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

© 2017 The Authors. Published by Elsevier Inc. on behalf of The Voice Foundation. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

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 that, based on the SVPA, the creation of an index of speaker similarity is possible, that this will reveal that MZ twins are—at least on average—more similar than non-twin speakers, but that it will still be useful to detect fine-grained VQ aspects between them.

METHODS

Participants and speech materials

Twenty-four male speakers of Standard Peninsular Spanish (SPS), the variety of European Spanish spoken in northern and central Spain,[38,39] participated in this study. The participants were aged 20–36 (mean: 26.83, standard deviation: 6.6) and they made up 12 pairs of MZ twins. They were selected from a larger corpus of Spanish speakers, including also dizygotic twins and non-twin siblings.[35,40] The subjects reported having no voice disorders or hearing difficulties.

The participants' conversations were recorded with omnidirectional condenser microphones (head-mounted device) with flat frequency response (*Countryman E6i Earset*, Countryman Associates, Inc., Menlo Park, California, USA) and a soundcard *Cakewalk by Roland UA-25EX USB Audio Capture* (Roland Corporation, Hamamatsu, Shizuoka, Japan). The software used for the recordings was *Adobe Audition CS5.5* (Adobe Systems Inc., San Jose, California, USA), and the operating system of the computer used was *Microsoft Windows XP Professional* (Version 2002; Microsoft Corporation, Redmond, Washington, USA). The following were the recording specifications: 44.1 kHz sample rate, 16-bit resolution, and mono channel. As for the data collection setup, each twin pair was recorded on the same day but separated in two different (acoustically isolated) rooms.

The speech materials for this study consisted of speech samples of spontaneous conversations of around 120 seconds produced by the participants. These were extracted from longer conversational exchanges (approximately 10 minutes), recorded in researcher-speaker informal conversations held over a landline telephone. Note, however, that the recordings are not telephone-degraded but high-quality recordings obtained through a microphone.[36] In this conversation, the researcher asks each twin individually about any of the topics that he had been discussing with his twin in the first task of the corpus described by San Segundo.[36,a]

Perceptual analysis

Perceptual assessment procedure

Two native Spanish phoneticians with over 5 years of experience listened to the 24 speakers of this study in random order (name in alphabetical order), thus ensuring that the twins were not evaluated consecutively. Using the SVPA introduced earlier, they rated the set of voices on two different occasions (two rounds), with a time lapse of one week. This rating procedure was blind (ie, each judge rated voices independently), and took place in a silent room and using AKG K 430 headphones (AKG Acoustics, Vienna, Austria). In the second round, the judges also rated voices independently from their first assessment session. Prior to these two evaluations, raters had been trained together by carrying out a joint listening of a small set of voices (eight speakers) belonging to the same corpus described earlier.[35,40] The joint listening of these voices by both analysts makes part of the calibration process. As explained in the next section, this was aimed at finding an acceptable working definition of the different settings and sharing a common understanding of the possible deviations from the neutral setting per category.

SVPA protocol

During the training meetings held by the two analysts, discussion about the interpretation of the different settings and their adaptation for SPS was key for the design of the SVPA proposed here (Appendix 2). In some cases, the VPA features are considered to be mostly language-independent. For example, *nasal* and *denasal* are considered to apply, respectively, to abnormally nasal or "twangy" voices (hypernasality) or abnormally denasal voices (hyponasality), typical of speech produced with a blocked nose during a cold.[41] However, some segments are more susceptible to the effects of specific VQ settings.[1] Consequently, the VPA protocol implies the identification of key speech segments in order to assess the effect of VQ settings on them.

Certain segments deserve some explanation in relation to the adaptation to SPS. Given that Spanish and English have different segmental inventories, differences in key segments were to be expected. For example, the original protocol focuses on alveolar consonants such as /t, d, n, s, z, l/ for the lingual tip/blade setting. In SPS, /n, s, l/ are also alveolar (alongside flap /ɾ/ and trill /r/), whereas /t, d/ are dental, and [z] is not a phoneme but an allophone of /s/. Moreover, it is common in SPS for retraction to be associated with a postalveolar articulation [ʃ] with variable degrees of lip rounding and groove width. Similarly, a key segment susceptible to the effects of tongue body settings is /s/. This is due to the tendency in SPS—particularly some language varieties around Madrid—to debuccalize coda /s/ as a voiceless glottal fricative or even replace it with a voiceless velar or uvular fricative (eg, *es que* [ɛhke]/[ɛxke]/[ɛχke]) "the thing is that. . .").[42]

Apart from adapting and redefining some settings for the language under investigation, sharing the same definition of the neutral setting was also of key importance. Research on the neutral setting of SPS is limited to sporadic references in general descriptions of Spanish.[43–45] In this literature, the neutral setting for SPS is described with the following characteristics: (1) relatively high muscular tension, (2) modal phonation, (3) neutral larynx height, (4) lax pharynx, (5) front-central resonance, with dental or alveolar articulatory anchorage, (6) considerable apical activity, (7) strong mandibular movement, (8) weak labialization, (9) weak (if any) nasalization, (10) relatively low pitch, and (11) low amount of airflow.

The main modifications toward simplification of the original VPA can be summarized as follows:

(1) reduction from 36 settings to 22
(2) 10 major "setting groups" with 22 possible settings within those groups, that is, two articulatory strategies as possible deviations from neutrality

[a] The task 1 of the corpus described in San Segundo[35] is a semi-structured conversation between twins. Several topics for conversation, adapted from Loakes,[32] were suggested to the speakers: (1) Speak with your partner about a situation in your life when you felt you were in serious danger of death. (2) What would you do if you had all the money in the world? (3) Speak with your partner about your favorite holidays.

(3) no scalar degrees; use of a binary (neutral/non-neutral) rating for each setting group
(4) no marking of intermittent settings
(5) possibility of including holistic descriptions regarding the settings being rated or any other VQ aspects

As pointed out earlier, within each major setting group, a decision must be made as regards the direction of the deviation from neutrality, whereas in the original protocol it is possible to select several options. For instance, in relation to phonation types, a rater could label a voice as both creaky and harsh, with the same or different scalar degrees. It is well known that combined phonation types exist, but usually one is predominant—which is the one that has to be rated in our SVPA—whereas the other appears only intermittently or is not as salient. For the rest of major settings, our simplified rating system is perfectly apt to the mutually exclusive nature of labels: for example, in relation to the vocal tract tension, if the speaker is non-neutral for that setting, he presents either tense vocal tract or lax vocal tract; or if he is non-neutral as concerns the lingual body, he will either tend to present a fronted and raised tongue body or a backed and lowered tongue body.[b]

The main modifications from the original settings were made for phonation types. We no longer distinguish between subgroups "voicing type", "laryngeal frication", and "laryngeal irregularity". All of them are merged into voice types; the neutral value standing for "modal voice", with only two deviations from neutrality: laryngeal irregularity, which can surface as "harsh" or "creak(y)" voice, and laryngeal friction, which can surface as "breathy" or "whisper(y)" voice. For the sake of simplification—and because the boundaries are sometimes blurred—there is no distinction between "creak" and "creaky" and "whisper" and "whispery", as in the VPA version described in Mackenzie Beck.[22]

Furthermore, we removed three settings deemed to be atypical in normophonic speakers of SPS: *labiodentalization*, *protruded jaw*, and *audible nasal escape*. In fact, the latter only admits scalar degrees 4–6 in Mackenzie Beck.[22] These deletions allowed us to obtain a simpler protocol with three options per setting group: the neutral configuration and a system of binary choices for non-neutral settings. This reduces the number of decisions taken by the analyst while it allows for a detailed description of typical articulatory configurations.

Finally, all the extensive and minimized range variants in Mackenzie Beck[22] (ie, extensive and minimized mandibular, labial, or lingual setting) were discarded, as they were deemed to be covered by other settings: "open jaw" can be used to describe all extensive configurations and "close jaw" the minimized configurations.

Rater agreement measurement

In this study, we used the following statistical tests to calculate both inter- and intra-rater agreement.

Overall percent agreement

It is the most popular method of computing a consensus estimate of inter-rater reliability, although it gives a rough estimate of reliability.[46] Because this measure does not take into account that agreement may occur solely based on chance, it is the least robust measure of reliability.

Cohen's kappa

This measure[47] is considered to be a better estimate of reliability than percentage agreement, as it estimates the degree of consensus between two judges after correcting the amount of agreement that could be expected by chance alone based upon the values of the marginal distributions.[48]

Linear weighted kappa

Weighted kappa partly compensates for a problem with unweighted kappa, namely that it is not adjusted for the degree of disagreement. When the categories are ordered, it is preferable to use weighted kappa,[49] which incorporates a notion of distance between rating categories. With linear weighted kappa, if there are k categories, the weights are proportional to the number of categories apart.

We used linear weights because the difference between the first and second category has the same importance as the difference between the second and third category, but the difference between the first and the third category is more important; this type of disagreement should weigh more, as it points to opposite directions of non-neutrality for each setting. In other words, the use of linear weighting with our specific data implies accounting differently for the disagreement between neutral ratings ("0" ratings, ie, second category) and any of the deviations from neutrality ("–1" or "+1"; first and third categories, respectively), and for the disagreement between the two opposing non-neutralities ("–1" and "+1").[c] Linear weighted kappa is calculated in this study with 95% confidence interval.[50]

The interpretation of kappa magnitudes is somehow arbitrary and heavily dependent on the type of study or scientific discipline. A value of 0 on kappa does not indicate that the two judges did not agree at all; it only indicates that the two judges did not agree with each other any more than would be predicted by chance alone. Landis and Koch[51] proposed some guidelines for the interpretation of kappa magnitudes: kappa values <0 indicate no agreement, 0–0.20 slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement, and 0.81–1 almost perfect agreement. It is generally accepted that kappa values below 0.2 indicate poor agreement and a kappa around 0.8 indicates very good agreement beyond chance. Fleiss et al[52] propose a similar interpretation of the magnitude of (unweighted and weighted) kappa: $\kappa \leq 0.75$ implies excellent agreement and $\kappa \leq 0.40$ poor agreement.

[b] Note that for the dorsal setting, fronted and raised have been merged; backed and lowered too.

[c] For illustration purposes, we explain two possible cases of disagreement between raters on judging labial settings. In the first case, the first rater (R1) disagrees with the second (R2) because R1 assigned the neutral label to a speaker whereas R2 judged him as lip-rounded. In the second case, raters disagreed because R1 considered that the speaker presented lip spreading, whereas R2 rated the same speaker as lip-rounded. Unweighted kappa takes both cases as exactly the same type of disagreement. Linear weighted kappa penalizes the second case more strongly.

TABLE 1.
Example of Calculation of Simple Matching Coefficients (SMC) for Twin Pair AGF-SGF

		Major Setting Groups										
		Labial	Mandibular	Apical	Dorsal	Velopharynx	Pharynx	Larynx Height	Vocal Tract Tension	Larynx Tension	Phon. Types	
Speakers	AGF	0	0	0	0	0	0	–1	1	1	1	
	SGF	0	1	0	0	1	0	–1	1	1	1	
Matches		1	0	1	1	0	1	1	1	1	1	0.8 SMC

Speaker similarity measurement

Among other reasons (cf, introduction), this simplification of the VPA protocol was envisaged to obtain a numerical measure of the distance between two speakers in terms of their VQ. Although in some scientific fields a qualitative description of VQ may suffice, other research areas typically require more quantitative approaches. For instance, in forensic phonetics, an index of similarity resulting from the comparison of two speakers is common. The use of Euclidean distances (EDs) for perceptual evaluation also allows comparing them with EDs calculated for acoustic features.[53] Considering that EDs for categorical data are best computed using the simple matching coefficient (SMC) method, we implemented this technique on our data.

If only one variable existed (for instance, labial setting), computing the distance between two speakers would be fairly trivial: for two speakers having the same configuration in certain setting (eg, lip rounding), their distance would be 1. If only one of them had lip rounding and the other lip spreading, their distance (i.e. similarity) would be 0. In addition, if one of them were neutral for that setting and the other had any type of deviation from neutrality—in this case, either lip rounding or lip spreading—the distance would be 0 as well. As several categorical variables (labial setting, mandibular setting, etc) exist for calculating the distance between two speakers, the simplest method is that of extending the "matching" idea and counting how many matches and mismatches there are between samples. As an example, in the case shown in Table 1, there are eight matches and two mismatches between twins AGF and SGF; hence, the distance between the two speakers is 8 divided by 10, the number of variables. Therefore, 0.8 is the SMC for speakers AGF and SGF. Differences between these speakers are due to dissimilarities in their mandibular and velopharyngeal settings; one member of the twin pair exhibits open jaw setting and nasality, whereas the other shows a neutral configuration for both aspects. They share the rest of setting options (SVPA numerical labels are available in Appendix 2).

RESULTS

Intra-rater agreement

Table 2 shows the intra-rater agreement results for each of the two raters (R1 and R2) on two different occasions. Internal consistency within each judge is almost perfect (Cohen's κ ranging between 0.81 and 1), regardless of the rater. According to the classification proposed in Landis and Koch,[51] "substantial agreement" (κ: 0.61–0.80) is obtained in just three settings: *velopharyngeal* (R1), *larynx tension* (R2), and *voice type* (both raters). Raters seem especially consistent when rating the setting *apical* (which refers to whether the speaker presents advanced or retracted tongue tip), with no speaker causing disagreement between the first and the second perceptual sessions.

Two further settings present the highest intra-rater agreement: *labial* and *dorsal*. In this respect, some of the speakers

TABLE 2.
Intra-Rater Agreement Results for the Two Raters (R1 and R2)

	Percentage Agreement		Cohen's Kappa (κ)		N Disagreements (Speaker/s)	
Setting ↓	R1	R2	R1	R2	R1	R2
Labial	95.83	100	0.91	1.00	1 (MHB)	0
Mandibular	95.83	91.67	0.93	0.80	1 (CAS)	2 (AGP, JHB)
Apical	100	100	1.00	1.00	0	0
Dorsal	91.67	100	0.81	1.00	2 (DCT, DSA)	0
Velopharyngeal	83.33	91.67	0.74	0.87	4 (**AMG**, CAS, DCT, ISA)	2 (**AMG**, RPJ)
Pharyngeal	91.67	95.83	0.87	0.93	2 (AGP, **APJ**)	1 (**APJ**)
Larynx height	87.50	87.50	0.81	0.81	3 (CGP, DSA, MHB)	3 (AMG, JCT, RPJ)
Vocal tract tension	87.50	95.83	0.81	0.93	3 (APJ, MHB, RPJ)	1 (SGF)
Larynx tension	91.67	87.50	0.84	0.71	2 (**DSD**, DSA)	3 (CGP, **DSD**, MML)
Voice type	83.33	87.50	0.70	0.78	4 (AMG, APJ, **DSD**, **MHB**)	3 (DCT, **DSD**, **MHB**)

Cases of "substantial agreement" (κ: 0.61–0.80) instead of "very good" or "almost perfect" (κ: 0.81–1) are gray-shaded. Bold and underlined are the speakers causing intra-rater disagreement in both raters.

TABLE 3.
Raw Agreement and Unweighted Kappa Results of Inter-Rater Agreement Between R1 and R2

Setting	Percentage Agreement	Cohen's Kappa	N Disagreements
Labial	75.00	0.55	6
Mandibular	50.00	0.06	12
Apical	54.17	0.11	11
Dorsal	91.67	0.78	2
Velopharyngeal	70.83	0.55	7
Pharyngeal	37.50	0.11	15
Larynx height	66.67	0.50	8
Vocal tract tension	41.67	0.13	14
Laryngeal tension	66.67	0.30	8
Voice type	66.67	0.42	8

Settings where less than moderate agreement (<0.41) was reached are gray-shaded.

who caused most of the intra-rater disagreements—shown in brackets in the last column—are recurrent in a single rater or in both. For instance, speaker APJ accounts for the disagreements in the setting *pharyngeal* in both raters, or speakers DSD and MHB are the main reason why better agreement is not achieved for *voice type*. Notably, speaker MHB seems to be causing most internal inconsistencies in the first rater (for the *labial*, *larynx height*, *vocal tract tension*, and *voice type* settings).

Inter-rater agreement

The results for the inter-rater agreement are shown in Tables 3 and 4. They are based on the ratings provided by the two raters in the second evaluation round. As each rater internal consistency was high (Table 2), any of the rounds of their perceptual assessment could have been used for inter-rater estimates; it seemed more logical, however, to use the ratings of the second round, where more confidence in the ratings was acknowledged by both raters. We first tested raw (percentage) agreement and unweighted Cohen's kappa. In a second step, linear weighted kappa was calculated to avoid the equal treatment of all types of disagreements.

Raw agreement and unweighted kappa

According to the results shown in Table 3, the overall inter-rater agreement is very high, especially in terms of percentage agreement. However, agreement seems to be strongly setting-dependent. This is especially clear in the kappa values. Out of the 10 settings, half of them achieve agreement values higher than 0.41 ("moderate agreement"), whereas for the other half raters attain less than moderate agreement. In other words, some settings seem to be easier to agree upon than others. In the first group, with κ values ranging from "moderate" (0.41–0.60) to "substantial" (0.61–0.81), we find the following settings, ranked from higher to lower kappa values: *dorsal* (0.78), *labial* and *velopharyngeal* (0.55), *larynx height* (0.42), and *voice type* (0.42). The second group of settings presents κ values ranging between "fair" (0.21–0.40) and "slight" (0.00–0.20) agreement: *laryngeal tension* (0.30), *vocal tract tension* (0.13), *apical* (0.11), *pharyngeal* (0.11), and *mandibular* (0.06).

Linear weighted kappa

Table 4 shows that the results improve for all settings when using linear weighted kappa. Standard errors are very similar across different settings. The last two columns provide information about (1) the maximum possible linear weighted kappa, given the observed marginal frequencies, and (2) a new observed kappa, proportional to the maximum possible. This is the best possible agreement and it shows a shift in the agreement level in all settings; for example, from "slight" to "fair" (*mandibular*) and from "slight" to "moderate" (*apical*). A considerable shift is also observed now in "larynx height", with almost perfect agreement. Sim and Wright[54] recommend reporting the magnitude of kappa to the maximum attainable kappa for the contingency table concerned, as this provides an indication of the effect of imbalance in the marginal totals on the magnitude of kappa. They also suggest constructing a confidence interval around the obtained value of

TABLE 4.
Linear Weighted Kappa Results of Inter-Rater Agreement Between R1 and R2

Setting	Observed Kappa (κ)	Standard Error	95% Confidence Interval Lower Limit	95% Confidence Interval Upper Limit	Maximum Possible κ†	Proportional κ to Maximum Possible*
Labial	0.53	0.17	0.20	0.86	0.80	0.66
Mandibular	0.11	0.15	0	0.41	0.56	0.20
Apical	0.14	0.12	0	0.39	0.28	0.50
Dorsal	0.79	0.13	0.52	1	0.79	1
Velopharyngeal	0.59	0.14	0.32	0.86	0.90	0.66
Pharyngeal	0.19	0.12	0	0.42	0.49	0.38
Larynx height	0.60	0.11	0.37	0.83	0.70	0.86
Vocal tract tension	0.21	0.16	0	0.51	0.82	0.25
Laryngeal tension	0.37	0.13	0.11	0.63	0.65	0.57
Voice type	0.43	0.17	0.09	0.76	0.94	0.45

* Maximum possible linear weighted kappa given the observed marginal frequencies.
† Observed kappa as proportion of maximum possible.

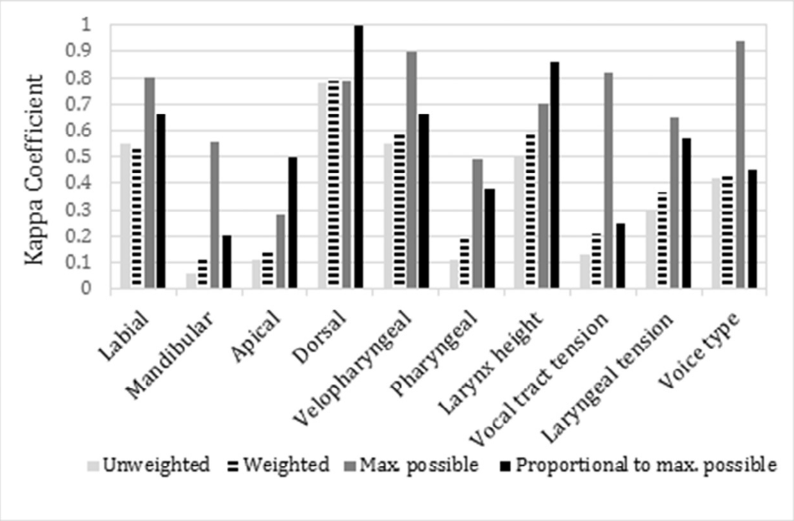

FIGURE 1. Kappa values per setting: unweighted Cohen's kappa (light gray), linear weighted kappa (lines), maximum possible linear weighted kappa given the observed marginal frequencies (dark gray), and observed kappa as proportion of maximum possible (black).

kappa to reflect sampling error. Both aspects can be observed in Table 4. Contingency tables for all settings can be found in Appendix 3, where an indication of the existence of bias and prevalence can be observed per setting.

Figure 1 shows the different kappa values obtained in each setting. Some leeway for improvement is possible in *mandibular*, *pharyngeal*, *velopharyngeal*, *vocal tract tension*, and *voice type*. However, the settings related to the activity of the tongue (both *apical* and *dorsal*) show a very high agreement in proportion to the maximum possible. *Larynx height*, together with *dorsal*, is the setting that benefits the most from the application of linear weighting: proportional kappa to maximum possible is better than the maximum possible.

In the case of *larynx height*, raters do not seem to be in strong disagreement by rating a voice as lowered larynx, one rater, and as raised larynx, the other rater. This can be clearly observed in the contingency table for this setting. Although all contingency tables can be found in Appendix 3, the contingency table of *larynx height* is also reproduced in Table 4 for illustration purposes. When R1 selects lowered larynx, R2 never selects raised larynx, so 0 appears in the upper right corner of the table. The same thing applies in the lower left corner of the table, as no cases of disagreement were found for R1 judging a voice as raised larynx. Kappa is affected by the presence of bias between observers and by the distributions of data across the categories, that is, prevalence.[55] As shown in Table 5, prevalence is on "neutral" and "lowered larynx"; R1 shows certain bias toward judging as raised larynx three voices that R2 considered neutral; conversely, R2's bias is toward rating as lowered larynx four voices that fall within the neutral configuration of this setting for R1.

TABLE 5.
Contingency Table for "Larynx Height" Setting Based on Judgments Made by R1 and R2

		R2		
		Lowered larynx	Neutral	Raised larynx
R1	Lowered larynx	6	0	0
	Neutral	4	6	1
	Raised larynx	0	3	4

Speaker similarity

The method for calculating EDs with categorical variables (ie, SMC), outlined in the Speaker Similarity Measurement section, allowed us to obtain a numerical index of similarity between pairs of speakers. These SMCs are based on the perceptual ratings made by R1 in the second evaluation round. Tables 6 and 7 show the results for twin pairs and unrelated pairs, respectively. On average, twin pairs obtained higher SMC (mean: 0.64) than unrelated pairs (mean: 0.35), indicating more similarity among the former.

DISCUSSION

Rater agreement

The results obtained allow us to provide an informed answer to the research questions formulated in this study. The first question was how reliable the proposed SVPA is in terms of agreement within and between raters. This implied, in turn, two derived research questions: whether the proposed SVPA can achieve satisfactory levels of intra- and inter-rater agreement, and whether intra- and inter-rater agreement is setting-dependent.

Intra-rater agreement

In terms of intra-rater agreement, both raters achieved excellent internal consistency for all settings, except for three where agreement is slightly lower, but still substantial (κ: 0.61–0.80): *velopharyngeal*, *larynx tension*, and *voice type*. Velopharyngeal disagreements mostly affect R1, whereas larynx tension disagreements are found in R2 to a greater extent. In contrast, there are several speakers whose voice type classification causes intra-rater disagreements equally for R1 and R2.

TABLE 6.
Similarity Matching Coefficients (SMC) as Distance Measure Between Twin Pairs

Speaker Pair	AGF-SGF	AGP-CGP	AMG-EMG	APJ-RPJ	ARJ-JRJ	ASM-RSM	CAS-PAS	CSD-DSD	DCT-JCT	DSA-ISA	JHB-MHB	MML-PML
SMC	0.8	0.7	0.8	0.4	0.5	0.7	0.6	0.8	0.5	0.6	1	0.3

TABLE 7.
Similarity Matching Coefficients (SMC) as Distance Measure Between Unrelated Pairs

Speaker Pair	AGF-AGP	AMG-APJ	ARJ-ASM	CAS-CGP	CSD-DCT	DSA-DSD	EMG-ISA	JCT-JHB	JRJ-MHB	MML-PAS	PML-RPJ	RSM-SGF
SMC	0.3	0.4	0.2	0.4	0.3	0.4	0.2	0.3	0.2	0.5	0.3	0.5

All in all, these results suggest that internal standards for setting assessment are clear within each rater (ie, they are consistent in their ratings because they have accurate definitions for each setting). On the other hand, disagreements in perceptual evaluations seem not to be completely speaker independent, as the same speakers frequently appear to be causing inconsistencies in the ratings of certain settings, regardless of the rater. Because *velopharyngeal*, *larynx tension*, and *voice type* are the settings making experts slightly less consistent—when we compare one perceptual evaluation with another—we will discuss some possible explanations for this.

As for *velopharyngeal* classifications, Mackenzie Beck[22] claims that velopharyngeal settings pose some of the most complex problems for phoneticians, possibly because neither the perceptual characteristics nor the physiological correlates are completely clear for the nasal and denasal setting or the cul-de-sac resonance. Because our SVPA forces the analyst to decide whether the abnormality in the speaker's velopharyngeal cavity is due to an excess of nasalization or rather to a lack of it (ie, hyper- or hyponasality), this compulsory binary distinction may induce internal inconsistencies in the rater, as some speakers may present a combination of those.[d] More investigations into the acoustic correlates of hyper- and hyponasality[56] could help analysts converge in their future ratings.

In terms of disagreements over *larynx tension*, these could be better explained when looking jointly at the voice type disagreements, as it is well known that some voice or phonatory types are typically associated with either a lax or tense configuration of the larynx. Prototypically, "breathy" phonation requires low (ie, lax) muscular tension, with minimal adductive tension, weak medial compression, and medium longitudinal tension of the vocal folds, whereas "harsh" occurs as a result of very strong tension in the vocal folds, medial compression, and adductive tension.[1] Interestingly, speaker DSD caused intra-rater disagreement in both raters for larynx tension and voice type, which further supports the dependence of both settings.

Voice type (ie, phonation features) is probably the setting for which SVPA is less suitable, or at least that for which more training is required to improve agreement. Combined phonatory qualities are frequent.[57] Laver[19] mentions some of them: "harsh whispery voice" or "harsh creaky voice", for instance. The latter does not cause any problem in our SVPA, as both harsh and creaky belong to the tense larynx typology. The former, however, can be problematic because some raters may categorize the voice as "tense"—considering that the harsh component is predominant—whereas some other raters may consider that the whispery aspect (airflow escape) prevails in perception, hence categorizing the voice as "lax." Therefore, for this *voice type* setting, some biases were expected in the raters due to both the nature of the task (binary decision) and to the existence of compound phonation types, probably more frequent in pathological speakers. Nevertheless, older versions of the VPA scheme had

[d] It is worth noting that the SVPA does not include the option of marking the presence of a setting as intermittent, which seems to be the proposed solution of Mackenzie Beck[22] for cases of occasional denasalization of nasal segments, as in some types of dysarthria. That is, in those instances, the appropriate scalar degree for "nasal" should be ticked on the protocol, while also marking "i" on the denasal scale.

to deal with this type of issues as well.[e] Other types of statistics for the measurement of intra-rater agreement that could be worth exploring in future studies are repeatability and test-retest reliability methods.

Inter-rater agreement

As regards inter-rater agreement, the results are strongly setting-dependent. Although there does not seem to be excellent agreement for any setting except the *dorsal*, the fact that none of the kappa values is negative means not only that the raters are never in disagreement but also that they agree more than would be predicted by chance alone. Some possible explanations of the excellent agreement for the dorsal setting are discussed below, although *labial* and *velopharyngeal*, *larynx height*, and *voice type* are also worth highlighting. Even with unweighted kappa, they yield agreement values above 0.41. These results, although labeled "moderate," are especially good considering the small number of calibration sessions, and the total size of the population perceptually assessed (n = 24).

When applying linear weighting, the results still show a division between half of the settings (*voice type*, *larynx height*, *labial*, *velopharyngeal*, and *dorsal*) attaining moderate or higher agreement ($\kappa < 0.41$) and the other half (*mandibular*, *apical*, *pharyngeal*, *vocal tract tension*, and *laryngeal tension*) ranging from slight to fair agreement. *Larynx height* and *vocal tract tension* are the settings that benefit the most from linear weighting. The former no longer yields moderate agreement but good, whereas the latter yields fair agreement instead of slight.

As for the setting *dorsal*, comparatively this is the highest agreement achieved for a setting, even with unweighted kappa. This could be due not only to its high perceptual salience but also to two further aspects. On the one hand, the calibration meeting held by the raters resulted in clear instructions on when to rate a speaker as presenting "backed and lowered tongue body". This configuration was reserved for speakers with a characteristic debuccalization, a well-known and perceptually salient sociolinguistic marker typically heard in some areas of Madrid (see Momcilovic[58] for a discussion). On the other hand, the prevalence of this non-neutral dorsal setting (ie, "backed and lowered tongue body" versus "fronted and raised tongue body") could also have favored the good inter-rater agreement obtained.

The linear weighted kappa results highlight at least four settings that would require further training to achieve better inter-rater agreement. The most difficult to agree upon is *mandibular* (unweighted $\kappa = 0.05$; weighted $\kappa = 0.11$; proportional κ to maximum possible = 0.20). This could be due to the fact that speakers' production varied throughout the recordings. Examined recordings were around 1.5 minutes per speaker, so different degrees of hyper- and hypoarticulation—correlates of open and close jaw—could appear in the speech of one and the same participant. Although VQ aspects need to be perceptually assessed on the basis of the speaker's long-term configurational tendencies, the mandibular setting could be one of the settings that depend more strongly on paralinguistic aspects. In view of the contingency table for this setting, there is a general prevalence of the "neutral" configuration with an important bias by R2 to judge as "close jaw" what R1 considers "neutral."

Apical is the second setting most difficult to agree upon (unweighted $\kappa = 0.11$; weighted $\kappa = 0.14$; proportional κ to maximum possible = 0.50). This is an expected result given that the neutral setting for SPS is characterized by dental-alveolar articulatory anchorage and considerable apical activity, with a number of sibilant sounds making a speaker differ from others, mainly in particular allophonic choices. Although /s/ in SPS has been described as apical in contrast with different varieties of predorsal and predorso-alveolar articulations in most of Andalusia and Central-South America,[38] a range of possible pronunciations can still characterize a speaker around Madrid and the center of Spain for a variety of cross-dialectal influences, migration context, speaker accommodation, or idiosyncratic factors (eg, physiological reasons). Although we can observe the prevalence of the "neutral" configuration in the relevant contingency table, there are also biases in the raters toward marking as "advanced tongue tip" or "retracted tongue tip" voices that the other rater considered neutral. This implies that a better definition of the non-neutral labels should be established in training sessions. Furthermore, *apical* seems to be a setting for which it would be recommended to conduct acoustic analyses, as acoustic correlates of the sounds involving apical activity are typically well known and could help analysts converge in their ratings.

The *pharyngeal* setting—especially the "expanded pharynx" articulation—was a recent addition to the protocol. However, scarce references can be found as to how to perceptually assess this aspect, which otherwise seems to be highly correlated with other settings in the protocol. For instance, expansion of the pharyngeal cavity could be due to lowering of the larynx, which makes a different setting on its own. As for pharyngeal constriction, descriptions are somehow impressionistic, suggesting that this type of constriction "lends a 'strangulated' quality to the voice, so that at high scalar degrees the empathetic listener is aware of considerable discomfort and obstruction of the pharynx"[22] (p. 12). Our agreement results show that this is a setting upon which it is difficult to agree (unweighted $\kappa = 0.11$; weighted $\kappa = 0.19$; proportional κ to maximum possible = 0.38). Voice experts would benefit from clearer descriptions of the pharyngeal setting and from a search for specific acoustic correlates.

Finally, agreement for *vocal tract tension* (unweighted $\kappa = 0.13$; weighted $\kappa = 0.21$; proportional κ to maximum possible = 0.25) is better when linear weighting is applied. However, it remains a subtle setting to evaluate perceptually. Unlike most other settings, fewer speakers were categorized as "neutral": eight speakers in the case of R1 and four in the case of R2; besides, raters only agreed in labeling one as "neutral". This makes the perception of this setting especially complex, probably due to the fact that vocal tract tension overlaps with a range of other dimensions. Mackenzie Beck[22] claims that "adjustments of overall muscle tension of the vocal tract tend to cause constellations

[e] Mackenzie Beck's manual[22] indicates the following instructions for rating phonation features: "Modal voice is marked simply as being present, intermittently present or absent on the protocol form. Where it occurs as a component of complex phonation types, it is described as 'voice' (e.g. in 'whispery in voice') and the auditory balance between it and other component(s) is indicated by the scalar degrees assigned to the accompanying component(s). For example, in a combination of voice with whisperiness, scalar degree 1–3 whisperiness would indicate that the voice component is perceptually more prominent; scalar degree 4–6 would indicate that the whisper component is perceptually most prominent"[22] (p. 16).

of changes in configurational and range settings" (p. 15). Indeed, the number of possible articulatory settings that would be associated with either lax or tense vocal tract is quite large (eg, different degrees of nasality and pharyngeal constriction). Furthermore, prosodic aspects seem to be associated with vocal tract tension, with faster tempo characterizing a high tense vocal tract and slower tempo a lax vocal tract. The number of acoustic correlates, although not all of them empirically tested yet, makes this a perfect candidate setting to increase agreement in future auditory evaluations, provided that perceptual assessment is aided by acoustic analysis.

In comparison with other perceptual protocols, there are few studies focusing on the reliability of VPA ratings with which we can contrast our results. Webb et al,[23] for instance, obtained much lower kappa values (ranging between 0.01 and 0.32) in the VPA ratings of seven judges (scalar degrees were reduced to 3 instead of a 6-point scale). Although it is not recommended to compare kappa results across studies because they are strongly influenced by the distribution of the data,[50] it is worth mentioning that—in view of their inter-rater agreement results—Webb et al[23] concluded that the greater scope of the VPA was at the expense of its reliability. Because these authors used the original VPA protocol, this brings us back to the question of the need for simplified protocols. Using the SVPA, our study shows kappa values overall higher than the study by Webb and colleagues. It seems, therefore, that the multidimensionality of the VPA scheme necessarily entails more rater discrepancies, and a setting reduction is justified. Furthermore, using the same set of experienced judges for both protocols, Webb et al[23] found that GRBAS was most reliable than VPA. The reliability of GRBAS has also been highlighted by Sellars et al[59], among others, even though they acknowledged that several studies report the highest kappa as no better than "moderate"—for overall grade.[29,59,60]

Mackenzie Beck[21] also tested inter-rater agreement between two skilled judges using the VPA scheme. Although the measures are not chance-corrected, the percent agreement is still informative; it shows that the stronger agreement (100%) is achieved in two rare settings (*protruded jaw* and *labiodentalization*). In fact, in San Segundo et al[13] none of these two settings were found in a normophonic population of 100 male speakers of Standard Southern British English, aged 18–25 (DyViS corpus[61]). Because of its low incidence also in Spanish, those non-neutral configurations were discarded from the *mandibular* and *labial* settings in the SVPA protocol. The strong agreement found in Mackenzie Beck[21]—given such a crude measure as percentage agreement—could be inflated due to the rare occurrence of the setting (ie, a high percentage agreement is expected when a setting is mostly absent, as it is easier for raters to agree on its non-presence).

To sum up this section, the results obtained show that the proposed SVPA is very reliable in terms of agreement within and between raters (RQ1), as satisfactory levels of intra- and inter-rater agreement are achieved, both in comparison with previous studies and taking into account the issues typically associated with agreement measures (ie, whether variables are weighted or not, whether there is bias or prevalence in the ratings, etc). We have shown that both intra- and inter-rater agreements are setting-dependent, and some possible explanations have been provided to discuss why certain settings are more difficult to agree upon than others.

Speaker similarity

Depending on the field where the SVPA protocol is to be used, different levels of agreement will be considered satisfactory. For example, in forensic phonetics, we can presume that inter-rater agreement is very relevant, as courts typically require the expert to provide a reliability measure or error rate of the method used. Equally important is, however, the potential of the technique to robustly capture the most idiosyncratic aspects of a speaker's voice, ideally those that can make him distinguishable from other speakers.

After testing the SVPA reliability, we applied a method for an easy quantification of VQ similarity between speakers. For that purpose, SMCs were used, calculated pairwise for twin speakers and unrelated speakers. This was aimed at testing the robustness of the proposed SVPA scheme, as it was hypothesized that a perceptual protocol for VQ assessment should reveal that twin pairs were more similar than non-twin pairs. Indeed, the results showed that higher SMCs occur in twin pairs than in unrelated speakers, indicating more similarity among the former: the average SMC is two times higher in the former than in the latter. This suggests that the proposed simplified method for VQ perceptual assessment is well designed and potentially useful for forensic applications: any similar speaker pair should be assessed as very similar in VQ terms, whereas dissimilar speaker pairs should show lower SMCs, thus reflecting VQ dissimilarity.

Although values can be pair-dependent in the case of twins (eg, JHB and MHB are completely similar with an SMC of 1; MML and PML are very different with an SMC of 0.3), twin pairs typically share more than half of their VQ characteristics. In the case of unrelated speakers, their SMC tends to be homogenously distributed around the mean, which indicates that most of them share only three or four setting configurations. They can be distinguished on average by more than seven settings, which shows the forensic discriminatory potential of the SVPA. Setting matches are based on shared accent features or coincidences on neutral configurations.

Although MZ twins are overall more similar than non-twin speakers in terms of VQ distances, the SVPA is still useful to detect fine-grained aspects of VQ, as twins do not exhibit an absolute match of settings. By way of example, twin pairs AGF and SGF have an SMC of 0.8, indicating their strong similarity in overall VQ. Nonetheless, they can be distinguished by two settings: SGF presents open jaw, whereas AGF has a neutral jaw. The same applies to velopharyngeal configurations: SGF deviates from neutrality, whereas AGF does not present either nasality or denasality. Typically, the same trend can be observed for the rest of twin pairs: even though their overall SMC indicates strong similarity, there are still particularities in the voice of each one that can tell them apart when we use this componential approach to VQ; it is possible to separate even very similar speakers on at least two components of our scheme.

The only exception seems to be twins JHB and MHB, who were judged completely similar with the SVPA protocol. These results are in good accordance with acoustic studies such as San

Segundo[35] or Loakes,[32] which showed that MZ twins do not make a homogenous group of speakers, with some pairs found to be strikingly similar and a minority of pairs found to be as different as two unrelated speakers. As a case in point, MML and PML obtained an SMC of 0.3, which lies in the mean value of SMC for non-twin pairs. Interestingly, this is the same pair that previous acoustic studies[53,55] found very dissimilar, especially in terms of phonatory aspects.

Summarizing the main points discussed in this section, the second research question of this study was whether an index or distance measure of speaker similarity could be extracted from the SVPA scheme; we have shown that it is possible to design a method that allows for a quantitative measure of speaker similarity. Related to that question, we have also shown that the SVPA scheme reveals that MZ twins are overall more similar than non-twin speakers, as expected, but at the same time it is a useful tool to detect fine-grained aspects of VQ that distinguish even very similar-sounding speakers (ie, MZ twins).

CONCLUSIONS

The main purpose of this study was to design an SVPA protocol for the assessment of VQ—reduced in number of settings and rating options—which could prove reliable in terms of intra- and inter-rater agreement, and from which an index of speaker similarity could be extracted.

First, the results of this investigation have shown that it is possible to achieve high intra-rater agreement and considerably good inter-rater agreement using the proposed SVPA scheme. The fact that inter-rater agreement seems to depend strongly on particular settings—only some showing certain improvement with linear weighting—makes it necessary to increase the number of training sessions between analysts. Furthermore, better agreement results could be achieved with the use of perceptual anchors, as it has been suggested in previous studies,[59,62] together with clearer definitions of the neutral baseline for the speaker population under evaluation. The search for acoustic correlates of some of the settings showing poorer agreement would be highly necessary as well. For the language variety of this investigation (ie, SPS), we suggest that *apical*, *pharyngeal*, and *vocal tract tension* are settings that require extra training to achieve better agreement.

Second, this study has shown that a distance measure of speaker similarity (ED or SMC) can be derived from the SVPA protocol, which improves on predominantly qualitative approaches to VQ and which could prove useful in areas such as FSC. Having selected MZ twins as subjects of our study, we were able to examine the degree of VQ similarity in speakers who represent the most extreme cases of anatomical similarity, both in vocal tract and vocal fold physiognomy. The comparison between the SMCs resulting from the perceptual assessment of twin pairs and the SMCs obtained when pairing non-twin subjects showed that the former are more similar in terms of VQ, as expected. This points to the adequate design of the SVPA. In other words, it can be argued that the SVPA must have preserved the most relevant settings from the original VPA, despite the simplification, given that it has yielded higher SMCs for the most similar speakers than for a random combination of two unrelated speakers from the same population (ie, sharing language variety, age range, etc). Nevertheless, the SVPA has also proved apt for detecting at least a few unshared settings in MZ twins with a very close VQ overall. When it allows for capturing fine-grained differences even in very similar-sounding speakers, the usefulness of this tool is revealed as a componential approach to the assessment of VQ.

Forensic phonetics is one of the research areas that can benefit from an index of speaker similarity based on a perceptual protocol that is not too difficult to implement and for which reliability estimates can be provided. Although the VPA protocol is already applied in FSC casework,[63] the SVPA could make its use more widespread, even in other forensic tasks such as the design and validation of voice lineups. Numerous methodologies have been recently proposed to assess the degree of similarity between speakers.[64–66] Although the main objective of these studies is typically to reduce subjectivity and increase efficiency in the selection of the suitable speakers for a voice parade (ie, foils or comparison speakers), other commercial applications of voice similarity assessment include voice casting or voice assignment.[67,68]

The current study has some limitations that should be acknowledged; some of them have already been mentioned in the discussion. For example, the compulsory binary choice that the rater must make for each setting group might not be the most appropriate to rate all VQ aspects, especially those that admit a combination of settings (eg, harsh-whispery in *voice type* or nasal-denasal in *velopharyngeal*). Although the dual nature of the SVPA seems essential for simplification purposes and in order to obtain an index of speaker similarity, it has to be noted that the SVPA is designed so that a holistic description of VQ can complement the featural analysis (Appendix 2). This can compensate the strictness of the binary criteria in research fields where it may not be so necessary to quantify speaker similarity and where qualitative feedback is deemed relevant and informative (eg, comments on VQ in the initial stages of traineeship in the protocol or during the process of learning the articulatory settings of a foreign language).

Despite these limitations, we suggest that a simplified protocol like the one proposed here, which is limited to 10 settings, with only three categories (one for neutral and two for opposing non-neutral configurations), will serve to characterize speakers of different language varieties and to achieve acceptable agreement within an analyst and between different analysts. Therefore, the SVPA can be a useful method not only in areas such as clinical therapy or forensic phonetics, but also in others such as sociophonetics or L2 (second language) phonology. Because the SVPA tool has only been validated in SPS so far, future studies will examine the potential of this tool in other languages.

Some of the questions that arise from this study are, first, whether rating normophonic speakers is more difficult than rating speakers who present some voice impairment, or at least whether the former require different (simplified) rating systems. Besides, further research seems necessary to explore whether different perceptual dimensions can be best measured using different scale resolutions, depending on the nature of the dimension (eg, visual analog scales or equal-appearing interval scales). This would be due to the existence of two basic types of perceptual continua: prothetic and metathetic continua.[2,69] Whereas a prothetic dimension is described as an additive, quantitative continuum—the dimension varies in magnitude or quantity—a metathetic dimension, also described as substitutive, qualitative continuum, would vary in terms of a change in quality. For instance, some studies have shown that hypernasality would be prothetic[70] and therefore the use of equal-appearing interval scales would not be recommended to rate hypernasality. Many other perceptual dimensions have not been investigated yet. Finally, as a description of the VQ of SPS, the settings described in this paper should ideally be checked against instrumental acoustic measures to further investigate the degree of correlation between perceptual and acoustic assessments.

Acknowledgment

This work was supported by the Arts and Humanities Research Council [grant number AH/M003396/1].

APPENDIX 1

Vocal Profile Analysis (VPA)

	First Pass		Second Pass						
				Moderate			Extreme		
	Neutral	Non-Neutral	Setting	1	2	3	4	5	6
A. Vocal tract features									
1.Labial			Lip rounding/protrusion						
			Lip spreading						
			Labiodentalization						
			Extensive range						
			Minimized range						
2. Mandibular			Close jaw						
			Open jaw						
			Protruded jaw						
			Extensive range						
			Minimized range						
3. Lingual tip/blade			Advanced tip/blade						
			Retracted tip/blade						
4. Lingual body			Fronted tongue body						
			Backed tongue body						
			Raised tongue body						
			Lowered tongue body						
			Extensive range						
			Minimized range						
5. Pharyngeal			Pharyngeal constriction						
			Pharyngeal expansion						
6. Velopharyngeal			Audible nasal escape	▨	▨	▨			
			Nasal						
			Denasal						
7. Larynx height			Raised larynx						
			Lowered larynx						
B. Overall muscular tension									
8. Vocal tract tension			Tense vocal tract						
			Lax vocal tract						
9. Laryngeal tension			Tense larynx						
			Lax larynx						
C. Phonation features									

		Present		Scalar Degree					
				Moderate			Extreme		
	Setting	Neutral	Non-Neutral	1	2	3	4	5	6
10. Voicing type	Voice			▨	▨	▨	▨	▨	▨
	Falsetto	▨		▨	▨	▨	▨	▨	▨
	Creak	▨		▨	▨	▨	▨	▨	▨
	Creaky	▨							
11. Laryngeal frication	Whisper	▨							
	Whispery	▨							
12. Laryngeal irregularity	Harsh	▨							
	Tremor	▨							

Table adapted from Beck (2007). Shaded cells mean that the corresponding setting does not admit the specified degree(s) or label.

APPENDIX 2

Simplified Vocal Profile Analysis (SVPA)

A. Featural *(tick the appropriate box)*				
Major Setting Groups	Settings	Numerical Labels for One Neutral (N) and Two Non-Neutral Configurations		
		–1	0	+1
Vocal tract settings	Labial	Spreading	N	Rounding
	Mandibular	Close	N	Open
	Apical	Retracted	N	Advanced
	Dorsal	Backed and lowered	N	Fronted and raised
	Velopharyngeal	Denasal	N	Nasal
	Pharyngeal	Constricted	N	Expanded
	Laryngeal height	Lowered	N	Raised
Overall muscular tension	Vocal tract tension	Lax	N	Tense
	Laryngeal tension	Lax	N	Tense
Phonation	Voice type	Whisper/Breathy	N	Creaky/Harsh
B. Holistic				
(fill with qualitative input; comments, etc)				

APPENDIX 3

Contingency Tables Showing the Frequency Distribution of Ratings (Per Setting) by Rater One (R1) and Rater Two (R2)

	Labial	R1		
		Spreading	*Neutral*	*Rounding*
R2	*Spreading*	**2**	1	1
	Neutral	0	**12**	1
	Rounding	0	3	**4**

	Mandibular	R1		
		Close	*Neutral*	*Open*
R2	*Close*	**3**	3	0
	Neutral	6	**9**	3
	Open	0	0	**0**

	Apical	R1		
		Retracted	*Neutral*	*Advanced*
R2	*Retracted*	**2**	5	1
	Neutral	1	**11**	4
	Advanced	0	0	**0**

	Dorsal	R1		
		Back and lowered	*Neutral*	*Front and raised*
R2	*Back and lowered*	**5**	0	0
	Neutral	1	**17**	1
	Front and raised	0	0	**0**

	Velopharyngeal	R1		
		Denasal	*Neutral*	*Nasal*
R2	*Denasal*	**7**	1	1
	Neutral	1	**7**	1
	Nasal	0	3	**3**

	Pharyngeal	R1		
		Constricted	*Neutral*	*Expanded*
R2	*Constricted*	**0**	1	0
	Neutral	9	**5**	3
	Expanded	1	1	**4**

	Laryngeal height	R1		
		Lowered larynx	*Neutral*	*Raised larynx*
R2	*Lowered larynx*	**6**	0	0
	Neutral	4	**6**	1
	Raised larynx	0	3	**4**

	Vocal tract tension	R1		
		Lax	*Neutral*	*Tense*
R2	*Lax*	**6**	2	2
	Neutral	2	**3**	5
	Tense	1	2	**1**

	Laryngeal tension	R1		
		Lax	*Neutral*	*Tense*
R2	*Lax*	**0**	0	0
	Neutral	3	**2**	2
	Tense	1	2	14

	Voice type	R1		
		Whisper/breathy	*Neutral*	*Creaky/harsh*
R2	*Whisper/breathy*	**1**	1	0
	Neutral	0	**7**	4
	Creaky/harsh	1	2	**8**

Diagonal cells represent agreement (bold) and off-diagonal cells represent disagreement; cases of remarkable bias (disagreements ≥5) are underlined.

REFERENCES

1. Laver J. *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press; 1980.
2. Kent RD. Hearing and believing: some limits to the auditory-perceptual assessment of speech and voice disorders. *Am J Speech Lang Pathol*. 1997;5:7–23.
3. Kreiman J, Gerratt BR, Kempster GB, et al. Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *J Speech Hear Res*. 1993;36:21–40.
4. Kreiman J, Gerrat BR, Ito M. When and why listeners disagree in voice quality assessment tasks. *J Acoustic Soc Am*. 2007;122:2354–2364.
5. Gerratt B, Kreiman J, Erman A, et al. Perceptual evaluation of voice quality: review, tutorial and a framework for future research. *J Speech Hearing Res*. 1993;36:21–40.
6. Kreiman J, Gerratt BR. Comparing two methods for reducing variability in voice quality measurements. *J Speech Lang Hear Res*. 2011;54:803–812.
7. Kreiman J, Sidtis D. *Foundations of Voice Studies: Interdisciplinary Approaches to Voice Production and Perception*. Boston: Wiley-Blackwell; 2011.
8. McGlashan J, Fourcin A. Objective evaluation of the voice. In: Gleeson M, Browning G, Burton M, eds. *Scott-Brown's Otorhinolaryngology, Head and Neck Surgery*. 7th ed. London: Hodder Arnold; 2008:2170–2191.
9. Ma EPM, Yu EML. Multiparametric evaluation of dysphonic severity. *J Voice*. 2005;20:380–390.
10. Wewers ME, Lowe NK. A critical review of visual analogue scales in the measurement of clinical phenomena. *Res Nurs Health*. 1990;13: 227–236.
11. Foulkes P, French P. Forensic speaker comparison: a linguistic-acoustic perspective. In: Solan L, Tiersma P, eds. *The Oxford Handbook of Language and Law*. Oxford: Oxford University Press; 2012:418–421.
12. Gold E, French P. International practices in forensic speaker comparison. *Int J Speech Lang Law*. 2011;18:293–307.
13. San Segundo E, Foulkes P, French P, et al. Voice quality analysis in forensic voice comparison: developing the vocal profile analysis scheme. 25th Conference of the International Association for Forensic Phonetics and Acoustics, York, July 24–27, 2016.
14. Esling JH. Voice quality in Edinburgh: a sociolinguistic and phonetic study [Unpublished PhD dissertation]. University of Edinburgh; 1978.
15. Stuart-Smith J. Glasgow: accent and voice quality. In: Docherty G, Foulkes P, eds. *Urban Voices: Accent Study in the British Isles*. London: Arnold; 1999:203–222.
16. Coadou M, Rougab A. Voice quality and variation in English. *Proc 16th Int Cong Phonetic Sci*. 2007;2077–2080.
17. Beck JM, Schaeffler F. Voice quality variation in Scottish adolescents: gender versus geography. *Proc 18th Int Cong Phonetic Sci*. 2015;0737.
18. Laver J, Wirz S, Mackenzie Beck J, et al. A perceptual protocol for the analysis of vocal profiles. Edinburgh: University of Edinburgh, Work in Progress. 1981;14:139–155.
19. Laver J. *The Gift of Speech*. Edinburgh: Edinburgh University Press; 1991.
20. Laver J. Phonetic evaluation of voice quality. In: Kent R, Ball M, eds. *Voice Quality Measurement*. San Diego, CA: Singular Publishing; 2000:37–48.
21. Mackenzie Beck J. Perceptual analysis of voice quality: the place of Vocal Profile Analysis. In: Hardcastle W, Mackenzie-Beck J, eds. *A Figure of Speech: a Festschrift for John Laver*. London/Mahwah, NJ: Lawrence Erlbaum Associates; 2005:285–322.
22. Mackenzie Beck J. *Vocal Profile Analysis Scheme: A User's Manual*. Edinburgh: Queen Margaret University College-QMUC, Speech Science Research Centre; 2007.
23. Webb AL, Carding PN, Deary IJ, et al. The reliability of three perceptual evaluation scales for dysphonia. *Eur Arch Otorhinolaryngol*. 2004;261:429–434.
24. Shewell C. *Voice Work: Art and Science in Changing Voices*. Oxford: Wiley-Blackwell; 2009.
25. Dejonckere PH, Bradley P, Clemente P, et al. A basic protocol for functional assessment of voice pathology. *Eur Arch Otorhinolaryngol*. 2001;258:77–82.
26. Hirano M. *Clinical Examination of Voice*. Vienna/New York: Springer-Verlag; 1981.
27. De Bodt MS, Wuyts FL, Van de Heyning PH, et al. Test-retest study of the GRBAS scale: influence of experience and professional background on perceptual rating of voice quality. *J Voice*. 1997;11:74–80.
28. Dejonckere PH, Remacle M, Fresnel-Elbaz E, et al. Reliability and clinical relevance of perceptual evaluation of pathological voices. *Rev Laryngol Otol Rhinol*. 1998;119:247–248.
29. Dejonckere PH, Obbens C, De Moor GM, et al. Perceptual evaluation of dysphonia: reliability and relevance. *Folia Phoniatr Logop*. 1993;45:76–83.
30. Gelfer MP. Perceptual attributes of voice: development and use of rating scales. *J Voice*. 1988;2:320–326.
31. Decoster W, Van Gysel A, Vercammen J, et al. Voice similarity in identical twins. *Acta Otorhinolaryngol Belg*. 2000;55:49–55.
32. Loakes D. A forensic phonetic investigation into the speech patterns of identical and non-identical twins [Doctoral dissertation]. University of Melbourne; 2006.
33. San Segundo E, Gómez-Vilda P. Evaluating the forensic importance of glottal source features through the voice analysis of twins and non-twin siblings. *Lang Law/Linguagem e Direito*. 2014;1:22–41.
34. San Segundo E, Künzel H. Automatic speaker recognition of Spanish siblings: (monozygotic and dizygotic) twins and non-twin brothers. *Loquens*. 2015;2:e021.
35. San Segundo E. Forensic speaker comparison of Spanish twins and non-twin siblings: a phonetic-acoustic analysis of formant trajectories in vocalic sequences, glottal source parameters and cepstral characteristics [Doctoral dissertation]. Menéndez Pelayo International University & Spanish National Research Council; 2014.
36. San Segundo E. Forensic speaker comparison of Spanish twins and non-twin siblings: a phonetic-acoustic analysis of formant trajectories in vocalic sequences, glottal source parameters and cepstral characteristics [Thesis abstract]. *Int J Speech Lang Law*. 2015;22:249–253.
37. Nolan F. Forensic speaker identification and the phonetic description of voice quality. In: Hardcastle W, Mackenzie Beck J, eds. *A Figure of Speech: A Festschrift for John Laver*. London/Mahwah, NJ: Laurence Erlbaum Associates; 2005:385–411.
38. Hualde JI. *The Sounds of Spanish*. Cambridge: Cambridge University Press; 2005.
39. Martínez-Celdrán E, Fernández-Planas AM, Carrera-Sabaté J. Castilian Spanish. *J Int Phonetic Assoc*. 2003;33:255–259.
40. San Segundo E. A phonetic corpus of Spanish male twins and siblings: corpus design and forensic application. *Procedia*. 2013;95:59–67.
41. Crystal D, Varley R. *Introduction to Language Pathology*. 4th ed. London: Whurr; 1998.
42. Quilis-Morales A. *Tratado de fonología y fonética españolas*. Madrid: Gredos; 1993.
43. Gil Fernández J. Implicaciones fonológicas de la base de articulación. In: Hernández Alonso C, ed. *Filología y Lingüística. Estudios ofrecidos a Antonio Quilis*, Vol. 1. Madrid: CSIC, UNED, Universidad de Valladolid; 2005:219–252.
44. Gili Gaya S. *Elementos de fonética general*. 5th. ed. Madrid: Gredos; 1966.
45. Kovachova Rivera-de-Rosales V. La base articulatoria del eslovaco y del español. *Eslavística Complutense*. 2002;2:302–321.
46. Multon KD. Interrater reliability. In: Salkind N, ed. *Encyclopedia of Research Design*. Thousand Oaks, CA: Sage; 2010:626–628.
47. Cohen J. A coefficient of agreement for nominal scales. *Educ Psycholog Meas*. 1960;20:37–46. doi:10.1177/001316446002000104.
48. Stemler SE. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Prac Assess Res Eval*. 2004;9.
49. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. *Am J Epidemiol*. 1987;126:161–169.
50. Uebersax J. Diversity of decision-making models and the measurement of interrater agreement. *Psychol Bull*. 1987;101:140–146.
51. Landis JR, Koch GG. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*. 1977;363–374.
52. Fleiss JL, Levin B, Paik MC. The measurement of interrater agreement. In: Fleiss JL, Levin B, Paik MC, eds. *Statistical Methods for Rates and Proportions*. 3rd ed. Hoboken, NJ: John Wiley & Sons, Inc.; 2003. doi:10.1002/0471445428.ch18.

53. San Segundo E, Tsanas A, Gómez-Vilda P. Euclidean distances as measures of speaker dissimilarity including identical twin pairs: a forensic investigation using source and filter voice characteristics. *Forensic Sci Int*. 2017;270:25–38.
54. Sim J, Wright CC. The kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther*. 2005;85:257–268.
55. Byrt T, Bishop J, Carlin JB. Bias, prevalence and kappa. *J Clin Epidemiol*. 1993;46:423–429.
56. Lee ASY, Ciocca V, Whitehill TL. Acoustic correlates of hypernasality. *Clin Ling Phonetics*. 2003;17:259–264.
57. Ni Chasaide A, Gobl C. Voice source variation. In: Hardcastle W, Laver J, eds. *The Handbook of Phonetic Sciences*. Oxford: Blackwell; 1997:427–461.
58. Momcilovic NB. *A Sociolinguistic Analysis of /s/-aspiration in Madrid Spanish. LINCOM Studies in Romance Linguistics 60*. Munich: Lincom GmbH; 2009.
59. Sellars C, Stanton AE, McConnachie A, et al. Reliability of perceptions of voice quality: evidence from a problem asthma clinic population. *J Laryngol Otology*. 2009;123:755–763.
60. Dejonckere PH, Remacle M, Fresnel-Elbaz E, et al. Differentiated perceptual evaluation of pathological voice quality: reliability and correlations with acoustic measurements. *Rev Laryngol Otol Rhinol (Bord)*. 1995;117:219–224.
61. Nolan F, McDougall K, de Jong G, et al. The DyViS database: style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *Int J Speech Lang Law*. 2009;16:31–57.
62. Galek KE, Watterson T. Perceptual anchors and the dispersion of nasality ratings. *Cleft Palate Craniofac J*. 2016.
63. Gil Fernández J, San Segundo E. La cualidad de voz en fonética judicial. In: Garayzábal E, Reigosa M, eds. *Lingüística forense. La lingüística en el ámbito legal y policial*. Madrid: Euphonía Ediciones; 2013:154–199.
64. McDougall K. Assessing perceived voice similarity using multidimensional scaling for the construction of voice parades. *Int J Speech Lang Law*. 2013;20:163–172.
65. Kelly F, Alexander A, Forth O, et al. Identifying perceptually similar voices with a speaker recognition system using auto-phonetic features. *Proc Interspeech*. 2016;1567–1568.
66. San Segundo E, Foulkes P, Hughes V. Holistic perception of voice quality matters more than L1 when judging speaker similarity in short stimuli. Proceedings of the 16th Australasian International Conference on Speech Science and Technology. 2016.
67. Obin N, Roebel A, Bachman G. On automatic voice casting for expressive speech: speaker recognition vs. speech classification. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2009:950–954.
68. Adachi Y, Kawamoto S, Yotsukura T, et al. Automatic voice assignment tool for instant casting movie system. 2009 IEEE International Conference on Acoustics, Speech and Signal Processing. 2009:1897–1900.
69. Stevens SS. *Psychophysics*. New York: Wiley; 1975.
70. Cheng TH. Direct magnitude estimation versus visual analogue scaling in the perceptual rating of hypernasality [BSc dissertation]. University of Hong Kong; 2007.