UNIVERSITAT DE BARCELONA

Estudios de Fonética Experimental
Journal of Experimental Phonetics

LF Laboratori de Fonètica

International survey on voice quality: Forensic practitioners versus voice therapists

Eugenia San Segundo[a]

[a]Universidad Nacional de Educación a Distancia (Spain), esansegundo@flog.uned.es

ARTICLE INFO

Article history
Received: 30/12/2020
Accepted: 08/02/2021

Keywords
Voice Quality
Forensic Phonetics
Methodologies
Reliability
International Survey

ABSTRACT

In recent years, numerous investigations have focused on voice quality (VQ) for forensic purposes. These studies notwithstanding, we lack an international picture of how VQ is generally understood by forensic practitioners and how it contrasts with the practices of voice therapists. To fill this gap, a survey was designed and sent to both cohorts: forensicists and clinicians. A total of 45 responses from 20 countries were received. Important differences were found between groups, particularly regarding perceptual assessment. One conclusion to be drawn is that more emphasis should be placed on calibration and error measurement in forensic approaches to VQ. Further collaborations with clinicians should also be encouraged.

1. Introduction

In Forensic Voice Comparison (FVC) different parameters can be analyzed by voice experts (typically forensic phoneticians) when they are requested to compare the voice recording of an offender with the recording(s) belonging to a suspect or several suspects. In recent years there has been a growing interest in evaluating the forensic performance of one vocal aspect in particular: voice quality (henceforth VQ). For instance, Hughes et al. (2017), San Segundo, Univaso and Gurlekian (2019) or Park, Afshan, Kreiman, Yeung and Alwan (2019) have delved into how to improve automatic systems using VQ parameters. Other investigations have focused on long-standing issues in the assessment of VQ, such as the measurement of inter-rater agreement and the need for calibration stages in the process of evaluation (San Segundo et al., 2019), the effect of telephone-degraded recordings on such perceptual evaluations (Passetti & Constantini, 2019), or the simplification and computer-based implementation (San Segundo & Mompeán, 2017; San Segundo & Skarnitzl, in press) of protocols such as the Vocal Profile Analysis (VPA) scheme (Laver, 1980; Beck, 2005). For the VPA in particular, some studies (San Segundo, Schwab, Dellwo, Le, & Mompeán, 2017) have explored possible acoustic correlates for under researched VQ dimensions (e.g. vocal tract tension). Other studies (San Segundo et al., 2018) have used different clustering methods to try to distinguish perceptually similar speakers on the basis of the VQ ratings given by three trained phoneticians.

What distinguishes the above-mentioned studies – whose publication spans just the past five years–

from previous approaches to VQ within Forensic Phonetics is that the research emphasis is now placed on the perceptual evaluation of VQ while previous investigations in the context of FVC typically focused on acoustic analyses of VQ. For example, San Segundo and Gómez-Vilda (2014) analyzed a large number of glottal parameters (e.g. glottal gap coefficients or biomechanical parameters related to the distribution of mass and viscoelasticity of the vocal folds' body and cover), extracted from the pause fillers that occur naturally in spontaneous conversations when speakers hesitate: when they are thinking of what they are going to say next, or when they are trying to remember something. The analysis of these fillers is considered very useful in FVC tasks because such units are longer than vowels in connected speech, so they are considered long enough for a robust glottal analysis (Gómez-Vilda, San Segundo, Mazaira, Álvarez, & Rodellar, 2014; Tsanas, San Segundo, & Gómez-Vilda, 2017).

The heterogeneity of approaches to VQ in FVC could be caused by the lack of consensus on how to define VQ. It could be hypothesized that the working definition of VQ depends on the professional background or the particular research interests of each researcher, which would have a bearing on their methodological decisions. The objective of this study is to present an international picture of how VQ is understood by forensic practitioners all over the world, and to compare their responses with those provided by another group of professionals for whom VQ is also very important, namely voice therapists. For this purpose, an online survey was designed and sent to two cohorts of participants: (a) forensic practitioners and (b) voice therapists. This investigation also aims to bridge the gap between state-of-the art research on VQ and real practices in current casework.

The definition of VQ differs substantially across researchers, so it is not easy to provide a simple definition which encompasses all the possible interpretations of this term. This is precisely one of the reasons why this survey has been designed: to shed some light into what VQ means for different experts. Nevertheless, it is widely accepted that there is a broad and a narrow definition of VQ, depending on whether laryngeal and supralaryngeal aspects are taken into account (broad definition) or only laryngeal features are considered (narrow definition). San Segundo et al. (2019) have recently explained these aspects in some more detail, while Gil and San Segundo (2014) summarized the main issues and challenges of analysing VQ in forensic reports. Furthermore, for those not familiar with the linguistic concept of VQ, it is worth highlighting what VQ is not. Above all, VQ is not a synonym for the quality of the acoustic signal or the sound quality of a recording (e.g. poor or low quality).

2. Survey design and participants

A survey was designed with *SurveyMonkey*, featuring 28 questions which covered a range of aspects related to VQ in forensic and clinical practice. Participants were targeted via mailing lists of the main international associations for forensic and clinical phonetics, and also through social media such as *Twitter* or academic social networking platforms such as *ResearchGate*. Potential participants were sent a link to the online survey and their responses were collected between July 2017 and September 2018. The questionnaire had been previously sent to two international experts, one in Germany and another one in Brazil, with experience in forensic and clinical phonetics, respectively. This phase was aimed at gathering initial feedback about the survey (e.g. in case that some important questions were missing or needed rephrasing).

In total, 45 answers were received from 20 different countries: 27 corresponding to forensic practitioners (henceforth forensicists) and 18 to voice therapists (henceforth clinicians). There were three participants who worked both in the clinical and forensic fields. In the design of the survey this possibility was foreseen, so those three participants answered: (a) common questions, (b) questions addressed to forensicists and (c) questions addressed to clinicians. Therefore, the total number of participants is 42 and the total number of valid answered questionnaires is 45. Only the responses of participants who completed the entire survey were considered valid responses. Because participants

were allowed to abandon the survey at any time, a high percentage of participants started the survey but did not continue until the end. Those who failed to complete the survey before reaching the final question were not included in this investigation, even if some answers could have been used. Note however that the survey allowed skipping questions. If participants skipped a few questions but reached the end of the survey, their submission was considered complete and their answers were taken as valid for this investigation. Those participants must have skipped those few questions knowingly –and the reasons why will be explored– and not just abandon the survey. This decision should result in a higher quality of the final data set to the detriment of a higher number of responses.

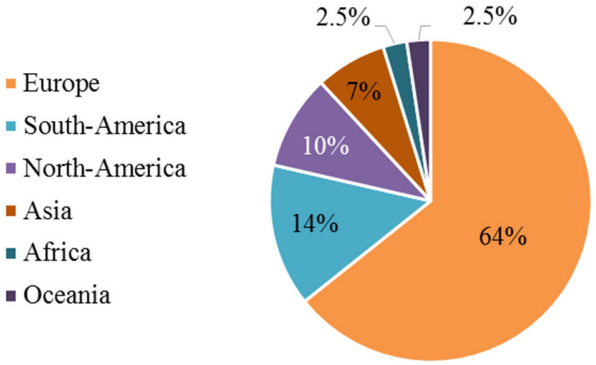

Figure 1. Survey participation per continent.

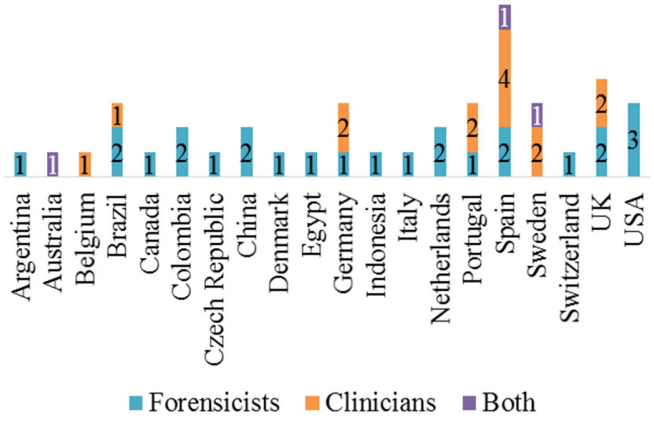

Figure 2. Survey participation per country (and divided by professional activity).

Figures 1-2 show the distribution of participants per continent and per country, respectively. Figure 1 shows that most participants are from Europe (64%), followed by South America (14%), North America (10%) and Asia (7%). There is only one participant from Africa and one from Oceania. Figure 2 shows that most participants are from Spain and the UK, followed by Brazil, Germany, Portugal, Sweden and the USA.

3. Results

Figure i (in Appendix A) shows all the survey questions grouped thematically, with their associated figures and/or tables, in the order that they appear in this section. Its aim is to make searching easier for the reader, particularly when cross-references are made.

3.1. Forensic practitioners versus voice therapists: common questions

Q1. *Do you consider voice quality in your professional activity?*

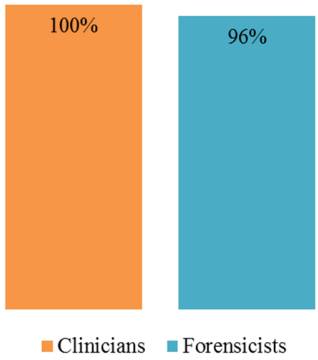

Figure 3. Respondents considering voice quality in their professional activity. Answered/Skipped: 18/0 clinicians; 27/0 forensicists.

The results show that all the clinicians consider VQ in their profession. The 4% not considering it in forensic practice corresponds to just one respondent. Three questions of the survey were designed in anticipation of a negative answer to the question of whether VQ was considered. Since only one participant responded 'no', I have not included any plots to explain the following three questions. His/her responses are simply commented below.

The first follow-up question was '*Please indicate which of the following reasons best suit your decision not to consider VQ in your casework*' with five options: (a) I use an automatic recognition system for speaker comparison which yields satisfactory enough results, so I do not feel the need to incorporate VQ analysis; (b) I lack the specific training for carrying out VQ analysis; (c) I think there is no consensus yet as to what VQ really means; (d) I think that the implementation of any existing VQ protocol nowadays is very difficult in casework, given several limitations like channel degradation or mismatch of the voice recordings; (e) Other (please specify). The chosen answer was (b).

The second follow-up question was '*Would you be willing to consider VQ in your casework in the future?'* Among three possible response options ('yes', 'no' and 'maybe'), this participant responded 'maybe'.

The third follow-up question was '*Which of the following might influence your decision to include VQ analyses in your casework in the future?*' with four options: (a) If they proved useful to complement the results of automatic speaker recognition systems; (b) If the existing perceptual protocols were modified / adapted for forensic purposes; (c) If I received some specific training; and (d) Other (please specify). The chosen answer was (d) and the participant specified that the type of forensic tasks in which his/her team is currently involved concern mostly environmental media (e.g. air, water, land, etc.) through direct data collection and simulation/modelling, with aural evidence aspects focusing only on the regulation of noise.

In view of these answers, it can be claimed that all the surveyed forensic speech experts do examine VQ in practice. It seems clear that the only outlier within the surveyed forensic group tackles aural evidence only peripherally. Nevertheless, the fact that he/she is still considering examining VQ in the future highlights the usefulness of this type of surveys. The kind of questions that this survey includes should be of interest for those forensic experts working in areas that deal with aural evidence only tangentially. This way, the survey will allow them to familiarize themselves with some key aspects of VQ before starting FVC analyses.

Q2. *How do you assess voice quality in clinical practice / casework?*

The results of this question can be observed in Figure 4. As with other questions in this survey, it is not necessary to repeat in the text what can be better shown through the diagrams or the figures.[1]

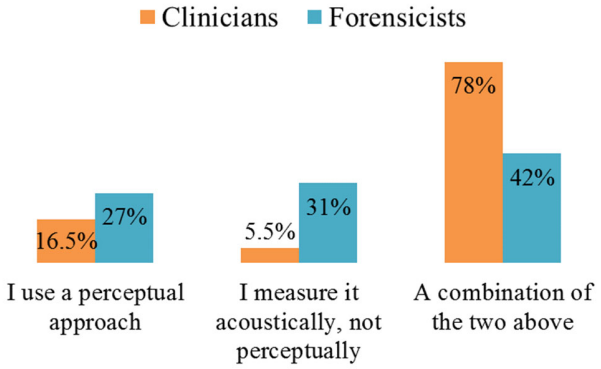

Figure 4. Approaches to voice quality. Answered/Skipped: 18/0 clinicians; 26/1 forensicists.

Q3. *What is your working definition of voice quality?*

For Q3 the answer choices in full form were as follows:

(1) VQ refers specifically to the effect resulting from the mode of vibration of a person's vocal cords; it is circumscribed to the phonatory activity.

(2) VQ refers to the characteristic auditory coloring of an individual's voice, derived from a variety of laryngeal and supralaryngeal features and running continuously through the individual's speech.

[1] An effort has been made in this investigation to depict the results of all the questions thorough data visualization techniques. The reader can expect some comments only if there are aspects that need clarification (e.g. Q1). Otherwise, the results are properly discussed in the discussion section.

(3) VQ refers to those characteristics which are present more or less all the time that a person is talking: it is a quasi-permanent quality running through all the sound that issues from his/her mouth. It encompasses more than laryngeal and supralaryngeal features.

The author or the source of the above definitions were not provided in the survey, but they were extracted from the following references: *definition 1* is one of the possible definitions of VQ that can be distinguished, according to Nolan (1982, p. 442) in his review of "The Phonetic Description of Voice Quality" by John Laver (Laver, 1980); *definition 2* is found verbatim in Trask (2004, p. 381); and *definition 3* is the one proposed by Abercrombie (1967, p. 91), as read in Beck (2005, p. 286).

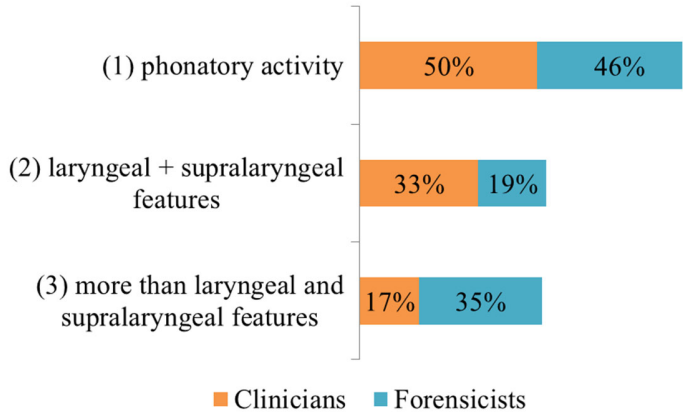

Figure 5. Definition of voice quality. Answered/Skipped: 18/0 clinicians; 26/1 forensicists.

Q4. *For the perceptual analysis of VQ, do you follow an established or known protocol, scheme or rating system (e.g. VPA or GRBAS)?*

The following note was provided with this question: "Even if you use a modified/in-house version of a well-known protocol and not the original version, you must select "yes" in this question (You can provide more details later on)". As Figure 6 shows, most clinicians and most forensicists use an established protocol for VQ perceptual analysis.

In this case, there is a relatively high number of forensicists that skipped the question. They correspond to the participants that responded that they use only an acoustic method for the analysis of VQ (see Q2) and the only forensic practitioner that replied s/he does not consider VQ in casework. In the case of clinicians, only one participant indicated using only acoustic methods in Q2. Therefore, from this question until Q9, the reader will find at least 1 skipped answer for clinicians and 8 skipped answers for forensicists. The rest of the skipped cases –in case there are more– truly correspond to participants that did not want to answer the question.

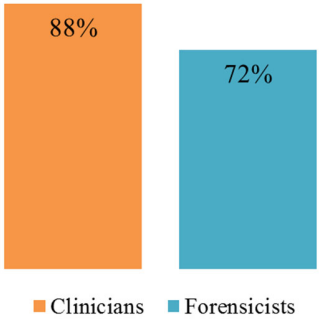

Figure 6. Use of established protocol for the perceptual analysis of voice quality. Answered/Skipped: 16/2 clinicians; 18/9 forensicists.

For those participants who responded that they do not follow an existing protocol, two further questions were posed: (1) '*Why don't you follow a protocol?*' and (2) '*How do you assess VQ perceptually*?' To the first question, three options were proposed: (a) I lack the specific training, (b) I don't see the need to follow an existing protocol' or (c) 'others'; with the option to provide their own reason. Among the clinicians, one marked 'the lack of training' option; the other one commented that, because she predominantly works on infants, she utilizes a protocol developed for them by herself. Among the forensicists, half of them acknowledged lack of training in a specific protocol, the other half responded in terms that suggest that they might either use established protocols in the future and/or that they already borrow elements from specific protocols without using them as such: "I might add in an existing protocol but I borrow elements from

many sources and I always start from the speakers' point of view" (forensicist #1); "Still researching the common methods and their merits and limitations" (forensicist #2).

As for the second question ('how do you assess VQ perceptually?'), only forensicists responded. A summary of their responses follow: "I triangulate how the speaker feels about their VQ, among other communication characteristics such as gesture, expression, signs, and how the 'native' listener feels, and my own assessment which includes shifts in language to accommodate unusual or 'impaired' VQ. Then I look at the behavioral outcome on natural conversation to see how VQ impacts meaning and non-verbal affect in communication. What do people do to 'manage' or 'accommodate' unusual VQ." (forensicist #1)[2]; "We use certain terms to describe the voice as best we can, e.g. harsh, breathy, lax etc." (forensicist #2); "Rough indication of perceptual features. More important is the description of the variation of VQ in relation to vocal effort and speaking style and attitude." (forensicist #3); "When possible, an analysis by a group of trained phoneticians. Otherwise, a personal listening to the details of phonetic content, prosodic regularities and deviations, and comparisons with published quality features." (forensicist #4)

Q5. *Please specify the name of the protocol, scheme or rating system that you use to analyze VQ perceptually (you can select more than one).*

Four clinicians chose more than one option while only one forensicist selected more than one. Two participants in each professional category selected "others", and they further specified the following remarks. One clinician said: "RBH-Scale by Nawka, Anders and Wendler (1994), which translates to the G, R and B of GRBAS (Hirano, 1981); and the other said: "GRBASH (Nemr & Lehn, 2010), an update of the original GRBAS that separates the Roughness and the Harshness". As for the forensic practitioners, one of them specified: "I use a scheme I built according to the methods used in Experimental Phonetics" and the second: "An in-house rating system for perceptual analysis in general (i.e., not only VQ); VQ dimensions are largely based on VPA and SVPA". [3]

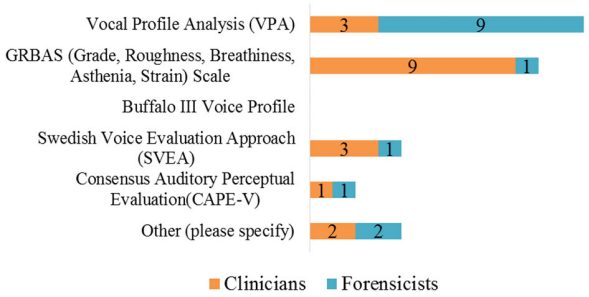

Figure 7. Protocol(s) used for the perceptual analysis of voice quality. Answered/Skipped: 14/4 clinicians; 13/14 forensicists.

Q6. *What is the main advantage that you see in the choice of that protocol over others?*

This was an open question with insightful information provided by the participants. For the sake of simplification, two wordclouds were created (Figure 8 a-b). These represent the most repeated words in the answers of the participants, with greater visual prominence (font size) given to words that appear more frequently in the source text. Full answers can be found in Appendix B (table i).

Table 1 includes the list of words used to generate the wordclouds in Figure 8, ordered from most to least frequently used. The information is based on 22 participants: 11 clinicians and 11 forensicists. Terms were simplified so that semantically similar words were grouped (e.g. 'ease of use' and 'easy to use' were merged into "easy").

[2] This means anonymous forensicist; 'forensicist #1' here is not necessarily the same respondent as 'forensicist #1' above.

[3] SVPA refers to San Segundo and Mompeán (2017). A variation of such protocol, using visual analog scales, followed later (San Segundo & Skarnitzl, in press).

Figure 8 a-b. Wordclouds with the keywords of the answers to the question "*What is the main advantage that you see in the choice of that protocol over others*". Left wordcloud: answers provided by clinicians; right wordcloud: answers provided by forensicists. Answered/Skipped: 11/7 clinicians; 11/16 forensicists.

Clinicians		**Forensicists**	
Words	*Occurrences*	*Words*	*Occurrences*
easy	4	well-established	2
used	3	standard	2
quick	3	easy	2
reliability	2	sociolinguistic-studies	1
acoustic	2	wide-range-of-features	1
correlate	2	relevant-to-FVC	1
simple	2	practicality	1
vocal-fold-physiology	1	flexibility	1
global-description	1	well-known	1
well-defined	1	report	1
scientific	1	court	1
evaluated	1	complicated	1
measures	1		
research	1		
validity	1		
studies	1		

Table 1. Words used to generate the wordclouds in Fig. 8; ordered from most to least frequently used.

Q7. *If you have marked more than one option, please indicate briefly the reasons why you use more than one protocol.*

Two clinicians and one forensicist responded to this question. Regarding the clinicians, one reported using GRBAS and VPA while the other would use GRBAS, VPA and occasionally CAPE-V (Kempster, Gerratt, Verdolini, Barkmeier-Kraemer, & Hillman, 2009). Their full responses were as follows. Clinician #1: "Useful as comparison - some scales can be used for patients" (the respondent provides the reason for using more than one protocol, and s/he specifically refers to GRBAS and VPA). Clinician #2: "I tend to use GRBAS most as it is very quick, but I would use VPA with clients who require further analysis/ whose issues are more supralaryngeal, often articulatory. I use CAPE-V

when there is inconsistency and/or I want to select other parameters, e.g. wetness." As for the forensicist, s/he also used the combination VPA and GRBAS, and his/her comment is: "VPA is more elaborated, but in the GRBAS framework the additional categories "grade" and "asthenia/strain" are helpful."

Q8. *Do you use the original version of the protocol (referred to in the previous question) or a modified version?*

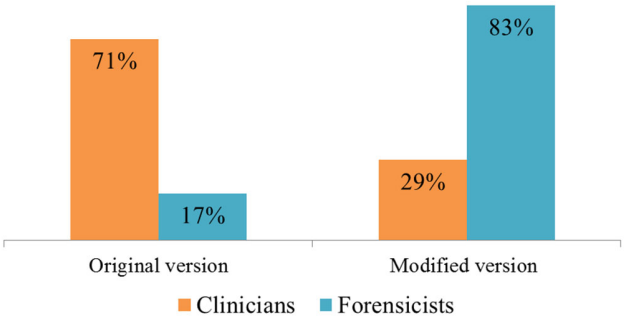

Figure 9. Use of original versus modified version of perceptual protocols. Answered/Skipped: 14/4 clinicians; 12/15 forensicists.

Q9. *How many years of experience do you have in the use of the protocol selected before?*

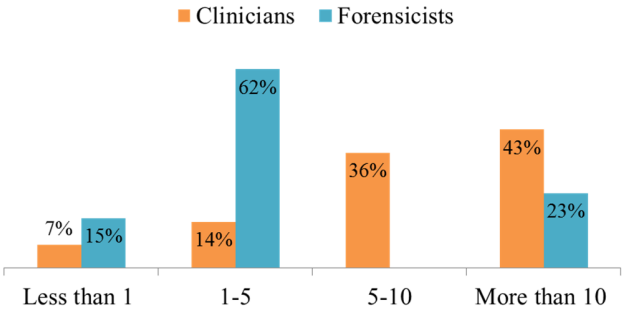

Figure 10. Years of experience. Answered/Skipped: 14/4 clinicians; 13/14 forensicists.

Q10. *Slight modifications have been made to the VPA protocol since it was created. Please select the VPA version that you are using. You can also select the field 'other' and specify which.*

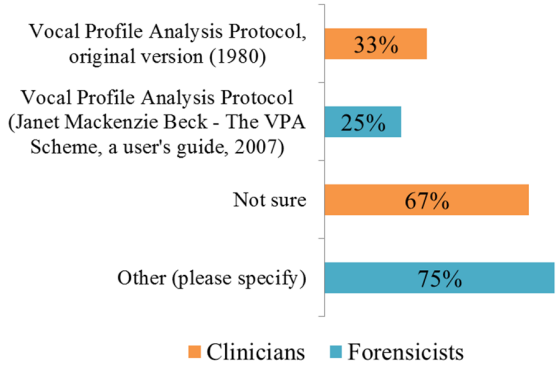

Figure 11. Version of the VPA protocol used. Answered/Skipped: 3/15 clinicians; 8/19 forensicists[4].

The high number of skipped answers by clinicians (from this question until Q14) is due to the smaller percentage of clinicians using the VPA protocol in comparison with forensicists (see Q5). In contrast, only one forensicist seems to have skipped this group of questions (VPA-specific questions) systematically. According to the information gathered in Q5, three clinicians and nine forensicists use the VPA. Therefore the skip rates shown in the figures' captions should not be interpreted as true skip rates (they are based on the total number of participants).

Six forensicists answered 'other' to this question. Two of them use a translation of the protocol into Brazilian Portuguese. Another two refer to a further modified version "closely aligned with Mackenzie-Beck's (2007) version, but with some modifications". A fifth respondent uses a simplified version specific for Chinese speakers and another one seems to use a mix between the original version and a further modified version, although claims that "s/he is not sure".

[4] For the two references cited in the figure, see Laver (1980) and Beck (2007).

Q11. *Depending on the version, the VPA scheme can include prosodic features, temporal organization and other features. Do you consider these aspects (e.g. pitch, loudness or respiratory support) within your VQ assessment/protocol or apart from it?*

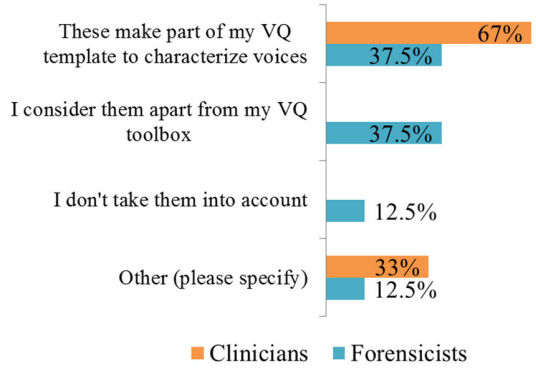

Figure 12. Do you consider prosodic aspects within your VQ assessment/protocol? Answered/Skipped: 3/15 clinicians; 8/19 forensicists.

The percentage of participants who answered 'other' to this question corresponds to one clinician and one forensicist. The former indicates that s/he also uses the "Christina Shewell voice skills framework (Shewell, 2013) which includes these features". The latter states: "Not considered as part of VQ analysis; I use VPA without these sections, but the analysis of these features is part of wider analysis of voice undertaken in casework".

Q12. *Are you or your team trained in the VPA scheme? Choose the answer which best fits your situation.*

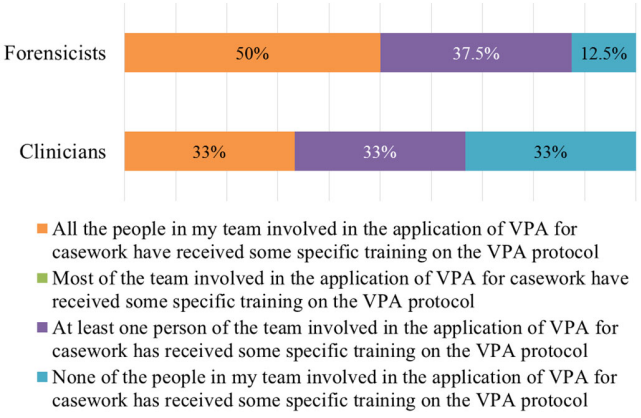

Figure 13. VPA training (analysts involved). Answered/Skipped: 3/15 clinicians; 8/19 forensicists.

Q13. *If applicable, choose the training method which best fits your situation.*

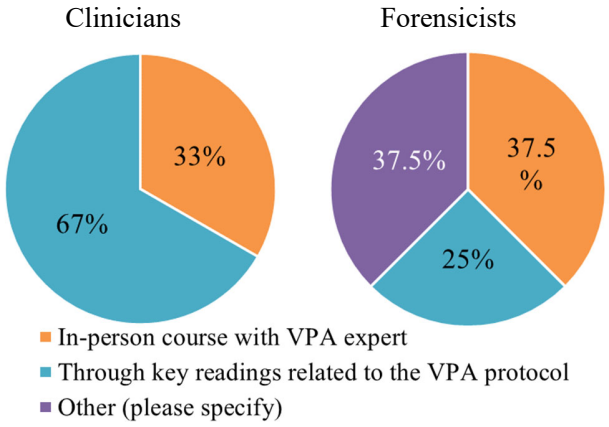

Figure 14. VPA training (method). Answered/Skipped: 3/15 clinicians; 8/19 forensicists.

The answers provided by the three forensicists who answered 'other' are: Forensicist #1: "My personal skills come from reading, studying specifically VQ, practical experience with voice quality (auditory analysis, production, acoustic analysis, and visual analysis/laryngoscopy), and special experience in voice therapy". Forensicist #2: "On the job training/shadowing, database practice and book-based learning using Laver's book and the User Manual from Beck". Forensicist #3: "I received instruction from a professor in Sweden, then learned it through reading and testing".

Q14. *How often do you use the VPA protocol?*

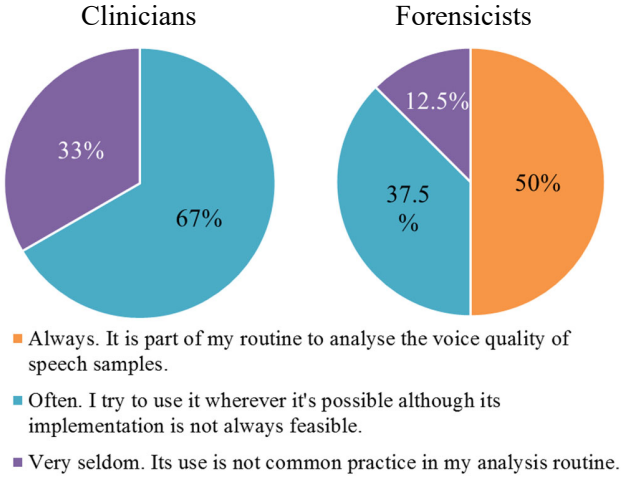

Figure 15. VPA use frequency. Answered/Skipped: 3/15 clinicians; 8/19 forensicists.

Q15. *If applicable, please indicate how strongly you agree or disagree with these statements (a-c) about the difficulty to implement the VPA protocol:*

(a) "In my work I usually have to compare high-quality recordings with telephone-filtered recordings, which makes the analysis of VQ very difficult".

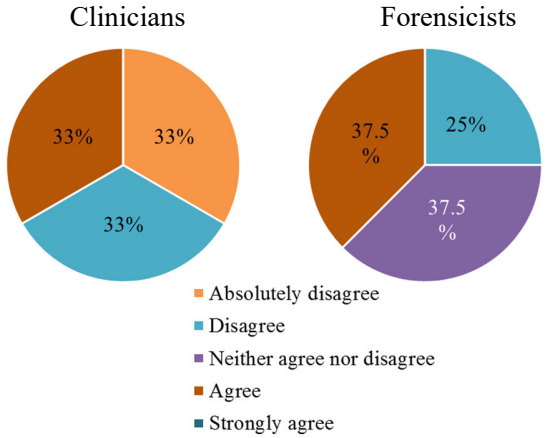

Figure 16. Degree of agreement with stated VPA difficulties (telephone effect). Answered/Skipped: 3/15 clinicians; 8/19 forensicists.

Taking the response options as a Likert scale where 'absolutely disagree' equals '1' and 'strongly agree' equals '5', the average agreement of clinicians with this statement is 2.33 and the agreement of forensicists is 3.125.

(b) "I find that the original VPA protocol includes too many labels and some of them seem confusing (e.g. because they are very similar and it's hard to tell when to use one or the other)".

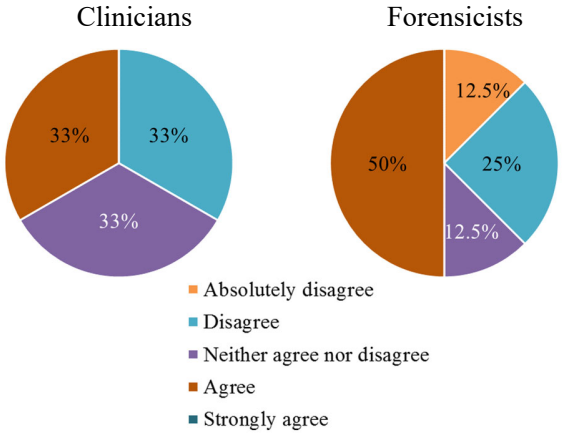

Figure 17. Degree of agreement with stated VPA difficulties (labels). Answered/Skipped: 3/15 clinicians; 8/19 forensicists.

Taking the response options as a Likert scale where 'absolutely disagree' equals '1' and 'strongly agree' equals '5', the average agreement of both clinicians and forensicists with this statement is 3, so 'neither agree nor disagree'.

Again for this question, taking the response options as a Likert scale where 'absolutely disagree' equals '1' and 'strongly agree' equals '5', the average agreement of clinicians with this statement is 2.33 and the average agreement of forensicists with this statement is 2.25. In other words, both groups rather disagree with the statement. The high percentage of respondents answering 'neither agree nor disagree' should correspond to those who have never tested VPA with languages other than English. It makes sense that only those who have actually analyzed voices in different languages answered 'disagree' or 'strongly disagree'.

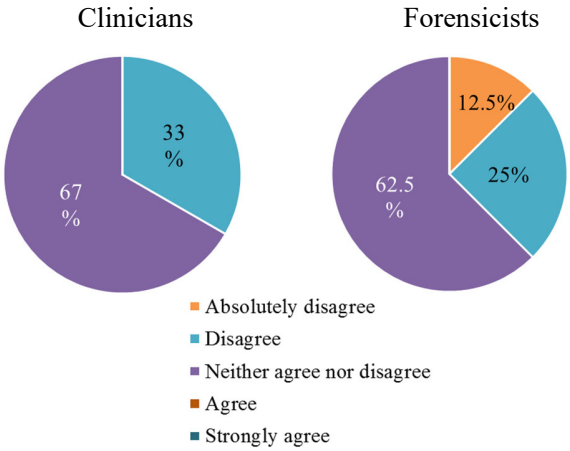

Figure 18. Degree of agreement with stated VPA difficulties (language). Answered/Skipped: 3/15 clinicians; 8/19 forensicists.

Q16. *If applicable, please specify the language/s of the speech samples you usually find in your forensic casework or clinical practice.*

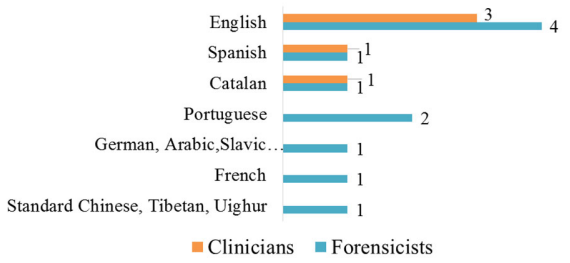

Figure 19. Languages found in forensic casework and clinical practice. Answered/Skipped: 3/15 clinicians; 8/19 forensicists.

In Figure 19 the full answer of a participant (forensic group), starting with "German, Arabic, Slavic…" was "German, Arabic, Turk languages, Slavic languages, languages from the Balkans, African languages, English, Baltic languages".

All the clinicians responded "English" on top of the other languages. In the survey they were instructed to write all the languages in order of frequency if the analyzed several languages. Their answers were: "English" (Clinician #1 and #2) and "English, Catalan, Spanish" (Clinician #3). As for the forensicists, half of them seem to carry out their casework in English. One in particular answered "English; sometimes (rarely) other languages with a native-speaker/phonetician". The one who responded "French", selected also "English" as the most frequent language typically found in casework.

Q17. *If the language/s that you work with in your casework is/are not English, do you find difficulties in the implementation of this protocol (especially with certain settings/labels)?*

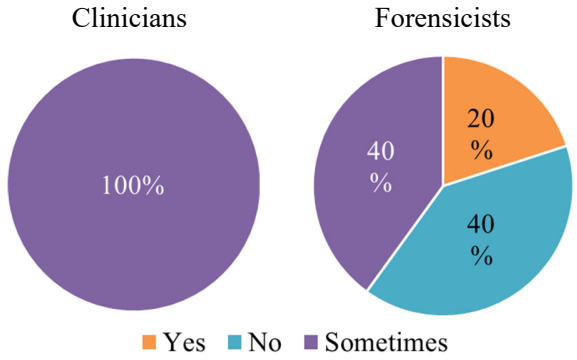

Figure 20. Difficulties using VPA protocol with languages other than English. Answered/Skipped: 1/17 clinicians; 5/22 forensicists.

Here the clinicians' answers are based on just one respondent. There were three clinicians reporting VPA use in this survey, so two of them skipped this question presumably because they use VPA with English patients.

Q18. *If applicable, select which VQ labels/settings are more difficult to relate to your language.*

The following note was added: "For instance, it has been suggested that if a given articulatory setting (e.g. nasality) corresponds to a linguistically distinctive feature in a given language (e.g. French or Portuguese), acoustic indicators of that articulatory setting would be primarily associated with the distinctive feature set of the language and would be more difficult to associate with the voice quality of the speaker" (Keller, 2005).

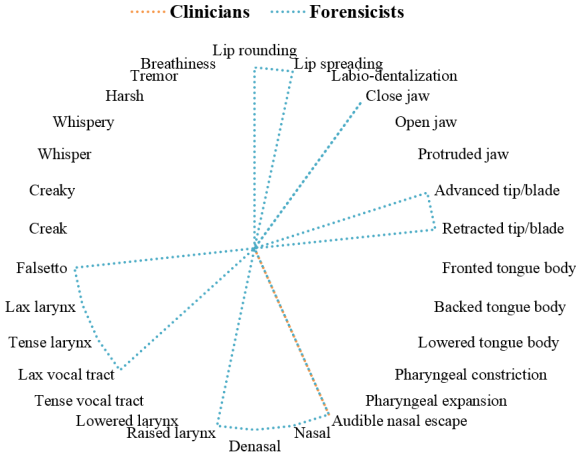

Figure 21. Radar plot (most difficult settings). Answered/Skipped: 1/17 clinicians; 5/22 forensicists. Each setting marked in the chart means that it was selected once.

Q19. *On the basis of your experience, which of these settings you seldom find in a speaker or which of these labels are seldom used to characterize a voice?*

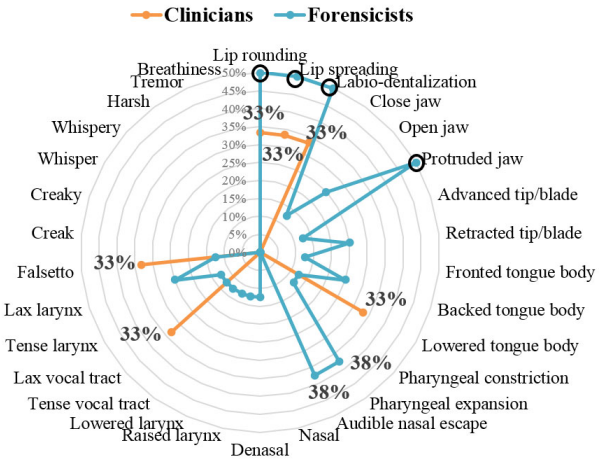

Figure 22. Radar plot (rarest settings found by forensicists and clinicians). Answered/Skipped: 3/15 clinicians; 8/19 forensicists.

In Figure 22 percentages represent the responses relative to the number of respondents per group. The higher the percentage, the more participants found that setting to be rare. The upper limit of the plot is 50%. Only settings with at least 33% agreement are highlighted in bold.

Q20. *Do you measure interrater agreement?*

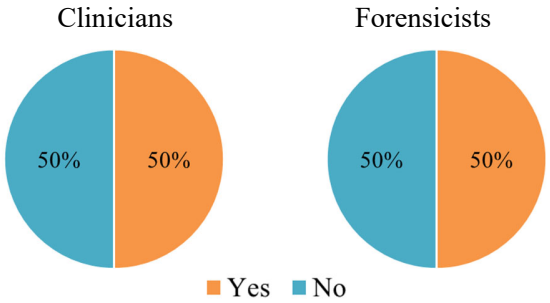

Figure 23. Interrater agreement measurement. Answered/Skipped: 16/2 clinicians; 18/9 forensicists.

Q21. *How many people conduct the VQ analysis?*

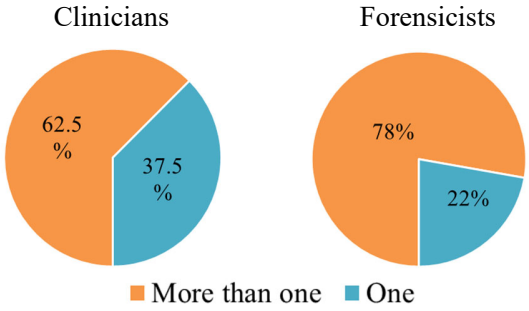

Figure 24. People involved in VQ perceptual assessment. Answered/Skipped: 16/2 clinicians; 18/9 forensicists.

Q22. *If you answered more than one, do you follow a blind procedure (each analyst carries out the analysis independently)?*

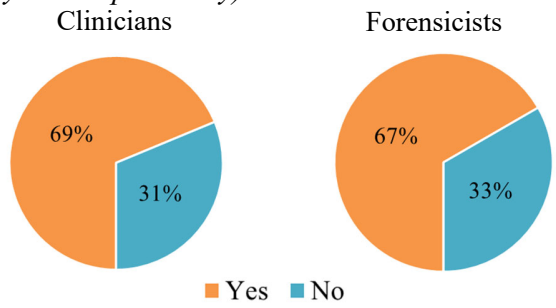

Figure 25. Use of blind procedures. Answered/Skipped: 16/2 clinicians; 18/9 forensicists.

Q23. *If applicable, how do you measure interrater agreement?*

Clinicians	**Forensicists**
"Cohen's kappa coefficient"	"Subjectively"
"Cronbach's alpha reliability coefficient"	"No particular method"
"Cohen's kappa coefficient (for 2 judges) and Fleiss's kappa coefficient (for more than 2 judges)".	"We have done this in research but not in respect of casework".
"We don't use interrater agreement in clinical practice, but we do in research studies. We use intraclass correlation coefficient or Kappa's coefficient based on blind procedures by 2 or 3 raters".	"Group consultation is made on the inconsistent aspects and finally the trade-off is made according to the value of these inconsistent characteristics in FVC".
"I've only done it in research. With 10 experts. I used the Interclass correlation coefficient".	"Recording of judgments about features, and determination of deviations in ratings".
"In the routine clinical work only one SLP rates the patient's voice. In research a group of experienced listeners listen and intra- and interrater reliability is measured (correlation or Cronbach's alpha)".	"By means of combined analysis and based on the UK position involving consistency, distinctiveness and probability scale".
	"It is not measurement, but comparison. Usually we come to an agreement. Sometimes we ask for a third opinion."
	"Kappa"
	"Euclidean distances"

Table 2. Summary of the responses given to the open question 'how do you measure interrater agreement'. SLP stands for Speech Language Pathologist. Answered/Skipped: 6/12 clinicians; 9/18 forensicists.

Q24. *You have indicated above that you consider VQ (also) acoustically: Which of the following methods do you use to measure it? (You can select more than one answer).*

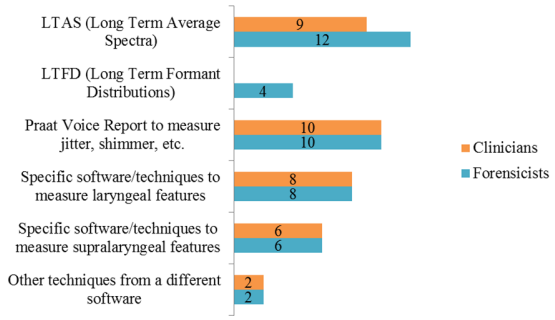

Figure 26. Methods used to evaluate VQ acoustically. Answered/Skipped: 16/2 clinicians; 18/9 forensicists.

Q25. *Please specify the name of the software or techniques that you use to measure VQ acoustically.*

An obligatory answer was required for this question although participants could state "confidential" if they did not want to disclose this information. As in Q6, wordclouds are provided in Figure 27 to visualize the most repeated words in the answers of the participants, with greater visual prominence given to words that appear more frequently.

Figure 27. Wordclouds with the answers to the question "Specify the name of the software or techniques that you use to measure VQ acoustically". Left wordcloud: answers provided by clinicians; right wordcloud: answers provided by forensicists. Answered/Skipped: 17/1 clinicians; 26/1 forensicists.

Clinicians		**Forensicists**	
Words	*Occurrences*	*Words*	*Occurrences*
Praat	7	Praat	9
Dr. Speech	2	confidential	4
Kay Elemetrics CSL (a)	2	EASERA (c)	1
Soundswell (b)	2	Kay Elemetrics CSL	1
Sopran	2	Matlab VoiceBox Toolbox	1
Voice Studio	1	Neurospeech	1
OpenSMILE	1	Smaart	1
Phog	1	STC SIS 2 (d)	1
laryngograph	1	VS-99 Voice Station (e)	1
own software	1	own software	1
confidential	1	custom measurement	1

Table 3. Words used to generate the wordclouds in Fig. 27. They are ordered from most to least frequently used. Full names of abbreviated software programs: (a) *Kay Elemetrics Computerized Speech Lab*; (b) *Soundswell Signal Workstation*, (c) *Electronic and Acoustic System Evaluation and Response Analysis*, (d) *SIS forensic audio analysis software of the Speech Technology Center* (e) *VS-99 Voice Station* developed by *Yangchen Electronic Company of Beijing*.

Figure 27 and Table 3 show that *Praat* (Boersma & Weenink, 2005) is the software that most clinicians and forensicists use. The only other software that both groups use is *Kay Elemetrics CSL (Computerized Speech Lab)* although to a lesser extent. All the other programs vary considerably between one professional group and the other. A final similarity between clinicians and forensicists is that at least one researcher of each group reported analyzing VQ acoustically using their own software. The forensicist that indicated 'own software' added that "it measures jitter, shimmer and harmonics-to-noise ratio"; as for the clinician, s/he gave the following details: "Prosody Module at FAU (Friedrich-Alexander-Universität Erlangen-Nürnberg)". Also within the group of clinicians, the one who gave the name Sopran actually defined it as 'custom written software' (Granqvist, 2020).

Interestingly, 'confidential' was the second most frequent response provided by forensicists. Only one clinician gave this answer and it actually corresponds to one of the three respondents that indicated that they work both in clinical and forensic domains.

Participants were instructed that they could include some basic references citing the software or techniques. Among clinicians, the users of *Voice Studio* mentioned that it is a Portuguese software program but no reference was provided. In terms of techniques or parameters used, one user of *Kay Elemetrics* provided the following details: *Multi-Dimensional Voice Profile* and *Voice Range Profile*; another user of *Praat* specified that the focus was placed on the smoothed cepstral peak prominence (CPPS) as well as the use of the *Acoustic Quality Voice Index version 03.01 (AVQI)* and the *Acoustic Breathiness Index (ABI)*. Among forensicists, only one participant gave specific details and a reference for the software used: *Neurospeech* (Orozco-Arroyave et al., 2018) "which measures phonation, articulation, prosody, and intelligibility-based features".

Q26. *Please provide any additional feedback which you think is relevant to this survey.*

Before tackling the responses to the questions which were specifically posed to forensicists and clinicians separately, a last open question was addressed to both forensicists and clinicians, which asked them to provide additional feedback about the survey. The full answers can be found in Appendix C (table ii). Participants' feedback was extremely positive, with many participants thanking the author for approaching the topics raised in the survey. Quite a few of the respondents also highlighted the need for their laboratories to embark in training programs in order to improve or enlarge their techniques in terms of VQ assessment.

3.2. Specific question for forensic practitioners

Q27. *In which type of forensic task do you analyze/consider VQ perceptually? (More than one option can be chosen).*

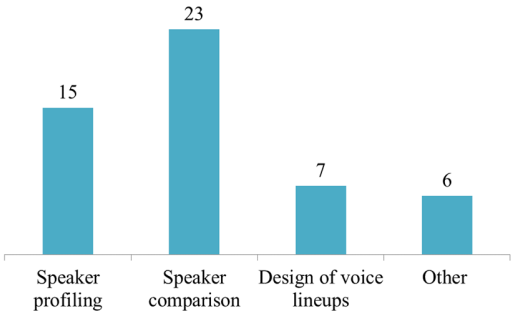

Figure 28. Types of tasks in which forensicists analyze VQ perceptually (count values, not percentages). Answered/Skipped: 26/1.

Most forensicists consider VQ in speaker comparison tasks, followed by speaker profiling, design of voice lineups and other tasks. In this last category, participants answered: "Biometrics, pathological speech assessment, automatic classification of voice disorders, speech recognition"; "transcripts and textualizations"; "blind grouping task" and "transcriptions". Two answers were considered invalid (off topic). As can be observed, most respondents chose more than one option: there were 51 total answers and 26 participants.

3.3. Specific question for voice therapists

Q28. *From a clinical point of view, which is the main use that perceptual protocols have for your assessment of voice quality? (Several answers are possible).*

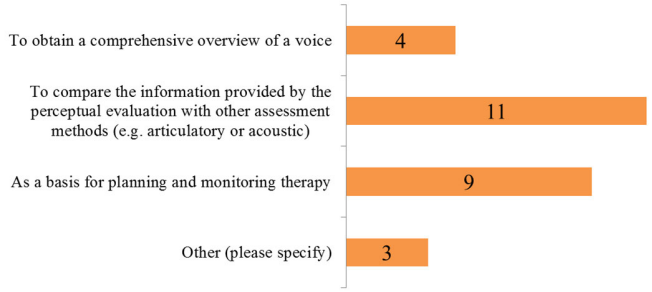

Figure 29. Main uses that perceptual protocols have for clinicians in their assessment of VQ (count values, not percentages). Answered/Skipped: 18/0.

The answer options offered in this question are based on what Beck (2005) and Carding, Carlson, Epstein, Mathieson and Shewell (2001) highlight as possible applications of perceptual assessment. A few respondents marked more than one option, usually a combination of two possible answers. There were 27 different answers by 18 clinicians. Three of them answered 'other'. Since they were required to provide a more specific answer, these were their comments: "It also helps in diagnosis. In VQ there are parameters that suggest or reinforce the anatomical findings (nodules versus intracordal cyst for example). This is very important for me as a Phoniatrician (medical doctor)"; "To compare different moments of therapy or exercises effect on VQ"; "Part of the procedure to diagnose a voice disorder".

4. Discussion

The results of this survey show important differences –as well as some similarities– between clinicians and forensicists as regards the assessment of VQ. First, I discuss their similarities and then I will focus on their divergent answers. Because the latter sometimes show very different methodological decisions taken in their respective professional activities, they will be discussed in more detail.

4.1. Similarities between the answers given by clinicians and those given by forensicists

i. All the clinicians consider VQ in their professional activity (Q1). The same could be said of forensicists because the only exception within this group actually corresponds to just one participant who had only recently started working on FVC tasks, having traditionally focused on other areas related to aural evidence. The high percentage of clinicians and forensicists involved in VQ analyses points to the importance of undertaking a survey on this topic at least for these two professional groups. It is true that a certain bias can be expected in the sense that only participants with an interest in VQ or analyzing VQ may have decided to participate in this survey. However, the survey was designed to consider the possibility that respondents answered 'I don't consider VQ in my professional activity'. In that case, a few follow-up questions inquired about the possible reasons for not considering VQ at all (cf. section 3.1, Q1).

ii. Most clinicians and most forensicists follow some kind of established or known protocol, scheme or rating system (e.g. VPA or GRBAS) for the perceptual analysis of VQ (Q2), even if they use a modified, in-house version of a well-known protocol and not the original version.

iii. As for the questions regarding specifically the VPA protocol, the answers of the respondents are equally heterogeneous within one group and the other. For example, either all of the members of a team who are involved in the application of VPA have received some specific training or at least one person has (Q12). Presumably the training strategy depends on each laboratory but it is not possible to define different trends for either clinicians or forensicists. Besides, the number of clinicians using the VPA is particularly low to draw any conclusions in this respect.

iv. Still within the VPA-specific questions, Q15 aimed to ascertain whether clinicians and forensicists agree with some statements about potential difficulties for the practical implementation of VPA analyses. The pictures in figures 16-18 show large individual variation in the responses within each group and yet on average the degree of agreement with the proposed statements is similar for both groups when taking into account the mean value of the Likert scale. In general, both groups tend to neither agree nor disagree with the criticisms of the VPA regarding (1) comparison of high-quality and telephone-filtered recordings, (2) use of too many labels in the protocol, and (3) bias towards the English language in the VPA design. For this last question in particular, both groups tend to disagree (mean value in the Likert scale equals 2.33 for clinicians and 2.25 for forensicists). This means that on a whole these professionals do not think that the VPA protocol was designed to be used with English speakers only and they do not think that the same settings cannot apply to other languages which they typically find in their professional activity. This is particularly encouraging since the respondents indicated that a large number of languages can be found in their forensic casework and clinical practice. Nevertheless, see point vii of section 4.2 below, where it is highlighted that forensicists work with more different languages than clinicians, and particularly point viii of section 4.2, which shows that some settings are still difficult to relate to particular languages.

v. There are key methodological questions (Q20, Q21, Q22) to which both cohorts responded in a similar way: 50% of clinicians and 50% of forensicists measure inter-rater agreement; 62.5% of clinicians and 78% of forensicists state that more than one person conducts the VQ analyses; 69% of clinicians and 67% of forensicists follow blind procedures. There are, however, differences in how they measure interrater agreement (see point xi in section 4.2).

vi. As to how clinicians and forensicists measure VQ acoustically (Q24, Q25), their answers are incredibly similar, with most of the given answer options ranking in the same order, from most to least used: Praat Voice Report and Long Term Average Spectra, specific software to measure laryngeal features and specific software to measure supralaryngeal features. Only forensicists, but not clinicians, seem to opt for Long Term Formant Distributions too. In terms of specific software, *Praat* ranks first for the acoustic analysis of VQ, which is clearly observable through the wordclouds in Figure 27. However, 'confidential' was the second most frequent response provided by forensicists. Only one clinician gave this answer and it actually corresponds to one of the three respondents that indicated that they work both in clinical and forensic domains. This should be discussed in combination with the large number of skipped answers that are found for forensicists in comparison with clinicians, for instance in Q6 or Q8. Since this is an important difference between both groups, it is further discussed below and in the conclusions.

4.2. Differences between the answers given by clinicians and those given by forensicists

i. Of the two surveyed groups, most clinicians and most forensicists prefer the combined method (perception + acoustics) to approach VQ assessment over any of the other methods (i.e. perceptual or acoustical) in isolation (see Q2). However, while there is a large majority of clinicians who prefer the combined method (78%), followed by the perceptual method (17%), a different trend is observed among forensicists. Their answers are more evenly distributed. Not even half of them prefer the combined method (42%), which is followed closely by the acoustic method (31%). In any case, the perceptual method is chosen by a similar percentage (27%) of the respondents. The low percentage of clinicians using the acoustic method in isolation could point to the fact that this professional group typically receive extensive training in one or several perceptual protocols (see Q5) throughout their professional career, the clinical context being the area where most of those rating schemes were born. In the case of forensicists, perceptual training in VQ protocols is not widespread and almost limited to one rating system, namely the VPA (see Q5 and Q9).

ii. Both groups prefer to circumscribe VQ to phonatory aspects, at least when given the choice among three possible options (simplified here as 'phonation', 'laryngeal and supralaryngeal aspects', and 'more than laryngeal and supralaryngeal aspects'; see full definitions in Q3). However, neither group is homogenous in this respect. Among forensicists, not even half of them (46%) agree on defining VQ as referring exclusively to the vibration of the vocal folds. Interestingly, the second most chosen option is the third definition (i.e. VQ refers to more than laryngeal and supralaryngeal aspects). The full definition is "VQ refers to those characteristics which are present more or less all the time that a person is talking: it is a quasi-permanent quality running through all the sound that issues from his/her mouth. It encompasses more than laryngeal and supralaryngeal features" by Abercrombie (1967, p. 91), as read in Beck (2005, p. 286). No wonder a large percentage of forensicists opt for this definition since Beck's book chapter is a key reference for those forensicists using the VPA protocol (see Q10).

iii. In terms of the specific protocols used to analyze VQ perceptually (Q5), two schemes stand out as the most used: VPA (Laver, 1980; Beck, 2005, 2007) is the preferred method for forensicists (used by 9 out of the 13 respondents who answered this question) and GRBAS (Hirano, 1981) is chosen by most clinicians (9 out of the 14 clinicians answering this question). Furthermore, forensicists and clinicians seem to differ in another aspect: it is not rare for clinicians to use more than one perceptual protocol while the opposite happens with forensicists (a single person selected more than one option).

iv. As for the advantages that the respondents found in the chosen protocols (Q6), we find some similarities and important differences in the most repeated words of their answers. For instance, the fact that it is an 'easy' protocol is mostly valued by clinicians, with four occurrences. Terms like 'simple' and 'quick' are also repeated by the respondents belonging to the clinician group. Forensicists, in turn, repeat terms like 'well-established (protocol)' or 'standard', although some of them seem to appreciate ease of use as well. Logically, the specificities of the professional practice of each group make them choose protocols which allow them to satisfy different needs. Hence, for clinicians it may be important that a perceptual protocol can easily correlate with acoustic measures or vocal-fold-physiology measures while forensicists may rather value that the protocol is well established or well-known, so that they can easily explain its use in court or describe it in a forensic report. What is surprising –and to a certain extent worrying– is that words like reliability or validity are not found among the forensicists' responses while clinicians seem to appreciate that the perceptual protocols have been 'used in scientific studies' or 'evaluated in research', 'with satisfactory reliability and validity' (See full answers in Appendix A).One could argue that the terms "well-established" or "standard" in response to this question do reflect these aspects as well, at least to some extent. However, a certain method may have been long established in a discipline but its validity and reliability may well have never been properly tested.

v. It is worth discussing that a large percentage of forensicists (83%) use a modified version of the perceptual protocol chosen for forensic practice while only 17% of the clinicians seem to need such modifications (Q8). This different trend may be explained by the fact that clinicians may use more than one protocol if only one does not fulfil their needs while forensicists might be simply more used to resort to modified or in-house versions of well-established protocols. It is also remarkable that this question was skipped by more than half of the forensicists. This is further discussed below (see point xii).

vi. The typical length in years of experience using perceptual protocols (Q9) also differs between clinicians and forensicists. The most chosen answer provided by clinicians is 'more than 10 years'; most forensicists chose 'between 1 and 5 years'. This could be indicative of the age of the participants in this survey. Furthermore, Forensic Speech Science is a particularly recent discipline in some of the countries where the respondents come from. While

the responses to this question surely depend on the proportion of junior and senior participants in each professional group, it is also true that the interest sparked by VQ perceptual schemes among forensicists is relatively recent, as it was commented on in the introduction.

vii. Several questions of this survey refer to one protocol in particular: the VPA scheme. Here I focus on the aspects in which clinicians and forensicists diverge. Because the VPA can include prosodic features, temporal organization and other features, depending on the version of the protocol, it was interesting to ask the respondents whether they actually consider this type of aspects (e.g. pitch, loudness or respiratory support) within their VQ assessment protocol or apart from it (Q11). While most clinicians (67%) seem to include these prosodic aspects within their VQ template to characterize voices, forensicists are quite divided in their answers: 38% consider them to make part of VQ; 38% think they are to be considered apart from VQ. Clinicians and forensicists differ, although slightly, in their VPA training methods. While most clinicians seem to be trained through key readings related to the VPA protocol, heterogeneous methods are highlighted by forensicists, who tend to mix reading with practical experience through job training/shadowing, etc. In terms of frequency of use (Q14), the answers of forensicists suggest that on average they use the VPA scheme more often than clinicians do. When asked about the specific languages found in casework or clinical practice, forensicists examine more different languages than clinicians (Q16).

viii. Although in point iv of the previous section (cf. 4.1) it was noted that most clinicians and forensicists do not have any problem assigning VPA settings to languages other than English (Q15-c), in Q18 five forensicists and one clinician pointed several settings that they find more difficult to relate to the particular languages examined in their clinical and forensic practice. Interestingly, the setting 'audible nasal escape' was selected by the clinician. This could be due to the fact that other techniques exist for nasal emission measurement, such as nasometry. One can wonder whether selecting the perceptual label 'audible nasal escape' in this question could actually be independent of the language examined and be related to the methodological decision to measure this aspect objectively and not perceptually. As for the responses of the forensicists, besides 'audible nasal escape' (which is also selected once) we can observe that many settings are marked with their respective 'opposite'. For example: lip rounding and lip spreading, advanced tip/blade and retracted tip/blade, nasal-denasal, lax larynx-tense larynx. In this respect, it is worth noting that some simplified versions of the VPA, such as the SVPA protocol (San Segundo & Mompeán, 2017), achieved a reduction of settings thanks precisely to pairing those settings considered to be 'mutually exclusive'.[5] Perhaps these are the settings for which it is difficult to conceptualize the whole setting dimension, understood as a continuum with two extremes (e.g. tongue fronting, nasality, laryngeal height, etc). See San Segundo and Skarnitzl (In Press) for a more extended discussion about the possible psychoacoustic nature of the different VQ settings.

ix. When inquired about the rarest (i.e. seldom found) settings (Q19), forensicists gave similar answers: 50% of them marked four supralaryngeal settings as rare: 'lip rounding', 'lip spreading', 'labiodentalization' and 'protruded jaw'. Seldom found in the population examined by forensic phoneticians are also: 'pharyngeal constriction' and 'pharyngeal expansion' (selected as 'rare' by 38%). Some clinicians seem to agree with the forensicist group that 'lip rounding', 'lip spreading' and 'labiodentalization' are rare (33% of them mark them as rare versus the 50% of forensicists). Other aspects not marked by any forensicist that clinicians noted as rare were: 'lowered tongue body', 'lax vocal tract' and 'falsetto' (33%). There are not enough clinician participants to talk about a trend;

[5] Note that there were also setting pairs for which only one pair member was considered 'difficult' by the respondents in Q18: close jaw but not open jaw, lax vocal tract but not tense vocal tract, and raised larynx but not lowered larynx.

instead, there are highly individual differences within this group, which would certainly depend on the type of patients and pathologies that clinicians may have found in their individual practice. All in all, the rarity of a setting is particularly useful in FVC when calculating the strength of the evidence. For instance, if both suspect and offender present labiodentalization –supposing this is very uncommon in the relevant population– the likelihood that both suspect and offender are the same person (prosecutor's hypothesis) will be higher than if both suspect and offender shared a VQ setting which also abounds among the relevant population.

x. In terms of methodologies, clinicians and forensicists seem to measure interrater agreement in very different ways (Q23). The six clinicians who replied to the question 'how do you measure interrater agreement', provided the name of a proper statistical method to measure reliability or consistency, sometimes giving a detailed explanation on when and why they may use one method or the other (mainly in research, not in clinical practice). However, among the eight forensicists who responded, only one replied "kappa" without further details. The responses of the other forensicists suggest either methodological subjectivity or lack of statistical awareness.

xi. Finally, a key difference between clinicians and forensicists is that the latter tended to skip more questions than the former. This happens, for instance, in the questions inquiring about the advantages found in their preferred protocol for perceptual evaluation (Q6) or asking whether the used protocol was the original or a modified one (Q8). Skip rates are particularly high again among forensicists when asked about how they measure interrater agreement (Q23). When asked about specific software or techniques used to measure VQ acoustically (Q25), skip rate is not so high because participants are allowed to answer 'confidential', and they choose this option more often than clinicians. The reason why skip rates are commoner among forensicists together with a reluctance to reveal in-house or personal methods should probably be investigated further. It is a trend that might signal a climate of certain mistrust among some professionals within this field or a lack of awareness of the importance of transparency and knowledge sharing.

5. Conclusions

According to the results of this survey, forensic practitioners analyze VQ perceptually more often in speaker comparison tasks (Q27); i.e. when the voice recording of an offender (unknown speaker) is to be compared with that of a suspect or several suspects (FVC tasks), although VQ seems to be useful also in speaker profiling tasks, the design of voice lineups or transcriptions. The purposes of VQ perceptual assessment for voice therapists are mainly two: to compare the information provided by auditory methods with other assessment methods such as articulatory or acoustic measures, and as a basis for planning and monitoring therapy; besides voice disorder diagnosis itself (Q28).

The type of tasks that one group of professionals and the other typically undertake can explain some of the differences found in their responses to many questions in this survey. For example, when asked about the advantages of their preferred perceptual protocols for the assessment of VQ (Q6), clinicians highlight aspects such as ease of use or the fact that the protocol can easily correlate with acoustic measures or vocal-fold-physiology measures while forensicists look for a standard method in the forensic arena. They seem to value that the protocol is well established or well known, which should facilitate its explanation in court or in forensic reports.

The results of this survey show other important differences between clinicians and forensicists as regards VQ assessment, for instance: preference for the GRBAS scale by clinicians and for the VPA scheme by forensicists; preference for original protocol versions by clinicians over modified versions, more often chosen by forensicists. Of all the differences, the most remarkable ones are those concerning methodologies. For example, this survey has revealed that statistical tests which measure

reliability or consistency (e.g. inter-rater agreement measures) are better known by clinicians –even if they highlight that they apply those measures in research and not so much in clinical practice–, while the responses of forensicists may suggest methodological subjectivity or lack of statistical awareness as regards inter-rater agreement. In this line, it is surprising that terms like reliability or validity are not found among any of the forensicists' responses in Q6 while some clinicians seem to highlight as an advantage of a perceptual protocol that it has been 'used in scientific studies' or 'evaluated in research', 'with satisfactory reliability and validity'. It is very important to contrast these aspects with the notable changes that FVC has undergone in the past decade.

The adoption of the likelihood-ratio framework and the quantitative evaluation of the reliability of results is more extended nowadays than it was twelve years ago (Morrison, 2009). Recent investigations show that likelihood-ratio based FVC with higher level features is feasible using very different phonetic features and in a number of languages and experimental conditions (French et al., 2015; Rose & Wang, 2016; San Segundo, Univaso, & Gurlekian, 2019; San Segundo & Yang, 2019). There is therefore an important contrast between the evolution of semiautomatic FVC methods and the evolution of methodologies in the so-called "traditional auditory-acoustic approach". For instance, error rates are commonly accepted and used in Automatic Speech Recognition (ASR). These concepts –together with the main parameters used in that field (namely, mel frequency cepstral coefficients) – were borrowed from ASR by Automatic Speaker Recognition (ASpR) experts. Forensic phonetic practitioners acknowledge and accept that ASpR systems have errors that should be reported so that systems can be improved. In the same way, auditory-acoustic methods with a strong human component have errors that should be measured and reported, but that does not seem to be widely acknowledged.

Therefore, one conclusion that can be drawn from this survey is that more emphasis should be placed on concepts like calibration and error measurement in auditory-acoustic approaches, such as those relying on VQ, particularly if such phonetic characteristics are to be combined with other features in ASpR systems. In the same way that ASpR or semiautomatic FVC have regarded and emulated Speech Technology methods, on the one hand, and likelihood-ratio frameworks widely adopted for DNA profile comparison, on the other hand –borrowing elements from both–, it seems logical that in the VQ arena (which is becoming increasingly popular in FVC) forensicists observe closely the professional practices of clinicians.

It is also remarkable that many questions of this survey were skipped by forensicists but not by clinicians; particularly those questions asking about specific software, name of the protocols used or details regarding methodological decisions. This should be discussed in the context of recent studies which highlight that cross-disciplinary collaborations in forensic sciences are rare and that knowledge sharing should be encouraged more frequently (Donnelly, 2019). There is no clear reason as to why some forensicists decided not to answer many of the posed questions in this survey or answered 'confidential'. One possibility could be that they are more reluctant to share methods because of fear of being judged, although this is quite unlikely since they were informed that their answers would remain anonymous.

All in all, the results of this survey should encourage more collaborations among experts from closely related fields of Applied Phonetics (namely, Clinical Phonetics and Forensic Phonetics), for instance in order to explore together different perceptual protocols or to be trained in the use of other acoustic measurements, statistical techniques, etc.

With this survey I have tried to offer a glimpse into aspects such as: how VQ is most frequently conceptualized, methodological preferences to evaluate it, or typical problems associated with its use in forensic cases versus clinical practice. This type of discussions about the critical issues of state-of-the-art VQ assessments in practice and research is

common among voice therapists (Barsties & De Bodt, 2015) but to the best of my knowledge there was no similar study focusing on the forensic aspects of VQ. On this occasion, my discussion has developed from the answers provided by international experts –mostly forensic practitioners– to a survey designed ad hoc for this investigation. This discussion seems very timely in the current forensic science context, where there is a strong interest in aspects such as cross disciplinary training and collaboration (Donnelly, 2019), bridging the gap between academia and practice (Beresford et al., 2020) and tackling cognitive bias in forensic judgments and peer review, which are typically the result of human assessment and consequently subjective (Mattijssen, Witteman, Berger, & Stoel, 2020). Last but not least, constructive self-criticism is of utmost importance in any forensic discipline, especially in view of the results of recent investigations (Kaplan, Ling, & Cuellar, 2020) showing that a large number of individuals in the United States are skeptical about the quality of forensic techniques, with voice analyses showing a particularly low level of perceived accuracy.

Acknowledgements

This research was partly funded by an IAFPA Grant entitled “How to deal with voice quality in forensic phonetics: a feasibility study towards a simplified perceptual protocol” (Award date: April 2016).

References

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh University Press.

Barsties, B., & De Bodt, M. (2015). Assessment of voice quality: Current state-of-the-art, *Auris Nasus Larynx*, *42*(3), 183-188.

Beck, J. M. (2005). Perceptual analysis of voice quality: The place of vocal profile analysis. In W. J. Hardcastle & J. M. Beck (Eds.), *A figure of speech: A Festschrift for John Laver* (pp. 285-322). Routledge.

Beck, J. M. (2007). *Vocal profile analysis scheme: A user’s manual*. Queen Margaret University College-QMUC, Speech Science Research Centre.

Beresford, D. V., Stotesbury, T., Langer, S. V., Illes, M., Kyle, C. J., & Yamashita, B. (2020). Bridging the gap between academia and practice: Perspectives from two large-scale and niche research projects in Canada, *Science & Justice*, *60*(1), 95-98.

Boersma, P., & Weenink, D. (2005). *Praat v.5.2.01*. Retrieved from www.praat.org

Carding, P., Carlson, E., Epstein, R., Mathieson, L., & Shewell, C. (2001). Re: Evaluation of Voice Quality, *Letters to Editor. International Journal of Language and Communication Disorders*, *36*, 127-143.

Donnelly, R. (2019). Cross disciplinary collaboration in the current market place, *Science & Justice, 59*(6), 678-679.

French, P., Foulkes, P., Harrison, P., Hughes, V., San Segundo, E., & Stevens, L. (2015). The vocal tract as a biometric: output measures, interrelationships, and efficacy, In M. Wolters, J. Livingstone, B. Beattie, R. Smith, Rachel, M. MacMahon […] & J. Scobbie (Eds.), *Proceedings of the 18th International Congress of Phonetic Sciences 2015, Glasgow, UK (ICPhS 18).*

Granqvist, S. (2020). *Tolvan Data 2020*. Retrieved from www.tolvan.com.

Gil, J., & San Segundo, E. (2014). La cualidad de voz en fonética judicial. In E. Garayzábal, M. Jiménez, & M. Reigosa (Eds.), *Lingüística Forense: La lingüística en el ámbito policial y judicial* (pp. 153-197). Euphonía Ediciones.

Gómez-Vilda, P., San Segundo, E., Mazaira, L., Alvarez, A., & Rodellar, V. (2014). Using dysphonic voice to characterize speaker’s biometry, *Language and Law/Linguagem e Direito*, *1*(2), 42-66.

Hammarberg, B. (1986). *Perceptual and acoustic analysis of dysphonia* (Unpublished doctoral dissertation). Karolinska Institute, Stockholm.

Hirano, M. (1981). *Clinical Examination of Voice*. Springer.

Hughes, V., Harrison, P., Foulkes, P., French, P., Kavanagh, C., & San Segundo, E. (2017). Mapping across feature spaces in forensic

voice comparison: The contribution of auditory-based voice quality to (semi-) automatic system testing, *Proceedings of Interspeech*, pp. 3892-3896.

Kaplan, J., Ling, S., & Cuellar, M. (2020). Public beliefs about the accuracy and importance of forensic evidence in the United States, *Science & Justice, 60*(3), 263-272.

Keller, E. (2005). The Analysis of voice quality in speech processing. In G. Chollet, A. Esposito, M. Faundez-Zanuy, & M. Marinaro (Eds.), *Nonlinear Speech Modeling and Applications*. *Lecture Notes in Computer Science* (pp. 54-73). Springer.

Kempster, G. B., Gerratt, B. R., Verdolini, K., Barkmeier-Kraemer, J., & Hillman, R. E. (2009). Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol, *American Journal of Speech Language Pathology, 18*, 124-132.

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge University Press.

Mattijssen, E. J., Witteman, C. L., Berger, C. E., & Stoel, R. D. (2020). Cognitive biases in the peer review of bullet and cartridge case comparison casework: A field study, *Science & Justice*, *60*(4), 337-346.

Morrison, G.S. (2009). Forensic voice comparison and the paradigm shift, *Science & Justice*, *49*, 298-308.

Nawka, T., Anders, L.C., Wendler, J. (1994). The auditory assessment of hoarse voices according to the RBH system, *Language Voice Hearing, 18*, 130-113.

Nemr, K., Lehn, C. (2010). Voz em câncer de cabeça e pescoço. In F. D. M. Fernandes, B. C. A. Mendes, & A. L. P. G. P. Navas, *Tratado de Fonoaudiologia* (pp. 787-802). Editora Roca.

Nolan, F. (1982). The Phonetic Description of Voice Quality by John Laver (Review), *Journal of Linguistics*, *18*(2), 442-454.

Orozco-Arroyave, J. R., Vásquez-Correa, J. C., Vargas-Bonilla, J. F., Arora, R., Dehak, N., Nidadavolu, P. S. ... & Vann, A. (2018). NeuroSpeech: An open-source software for Parkinson's speech analysis, *Digital Signal Processing, 77*, 207-221.

Park, S. J., Afshan, A., Kreiman, J., Yeung, G., & Alwan, A. (2019). Target and Non-target Speaker Discrimination by Humans and Machines. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6326-6330).

Passetti, R., & Constantini, A.C. (2019). The effect of telephone transmission on voice quality perception, *Journal of Voice, 33* (5), 649-658.

Rose, P., & Wang, B. X. (2016). Cantonese forensic voice comparison with higher-level features: Likelihood ratio-based validation using F-pattern and tonal F0 trajectories over a disyllabic hexaphone. In *Proceedings of Odyssey* (pp. 326-333).

San Segundo, E., Foulkes, P., French, P., Harrison, P., Hughes, V., & Kavanagh, C. (2018). Cluster analysis of voice quality ratings: identifying groups of perceptually similar speakers. In *Proceedings of the Conference on Phonetics & Phonology in German-speaking countries (P&P 13)*, Humboldt Universität zu Berlin (pp. 173-176).

San Segundo, E., Foulkes, P., French, P., Harrison, P., Hughes, V., & Kavanagh, C. (2019). The use of the Vocal Profile Analysis for speaker characterization: Methodological proposals, *Journal of the International Phonetic Association*, *49*(3), 353-380.

San Segundo, E., & Gómez-Vilda, P. (2014). Evaluating the forensic importance of glottal source features through the voice analysis of twins and non-twin siblings, *Language and Law = Linguagem e Direito*, *1*(2), 22-41.

San Segundo, E., & Mompeán, J. (2017). A simplified vocal profile analysis protocol for the assessment of voice quality and speaker similarity, *Journal of Voice*, *31*(5), 644-e11.

San Segundo, E., Schwab, S., Dellwo, V., He, L., & Mompeán, J. (2017). Perception of vocal tract tension: Exploring possible prosodic correlates, in *Current trends in experimental phonetics: Cross-disciplines in the hundredth anniversary of Manual de Pronunciación*

Española (Tomás Navarro Tomás), pp. 79-82, 2017.

San Segundo, E. & Skarnitzl, R. (in press). A computer-based tool for the assessment of voice quality through visual analogue scales: VAS-Simplified Vocal Profile Analysis, *Journal of Voice*.

San Segundo, E., Univaso, P., & Gurlekian, J. (2019). Sistema multiparamétrico para la comparación forense de hablantes, *Estudios de Fonética Experimental*, *28*, 13-45.

San Segundo, E., & Yang, J. (2019). Formant dynamics of Spanish vocalic sequences in related speakers: A forensic-voice-comparison investigation, *Journal of Phonetics, 75*, 1-26.

Shewell, C. (2013). *Voice work: art and science in changing voices*. John Wiley & Sons.

Trask, R.L. (2004). *A dictionary of phonetics and phonology*. Routledge.

Tsanas, A., San Segundo, E., & Gómez-Vilda, P. (2017). Exploring pause fillers in conversational speech for forensic phonetics: Findings in a Spanish cohort including twins. In *Proceedings of ICPRS 2017: 8th International Conference on Pattern Recognition Systems* (pp. 32-37).

Appendix A

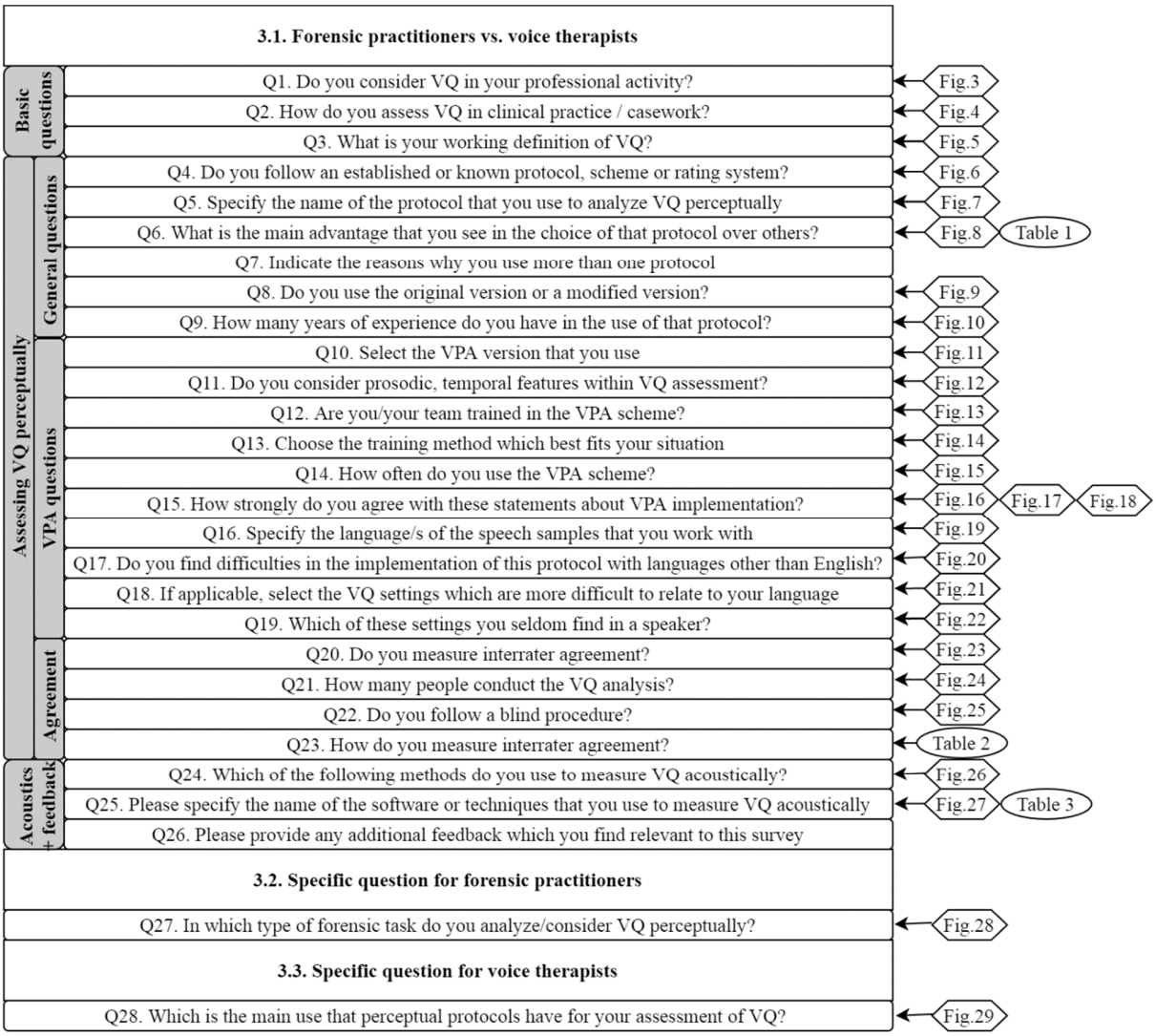

Figure i. Survey questions grouped thematically, with their associated figures and/or tables.

Appendix B. Answers to Q6 (advantages of the perceptual protocol used)

Clinicians	Forensicists
-"Easy to use". [GRBAS] -"Ease of use". [GRBAS] -"It has been evaluated in research, it is easy now that I can do it, it's fast, it is used all over Germany, so there is the possibility to compare results with other clinicians". [RBH-Scale] -"Global description ('G' in GRBAS)" [GRBAS] -"Simple. Quick". [GRBAS+GRBASH] -"Easy, quick, simple, reliable" [GRBAS] -"The Stockholm/Swedish Voice Evaluation Approach was first developed in my doctoral dissertation in 1986 and has since then been used in a number of scientific studies, esp. in Sweden, and proven to reach satisfactory reliability and validity. The SVEA parameters have been proven to correlate well with acoustic measures, such as F0 and F0-range, Long Time Average Spectrum (LTAS) analysis, aperiodicity measures such as waveform perturbation measures, correlogram analysis among others." [SVEA] -"GRBAS is widely used amongst therapists as well as other medical professionals". [VPA+GRBAS] -"It is the only one I know. In fact, I am not a voice therapist, I am a fluency therapist (stuttering). Voice quality is a secondary aspect of my work." [GRBAS] -"The perceptual variables are well-defined and shown to correlate to voice acoustics and vocal fold physiology which is highly relevant and useful in a clinical setting". [SVEA]	-"Its practicality. The complexity of analysis in real cases requires a complicated method which should cover many elements. Only VPA has the potential". [VPA] -"It is very well established". [VPA] -"I believe it captures the dimensions and scalar degrees that are relevant in FVC". [In-house rating system based on VPA and SVPA] -"The standardization of choices". [VPA] -"VPA is well established and used by practitioners around world. It is almost a standard to be used in VQ." [VPA] -"I have not compared different protocols. I trust that the J.P. French protocol is appropriate." [VPA] -"Wider range of features than many scales, flexibility in choosing degrees of presence where non-neutral, sensible division of features into phonation/tension/tract categories". [VPA] -"Not familiar with the other protocols listed". [VPA] -"It is the most well-known system in the UK and there are some sociolinguistic studies which have adopted this and thus there is a wider consensus on the method in the UK. Also the protocol used within the firm, thus there is agreement and understanding within practitioners of how to use and apply the protocol". [VPA] -"Well established and easy to explain/demonstrate in court, easy to describe in a report". [VPA+GRBAS]
Experts working both as clinicians and forensicists	
- "It gives me a fairly wide picture of voice quality aspects, but mostly assesses laryngeal aspects +nasality" [SVEA]	

Table i. Full answers to the question "What is the main advantage that you see in the choice of that protocol over others" (Q6). In brackets, the name of the protocol. GRBAS stands for Grade Roughness, Breathiness, Asthenia and Strain; SVEA for Stockholm/Swedish Voice Evaluation Approach (Hammarberg, 1986); VPA for Vocal Profile Analysis, VQ means voice quality; FVC is Forensic Voice Comparison.

Appendix C. Answers to Q26 (general feedback)

Clinicians	Forensicists
- "Not only VQ is important. Equally important, and have to be measured, are: F0, F0 range, intensity and intensity range". - "I would love to use acoustic measures of VQ, but there is seldom any equipment in German practices. Also I feel not yet confident to measure acoustic parameters". -"In the clinical field the trend is to use short forms such as GRBAS. I find this too restrictive to detect VQ variables for very heterogenic dysphonic voices. The same trend is found in acoustic measures where index measures are more and more common, which includes many acoustic measures. I think that may result in not meaningful measures and that it is not possible to relate VQ measures to specific acoustic variables. I am very interested to know the result of this survey. In Sweden the SLP-students are trained using SVEA and have workshops performing acoustic analyses of dysphonic voices".	-"This is an outstanding topic to survey. The questions do not disclose how much actual research or forensic testing is done by the respondents. I have methods, but limited recent work; others may be more active." -"Our actual work is about quality voice differences between known and unknown forensic voices. Some that we found it is not usually investigated". -"The seldom use of some features is as a consequence of two factors: These particular features are difficult to determine without visual information unless extreme (an area in which the VPA could be developed) and also, that they do not form part of any English variety". -"The acoustic analyses are of course "monitored" by the ears". -"Thank you for the chance to participate! I am intrigued by the evidence generated through this work and will continue to use it in judicial training". -"Thanks for asking. We plan to do a training for VQ". -"Our office is considering to have a training in VPA, since none of us have had this specific type of training".

Table ii. Feedback about the survey provided by the participants. Two answers containing personal information (i.e. names) have not been included. VQ stands for voice quality; F0 for fundamental frequency; GRBAS for Grade Roughness, Breathiness, Asthenia and Strain; SLP for Speech Language Pathology; SVEA for Stockholm/Swedish Voice Evaluation Approach; and VPA for Vocal Profile Analysis.