A Computer-Based Tool for the Assessment of Voice Quality Through Visual Analogue Scales: VAS-Simplified Vocal Profile Analysis

*Eugenia San Segundo, and †Radek Skarnitzl, *Taiyuan, China, and †Prague, Czech Republic

Summary: In this study we propose a new tool for the perceptual assessment of voice quality. For its design, we have adapted the Simplified Vocal Profile Analysis so that the new tool features two main characteristics: (1) the ordinal scalar degrees of the original protocol are turned into a visual analog scale; and (2) the original paper-based version of the protocol is now a computer-based implementation. In order to assess the reliability of the new tool, five phoneticians listened to 12 different speakers and evaluated their voice quality using the proposed tool. Inter-rater agreement was then calculated using the Intraclass Correlation Coefficient. The results show that high agreement was reached for most of the perceptual settings of the protocol. Yet more investigations seem necessary into the continuous nature of the perceptual dimensions making up the voice quality of a speaker. As a preliminary approach to the graphical possibilities that the visual analog scale offers to the Simplified Vocal Profile Analysis, we explore the usefulness of multiple dot plots and propose an adaptation of the Bland-Altman plot to be used in pairwise comparisons. In this study, these visualization techniques are tested on two pairs of identical twins.

Key Words: Inter-rater agreement−System reliability−Tool design/development−Perceptual assessment−Voice quality−Twins.

INTRODUCTION

Issues with the perceptual evaluation of voice quality

A range of definitions for voice quality (henceforth VQ) exist in the specialized literature[1,2]. A recent study emphasizes the hybrid nature of VQ as the combination of long-term, quasipermanent laryngeal and supralaryngeal adjustments in a speaker's production, which are evaluated and classified by a listener through different perceptual processes.[3]

The auditory-perceptual assessment of VQ is necessary in different areas of Applied Linguistics. For example, voice therapists are trained in the application of one or more protocols that describe a patient's VQ and help monitor its changes.[4] In forensic applications of voice comparison, most experts place a great discriminatory potential on VQ, which is considered a key phonetic parameter for the characterization of a speaker's voice.[5,6] VQ is such an inextricable part of speakers' identity that it is not surprising that several 'general' synonyms exist for VQ, such as 'timbre' or the characteristic 'colouring' of a voice.[7,8]

The potential of VQ notwithstanding, its use is not devoid of practical problems. A brief summary of these issues follows:

Multidimensionality of VQ:

Multidimensionality is considered an important challenge in the perceptual assessment of complex stimuli, and human voices are such complex sounds. In studies focusing on how perceptual dimensions overlap and group[9], factor analyses have often been used for dimension reduction. For instance, the work by Isshiki et al.,[10] using factor analysis, gave rise to the GRBAS scale, developed by a committee within the Japanese Society of Logopedics and Phoniatrics.[11]

Voice labelling:

Listeners sometimes lack a common understanding of the labels used in perceptual protocols[12] or they may be biased toward using different verbal descriptors to rate the same voice characteristic. San Segundo et al.[3] give some examples of this phenomenon, which seems particularly frequent in auditory schemes with correlated dimensions (e.g., the Vocal Profile Analysis, henceforth VPA;[13]) and propose a two-stage calibration method to alleviate this kind of labelling issues.

Rating normophonic voices:

It is not clear whether normophonic and pathological voices can be assessed using the same rating system. Gelfer[14] (in Kent[12], page 11) noted that "perceptual rating systems designed for use with disordered speech or voice may not be effective in distinguishing among normal variants of speech and voice." This idea has supported recent research efforts toward simplifying perceptual schemes which were originally created for − or more commonly used in − clinical contexts.[15]

Accepted for publication October 11, 2019.

From the *Department of Criminal Science and Technology, Shanxi Police College, Taiyuan, Shanxi, China; and the †Institute of Phonetics, Faculty of Arts, Charles University, Prague, Czech Republic.

Address correspondence and reprint requests to No. 799, North-west Section, Qing Dong Road, Qingxu County, Taiyuan, Shanxi, China.

E-mail addresses: Eugenia@sxpc.edu.cn radek.skarnitzl@ff.cuni.cz

Journal of Voice, Vol. ■■, No. ■■, pp. ■■−■■

0892-1997

https://doi.org/10.1016/j.jvoice.2019.10.007

Nevertheless, it seems more sensible to think of pathological and normophonic speakers as overlapping groups.[16,17]

Holistic versus componential approaches:

Several studies[18,19] (see also San Segundo, Foulkes and Hughes[20]) suggest that the perception of VQ cannot be explained as the sum of separate features because VQ perception involves a significant component of holistic, gestalt-like pattern processing. However, the different perceptual assessment schemes (e.g., VPA, GRBAS, CAPE-V, SVEA; cf. summary in[21]) which are more commonly used rely on the description of a voice in terms of a variable number of settings or perceptual dimensions. They are thus componential approaches. It should be noted, however, that holistic processing is applied especially in listening to voices of familiar speakers,[19] a scenario which is less frequent in forensic or clinical contexts (see Nolan[22] for some forensic exceptions).

Inter- and intra-rater agreement:

Intercoder agreement in Computational Linguistics has been extensively debated, for instance in relation to corpus annotation tasks[23,24]. Likewise, the question of consensus and agreement has been a long-standing issue in the evaluation of VQ.[25] Similar conclusions have been drawn in the context of VQ assessment as in other fields of Applied Linguistics; namely, that reporting chance-corrected measures is preferable over simple percent agreement; and that increasing the number of annotators/raters is the best strategy to reduce personal biases.

However, these recommendations are not followed equally across all VQ protocols. For instance, there are scarce examples of the application of chance-corrected measures using the VPA scheme[3] and even fewer investigations using weighting techniques.[15]

Other issues:

The above-mentioned issues are long-standing problems in the assessment of VQ. The aim of this study is to highlight two further issues that have been debated to a lesser extent, if not completely neglected in some specific VQ protocols: (1) the nature of perceptual dimensions (discrete vs continuous) and (2) the existence of random errors linked to the use of paper protocols (see Research objectives).

Firstly, the quantification procedures of VQ protocols differ both in the number of degrees used to indicate the extent to which a feature is present and also in the type of scale used to measure different degrees. For instance, the GRBAS scale[11] is an ordinal scale with three grades (mild, moderate, and severe) while the VPA scheme has six grades to mark the non-neutrality of a setting. In contrast, the Consensus Auditory Perceptual Evaluation − Voice (CAPE-V) protocol[26] adopts a visual analog scale (VAS) with an asymmetric distribution of the mild, moderate, and severe degrees.

Some studies have compared VAS and ordinal scales for the evaluation of dysphonia.[27,28] In particular, Wuyts, De Bodt and Van de Heyning[27] compared two versions of the GRBAS scale: the original 4-point scale and a VAS, while Nemr et al.[28] compared the GRBAS scale with the CAPE-V scale. In the first study[27] the results showed that a VAS enables finer judgments of VQ but interrater agreement decreases considerably. The second investigation[28] showed the same reliability and consensus between both scales at least when applied to the same vocal sample at different times. Interestingly, studies such as Wuyts, De Bodt and Van de Heyning[27] show that it is possible to implement VAS for protocols that have traditionally employed equal appearing interval (EAI) or ordinal scales. To the best of our knowledge, no studies so far have reported the use of VAS for the VPA protocol.

All in all, the main research question remains unanswered: can different perceptual dimensions be best measured using a different scale resolution depending on the nature of each specific dimension? According to Stevens[29] (cf.[12]), there are two basic types of perceptual continua: prothetic and metathetic continua. While a *prothetic* dimension is described as an additive, quantitative continuum − the dimension varies in magnitude or quantity − a *metathetic* dimension, also described as a substitutive, qualitative continuum, would imply a change in quality.[12] As explained by Bettens et al[31], methatetic dimensions should be rated using EAI or ordinal scales, which present a finite number of categories, while prothetic dimensions would be better suited to VAS or direct magnitude estimation, which provide a measure in proportion to the magnitude of the dimension or stimulus. For instance, some studies have shown that hypernasality would be rather prothetic than metathetic[30] or that VAS ratings are a reliable and valid alternative for ordinal ratings in the perceptual assessment of hypernasality.[31] In contrast, Yiu and Ng[32] suggest that the psychoperceptual characteristics of breathy and rough qualities can be captured equally well by EAI and VAS. These results are in agreement with Sewall et al.[33] who found that the ratings of breathiness in normal speakers are metathetic. However, the nature of other perceptual attributes remains largely unknown.

Secondly, an aspect that has been considerably underresearched in VQ assessment is related to the random errors induced by the use of paper protocols versus computer-based protocols. Most sciences report two types of errors in experimental measurements: random errors and systematic errors. The former tend to occur because of environmental conditions while the latter usually come from the measuring instrument. This twofold nature of errors also applies to perceptual ratings.[34]

Random errors could be caused by the listener's lapses in attention or fatigue as well as by the characteristics of the testing situation.[35] For instance, VQ assessments tend to be performed by filling out paper protocols with ratings that are then transposed manually to other formats for further processing (e.g., statistics). These errors are clearly different

from systematic or criterion errors which usually correspond to biases in the rater. For example, one rater may tend to rate voices toward the higher end of the scale relative to another rater.[2] San Segundo et al.[36] noted the existence of a number of random errors when using a paper version of the VPA protocol. However, investigations aimed at exploring how to minimize this type of errors are rare.

Identical twins in voice studies

The interest sparked by twin pairs in voice studies lies in their extreme physical similarity due both to genetic and to environmental factors. There are two main types of twins: monozygotic (MZ) and dizygotic (DZ). The former − also called identical twins − develop from a single ovum, fertilized by a sperm cell and forming one zygote which is then divided into two separate embryos. It is commonly assumed that these twin pairs share 100% of their genes[37-39] even though nowadays it is more and more frequent to explain the possible differences between twins by highlighting the role of epigenetics, which would account for the alteration in the expression of specific genes caused by mechanisms other than changes in the underlying DNA sequence. As for the second type, DZ twins (i.e., non-identical, or fraternal twins), they develop from two separate eggs that are fertilized by two separate sperm cells. They share an average of 50% of their genetic information, although a more realistic percentage range for same-sex pairs seems to be 25−75.[40]

While some phonetic studies on twins have focused only on identical twins,[41-44] most voice investigations typically recruit both identical and non-identical twins,[45-48] usually with the aim of comparing vocal performance or certain speech patterns between the two types of twins. While such studies follow very heterogenous experimental designs, ideally the goal of comparing MZ and DZ twins is to "provide a useful indication of the relative contribution of genetic and environmental factors on individual differences in measured traits" (Haworth et al[49], page 1). Scientists of different disciplines refer to this as the 'nature-nurture dichotomy.'[50,51] In this context, the most common twin research design is called the 'classic twin method,' which compares reared-together pairs of MZ and DZ twins.

A few investigations have recently undertaken the joint investigation of MZ and DZ twins with nontwin siblings. For instance, San Segundo and Yang[48] showed that some nontwin brothers can be more similar than some pairs of MZ twins in terms of formant dynamics. It is further claimed in that study that the investigation of nontwin brothers and other type of related speakers should be encouraged, particulary in fields such as Forensic Voice Comparison (FVC). In FVC, phonetic knowledge is applied to solve legal issues arising out of police work,[52] such as comparing the voice recording of an offender with the voice of one or several suspects. On the one hand, it is not uncommon that members of a family collaborate together in crimes or offences which subsequently involve the analysis of their voices. On the other hand, nontwin brothers are easier to recruit than twins, due to the lower incidence of the latter, particularly MZ twins, as this kind of birth seems to occur at a low rate of 3.5−4 per 1000 births, a rate that is relatively constant worldwide.[53]

In this investigation, two pairs of MZ twins were included in the set of voices that the raters were asked to listen to and rate in terms of their VQ (see Materials and Methods). The reason for selecting at least two MZ pairs was to compare the VQ ratings given to each member of the twin pair and explore the usefulness of our rating tool in such cases of strong voice similarity. As explained before, MZ twins represent extreme examples of similarity, expected to apply both in terms of vocal tract anatomy and in relation to phonatory dynamics. Both aspects have a bearing on the VQ of a speaker. As raters did not know beforehand that the listening experiment included some pairs of twins, we aimed to observe if the intrapair results were very similar through different data visualization techniques and then draw some conclusions about the relevance of such visual exploration for general voice studies.

RESEARCH OBJECTIVES

This study aims to create a new tool for the perceptual assessment of VQ. For its design, we have adapted the Simplified Vocal Profile Analysis (henceforth SVPA), described in detail in San Segundo and Mompeán.[15] The new tool features two main characteristics:

(1) the ordinal scalar degrees of the original protocol are turned into a visual analog scale.
(2) the original paper-based version of the protocol is implemented in a web-based environment.

As a first attempt to evaluate the reliability of the new tool, we aim to provide some measures of inter-rater agreement, reached by five phoneticians who listened to 12 different speakers and evaluated their VQ using the proposed tool.

A second research objective is to explore whether the tool can be useful in contexts where two voices are to be compared, for instance in FVC applications (see Identical twins in voice studies). With this aim, several data visualization techniques are proposed which allow to compare pairs of speakers in terms of their VQ components. The focus of the speaker similarity analysis is placed on the two pairs of MZ twins selected for this study (see Subjects).

MATERIALS AND METHODS

Subjects

Twelve male speakers were selected from the corpus collected by San Segundo.[54] All were native speakers of Standard Peninsular Spanish and none reported any voice pathology. Their age ranged between 18 and 36 years (mean: 26.67).

Speakers were recorded with an omnidirectional condenser microphone with flat-frequency response (20 Hz to 20 kHz), a sensitivity of 2.0 mV/Pascal, Equivalent Acoustic Noise 29 dBA SPL and Overload Sound Level 130 dB SPL. The microphone was connected to a soundcard (*Cakewalk by Roland UA-25EX USB AudioCapture*) with the following specifications selected for the recording: 44 100 Hz sample rate, 16 bits resolution, and mono channel.

As introduced in the section Identical twins in voice studies, there were two MZ twin pairs among the 12 speakers participating in this investigation. Twin Pair 1 was made up of Speaker #2 and Speaker #8; Twin Pair 2 was made up of Speaker #9 and Speaker #12. For the sake of simplification, they have been renamed 'Speaker A' and 'Speaker B' (Twin Pair 1), and 'Speaker C' and 'Speaker D' (Twin Pair 2). Both pairs had been raised together (lived in the same house and went together to the same primary and secondary school). At the time of the recording, Speakers A&B were 28 years old and had been living together for 27.5 years. Speakers C&D were 33 years old and had been living together for 30 years.

As regards the original nomenclature given to the participants, the speakers' numbers corresponded to the order in which they appeared in the listening and rating test. Since the speakers were selected from a larger database, the way that we chose to randomize their order in the test was to follow the alphabetic order of their first names. For example, the names of the brothers in Twin Pair 1 begin with A and with C, respectively, which makes them occupy places #2 and #8 in the list of subjects.

Stimuli and listeners

One voice sample (90−120 seconds) was extracted from semi-directed spontaneous conversations for each of the 12 subjects described above. These make up the 12 stimuli that the raters had to listen to and then evaluate in terms of VQ aspects. All the raters listened to the stimuli in the same order. The perceptual and assessment process took place through a computer interface (see Procedures) embedded in the first author's website: https://eugeniasansegundo.github.io/vas/

Five listeners participated in the perceptual experiment as raters. They were all native speakers of Czech with their knowledge of Spanish ranging from zero to lower intermediate. On the one hand, using native speakers of another language with only basic knowledge of Spanish was a way of ensuring that the listeners focused exclusively on VQ when evaluating the voices. On the other hand, however, insufficient knowledge of the target language makes the task more challenging, as VQ is to a certain extent language specific[55]. They were all phoneticians (two senior academics and three doctoral students). None of the raters reported any hearing difficulty. They used headphones during the test and had the opportunity to listen to each stimulus several times if necessary. Prior to the execution of the perceptual test, which took place in a silent room at the Institute of Phonetics in Prague, the five raters had received a two-day training session in VQ evaluation, with an emphasis on the SVPA.[15] The training consisted in the gradual introduction of different VQ settings, from those which are easier to conceptualize (e.g., nasality or phonation type) to those which tend to be slightly less accessible to raters (e.g., pharyngeal expansion). The VQ training featured speakers of several languages, most notably English, Czech and Spanish.

Procedures

Computer-based protocol

The protocol proposed here (VAS-SVPA henceforth) draws on the SVPA designed by San Segundo and Mompeán[15] but adds a number of improvements. The SVPA basically transformed the VPA scheme[13,56] from a high-dimensional scheme (e.g., thirty-two settings in the version used by[3]) to a simplified one, limited to 10 VQ settings and only three rating categories: one for the 'neutral' configuration and two for opposite 'non-neutral' configurations. For instance, for the labial configuration, a rater has to choose whether a certain speaker presents a neutral configuration or not. If not, the subsequent decision concerns the direction of the deviation from neutrality: lip spreading or lip rounding. Three main decisions made it possible to distill the original dimensions to just 10 settings: (1) only one setting is distinguished for 'phonation type', with the deviation from neutrality corresponding to harsh or creak(y) and breathy or whisper(y); (2) three settings which rarely occur in normophonic speakers were removed; and (3) the remaining settings were sorted in opposite pairs, as listed below (see San Segundo and Mompeán[15] for more detailed information).

These three measures reduce the number of perceptual decisions taken by the rater while the resulting simplified protocol still allows for a detailed description of typical articulatory configurations. Therefore, both the SVPA and the VAS-SVPA comprise the following dimensions or settings: (1) *phonation type*, with the opposite non-neutral configurations corresponding to whisper or breathy versus creaky or harsh; (2) *larynx tension*, with the non-neutral configurations being lax and tense; (3) *vocal tract tension*, with the non-neutral configurations being lax and tense; (4) *larynx height*, with lowered and raised as non-neutral configurations; (5) *pharynx expansion*, with constricted versus expanded configurations; (6) *velopharynx* or *nasality*, with denasal versus nasal configurations; (7) *tongue body fronting*, with backed and lowered versus fronted and raised positions; (8) *tongue tip fronting*, with retracted versus advanced positions (9) *mandible openness* with close and open configurations; and (10) *labial protrusion*, with the opposite non-neutral configurations corresponding to lip spreading versus lip rounding.

Here follows a summary of the new features of the computer-based SVPA protocol:

Ratings along a VAS:

Raters no longer need to make a hard decision on categories (e.g., 'slight', 'marked', and 'extreme') for each VQ

setting. Instead, a line with a slider is provided so that they can place it at any point along a continuum to indicate the extent to which each dimension is perceived in a particular voice. Positions along the VAS were subsequently converted into values between 0 and 100.

Bottom-up design:

Beck[56] suggests that VPA protocols should be filled out following anatomical progression down the vocal tract from the lips to the larynx. The VQ settings in the VPA template are displayed following that order. San Segundo and Mompeán[15] proposed marking first what is more remarkable for the rater and then trying to decide on the rest of settings. This seems to reflect better how our brain works: perceiving first the most salient VQ aspects in a speaker and not necessarily labial aspects first. Going one step further, here we have arranged the 10 VQ settings following anatomical progression up the vocal tract from the larynx. The reason for this is that laryngeal aspects (e.g. phonation type) continue to be what most experts associate with VQ[1] (cf. narrow definition of VQ[57]) and what they are trained to perceive first.

Left-right order of dimension extremes:

Pairwise VQ labels occupy the left or right extremes of the VAS line following an intuitive semantic logic. Labels placed on the right signify: 'more' (e.g., tension, intensity, or adduction), or 'increased, advanced', 'bigger', or 'more open'. Since the slider is placed by default in the midpoint of the scale and the rater has to move it rightwards or leftwards, movements to the right would be associated iconically with the (+) symbol. Likewise, labels placed on the left of the VAS are the opposite labels of their right counterparts, associated with (−) aspects: 'less' tension or adduction, 'downward' movements, 'smaller' cavities (e.g. constricted pharynx vs expanded pharynx), 'closure' or 'back and low' aspects.

Computer-based protocol:

Last but not least, a key feature of the proposed protocol is that it allows for online implementation. Getting access to the experiment platform through a link, the raters can both (1) listen to each voice by simply clicking the 'play' symbol and 'next' when they want to proceed to the next speaker; and (2) rate the VQ of the speakers by sliding the cursor along a VAS.

Figure 1 shows the Graphical User Interface of the tool designed ad hoc for the perceptual evaluation of VQ in this experiment. The 12 voices that raters had to listen to and evaluate were already included in the online tool. Since one of the aims of this study is to make this tool available for voice scientists, this Graphical User Interface has changed slightly to enable future users to upload their own sound files and be able to save the results after rating.[2]

Statistical analysis

Reliability means the extent to which measurements can be replicated.[58] As stated by Koo and Li,[59] a good reliability measure should reflect both degree of correlation and degree of agreement. The Intraclass correlation coefficient (ICC) meets these criteria[59,60] and has been widely used in voice studies to evaluate interrater reliability.[61,62] Reliability value ranges between 0 and 1, with values closer to 1 representing stronger reliability.

In order to calculate the ICC index, we used *MedCalc* statistical software (v. 18.10.2). Results are based on a mean-rating ($k = 5$), consistency, 2-way mixed-effects model: ICC (3, 5).

RESULTS

Inter-rater consistency

The results of the statistical test show ICC values between 0.448 and 0.836, with the notable exception of labial protrusion (ICC = −0.156) indicating poor to good reliability, depending on the specific VQ setting under consideration.

As shown in Table 1, the best results are obtained when assessing phonatory and laryngeal aspects (ICC = 0.836 for larynx tension and 0.786 for phonation type), followed by the velopharyngeal (nasality) dimension (ICC = 0.764), as well as the assessment of laryngeal height (ICC = 0.676) and pharyngeal expansion (ICC = 0.660).

The settings showing lower consistency among this group of raters relate to vocal tract aspects, from vocal tract tension (ICC = 0.588) to the remaining settings forward from the velopharynx. Moderate agreement is achieved in terms of the degree to which the jaw is open/close (ICC = 0.552) although the lower limit of the 95% CI is very low for all the last four settings: mandible openness, tongue body fronting, tongue tip fronting, and labial protrusion; the latter with a particularly poor consistency (ICC = −0.156).

Appendix 1 shows the different patterns of the distribution of ratings. On the one hand, we can observe that the ratings differ notably from one VQ setting to another. Basically, there are two main groups of settings: (1) those with ratings dispersed along the VAS scale and where most points of the scale have been used (e.g. min. ratings are between 10 and 20 and max. ratings between 80 and 90), and (2) those with most ratings concentrated around the midpoint of the scale and where extreme values have been seldom assigned. Interestingly, among the settings belonging to the first group we find phonation type, larynx tension,

[1]As Kent[12] explains, a laryngeal aspect (*hoarseness*), together with *nasality,* were the only VQ aspects repeated across multiple studies: "Perkins (1971) identified 27 different terms that were used in nine studies of abnormal voice quality. Of the total 27 terms, only two (*hoarse* and *nasal*) were used in all nine studies".

[2]Please refer to the following URL to access the VAS tool (v.1.0 as of November 2019) with detailed instructions about the uploading of files as well as conditions and user terms (licence type): https://eugeniasansegundo.github.io/vas/tool.html

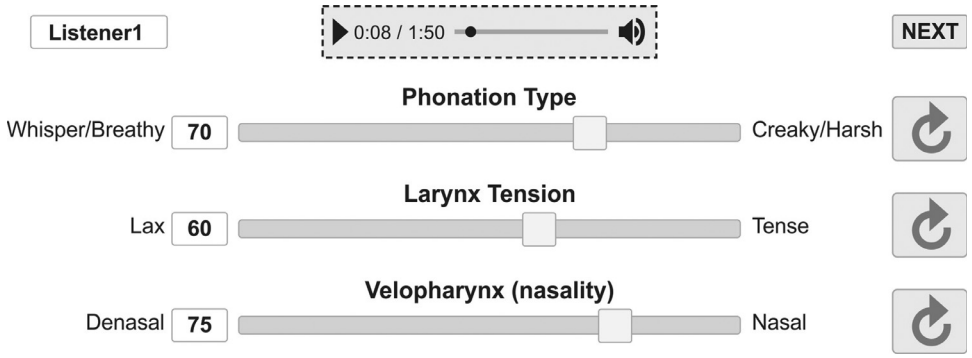


FIGURE 1. Graphical User Interface of the VAS tool, showing the following information (left to right and top to bottom): text box where the listener can type her name, audio player with duration information, play and volume buttons; rightmost: 'next' button to proceed with the test. For each of the three example settings showed in this figure, we can see: label of the left extreme of the VAS (e.g., whisper/breathy), the assigned value along the VAS corresponding to the slider's position and the label of the right extreme of the VAS (e.g., creaky/harsh). Rightmost, the 'reset' button can be used to return to the default value '50'.

larynx height, and pharynx expansion. Rating consistency for these settings was found to be relatively good (ICC above 0.660). In contrast, among the settings of the second group we find labial protrusion, tongue tip fronting, and tongue body fronting, with poor consistency among raters (ICC below 0.458).

A correlation test showed that the percentage of neutral ratings correlate negatively ($r = -0.843$) with the ICC value. For this calculation we considered the sum of the neutral ratings across raters per setting. This means that the more ratings are assigned to the value '50' in a setting (i.e., the neutral value on the 0−100 scale), the lower its ICC.

On the other hand, the distribution of ratings also seems to depend somehow on the specific rater. For example, Rater 3 shows a bias towards discretizing the continuous scale, as he only uses multiples of five to rate voices. Raters 1 and 5 are characterized by using the whole range of the scale values quite often. This is more evident in some settings such as larynx height (Appendix 1d) or pharynx expansion (Appendix 1e). It is less evident in the second-group settings (Appendix 1g-j) where, as we have said, most raters agree on assigning the value '50' to a good number of speakers.

TABLE 1.
Intraclass Correlation Coefficient (ICC) and 95% Confidence Interval (CI) for the 10 VQ Settings

VQ setting	ICC	95% CI (Lower Limit to Upper Limit)
Phonation type	0.786	0.507 to 0.930
Larynx tension	0.836	0.624 to 0.946
Vocal tract tension	0.588	0.052 to 0.865
Larynx height	0.676	0.255 to 0.894
Pharynx expansion	0.660	0.218 to 0.888
Velopharynx (nasality)	0.764	0.457 to 0.923
Tongue body fronting	0.448	−0.270 to 0.819
Tongue tip fronting	0.458	−0.247 to 0.822
Mandible openness	0.552	−0.031 to 0.853
Labial protrusion	[3]−0.156	−1.661 to 0.620

[3] According to Nunnally and Bernstein,[63] negative ICC values occur when the between-subject variation is relatively small compared to the within-subject variation, e.g., due to different raters. If that is the case the negative ICC estimate should not be quoted but one can say that the scale is not reliable.

Speaker similarity

A second research objective of this investigation was to explore whether the proposed online tool for the perceptual evaluation of VQ can be useful in contexts where the expert needs to perform pairwise comparisons of the VQ of two speakers or of the same speaker in two different recording sessions.

For this purpose, we implemented two data visualization techniques. The first one is a variant of the multiple dot plots (Appendix 1) that were described in the previous section. On this occasion the dot plot (Figure 2) shows the mean ratings (for all the five raters) per VQ setting and only for four speakers; that is, the two twin pairs. On the one hand, the figure reveals an important number of VQ configurations that are shared by each pair of twins, as they are members of the same speech community. Since the rating '50' is the neutral value for each setting, we have given a ±10 interval to what raters may have considered neutral (see yellow dashed lines). This makes sense given that the dot values in Figure 2 are the average ratings of the five listeners. All the speakers, therefore, seem to have a neutral VQ in terms of the following settings: vocal tract tension, tongue body fronting, tongue tip fronting, mandible openness, and labial protrusion. The figures in Appendix A already showed that the ratings given to this group of settings gathered around '50' quite homogenously for this group of raters when considering all 12 speakers.

On the other hand, Figure 2 allows us to easily detect the main VQ trends in terms of intrapair twin characterization. Twins A&B are close to each other in settings #1, #2, and #6. This means that they are both characterized by creak/harsh phonation, tense larynx, and nasality. In contrast, twins C&D are jointly characterized by whisper/breathy phonation (setting #1) and raised larynx (setting #4). For

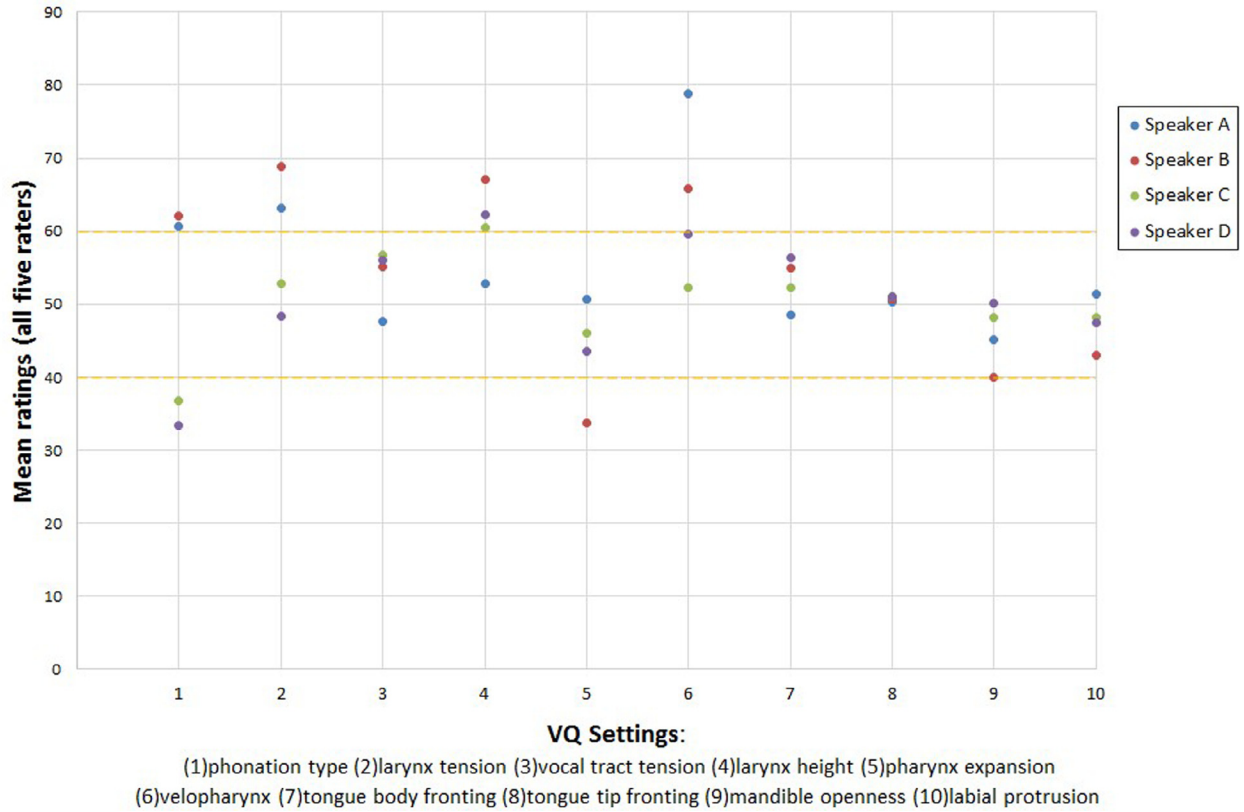


FIGURE 2. Multiple dot plot for Twin Pair 1 (Speakers A&B) and Twin Pair 2 (Speakers C&D). The points represent the average ratings given by the five raters for each voice quality (VQ) setting.

all other settings, they seem to be quite neutral. Nevertheless, it is worth noting that the average value that Speaker B gets for setting #4 is markedly high and this is precisely one of the aspects of his VQ in which he differs from his brother, together with pharynx expansion (setting #5). All in all, this kind of plot allows the voice scientist to detect in a straightforward way the sources of similarity and dissimilarity in a reduced number of subjects.

The second type of graphical representation that we propose is a variant of the Bland-Altman plot.[64-66] Originally, this type of graphical method, also called the difference plot, is used to compare two measurement techniques. The differences between them are plotted against their averages. The goal of this type of graphical representation is to reveal possible relationships between differences and averages, to look for potential systematic bias and to identify outliers. Other possible uses include evaluating the repeatability of a method by comparing repeated measurements using one single method on a series of subjects.[67]

Here we propose another use of the Bland-Altman plot, still aimed at plotting averages against differences but in this case between pairs of speakers instead of techniques. All the characteristics of this graphical representation remain the same as in the original plot. Horizontal lines are drawn at the mean difference and at the limits of agreement. The latter are defined as the mean difference plus and minus 1.96 times the standard deviation of the differences. For voice research purposes, this plot is useful to compare the differences between the ratings given to two particular voices by *n* number of raters. Because it is possible to distinguish (with different symbols and colours) among the 10 different VQ settings in our protocol, this use of the Bland-Altman plot also serves to find out if there are recurrent patterns in different settings or setting groups. In comparison with the multiple dot plot shown in Figure 2, the Bland-Altman plots (Figures 3 and 4) show the ratings provided by each of the five raters – and not just the average – which opens up the possibility to detect potential biases in a particular rater.

Figure 3 shows different setting trends in terms of the distribution of the ratings. For instance, pink squares are found towards the upper left corner of the plot. There are five, one per rater, and represent 'pharynx expansion'. The lower the mean, the lower both speakers ranked in terms of pharynx expansion (i.e., the more constricted their pharynxes). The average values (means of Twin A&B) for this setting are below '60' with a difference of less than 30 (i.e., between the rating given to Twin A with respect to Twin B). Because most of these points are above 0 in the y-axis we know that raters agreed that the pharynx was more expanded for A than for B, or what is the same, it was more constricted for B than for A. There was just one rater who gave a comparatively high rating to A in comparison with B and that can be considered an outlier because it goes beyond the upper limit of agreement. Correspondingly, all raters agreed that B had a more raised larynx than A, as all the values for this setting (green triangles) are below 0 on the

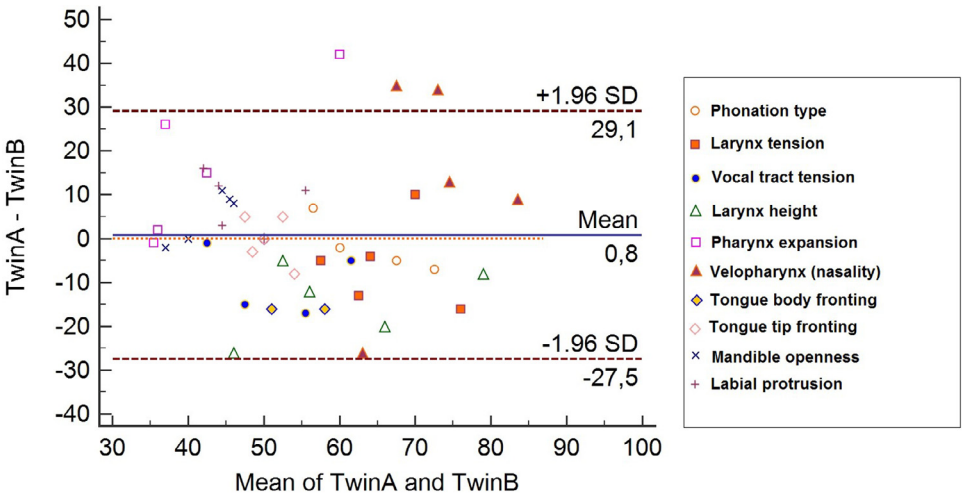


FIGURE 3. Bland-Altman plot showing the difference in ratings between Twin A and Twin B against their means. For each of the 10 VQ settings (legend on the right), each point represents a rater.

y-axis. Previous studies[3] have observed that constrincted larynx occurs together with raised larynx position in many speakers. The dot plot in Figure 2 already showed this VQ difference in this twin pair but in this new plot (Figure 3) we can observe in more detail whether a certain rater was particularly biased to rate a twin's voice comparatively higher than the other raters. In such cases, the analyst can decide whether to take into account that particular rating for the calculation of a mean value or whether to consider it an outlier and discard it.

Clearly, this kind of plots are more interesting when the analyst wants to detect VQ differences rather than similarities between pairs of speakers, and particularly to detect subtle biases in raters. As a case in point, Twins C&D were perceived to be very similar (cf. Figure 2), with ratings gathering around the neutral '50' for most of their settings. It does not really matter whether one is rated ±10 points above or below '50' for a particular setting, since on a 0–100 continuous scale, a VQ setting of 40–60 will not be considered particularly 'marked' or 'extreme' (to mention some of the labels typically used in ordinal scales). Notably in Figure 4, two values (one orange circle and one pink square) in the upper left corner seem to be beyond the upper limit of agreement. Interestingly, both correspond to ratings provided by Rater 5. This rater seems particularly prone to perceive VQ differences strikingly higher for one speaker than for the other. Orange circles are used for phonation type and pink squares for pharynx expansion. Bearing in mind that all the other points for those settings are quite closely clustered together – and taking into account again that it is a case of particularly similar-sounding speakers – at least these results should make the researcher wonder whether all the raters are equally reliable.

DISCUSSION

The results of this investigation have shown that it is possible to achieve moderate to good agreement in the perceptual

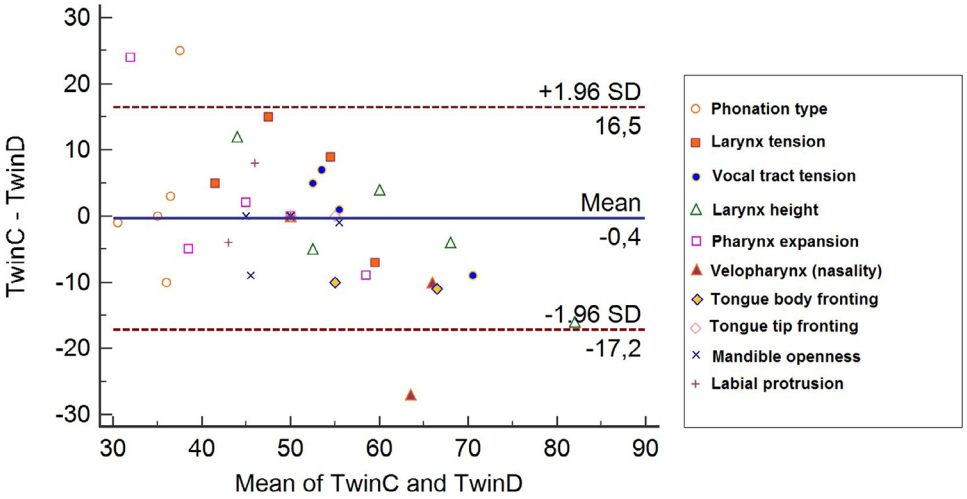


FIGURE 4. Bland-Altman plot showing the difference in ratings between Twin C and Twin D against their means. For each of the 10 VQ settings (legend on the right), each point represents a rater.

assessment of most of the settings of the SVPA when using the proposed protocol (VAS-SVPA). Seven out of 10 settings obtain an ICC of 0.5 or higher. These results are particularly good taking into account the short training received by the raters and the fact that they were not native speakers of Spanish.

The results have also revealed that, for the current raters, the scale used was not reliable for at least one of the settings: labial protrusion. Moreover, agreement was especially poor in the settings referring to the configuration of the tongue body and tip. This may be due to the strong dependence between the long-term configuration of these settings and the key segments (i.e., those segments that are most susceptible to the effects of a setting[56]) of the language considered. For instance, Standard Peninsular Spanish has been described as presenting considerable apical activity in terms of the basis of articulation.[68] It is therefore not surprising that non-native raters with a varying command of Spanish (cf. Stimuli and Listeners), are not consistent in their assessment of whether the speaker is neutral or non-neutral in these settings. This may be particularly relevant in the case of labial protrusion but could also explain lingual results, as [s] is a key segment for those settings too.[15] As for labial protrusion in particular, the strident apico-alveolar [s] of some varieties of Spanish has a lower centre of gravity of the noise[69] than the Czech lamino-alveolar [s], which results – for Czech listeners at least – in a clear impression of labialized pronunciation. It is conceivable, then, that this characteristic feature of the "sound of Spanish" will be assessed differently by L2 raters depending on their knowledge of Spanish: the speakers may sound more labialized to those who are not familiar with Spanish and more neutral to those whose command of Spanish is higher. This perceptual process may be regarded as equivalent to compensation for coarticulation[70,71] whereby, for example, a nasalized vowel is not heard as nasalized in a word like *seam* where coarticulatory nasalization is expected but it would be clearly audible in a word like *seed*.

One of the conclusions to be drawn from this study thus concerns the fact that listeners are capable of assessing VQ in a language they have little experience with (cf. also[72]). However, caution is necessary when interpreting the results and comparing them with other studies: there is some support to the claim that L1 and L2 listeners' rating differs to some extent.[73] While a native advantage has been suggested by some researchers[74,75] for tasks such as speaker identification, other investigations evaluating the performance of different groups of listeners in a task asking them to rate speaker similarity found that all listeners judged speaker similarity in a comparable way irrespective of their L1.[19] This was thought to be due to shared strategies when evaluating a speaker's VQ based on holistic approaches. Reaction times, however, can differ between L1 and L2 listeners in this kind of experiments.[76] This is a variable which should be taken into account in future studies.

It has been mentioned before that our investigation shows poor inter-rater agreement for labial protrusion, tongue body, and tongue tip. A possible explanation in relation to the raters' L1 has just been put forward. A further question arises here as to whether this low agreement could be due to the metathetic rather than prothetic nature of those particular settings, which would make the use of VAS inefficient for their measurement. This is particularly noticeable if we contrast the dot plots of labial protrusion, tongue body and tongue tip with the rest of the settings (see Appendix). The low dispersion of the points in the plot for the former settings suggests that the raters can only notice whether the setting is present or absent, and in which direction it is present. The ordinal scale of the SVPA[15] seems to be better suited for such a rating pattern, as it provides just three rating options or categories. In contrast to labial protrusion, tongue tip and tongue body, aspects related to the activity of the vocal folds, together with 'nasality', tend to obtain higher agreement from our raters (see Table 1). On the one hand, this points to the ease of perception of these dimensions, regardless of the mother tongue of the listener[4]. Those settings in particular also obtained good agreement in previous studies with Spanish native raters.[15] On the other hand, the dispersion of the points in their corresponding plots (see Appendix) suggests that a VAS may be suitable to measure them. In other words, raters tend to make use of most of the scale values in order to rate speakers on those dimensions and do not confine their ratings to just three categories, as explained before. Nevertheless, we would need specific tests[78] to give a clear answer to the question of whether 'phonation type' and 'nasality' are prothetic rather than metathetic continua. This would undoubtedly imply a direct comparison between EAI or ordinal scales and VAS or direct magnitude estimation.

The correlation analyses have shown that the more ratings assigned to the value '50', the lower the ICC is for that setting. This affects many settings for which we can consider that the deviation from the neutral configuration is rare. Indeed, the neutral configuration for the labial, apical and dorsal settings was found by San Segundo and Mompeán[79] to have a high occurrence in the same population that is examined in this study: 67%, 67%, and 71% of occurrence, respectively. As directions for future research, more investigations seem necessary into how to deal with this type of settings, as either the VAS scale seems to be unfit for measuring them perceptually or the ICC index might not be appropriate for calculating inter-rater agreement. Weighting alternatives could be explored in upcoming studies.

Finally, we have undertaken a preliminary exploration into new visualization techniques for plotting similarities and differences between speaker pairs in terms of their whole set of VQ components. While the multiple dot plot in Figure 2 was useful to reveal those VQ settings of the VAS-SVPA in which speakers are closer to or further away from each other, the Bland-Altman plots of Figures 3 and 4 added some extra

[4]It would be interesting, however, to have speakers of, for instance, American English assessed for VQ by raters who are not familiar with the extent to which nasality is an integral part of this variety's setting.[77]

information about the rating behaviour of individual raters. The potential of both visualization techniques remains to be fully explored, for example with speakers different than MZ twins. The main uses that these plots have shown so far are: detection of subtle differences between speakers depending on the setting under consideration, and detection of possible biases of a particular rater. The fact that the VQ protocol used in this investigation is based on a simplified protocol makes the Bland-Altman plot particularly apt for visual representation, since only 10 different combinations of colors and symbols need to be used in order to characterize the 10 settings making up the VQ of a speaker. In this respect, it is worth mentioning that the SVPA[15] reduced the number of settings in the original protocol (ranging from 30 to 40 settings, depending on the version) into only 10 main setting groups.

CONCLUSIONS

VAS is devised to capture the idea of an underlying continuum. To the best of our knowledge, this is the first study that explores the use of a VAS with the VQ settings of the SVPA protocol. This protocol for the perceptual assessment of VQ[15] was designed with the aim of simplifying a rather complex scheme[13,54] and enabling easier pairwise comparisons, for instance in the context of FVC. With such forensic purposes, the SVPA has inspired numerous studies so far.[44,80,81] For example, in their adaptation of the protocol to Chinese, Jintao et al.[80] also reduced the total number of VPA settings and the range of the scalar degrees. In Passetti and Constantini,[81] the authors used the measurement procedure proposed in San Segundo and Mompeán[15] to compute similarity between voice samples. However, the dual or binary nature of the SVPA could be considered too simplistic depending on the particular linguistic case or the specific purpose of use. Consequently, we have devised the new protocol presented in this paper, which combines some of the characteristics already present in the SVPA[15] with the novel idea of allowing listeners to rate each VQ setting along a VAS in a computer-based interface.

Although the purpose of this investigation has not been on this occasion to compare the results of VQ assessment using the SVPA and using the VAS-SVPA, the current study has shed some light into which perceptual dimensions could be assessed using a VAS with an acceptable inter-rater reliability and which ones might present more problems, at least in the current state of knowledge about the nature of VQ dimensions. What seems clear so far is that the SVPA (even if it is already a simplified version of the original VPA) is a heterogenous protocol comprising multiple dimensions which deserve much more detailed psychoacoustic investigations than they have received so far. Alternative procedures for multidimensional evaluation are worth exploring for normophonic speakers, in the same way that previous investigations started suggesting that the rating of dysarthric speech require a combination of prothetic and metathetic scales.[82]

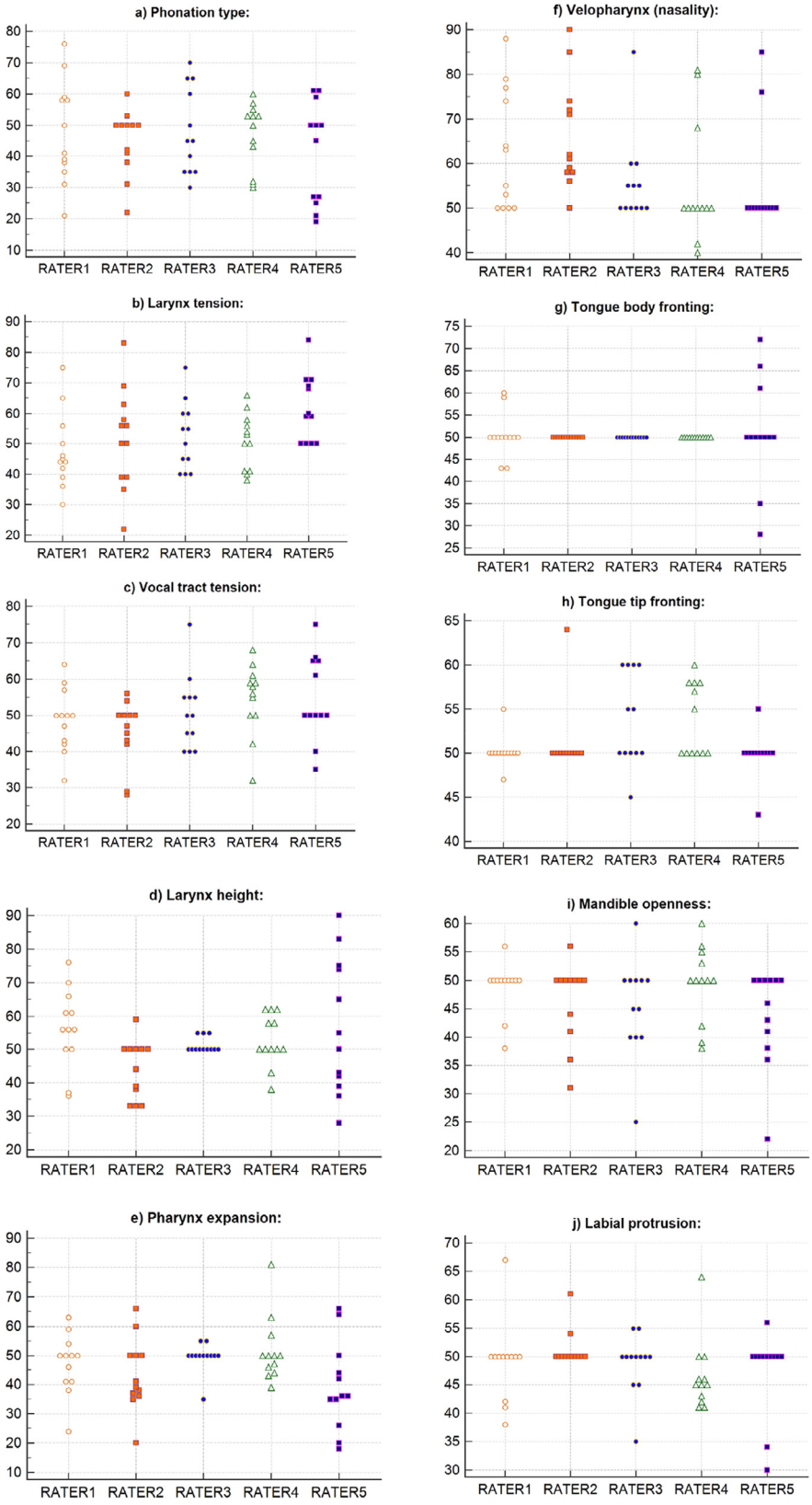
To conclude, this study has proposed a computer-based protocol as a new tool publicly available at https://eugeniasansegundo.github.io/vas/tool.html for researchers and practitioners of different fields of Applied Phonetics. In a preliminary study on the nature of raters' disagreements,[36] it was found that besides proper disagreements due to the raters' biases, lack of training etc., there were also labelling reassignments (i.e., disagreements that could be solved in a calibration meeting held by the raters), but more importantly, there were also data entry errors. Little research has been made into the impact of this type of random errors. We trust that this new tool encourages researchers to compare their results derived from paper-version and computer-version protocols in the hope that – if the latter outperforms the former – they favour the computer-based tools so that they keep random errors at bay and can concentrate efforts on the analysis and improvement of criterion errors.

As far as the use of visualization techniques concerns, we are convinced that the development of new types of graphical representations for perceptual analyses should be encouraged in future studies. Without neglecting the importance of proper analyses with adequate statistical evaluations, visual representations of perceptual data, such as VQ ratings, seem to lag behind the numerous ways in which acoustic data of all sorts can be presented (e.g., F1/F2 plots or Long-term Average Spectrum plots). From a clinical perspective, visual representation of data can be very useful for both researchers and practitioners in nearly all possible applications of perceptual assessment that were highlighted by Carding et al.,[83] from monitoring patients' evolution of VQ (e.g. mapping their VQ ratings in the first session against their second and following sessions); to obtaining a comprehensive overview of the characteristics of the patient's voice, or enabling exchange of opinions with other professionals based on the visual representations. In forensic terms, this need for visual techniques applies as well. For example, in forensic reports the expert may need to present VQ analyses of the known and unknown speakers to the trier of fact. Given the componential nature of VQ and the importance of undertaking analyses by more than one person[84,85] (cf.[3]), plots which can capture the different dimensions of VQ and which, at the same time, provide information about several raters should be particularly welcome.

ACKNOWLEDGMENTS

The study was supported by the European Regional Development Fund-Project "Creativity and Adaptability as Conditions of the Success of Europe in an Interrelated World" (No. CZ.02.1.01/0.0/0.0/16_019/0000734).

APPENDIX: DOT PLOTS SHOWING THE RATING DISTRIBUTION (RANGING FROM 0 TO 100) FOR THE TEN VQ SETTINGS BY THE FIVE RATERS



REFERENCES

1. Laver J. Phonetic evaluation of voice quality. In: Kent RD, Ball MJ, eds. *Voice Quality Measurement*. San Diego, CA: Singular Publications; 2000:37–48.
2. Kreiman J, Gerratt BR. Perceptual assessment of voice quality: Past, present, and future. *Perspect Voice Voice Disord*. 2010;20:62–67. https://doi.org/10.1044/vvd20.2.62.
3. San Segundo E, Foulkes P, French P, Harrison P, Hughes V, Kavanagh C. The use of the Vocal Profile Analysis for speaker characterization: Methodological proposals. *J Int Phon Assoc*. 2018:1–28. https://doi.org/10.1017/S0025100318000130.
4. Dejonckere PH, Remacle M, Fresnel-Elbaz E, Woisard V, Crevier L, Millet B. Reliability and clinical relevance of perceptual evaluation of pathological voices. *Rev Laryngol Otol Rhinol*. 1998;119:247–248.
5. Gold E, French P. International practices in forensic speaker comparison. *Int J Speech Lang Law*. 2011;18:293–307. https://doi.org/10.1558/ijsll.v18i2.293.
6. San Segundo E, International survey on voice quality practices: forensic practitioners versus voice therapists (submitted).
7. Laver J. *Individual Features in Voice Quality*. Edinburgh: University of Edinburgh; 1975.
8. Trask RL. *A Dictionary of Phonetics and Phonology*. London: Routledge; 1996.
9. Bele IV. Dimensionality in voice quality. *J. Voice*. 2007;21:257–272. https://doi.org/10.1016/j.jvoice.2005.12.001.
10. Isshiki N, Okamura H, Tanabe M, Morimoto M. Differential diagnosis of hoarseness. *Folia Phoniat*. 1969;21:9–19. https://doi.org/10.1159/000263230.
11. Hirano M. *Clinical Examination of Voice*. New York: Springer Verlag; 1981.
12. Kent RD. Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *Am J Speech-Lang Pathol*. 1996;5:7–23. https://doi.org/10.1044/1058-0360.0503.07.
13. Laver J. *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press; 1980.
14. Gelfer MP. Perceptual attributes of voice: development and use of rating scales. *J Voice*. 1988;2:320–326. https://doi.org/10.1016/S0892-1997(88)80024-9.
15. San Segundo E, Mompeán JA. A simplified vocal profile analysis protocol for the assessment of voice quality and speaker similarity. *J Voice*. 2017;31:644.e11–644.e27. https://doi.org/10.1016/j.jvoice.2017.01.005.
16. Gómez Vilda P, San Segundo E, Mazaira Fernández LM, Álvarez Marquina A, Rodellar Biarge M. Using dysphonic voice to characterize speaker's biometry. *Lang Law / Linguagem e Direito*. 2014;1:42–66.
17. Delvaux V, Pillot-Loiseau C, Perceptual judgment of voice quality in non-dysphonic French speakers: effect of task-, speaker- and listener-related variables, J Voice (in press). https://doi.org/10.1016/j.jvoice.2019.02.013.
18. Kreiman J, Gerrat B. Comparing two methods for reducing variability in voice quality measurements. *J Speech Lang Hear Res*. 2011;54:803–812. https://doi.org/10.1044/1092-4388(2010/10-0083).
19. Kreiman J, Sidtis D. *Foundations of Voice Studies: Interdisciplinary Approaches to Voice Production and Perception*. Boston: Wiley-Blackwell; 2011.
20. San Segundo E, Foulkes P, Hughes V. Holistic perception of voice quality matters more than L1 when judging speaker similarity in short stimuli. In: *Proc. 16th Australas. Int. Conf. Speech Sci. Technolog*. 2016.
21. Gil 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.
22. Nolan F. Forensic speaker identification and the phonetic description of voice quality. In: Hardcastle WJ, Mackenzie Beck J, eds. *A Figure of Speech: A Festschrift for John Laver*. Mahwah, NJ: Lawrence Erlbaum; 2005:385–411.
23. Carletta J. Assessing agreement on classification task: the kappa statistic. *Comput Ling*. 1996;22:249–254.
24. Arstein R, Poesio M. Inter-coder agreement for computational linguistics. *Comput Ling*. 2008;34:555–596. https://doi.org/10.1162/coli.07-034-R2.
25. Xie Z, Gadepalli C, Jalalinajafabadi F, Cheetham BM, Homer JJ. Measurement of rater consistency and its application in voice quality assessments. In: *Proc 10th Int Cong Image Signal Proc, BioMed Eng Inform (CISP-BMEI)*. 20171–6.
26. Kempster GB, Gerratt BR, Verdolini Abbott K, Barkmeier-Kraemer J, Hillman RE. Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *Am J Speech-Lang Pathol*. 2009;18:124–132. https://doi.org/10.1044/1058-0360(2008/08-0017).
27. Wuyts FL, De Bodt MS, Van de Heyning PH. Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of dysphonia. *J Voice*. 1999;13:508–517. https://doi.org/10.1016/S0892-1997(99)80006-X.
28. Nemr K, Simoes-Zenari M, Cordeiro GF, Tsuji D, Ogawa AI, Ubrig MT, Menezes MHM. GRBAS and Cape-V scales: high reliability and consensus when applied at different times. *J Voice*. 2012;26. 812.e17−22; https://doi.org/10.1016/j.jvoice.2012.03.005.
29. Stevens SS. *Psychophysics*. New York: Wiley; 1975.
30. Cheng TH. *Direct Magnitude Estimation Versus Visual Analogue Scaling in the Perceptual Rating of Hypernasality*. Hong Kong: University of Hong Kong; 2007.
31. Bettens K, Bruneel L, Maryn Y, De Bodt M, Luyten A, Van Lierde KM. Perceptual evaluation of hypernasality, audible nasal airflow and speech understandability using ordinal and visual analogue scaling and their relation with nasalance scores. *J Commun Disord*. 2018;76:11–20. https://doi.org/10.1016/j.jcomdis.2018.07.002.
32. Yiu EML, Ng C-Y. Equal appearing interval and visual analogue scaling of perceptual roughness and breathiness. *Clin Ling Phon*. 2004;18:211–229. https://doi.org/10.1080/0269920042000193599.
33. Sewall A, Weglarski A, Metz DE, Schiavetti N, Whitehead RL. A methodological control study of scaled vocal breathiness measurements. *Cont Issues Comm Sci Dis*. 1999;26:168–172. https://doi.org/10.1044/cicsd_26_F_168.
34. Shrivastav R, Sapienza C, Nandur V. Application of psychometric theory to the measurement of voice quality using rating scales. *J Speech Lang Hear Res*. 2005;48:323–335. https://doi.org/10.1044/1092-4388(2005/022).
35. Bele IV. Reliability in perceptual analysis of voice quality. *J Voice*. 2005;19:555–573. https://doi.org/10.1016/j.jvoice.2004.08.008.
36. San Segundo E, Foulkes P, French P, Harrison P, Hughes V. Voice quality analysis in forensic voice comparison: developing the vocal profile analysis scheme. In: *Proc. 25th IAFPA*. 2016.
37. Abril A, Ambrosio E, de Blas M, Caminero A, García C, de Pablo J. *Fundamentos de psicobiología*. Madrid: Sanz y Torres; 2009.
38. Segal N. The importance of twin studies for individual differences research. *J Couns Dev*. 1990;68:612–622. https://doi.org/10.1002/j.1556-6676.1990.tb01425.x.
39. Stromswold K. Why aren't identical twins linguistically identical? Genetic, prenatal and postnatal factors. *Cognit*. 2006;101:333–384. https://doi.org/10.1016/j.cognition.2006.04.007.
40. Pakstis A, Scarr-Salapatek S, Elston R, Siervogel R. Genetics contributions to morphological and behavioral similarities among sibs and dizygotic twins: Linkages and allelic differences. *Soc Biol*. 1972;19:185–192. https://doi.org/10.1080/19485565.1972.9987983.
41. Nolan F, Oh T. Identical twins, different voices. *Int J Speech Lang Law*. 1996;3:39–49. https://doi.org/10.1558/ijsll.v3i1.39.
42. Van Gysel WD, Vercammen J, Debruyne F. Voice similarity in identical twins. *Acta Oto-Rhino-Laryngol Belg*. 2000;55:49–55.
43. Van Lierde K, Vinck B, De Ley S, Clement G, Van Cauwenberge P. Genetics of vocal quality characteristics in monozygotic twins: a multiparameter approach. *J Voice*. 2005;19:511–518. https://doi.org/10.1016/j.jvoice.2004.10.005.
44. 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. https://doi.org/10.1016/j.forsciint.2016.11.020.
45. San Segundo E, Tsanas A, Gómez-Vilda P. Evaluating the forensic importance of glottal source features through the voice analysis of twins and non-twin siblings. *Language and Law / Linguagem e Direito*. 2014;1:22–41.
46. Przybyla B, Horii Y, Crawford M. Vocal fundamental frequency in a twin sample: looking for a genetic effect. *J Voice*. 1992;6:261–266. https://doi.org/10.1016/S0892-1997(05)80151-1.
47. Debruyne F, Decoster W, Van Gijsel A, Vercammen J. Speaking fundamental frequency in monozygotic and dizygotic twins. *J Voice*. 2002;16:466–471. https://doi.org/10.1016/S0892-1997(02)00121-2.
48. San Segundo E, Yang J. Formant dynamics of Spanish vocalic sequences in related speakers: a forensic-voice-comparison investigation. *J Phon*. 2019;75:1–26. https://doi.org/10.1016/j.wocn.2019.04.001.
49. Haworth C, Asbury K, Dale P, Plomin R. Added value measures in education show genetic as well as environmental influence. *PLoS One*. 2011;6:16006. https://doi.org/10.1371/journal.pone.0016006.
50. Segal N. Implications of twin research for legal issues involving young twins. *Law Hum Behav*. 1993;17:43–58. https://doi.org/10.1007/BF01044536.
51. Plomin R, DeFries J, McClearn G, McGuffin P. *Behavioral Genetics*. fifth ed. New York: Worth; 2008.
52. Jessen M. Forensic phonetics. *Lang Ling Compass*. 2008;2:671–711. https://doi.org/10.1111/j.1749-818X.2008.00066.x.
53. Smits J, Monden C. Twinning across the developing world. *PLoS One*. 2011;6:e25239. https://doi.org/10.1371/journal.pone.0025239.
54. 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*, Spain: Consejo Superior de Investigaciones Científicas – Universidad Internacional Menéndez Pelayo; 2014.
55. Mennen I, Scobbie JM, de Leeuw W, Schaeffler S, Schaeffler F. Measuring language-specific phonetic settings. *Second Lang Res*. 2010;26:13–41. https://doi.org/10.1177/0267658309337617.
56. Beck J. Perceptual analysis of voice quality: the place of vocal profile analysis. In: Hardcastle WJ, Mackenzie Beck J, eds. *A Figure of Speech: A Festschrift for John Laver*. Mahwah, NJ: Lawrence Erlbaum; 2005:285–322.
57. Kreiman J, Vanlancker-Sidtis D, Gerratt BR. Defining and measuring voice quality. In: *Proc. VOQUAL'03*. 2003115–120.
58. Daly L, Bourke GJ. *Interpretation and Uses of Medical Statistics*. Oxford: John Wiley & Sons; 2008.
59. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15:155–163. https://doi.org/10.1016/j.jcm.2016.02.012.
60. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86:420–428. https://doi.org/10.1037/0033-2909.86.2.420.
61. Schindler A, Ginocchio D, Atac M, Maruzzi P, Madaschi S, Ottaviani F, Mozzanica F. Reliability of the Italian INFVo scale and correlations with objective measures and VHI scores. *Acta Otorhinolaryngol Ital*. 2013;33:121–127.
62. Bettens K, Bruneel L, Maryn Y, De Bodt M, Luyten A, Van Lierde K. Perceptual evaluation of hypernasality, audible nasal airflow and speech understandability using ordinal and visual analogue scaling and their relation with nasalance scores. *J Commun Disord*. 2018;76:11–20. https://doi.org/10.1016/j.jcomdis.2018.07.002.
63. Nunnally JC, Bernstein IH. *Psychometric Theory*. New York, NY: MacGraw-Hill; 1994.
64. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;327:307–310. https://doi.org/10.1016/S0140-6736(86)90837-8.
65. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8:135–160. https://doi.org/10.1191/096228099673819272.
66. Hanneman SK. Design, analysis, and interpretation of method-comparison studies. *AACN Adv Crit Care*. 2008;19:223–234. https://doi.org/10.1097/01.AACN.0000318125.41512.a3.
67. MedCalc Manual, MedCalc Software. https://www.medcalc.org/download/medcalcmanual.pdf, 2018(accessed 10 February 2019).
68. Gili Gaya S. *Elementos de fonética general*. fifth ed. Madrid: Editorial Gredos; 1956.
69. Real Academia Española. *Nueva gramática de la lengua española: Fonética y fonología*. Barcelona: Espasa Libros; 2011.
70. Mann VA, Repp BH. Influence of vocalic context on perception of the [sh]–[s] distinction. *Perception Psychophysics*. 1980;28:213–228. https://doi.org/10.3758/BF03204377.
71. Fowler CA. Compensation for coarticulation reflects gesture perception, not spectral contrast. *Perception Psychophysics*. 2006;68:161–177. https://doi.org/10.3758/BF03193666.
72. Järvinen K, Laukkanen A-M, Geneid A. Voice quality in native and foreign languages investigated by inverse filtering and perceptual analyses. *J Voice*. 2017;31:261.e25–261.e31. https://doi.org/10.1016/j.jvoice.2016.05.003.
73. Yiu EM-L, Murdoch B, Hird K, Lau P, Ho EM. Cultural and language differences in voice quality perception: a preliminary investigation using synthesized signals. *Folia Phoniatr Logop*. 2008;60:107–119. https://doi.org/10.1159/000119746.
74. Köster O, Schiller NO. Different influences of the native language of a listener on speaker recognition. *Forens Ling*. 1997;4:8–28.
75. Perrachione TK, Pierrehumbert JB, Wong PCM. Differential neural contributions to native- and foreign-language talker identification. *J Exp Psychol Hum Percept Perform*. 2009;35:1950–1960. https://doi.org/10.1037/a0015869.
76. San Segundo E, Braun A, Hughes V, Foulkes P. Speaker-similarity perception of Spanish twins and non-twins by native speakers of Spanish, German and English. In: *Proc. 26th IAFPA Conf*. 2017.
77. Chen NF, Slifka JL, Stevens KN. Vowel nasalization in American English: acoustic variability due to phonetic context. In: *Proc. 16th Int. Cong. Phon. Sci*. 2007905–908.
78. Eadie TL, Doyle PC. Direct magnitude estimation and interval scaling of pleasantness and severity in dysphonic and normal speakers. *J Acoust Soc Am*. 2002;112:3014–3021. https://doi.org/10.1121/1.1518983.
79. San Segundo E, Mompeán JA. Voice quality similarity based on a simplified version of the Vocal Profile Analysis: a preliminary approach with Spanish speakers including identical twin pairs. *Socioling Symp*. 15-18 June 2016;21.
80. Jintao K, Jingyang L, Li W, Xiaodi W. Development of a modified VPA protocol for Chinese speakers in Forensic Speaker Comparison. In: *Proc. 27th IAFPA Conf*. 201841–43.
81. Passetti RR, Constantini AC, The effect of telephone transmission on voice quality perception, J Voice (in press). https://doi.org/10.1016/j.jvoice.2018.04.018.
82. Darley FL, Aronson AE, Brown JR. Differential diagnostic patterns of dysarthria. *J Speech Hear Res*. 1969;12:249–269. https://doi.org/10.1044/jshr.1202.246.
83. Carding P, Carlson E, Epstein R, Mathieson L, Shewell C. Re: Evaluation of voice quality; Letters to Editor. *Int J Lang Comm Disord*. 2001;36:127–143.
84. Forensic Science Regulator. *Codes of Practice and Conduct for Forensic Science Providers and Practitioners in the Criminal Justice System, Issue 3*. Birmingham: Forensic Science Regulator Publications; 2016.
85. Forensic Science Regulator. *Codes of Practice and Conduct for Forensic Science Providers and Practitioners in the Criminal Justice System, Appendix: Speech and Audio Forensic Services, FSR-C134, Issue 1*. Birmingham: Forensic Science Regulator Publications; 2016.