vocal Profile Analysis (VPA), created in the early 1980s by John Laver and colleagues[18,19] as a means to identify and rate a speaker's VQ features. One of its key characteristics is its comprehensive scope, as it considers not only phonatory but also supralaryngeal features.[20,21] VPA analyses are based on recordings of at least 40 seconds of connected speech in spontaneous recordings, as these are said to provide the most realistic representation of a speaker's habitual VQ.[21] The analytic unit of the protocol is the setting, or long-term articulatory, phonatory, or muscular tendency. In one of the most common versions of the protocol,[22] there are 36 settings: 25 describe vocal tract

Accepted for publication January 11, 2017.

From the *Department of Language and Linguistic Science, University of York, York, UK; and the †English Department, University of Murcia, Murcia, Spain.

Address correspondence and reprint requests to Eugenia San Segundo, Department of Language and Linguistic Science, University of York, Heslington, York YO10 5DD, UK. E-mail: eugenia.sansegundo@york.ac.uk

Journal of Voice, Vol. 31, No. 5, pp. 644.e11–644.e27
0892-1997

http://dx.doi.org/10.1016/j.jvoice.2017.01.005

(supralaryngeal) features, 7 describe phonation features, and 4 describe overall muscular (laryngeal and vocal tract) tension features. Depending on the version, the VPA protocol may also include some extra features, mostly referring to prosody and temporal organization.[22] Appendix 1 shows the list of settings included in the VPA version described in Mackenzie Beck,[22] without the extra features.

As far as the rating of settings is concerned, each VPA setting is described as a deviation from a clearly defined "neutral" or standard condition. This implies that there are, for the vocal tract dimension, no constrictive or expansive effects in the vocal tract cavities and no shortening or lengthening of the extension of the vocal tract between vocal cords and lips. The neutral setting also implies, for the phonatory dimension, no extreme variations in terms of muscular tension activity in the supralaryngeal and laryngeal parts of the vocal tract, and balance in terms of the adduction forces and longitudinal tension of the vocal folds without audible whispering. The first step in the perceptual evaluation using the VPA is to identify the presence of neutral and non-neutral settings. In the second step, the judge is asked to rate only the non-neutral settings using a scalar degree ranging from 1 to 6, where 1–3 are classed as "moderate" and 4–6 as "extreme" (Appendix 1).

One of the advantages of the VPA scheme is its completeness, although some authors consider it to be "too complex"[8] (p. 2175). In the same line, Webb et al[23] claim that "its greater scope is at the expense of reliability"[23] (p. 429). The complexity of this protocol is understood both as comprising a very large number of settings and as making use of too many scalar degrees in order to mark to which extent the setting is present. A typical way of alleviating common problems associated with comprehensive and somewhat complex protocols like the VPA has been to develop simpler perceptual assessment methods. This is the principle behind proposals such as Shewell's Voice Skills Perceptual Profile,[24] targeted at voice practitioners other than speech and language therapists, such as voice teachers and singing teachers. An alternative approach is to simplify existing protocols by reducing, for example, the number of categories or settings. The GRB protocol,[25] a simplified version of the GRBAS protocol,[26] is a case in point. It consists of G (grade), R (roughness), and B (breathiness), and it originated as a response to the fact that measurements of inter-rater reliability using GRBAS had shown that the reliability was moderate (eg, Webb et al, De Bodt et al, and Dejonckere et al[23,27,28]) for A (asthenia) and S (strain).[29]

A simplification of an existing protocol is also the approach taken in this study. Here, VPA was chosen instead of GRBAS. Thus, a simplified version of the VPA scheme is proposed below with a reduction of the number of settings in the original protocol and using no scalar degrees. The decision of reducing the number of settings and using binary judgments rather than scalar degrees is based on a number of issues relevant to VQ perceptual assessment:

(1) Multidimensionality and isolation of dimension. The highly multidimensional nature of VQ is often considered a problem in perceptual evaluations. Raters usually find it difficult to isolate specific dimensions[2] as they tend to be interrelated.

(2) Labeling. Raters can fail to agree on definitions of a voice feature, which can lead to different assessments for specific dimensions based on different understanding of the labels that should be assigned to a voice feature. In this respect, a simplified protocol with fewer labeling options may reduce this problem.

(3) Normal versus pathological VQ rating. Although the perceptual assessment of pathological voices may require complex protocols, the latter may be less effective with non-pathological VQ.[30] This suggests that when normal voice is under study, a protocol that leaves out clearly pathological settings (eg, audible nasal escape) may suffice.

(4) Cognitive processing constraints. Perceptual assessment is a cognitively demanding task. Given this, a simpler protocol may impose fewer cognitive demands on raters, especially because the process of rating voices not only implies the assessment itself but a previous process of identifying and isolating the different aspects of the stimuli.[6]

Rationale for the analysis of monozygotic twins

The rationale for using monozygotic (MZ) in this study is their strong similarity. Previous investigations have shown that MZ twin pairs can be distinguished perceptually[31] and also acoustically,[32–34] although some exceptions are possible due to a number of sociolinguistic reasons.[35,36] Yet little is known about how speaker similarity is affected by VQ in particular, and more accurately using a componential approach to the perceptual assessment of VQ, like the VPA scheme. Selecting MZ twins as subjects is an opportunity to explore VQ closeness in speakers who represent the most extreme examples of vocal tract similarity. In this respect, we could compensate for one of the shortcomings that Nolan[37] mentions for VQ assessment protocols: the lack of vocal tract isomorphism across speakers. In other words, the fact that different speakers typically present isomorphic but not identical vocal tracts implies that the small differences in size or shape that two speakers have make them sound different even if they choose the same articulatory options.[37] Therefore, investigations with MZ twins—presenting identical vocal tracts, or at least the most similar possible—can be of great use for VQ research, as they can prove useful to test to what extent even a simplified protocol allows for detection of fine-grained differences in very similar-sounding speakers.

OBJECTIVES AND RESEARCH QUESTIONS

The main purpose of this study is to design a simplified VPA (henceforth SVPA) that researchers and voice professionals can use to rate VQ. In particular, this study addresses two main research questions (RQ): (1) How reliable is the proposed SVPA in terms of intra- and inter-rater agreement?—and to which extent this agreement is setting-dependent; and (2) can an index (distance measure) of speaker similarity be extracted from the SVPA assessment method?

For RQ1, we hypothesize that the SVPA will yield satisfactory values of intra- and inter-rater agreement and that agreement will depend strongly on each setting. For RQ2, we hypothesize