Article:
San Segundo Fernandez, Eugenia, Foulkes, Paul, French, John Peter et al. (2019) The use of the vocal profile analysis for speaker characterization: methodological proposals. Journal of the International Phonetic Association. pp. 353-380. ISSN: 0025-1003

https://doi.org/10.1017/S0025100318000130

The use of the Vocal Profile Analysis for speaker characterization: Methodological proposals

Eugenia San Segundo, Paul Foulkes, Peter French, Philip Harrison, Vincent Hughes & Colleen Kavanagh

Department of Language and Linguistic Science, University of York

eugenia.sansegundo@york.ac.uk
paul.foulkes@york.ac.uk
peter.french@york.ac.uk
philip.harrison@york.ac.uk
vincent.hughes@york.ac.uk
colleen.kavanagh@york.ac.uk

Among phoneticians, the Vocal Profile Analysis (VPA) is one of the most widely used methods for the componential assessment of voice quality. Whether the ultimate goal of the VPA evaluation is the comparative description of languages or the characterization of an individual speaker, the VPA protocol shows great potential for different research areas of speech communication. However, its use is not without practical difficulties. Despite these, methodological studies aimed at explaining where, when and why issues arise during the perceptual assessment process are rare. In this paper we describe the methodological stages through which three analysts evaluated the voices of 99 Standard Southern British English male speakers, rated their voices using the VPA scheme, discussed inter-rater disagreements, and eventually produced an agreed version of VPA scores. These scores were then used to assess correlations between settings. We show that it is possible to reach a good degree of inter-rater agreement, provided that several calibration and training sessions are conducted. We further conclude that the perceptual assessment of voice quality using the VPA scheme is an essential tool in fields such as forensic phonetics but, foremost, that it can be adapted and modified to a range of research areas, and not necessarily limited to the evaluation of pathological voices in clinical settings.

1 Introduction

1.1 The perceptual assessment of voice quality

Voice quality (hereafter VQ) is broadly defined as the combination of long-term, quasi-permanent laryngeal and supralaryngeal parameters or settings and their associated perceptual effects. Many authors have adopted this broad sense of voice quality, which draws on the view that each of the organs of the vocal apparatus has a bearing on a speaker's VQ (e.g. Laver 1980, Klatt & Klatt 1990, Beck 2005). This interpretation differs from narrower definitions

doi:10.1017/S0025100318000130

which restrict VQ to effects derived from vocal fold activity alone. In this study we have adopted the broad definition of VQ.

Two main ideas underlie the broad concept of VQ. First, VQ is defined as quasipermanent: Abercrombie (1967: 91) points out that VQ refers to 'those characteristics that are present more or less all the time that a person is talking'. Thus many authors also refer to VQ as 'timbre' or the characteristic 'colouring' of a voice (Laver 1975, Trask 1996). Secondly, its definition lies at the intersection between physiologically-grounded qualities and psychoacoustic phenomena. This is so because, even though VQ dimensions derive from vocal production (i.e. states of the larynx and the vocal tract), the resulting qualities are necessarily evaluated and classified through perceptual processes in an intertwined process that remains to be fully understood. Thus, Kreiman & Sidtis (2011: 9) summarize VQ as 'an interaction between a listener and a signal'.

Our aim in conducting the present investigation is twofold:

1. To assess the reliability of Vocal Profile Analysis (VPA) ratings across three analysts by testing different measures of inter-rater agreement.
2. To evaluate the extent to which VPA settings are independent, i.e. to explore the degree of correlation between them.

More generally, our objective is to provide a model for how to make the VPA protocol a workable tool for use by those involved in a range of domains including speech therapy, dialectology, and forensic phonetics. While the study was undertaken as part of a larger, forensically-oriented research project, we do not claim that the end product of the study – a version of the VPA to which various modifications have been made – is one best suited for use by forensic phoneticians. As we explain below, the data set on which it was developed lacks the diversity found in typical forensic recordings. However, the methodology set out here may serve as an example for those wishing to make adaptations of the VPA for forensic applications.

In the next sections we briefly discuss different methods for the perceptual assessment of VQ (Section 1.2), and the main applications, issues and challenges of the VPA scheme – the most commonly used protocol in forensic speech science (Section 1.3). Although these sections serve as a necessary background for the rest of the investigation, readers familiar with VQ analysis may wish to move directly to Section 2 for a description of the materials and methods, leading to the results presented in Section 3. We specifically tackle aims 1 and 2 in Sections 3.2 and 3.3, respectively.

1.2 Systems for VQ analysis

In terms of methodology, VQ can be approached from different perspectives: articulatory, acoustic, and perceptual. Hybrid approaches, however, are also widespread. Articulatory studies typically involve laryngoscopy (Abberton & Fourcin 1978), electroglottography (Kitzing 1982) or measurements of nasal and oral airflow (Hirano, Koike & Von Leden 1968). Acoustic investigations include analyses of different dimensions of the speech signal, such as the Long Term Average Spectrum (Löfqvist 1986, Harmegnies & Landercy 1988). Most investigations, however, are concerned with the search for acoustic correlates of perceptual categories (Eskenazi, Childers & Hicks 1990, Hanson 1997, Keller 2004; see further Maryn et al. 2009, Gil & San Segundo 2013). Correlations between some acoustic measures and a certain – perceptually rated – VQ setting are usually considered an important aid for clinicians as well as a possible solution to the subjectivity implied in purely auditory evaluations of VQ.

Although a range of instrumental and objective measures exists to evaluate VQ, perceptual approaches continue to be the 'gold standard' to which other analyses are compared (Ma & Yu 2005). This is because there are few, if any, single, independent acoustic correlates of individual VQ settings. Hence the importance of understanding what the perceptual assessment of VQ exactly is, and how and why it is used by different voice professionals and researchers.

A wide range of protocols has been proposed for perceptual assessment of VQ. These formal evaluation schemes require the listener to document diverse aspects of a speaker's VQ in some systematic way (Carding et al. 2001). The number of specific VQ features – as well as the rating system used to assess their degree – varies across protocols, and sometimes in accordance with the purpose of the evaluation. Most of them, however, originated in a clinical context with the aim of characterizing different forms of dysphonia and quantifying their severity. Kreiman et al. (1993) identified around 60 different perceptual voice protocols in the USA, the Buffalo III Voice Profile being the most widely used (Wilson 1987). Three main schemes are in use in the UK: the Vocal Profile Analysis scheme (Laver et al. 1981, Laver 1991), the GRBAS Scale (Hirano 1981), and the Buffalo III Voice Profile (Wilson 1987). Two further protocols have been designed more recently and are also widely used: The Stockholm Voice Evaluation Approach (SVEA) (Hammarberg 2000), and the Consensus Auditory Perceptual Evaluation (CAPE-V) (Kempster et al. 2009).

The main reasons why these protocols were created are: (i) to obtain a comprehensive overview of the characteristics of a voice; (ii) to compare the information provided by the perceptual VQ evaluation with other voice assessment methods; (iii) as a basis for planning and monitoring therapy; and (iv) as a way of facilitating communication with other professionals (see Carding et al. 2001, Beck 2005). While these represent the most typical uses of perceptual VQ assessment by a clinician, in the next section we comment on some of the specific uses of these auditory tools – particularly the VPA – which have recently appeared in a wider range of linguistic fields, including forensic phonetics.

1.3 The Vocal Profile Analysis: Applications, issues and challenges

In this section we briefly introduce the main characteristics of the Vocal Profile Analysis (VPA), the protocol used for our study (see also Section 2.2.1). We then outline the main research areas that have shown an interest in using the VPA beyond clinical applications, and finally we highlight some of the problems that its use typically entails.

The main characteristics of the VPA can be summarized as follows (see Laver 2000, Beck 2005; see also Wirz & Mackenzie Beck 1995 for fuller descriptions):

(a) The VPA is a componential (i.e. featural) approach to VQ, involving various components or settings. There is some debate over the degree of independence of these settings. Carding et al. (2001) claim that they are 'independent', while other commentators acknowledge that this is not always the case. Beck (2005) describes the VPA settings as 'more or less independent', while Kreiman & Sidtis (2011) use the term 'quasi-independent'.

(b) Settings have an articulatory basis that gives rise to the perceptual effect that is captured in the VPA analysis. Drawing on the concept introduced by Honikman (1964) in the context of teaching the pronunciation of a foreign language, a 'setting' is defined as 'the common articulatory tendency underlying (or arising from) the momentary actions of segmental performance ... with its auditory consequences' (Laver 2000: 39).

(c) A variable number of settings make up the VPA, depending on the version of the protocol, as it has undergone slight modifications since its inception in the 1980s. Typically between 30 and 40 settings are included. Whether the protocol includes prosodic and other global features (e.g. speaking rate or loudness), as in Beck (2007), or it is limited to supralaryngeal and laryngeal features, its most distinctive feature is its comprehensiveness and exhaustiveness in the physiological domain. Recent versions of the protocol include muscular tension settings besides supralaryngeal and laryngeal settings (Beck 2007), although these settings of overall degrees of muscular tension exercising their effect throughout the vocal system were introduced by Laver (1980).

(d) Individual settings are defined in relation to a 'neutral setting'. This serves as a reference and baseline for raters to judge whether a voice is non-neutral and, if so, to what extent.

Laver & Hanson (1981), Laver (2000) and Beck (2007) provide detailed articulatory and acoustic characterizations of the neutral setting, which is considered purely a 'convenient theoretical construct' (Laver 2000: 40).

(e) A six-point rating scale is used to mark potential deviations from the neutral setting. Scalar degrees 1–3 are considered typical of non-pathological populations, where 4–6 are considered extreme degrees, typically found in speakers with some voice or speech disorder. According to Beck (2007), the rating of voices follows a two-stage process. In a 'first pass', raters note if the voices are neutral or non-neutral for each setting; in the 'second pass', raters specify the exact nature of the deviation from 'neutral'.

Besides its clinical applications (e.g. Scott & Caird 1983, Mackenzie et al. 1998, Webb et al. 2004), the VPA has been used in a number of dialectological, sociolinguistic and forensic investigations. Esling (1978) and Stuart-Smith (1999) describe the VQ of English speakers from Edinburgh and Glasgow, respectively, taking into account a wide range of variables (e.g. age, sex/gender and social background). More recently, Coadou & Rougab (2007) and Wormald (2016) complement their VPA-based assessments of English varieties with either acoustic (Long Term Average Spectrum, LTAS) or articulatory analyses (ultrasound techniques).

VQ is also regularly analysed in forensic phonetics, particularly in forensic speaker comparison (FSC) tasks (Nolan 2005). Note, however, that Nolan (2005) specifies that it is VQ in a holistic sense –not necessarily following a protocol such as the VPA– that is commented on in most forensic reports. All in all, in FSC tasks the voice recording of an offender is compared with that of the suspect using a range of methods (e.g. auditory and acoustic phonetic analyses, automatic speaker recognition systems, or a combination of these; see Foulkes & French 2012 and French & Stevens 2013). It seems logical that VQ plays an important role in speaker identification since classical descriptions of VQ suggest its speaker-specific nature. For example, Laver (1980: 2) states that 'voice quality is one of the primary means by which speakers project their identity – their physical, psychological, and social characteristics to the world'. Other studies highlight the fact that listeners often use voice quality cues to judge personality factors (Scherer 1979, as cited in Kreiman & Sidtis 2011).

Some recent studies have explored the speaker discriminatory potential of VQ. Here we focus only on those investigations applying the VPA protocol. For instance, Stevens & French (2012) explored how telephone transmission affects VPA settings, indicating whether perceived VQ (e.g. nasality, creakiness) increased, decreased or remained the same when comparing high quality recordings and telephone-degraded recordings. We have investigated possible correlations between different long-term vocal tract output measures – including supralaryngeal settings – with the aim of finding how VQ analysis can complement long-term formant distributions (LTFDs) and Mel frequency cepstral coefficients calculated across entire speech samples (MFCCs; French et al. 2015, Hughes et al. 2017).

San Segundo, Tsanas & Gómez-Vilda (2017) propose a preliminary simplification of the VPA which enables the quantification of speaker similarity using Euclidean distances. San Segundo & Mompeán (2017) elaborate on the idea of extracting a VQ similarity index to compare speaker pairs, and also show that it is possible to obtain high intra- and inter-rater agreement using a simplified VPA version.

In addition, earwitness evidence can greatly benefit from research on VPA. As a case in point, San Segundo, Foulkes & Hughes (2016) used VPA ratings by experts to select the most similar-sounding voices in a perceptual experiment aimed at testing perceived speaker similarity by naïve listeners.

All in all, the range of possibilities that the VPA offers is wide. Nonetheless, it has some clear limitations. Nolan (2005) highlights the main obstacles to a more widespread use of the VPA in forensic phonetics: (i) lack of training in the use of the protocol, (ii) practical considerations of time in casework due to the magnitude of the assessment task (although depending on the number of speakers or the length of the samples this may not be the case),

(iii) high variability of VQ settings within a sample (e.g. VQ can be manipulated for pragmatic effects on a temporary basis) and largely unexplored intra-speaker variability between non-contemporaneous recordings and (iv) VQ can be compromised by the distorting effects of telephone transmission, background noise, and emotional speech that often characterize forensic recordings.

Here we are concerned with the issues – and challenges – that emerge from the methodological process involved in the perceptual evaluation by independent raters:

(a) ISOLATION OF SETTINGS. The assumption that listeners are able to separate a perceptual dimension, or setting, from several co-occurring dimensions has been often questioned (Kent 1996, Kreiman, Gerratt & Ito 2007). This problem is shared with other componential analyses of VQ. The added difficulty of the VPA is that the settings are based on the speaker and his / her articulatory possibilities, and not on the perceptual processing of such settings by the listener. As Kreiman & Sidtis (2011: 19) point out, there is no indication of 'whether (or when, or why) some features might be more important than others, or how dimensions interact perceptually'.

(b) SCOPE AND NUMBER OF SETTINGS. Beck (2005: 293) comments on the trade-off between the scope and number of parameters included in the VPA scheme, and its ease of use. The more settings added to the protocol, the more sensitive this is. Unfortunately, that runs parallel with an increased complexity and a potentially decreased inter-rater agreement.

(c) SETTING INTERDEPENDENCE. Kent (1996) suggests that the high degree of interdependence between perceptual dimensions justifies attempts to simplify complex and multidimensional protocols. He argues that 'multiple dimensions not only require more time from the clinician, but they may also yield diminishing returns if the various dimensions are highly correlated' (Kent 1996: 17), a point that also applies to FSC since forensic analyses should rely on non-correlated features in order to avoid overweighting evidence (Gold & Hughes 2015, Rose 2006).

In order to reduce dimensionality, some studies have used factor analyses to investigate how perceptual dimensions overlap and group (e.g. Bele 2007), although those approaches are not without limitations. For instance, the results of factor analyses are strongly dependent on the type of stimuli and rating scales (see Kreiman & Sidtis 2011).

(d) VERBAL DESCRIPTORS AND TERMINOLOGICAL ISSUES. Kent (1996: 16) claims that 'it cannot be taken for granted that any given term used in perceptual assessment will have the same meaning for any two judges, or that two judges will share a verbal description of a clinical speech sample'. Even though the VPA verbal descriptors are mostly based on physiological terms – that is, they are not impressionistic, aesthetic terms (e.g. *bright* or *thin*) – the need for standard and well defined terms across different raters is still key in the use of this protocol.

(e) REACHING AGREEMENT. The question of how to best measure intra- and inter-rater reliability is a longstanding issue in VQ research. It is not uncommon to find low inter-rater agreement reported (see Kreiman & Gerratt 1998), so several methodological proposals have been implemented to solve what Kreiman & Gerratt (2010) call the 'unreliable rater problem': from using fewer scale values to increasing the training sessions of raters.

The few studies which have tested inter-rater agreement using the VPA protocol have failed to generate a consensus over how to best measure it. The kappa statistic is the most widely used measure when ratings are treated as nominal (Webb et al. 2004). However, Beck (2005) provides only raw percentage agreement results, which is a less robust and cruder measure of reliability, as it does not correct for the degree of agreement that could be expected by chance (Stemler 2004, Multon 2010). For instance, when rating a population of 25 non-pathological adult speakers, the two raters in the study mentioned in Beck (2005)

reached maximum agreement (100%) in two settings: *protruded jaw* and *labiodentalization*. Due to the rare occurrence of those settings in a normophonic population, agreement could be unintentionally inflated: it is easier for raters to agree when a setting is rare.

(f) KEY SEGMENTS. One of the advantages of the VPA – and potentially also one of its disadvantages – is that it relies strongly on the existence of certain 'key segments' that would be affected by each setting. They are defined as the 'most susceptible [segments] to the performance effects of a given setting or those on which the auditory perceptual effects are most perceptually salient' (Beck 2005: 297). Beck (2007) provides a detailed account of the main key segments per setting. For instance, lip spreading has a significant effect on segments which are normally rounded (e.g. /u/), whereas /i/ is expected to be marked by lip spreading in any case and hence not susceptible to lip spreading. The use of key segments is intended to serve as an economical listening strategy.

However, it can often be difficult to disentangle what is a long-term setting and what is better described as segmental (Abercrombie 1967). While they can be different outcomes of the same articulatory configuration, segments are short-lived while true settings are long term (Laver 2000).

In summary, then, the use of the VPA presents a number of difficulties. Acknowledging them is a first step towards trying to propose some methodological solutions, which is what concerns us in this investigation.

2 Materials and methods

2.1 Materials

Data for analysis were drawn from the DyViS corpus (Dynamic Variability in Speech; Nolan et al. 2009). This corpus was constructed specifically as a publicly-available tool for experimental work in forensic speech science. It contains recordings of 100 male speakers of Standard Southern British English (SSBE), aged 18–25 years. The corpus is divided into various sections or speaking tasks, including samples of spontaneous speech, read text, and speech transmitted via telephone. For the purposes of the research reported here, recordings from Task 2 were used. Task 2 is a telephone conversation, with the target speaker recorded at the near end of the telephone line (i.e. the acoustic signal was not transmitted through the telephone line). These are high-quality recordings (44.1 kHz sample rate, 16-bit resolution) with around seven minutes net speech, after the audio files were manually edited to remove overlapping speech, background noise, and long portions of silence. According to Beck (2005), VPA analyses should be based on recordings of at least 40 seconds of connected speech, and spontaneous speech provides the most realistic representation of a speaker's habitual VQ.

None of the speakers reported having any speech pathology or hearing difficulty. A few were bilingual in English and another language, but they were all native SSBE speakers. They form a largely homogeneous group of speakers, with the majority of them being or having been students of the University of Cambridge. Therefore, the DyViS population is much more homogeneous than the populations encountered in, for example, forensic or clinical casework. The inter-rater results presented here need to be considered in light of this (discussed below).

Our analyses were based on 99 speakers. One speaker (#080) was excluded because of technical problems with his recording. (The recordings are publicly available, and we therefore refer throughout to speaker numbers as shown in the original DyViS corpus.)

2.2 Methods

2.2.1 Adapted VPA scheme

It was mentioned in the introduction that there are several versions of the VPA scheme. The one adopted here is based on Beck (2007) and comprises 32 features: 21 supralaryngeal, seven laryngeal and four referring to muscular tension (see Figure A1 in the appendix). This version was developed in part at JP French Associates, a forensic speech and acoustics laboratory in the UK, and modified for this investigation in light of a calibration exercise (see Section 2.2.2).

In terms of setting repertoire, the main difference between the VPA protocol used in this study and the one described in Beck (2007) lies in the reduction of settings. For example, both *protruded jaw* and *audible nasal escape* were removed, as they are considered rare or associated with some speech disorder (e.g. *audible nasal escape* only admits the extreme scalar degrees in Beck 2007). Several separate settings in Beck (2007) were merged: *fronted tongue body* and *raised tongue body*; *backed tongue body* and *lowered tongue body*; *creak* and *creaky*; *whisper* and *whispery*. The justification for these mergers lies in the very slight differences in the articulatory strategy necessary to achieve the settings (e.g. 'general laryngeal characterization of creak and creaky voice', as described by Laver (1980: 126); see also Ladefoged 1971: 15). By contrast, *murmur* was included because we considered that this was necessary to complete the range of phonatory possibilities (see Esling & Harris 2005), and this setting is included in the modified version used by JP French Associates.

Besides the division between (i) vocal tract features, (ii) overall muscular tension features, and (iii) phonation features, another common distinction is sometimes made between CONFIGURATIONAL SETTINGS – the majority of settings in the VPA scheme, describing the long-term configuration of the vocal tract – and ARTICULATORY RANGE SETTINGS, a smaller class of settings which relate to the habitual range of articulatory movement. There are three settings of this sort in Beck (2007): range of lip, jaw, and tongue movement. In all cases, the two possible deviations from neutral are minimized and extensive (lip/jaw/tongue) range.

Beck (2007) includes three further setting groups in a supplementary page of the VPA: (iv) prosodic features, including pitch and loudness; (v) temporal organization features, including continuity and range; and (vi) other features, comprising respiratory support or diplophonia. These were not included in the adapted VPA we used since they generally either refer to acoustically measurable features (f0, loudness), or are relevant for pathological speech (diplophonia).

Apart from the differences in the setting repertoire, an important characteristic of the VPA that we used for this investigation lies in the reduction of scalar degrees. As points 4–6 in the VPA are defined as pathological settings, these were removed from our adapted version. Deviation from neutral was thus marked on a three-point scale, where 1, 2 and 3 are defined as SLIGHT, MARKED AND EXTREME, respectively.

An important decision taken for this adapted VPA scheme was not to mark intermittent features. This is a scoring convention useful to characterize speakers who adopt a setting only sporadically. Sometimes a percentage of frequency of occurrence can be written alongside the scalar degree (Beck 2007). However, Ball, Esling & Dickson (1995: 72) note that 'any particular setting will only have an intermittent effect throughout an utterance'. When raters indicate a certain setting in the protocol, it does not mean that all segments, or even all examples of key segments (Section 3.1), are uttered with that setting configuration. As exemplified by Ball et al. (1995: 72), 'a nasalized voice does not mean that all sounds are uttered with a lowered velum; rather the term suggests a perceptually greater use than normal of nasal and nasalized articulations'. Although some raters opted for marking intermittent features in their notes, during the calibrations sessions a firm decision on whether the setting was absent or present, and to which degree, was taken to enable cross-rater agreed profiles and inter-rater agreement measurements. For our final agreed VPAs we retained settings that

were impressionistically in evidence throughout the majority of a recording, but eliminated any that had been initially marked as intermittent if they occurred only occasionally.

A section for 'notes' was added to the protocol in order to admit a first impressionistic and holistic assessment of voices, to be completed while the raters were listening to voices individually. Any perceptual label could be used in this first stage, like 'bright' or 'lively'. In a second step, the raters attempted to establish which VPA setting(s) best suited that first perceptual impression. This part was also useful during the calibration sessions, as raters often found that they were referring to the same perceptual impression, but had then conceptualized it differently within the set of VPA pre-established labels (see also Section 3.1 below). Note that we did not follow the strict two-stage process described by Beck (2007). That is, if raters decided the voice was non-neutral for a particular setting, they generally at the same time decided on the extent to which the voice deviated from neutral.

2.2.2 Perceptual assessment

Three analysts conducted the perceptual evaluation of voices: authors San Segundo (ES), Foulkes (PF) and French (JPF). All were trained in the use of the VPA protocol and had used this protocol for forensic purposes before, in casework, forensically-oriented research, or both. PF and JPF had gathered most of their experience in rating British English speakers, while ES was trained in the perceptual assessment of Spanish speakers. The length of experience was variable, with only JPF having used it extensively in forensic casework. Beck (2007) was used as the main guideline for the assessment.

The perceptual assessment was blind; that is, the raters conducted the listening and rating procedure independently. They could listen to the voices as many times as they wished. Prior to the evaluation, they agreed to focus on the middle of each recording, as it is well known that at the beginning of the recording the speaker may not have settled to his ordinary speaking style. The analysts used the same audio software (SONY Sound Forge Pro 11.0), soundcard (Focusrite Scarlett Solo) and high quality headphones (Sennheiser HD 280 Pro).

The stages of assessment are summarized in Table 1 below. The 99 speakers were analysed in numerical order, with the exception that the first ten rated speakers were randomly selected from the corpus (speakers #006, #009, #022, #025, #028, #032, #046, #063, #072 and #105). These ten were used to pilot the procedure and to enable calibration and procedural consistency between the analysts before embarking on the rest of the corpus. After each analyst had completed the ten VPAs independently, a calibration meeting was held at which the results were compared, problematic perceptual labels were discussed, and differences in analytic strategy were identified (see Section 3.1). The ten samples were later analysed afresh, and the second pass results are included in the overall data set reported below.

Table 1 Perceptual assessment stages and number of speakers evaluated.

Methodological procedure stages	Assessed speakers
1. Pilot experiment: Blind perceptual assessment 2. Pilot experiment: Calibration meeting (a) Joint listening (b) Discussion of ratings	10
3. Complete corpus: Blind perceptual assessment 4. Complete corpus: Calibration meeting (a) Disagreement typology consensus (b) Production of agreed profiles	89

The remaining 89 speakers were analysed independently, and again discussed at a calibration and agreement meeting. This entailed both a process of documenting the different

types of disagreements encountered, and a cross-coder calibration process in order to produce agreed profiles. While the former was necessary for measuring inter-rater agreement, the latter was essential for the calculation of correlations between settings.

2.2.3 Statistical analyses

In view of the two main aims of this study, two types of statistical analyses were conducted. Firstly, in order to assess the convergence over VPA ratings across the three analysts, we used several measures of inter-rater agreement. Secondly, in order to evaluate to what extent VPA settings are independent, we explored the degree of correlation between them.

2.2.3.1 *Inter-rater agreement*

Different inter-rater agreement measures exist to provide an indication of the extent to which two or more judges make the same decisions. These measures are also known as consensus estimates of reliability. For this study we tested two types of percentage agreement as well as Fleiss' kappa:

(a) ABSOLUTE PERCENTAGE AGREEMENT. Measures of overall percentage agreement are considered rough estimates of reliability, as they do not take into account the possibility that agreement may occur by chance. Around 70% is generally considered good agreement (Multon 2010). As there were three raters, we provide pairwise percentage agreement results as well as a mean agreement per setting.

(b) AGREEMENT WITHIN ONE SCALAR DEGREE. This is a popular method of measuring agreement when using the VPA protocol (Beck 1988, 2005; Stevens & French 2012). It represents a more realistic definition of disagreement, occurring in two instances: (i) raters' disagreement on the presence or absence of a setting (i.e. a rating of 0 versus 1/2/3), and (ii) raters' disagreement beyond one scalar degree. In practical terms, this disagreement accounts for differences of '1' vs. '3' in the ratings (since there were no higher scalar degrees in this study, and the situation '0' vs. '2' is already covered as an 'absent' vs. 'present' disagreement). As in the case of absolute agreement, we calculated all pairwise agreements between raters and then averaged the results to provide a single (mean) percentage agreement per setting. This is of interest for the discussion, where we provide some possible explanations for the different or similar pairwise results.

(c) FLEISS' KAPPA (Fleiss 1971). This statistical measure is a generalization of Cohen's kappa for multiple raters. It assesses the agreement between a fixed number of raters assigning categorical ratings to a fixed number of items. It can be interpreted as expressing the extent to which the observed agreement among raters exceeds what would be expected if all raters made their ratings randomly.[1]

The interpretation of the kappa statistic differs considerably from the interpretation of the percentage agreement figure. A value of zero on kappa does not indicate that the raters did not agree at all; it just indicates that they did not agree with each other any more than would be predicted by chance alone. Although the kappa is not devoid of criticism (e.g. Powers 2012), it is recognized as a highly useful statistic when one is concerned that the percent-agreement statistic may be artificially inflated because most observations fall into a single category, for instance when a setting is rare and most ratings fall within the neutral label.

[1] Cohen (1960) originally formulated kappa for the case of two raters:

$$\kappa = (p_a - p_c) / (1 - p_c)$$

where p_a is the proportion of times the raters agree (observed agreement) and p_c is the proportion of agreement we would expect by chance (expected agreement).

Landis & Koch (1977) provide some guidelines for the interpretation of kappa magnitudes: values < 0 indicate no agreement, 0–0.20 slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement, and 0.81–1 almost perfect agreement.

We also carried out a preliminary intra-rater study with the ten voices that were rated twice. However, the subset of ten voices analysed twice in this investigation were, above all, considered part of a calibration exercise. Intra-rater consistency was not deemed a research objective for this investigation and fuller studies on this research line are under way (Klug 2017). Therefore, we just include a brief description of results in Section 3.2.

2.2.3.2 *Correlation measures*

The variables (i.e. the VPA settings) analysed in this study produce nominal data. Nominal measures of correlation are: Phi, Cramer's V and Contingency Coefficients. All three measures are based on the prior calculation of the Chi-Square statistic and all three are used to determine the degree of association that exists between variables. In this case we used Contingency Coefficients (*C*), as there are three or more values for each nominal variable, with an equal number of possible values: 0, 1, 2 and 3.

Given the nature of nominal data, the obtained values for *C* always fall along a range from 0 to 1 (i.e. negative correlations are mathematically impossible). This is the reason why alongside *C* we also provide Spearman's rank correlation coefficient (Spearman's *r*). This statistic gives a value between +1 and −1 inclusive, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation. The difference between using one statistic and the other was negligible for most settings (see Section 3.3), but knowing whether a correlation is positive or negative is of interest for this study.

We considered only 13 settings for the analysis of correlations. This subset corresponds to the settings that occur with higher frequency in the examined corpus ($\geq$ 10% of occurrence): *advanced tongue tip*, *fronted tongue body*, *nasal*, *raised larynx*, *lowered larynx*, *tense vocal tract*, *lax vocal tract*, *tense larynx*, *lax larynx*, *creaky*, *whispery*, *breathy* and *harsh*.

This decision was taken because including all the 32 settings would result in a high number of spurious correlations, for instance due to the fact that numerous settings were never or seldom present (e.g. *tremor* or *falsetto*), which would increase their likelihood of being correlated (i.e. these settings have just '0' values for all examined speakers). Besides, it is well known that performing a large number of correlations increases the chances of finding a significant result by pure chance (Type I error). For studies with multiple correlations, focusing on the most relevant variables or the use of multivariate statistics is advocated in Curtin & Schulz (1998), together with Bonferroni correction. In this investigation we avoided spurious results by both (i) selecting a suitable subset of settings with enough variability of ratings and thus sufficiently present in this population; and (ii) applying Bonferroni correction. This type of correction implies lowering the alpha value to account for the number of comparisons being performed; in this case, 78 correlation analyses (each setting with every other setting). The resulting alpha values ($\alpha_{altered}$) considered to determine if a correlation was significant in this study were: .0006 (for $\alpha_{original} = .05$), .0001 (for $\alpha_{original} = .01$) and .000012 (for $\alpha_{original} = .001$).

3 Results

3.1 Results of the calibration sessions

A number of practical issues emerged from the calibration sessions. Some of them are closely related to the type of problems described by previous authors and summarized in Section 1.3

above. Nonetheless, we list these issues here with more detail and provide examples extracted from our own investigation where possible:

(a) SEGMENTS VS. SETTINGS. It was noted by the raters that sometimes a feature was perceptible, but restricted to few segments. For instance, fronted or retracted tongue tip was perceived on /s/, or fronting of tongue body on GOOSE and GOAT vowels (Wells 1982). Likewise, labiodentalization was most likely perceived just for /r/ (the variant [ʋ] is common in British English – Foulkes & Docherty 2000). Beck clarifies how the VPA is intended to be used in such cases:

> [T]he task is to identify those features which are common to all, or at least SOME SIZEABLE SUBSET, of the segments in a sample of speech ... The auditory effects of advanced or retracted tip are often most marked on /s/, but the judge should check that any deviation from alveolar placement in generalized THROUGHOUT the whole set of segments. It is not uncommon for an accent, or an individual, to be characterized by non-alveolar pronunciation of only one of the set of susceptible segments. In this case it would be more appropriate to view the non-alveolar production as a segmental feature rather than a habitual setting adjustment. (Beck 2007: 5, our emphasis)

Following this advice, we agreed to mark a perceptual effect as a setting only if most or all relevant segments were affected by the proposed setting. Labiodentalization, for example, was barely used in the final agreed versions, as it almost always affected /r/ alone and was therefore classified as a segmental effect.

(b) SCALAR DEGREES. Some difficulty arose from trying to adapt our three-point scalar degree scale from the original six-point scale, as some of the guidelines in Beck (2007) seem inconsistent (e.g. degrees 4 and 6). For instance, in Beck (2007: 7) we find that degree 3 means 'the strongest setting that may reasonably be expected to act as a regional or sociolinguistic marker', degree 4 means 'beyond the limits of any normal population' but degree 6 is again 'the most extreme adjustment of which the normal, non-pathological voice is capable'. For this reason, we found it helpful to define our own three labels precisely, and to map them against Beck's. As a case in point, raters could aim to avoid what Beck defines as 1 ('just noticeable') as it is probably useless for e.g. forensic purposes. Our degree 1 (SLIGHT) equates to Beck's 2 ('confident that the setting is audible'), our 2 (MARKED) would equate to Beck's 3, and our 3 (EXTREME) would be anything in her 4–6 range. In any case, the DyViS corpus is quite homogenous, as mentioned above, so any marked speakers were likely to have been screened out during participant recruitment. Therefore, degree 3 was seldom used in this investigation.

(c) CORRELATIONS. As some authors have noted before, certain settings often – but not always – occur together in the perceptual realm, and logically ought to from an articulatory point of view. Beck (2007) notes the clustering of:

- lax larynx and lowered larynx, low pitch, and breathy/whisper
- tense larynx and raised larynx, high pitch, and harshness

In the course of our assessment, we also observed the patterns Beck lists above, and we further noted, among others, the following main correlations:

- lowered larynx – pharyngeal expansion
- advanced tongue tip – fronted tongue body

(d) DEFINITION OR PERCEPTUAL ISSUES WITH SPECIFIC SETTINGS.

(i) *Breathy and whispery*. The distinction between breathy and whispery is somewhat problematic, as their perceptual effects sometimes overlap, or there is no clear consensus on how to define them. Laver (1980), for instance, points out that there is no clear perceptual boundary between whispery and breathy voices, which suggests that there

may be a continuum between phonation types, in turn making it difficult for raters to decide between one label or the other (sometimes using both). However, Laver (1980: 134) provides some guidelines for distinguishing between them:

> [F]rom an auditory point of view, it is practical to use the label 'breathy voice' for the range of qualities produced with a low degree of laryngeal effort, and where only a slight amount of friction is audible. If one thinks of the friction component and the modal voice component as being audibly co-present but able to be heard individually, then the balance between the two components in breathy voice is one where the modal voice element is markedly dominant. 'Whispery voice' can then be used for phonations produced with a greater degree of laryngeal effort, and where a more substantial amount of glottal friction, from a more constricted glottis, is audible. The audible balance between the friction component and the periodic component is different from that in breathy voice; the friction component is more prominent than in breathy voice, and on occasion may even equal the periodic component, (and sometimes dominate it strongly, as in 'extremely whispery voice').

Instrumental studies indicate a difference in epilaryngeal activity between whisper/whispery voice and breath/breathy voice, with greater constriction for whisper (Esling & Harris 2005, Honda et al. 2010).[2]

In this investigation we used a perceptually high rate of glottal friction as a key criterion for evaluating a voice as whispery, as opposed to breathy voice, since the criterion of the greater degree of adductive tension of the vocal folds seems more subjective to assess on an auditory basis. However, we still acknowledge difficulties in rating these settings. This could also be related to the existence of different types of whisper, as suggested by several authors. Catford (1977), for example, notes that some types of whisper may be made with the arytenoidal end of the glottis. Likewise, Laver (1980, 1994) mentions that the glottis may be narrowed across most or all of its length, even though it is said that whisper is generated with a posterior triangular opening of the glottis and with the anterior portion adducted. While discussing the use of VQ diacritics in the ExtIPA (extensions to the International Phonetic Alphabet), Ball et al. (1995) recognized that anterior and posterior locations should be marked for this setting, which entails the use of specific diacritics. For our purposes here, these discussions mean that the whispery voice does not represent a homogenous phonation type and can entail worse inter-rater agreement.

(ii) *Murmur*. Ball et al. (1995) discuss whispery voice as synonymous with murmur. This is perhaps the reason why the latter does not appear in the VPA protocol described in Beck (2007) or older versions of the scheme. In Esling & Harris (2005) we find another reference to 'vocal murmur' as 'Bell's category of "whisper and voice heard simultaneously"' (Bell 1867), which in Esling & Harris' words 'persists throughout the literature as "whispery voice"' (Esling & Harris 2005: 367). We found it useful to add this setting into the protocol as it seemed to fill a gap in the repertoire of phonatory settings.

The three-way distinction (*whispery*/*breathy*/*murmur*) used in casework at JP French Associates is as follows: *Whispery* refers to episodes of voicelessness in normally voiced portions, with audible abrasion of airflow at the larynx – which is tense – and in the supralaryngeal tract. *Breathy* refers to normally distributed voicing, but with audibly high airflow accompanying voiced and voiceless portions. In *murmur* voicing is perceived as 'buzzy' owing to very weak adduction and some abduction of the vocal folds, which are very lax. Beck (p.c) has described this as 'come to bed' voice.

(iii) *Creaky and harsh*. Almost all the speakers in our data had some sort of creakiness, which is likely indicative of a social group pattern, but there were also clearly different types of creak. An interesting case is #032. Two raters classified this speaker as creaky, while the third rater noted 'hard modal akin to creak'. Very strong glottal pulses could be perceived,

[2] We are grateful to an anonymous reviewer for alerting us to these observations.

which were indeed clear in the acoustic signal, but no label quite captured it for this rater prior to the calibration meeting (creaky was then agreed as the best available label).[3]

Likewise, some raters found it difficult to decide whether some voices fell within the creaky or harsh category. This happened when the type of perceived creak was one involving a particularly tense larynx (e.g. speaker #018). According to Laver (1980: 122), low fundamental frequency distinguishes creaky from harsh voice, 'which is otherwise somewhat similar'. On the one hand, this is evidence that (some types of) creak can be confused with (some types of) harsh phonation, which would explain lower inter-rater agreement results for this dimension (see Table A1 in the appendix). On the other hand, it provides a clue for disentangling when to use one label or the other: harsh voices seem to have fundamental frequencies consistently above 100 Hz, and creaky ones consistently below 100 Hz for men (Michel 1968, as cited in Laver 1980). See Keating & Garellek (2015) for an acoustic typology of creaky phonation modes.

Harsh phonation may also be differentiated from creak via greater intensity, which would explain the correlation between these phonation qualities and pitch.[4]

We also avoided rating creakiness only on the basis of hesitation points or ends of inter-pause stretches, where it is well known that creak appears regularly. For instance, as a paralinguistic regulator of interactions, some speakers use creaky voice to signal completion of their turn and thus yield the floor to an interlocutor (Laver 1994).

(iv) *Lax* and *tense vocal tract*. Several speakers (e.g. #006) had notably lax articulation, manifesting as marked undershoot on obstruents and extreme reduction in unstressed syllables. We first considered whether this should count as part of the VPA settings, and if so, as what? We decided to use the label *lax vocal tract* to mark this laxity in articulation, even though this is not exactly the meaning implied by Beck (2007), where laxness is basically associated with *open jaw*, *nasal*, and *minimised range* of lips, jaw and tongue. It is also impressionistically described by Beck as a 'muffled quality of the voice'. Noting that Beck suggests considerable correlation with other settings, and that 'muffled' is a rather subjective and vague description, we found it useful to keep our ad hoc usage of *lax vocal tract* as a single label to refer to markedly reduced articulation.

By extension, tense vocal tract was adapted to capture very precise articulation. In other words, this setting corresponds to speakers who fully articulate vowels and consonants for which the norm (in this dialect) admits weaker allophonic articulations. The label was also used to capture fortitions, such as strong stop releases.

One problem that arose from our usage of *tense*/*lax vocal tract* is that some speakers showed combinations of both: typically they produce consonants and stressed syllables rather precisely or in citation-like form, whereas unstressed syllables were heavily reduced. This appears again to be a group pattern for young SSBE speakers. A subsidiary problem – resulting from the assessment method itself – is that the presence or absence of these settings should not be evaluated on segments where the speaker is spelling or confirming elicited words. In those cases, the speaker tends to hyperarticulate. The focus of the rater should instead be on spontaneous segments.

[3] Following the recommendation of an anonymous reviewer, we acknowledge the 'problematic' definition of creaky, as it combines (i) low longitudinal tension of the vocal folds, making vibration slow and irregular (Eckert & Laver 1994) with (ii) constriction of the aryepiglottic sphincter (Esling & Harris 2005). Despite the clear link between low fundamental frequency and creaky, it is worth mentioning that Laver (1980: 31) actually highlights the natural link between lowered larynx and other type of phonation: breathy voice. Furthermore, we find appropriate to repeat here that our calibration session results (and relatively low inter-rater reliability scores) provide evidence that (some types of) creak can be confused with (some types of) harsh phonation.

[4] We thank an anonymous reviewer for this useful information about the role of intensity in differentiating creaky and harsh phonation qualities.

Finally, as all minimized and extensive range labels seemed to also represent the tense/lax vocal tract dichotomy, we did not considered it a disagreement if one rater had ticked the box of any of the extensive range settings and another had opted for one of the vocal tract tension settings.

(v) *Lowered larynx* and *pharyngeal expansion*. With some speakers (e.g. #006 and #032) raters also used impressionistic labels such as 'booming' or 'resonant' voice (in the first pass). During the calibration sessions, all raters agreed that it was partly due to low f0, and partly to lowered larynx or pharyngeal expansion, but we did not feel confident in determining whether the enlargement of the pharyngeal cavity was achieved by its distension in the vertical or the horizontal dimension.

(e) SYSTEMATIC USE OF OPPOSITE LABELS. Our main concern when evaluating inter-rater agreement during our calibration meetings were those cases where opposite settings were marked by different raters. For example, when one marked *constricted pharynx* and one of the other analysts marked *expanded pharynx*, there were two possible causes:

(i) *Instability of the setting within the speaker*. We were listening to sound samples of several minutes' duration, in contrast with Laver's recommendation to conduct VPA on around 40 seconds speech (Beck 2005). It could thus be the case that both settings occurred.

(ii) *Different understanding of what the setting means*. In the following section we explain that many of these disagreements were actually due to labelling issues, and we propose not to treat them in the same way as proper disagreements.

(f) TYPES OF DISAGREEMENTS. In order to conduct the correlation and inter-rater tests we agreed on a typology of disagreements. This would be useful for (i) creating a single and agreed version of VPA profiles, necessary for testing correlation (where maximum agreement needed to be achieved); and (ii) keeping individual ratings, after calibration sessions, as close as possible to original ratings in order to provide truthful inter-rater agreement results.

To this end, we distinguished two main types of disagreements: (i) PROPER DISAGREEMENTS, which reflect an agreed error on the part of a rater (i.e. during the calibration sessions the rater agreed that the use or non-use of a label was incorrect); and (ii) LABELLING ISSUES, to account for those cases where there were good reasons to think labels were used differently by different raters while they refer to the same perceptual aspect. In other words, we had systematic agreement and within-rater consistency in perceptual judgments, but differences in labelling. For example, ES regularly used *tense larynx* as equivalent to *harsh phonation* for PF and JPF. Likewise, there was often systematic use of different labels known to correlate in a given perceptual category, such as *retracted tongue body* and *constricted pharynx*.

These labelling disagreements were corrected or adjusted in the calibration sessions so that they did not count either for correlation or inter-rater tests (i.e. after discussion, the diverging rater converged with the others in the use of labels). Type (i) differences between raters were counted as proper disagreements in order to calculate inter-rater agreement (i.e. no amendment was made in the individual ratings). However, for the creation of a cross-rater agreed version, any clear errors of this sort were amended after discussion so that a consensus-based agreed VPA profile could be created per speaker.

Essentially, the proposed methodology thus consisted of two phases:

- If any analyst registered a score > 0 for a speaker (i.e. setting presence), we listened collectively to the recording, with the rater(s) who had missed that setting focusing on it.
- If analysts agreed within one scalar degree on the presence of a certain setting, we accepted the mode score for the creation of a single VPA profile version.

3.2 Inter-rater agreement results

The first test for measuring inter-rater agreement was percentage agreement, both absolute and within one scalar degree (Table A1 in the appendix). As expected, results improve in most of the settings when it is measured within one scalar degree. However, it is considerably better in seven out of the 32 settings: *advanced tongue tip*, *fronted tongue body*, *nasal*, *tense vocal tract*, *lax vocal tract*, *creaky* and *breathy*. Overall, agreement is very good (> 70%), although it is strongly setting-dependent, as we will explain in more detail later. Table A1 also provides information about the frequency of occurrence (%) of the settings in the examined corpus (first column). Note that the more frequent the setting, the lower the inter-rater agreement, as it was expected.

In a second step, a chance-corrected measure was tested: Fleiss' kappa (κ; Table A2 in the appendix). This statistic presented some issues in settings where two conditions applied simultaneously: (i) all of the raters had attained 100% agreement, and (ii) they all selected the same variable value for every speaker. For instance, *falsetto* was never recorded for any of the speakers (although it was noted as intermittent for a few speakers) and therefore all raters agreed on assigning '0' to all of the speakers. This is called the 'invariant value' scenario, explained in Freelon (2010).[5]

If observed and expected agreement was the same (when rounded to two decimal places),[6] even if not all raters agreed on the use of the same variable value (for instance in *close jaw*), the resulting κ was too low (even negative) to be considered reliable. Given the effect of invariant values on the calculation of this statistic, κ appears as undefined in these cases.

When the percentage agreement between at least two raters is perfect (100%) but this is not always due to the use of the same variable value (e.g. always neutral) but to the use of other possible values, κ values are among the highest. This is the case for *extensive lingual range* (0.77) and *murmur* (0.83).

All in all, most of the settings reached very good agreement: two settings obtained values between 'almost perfect' (κ = 0.81–1) and 'substantial' (κ = 0.61–0.80) agreement: *extensive lingual range* and *murmur*; six of them fell within the 'moderate' agreement category (κ = 0.41–0.60): *minimized mandibular range*, *minimized lingual range*, *pharyngeal constriction*, *pharyngeal expansion*, *raised larynx* and *lowered larynx*; six further settings were classified as 'fair' agreement (κ = 0.21–0.40): *advanced tongue tip*, *denasal*, *tense vocal tract*, *lax vocal tract*, *tense larynx* and *lax larynx*; only three attained 'slight' agreement (κ = 0–0.20): *lip spreading*, *fronted tongue body* and *nasal*. The ten remaining settings obtained undefined κ values, which is due to their high absolute percentage agreement (giving rise to invariant values). Therefore, they also represent good agreement results, but a chance-corrected measure cannot be provided in those cases.

Despite the small number of voices available for intra-rater consistency testing, the preliminary results are promising. Agreement within a rater ranged between 93% and 96% when all settings were considered. Results considering only the settings used more frequently across the corpus as a whole (here taken as > 60%) were also very good. Percentage agreement

[5] When this situation happens, Fleiss' kappa values are undefined. According to the kappa formula (see footnote 1), expected agreement in the above cases becomes 1. The kappa equation would then be: (observed – expected) / (1 – expected) = (1 – 1) / (1 – 1). This leads to division by zero, which basic arithmetic tells us is undefined. This happened in the following settings, which obtained 100% agreement, because they were not found in any of the speakers by any of the raters: *extensive labial range*, *minimized labial range*, *open jaw*, *falsetto* and *tremor*.

[6] Observed agreement and expected agreement were almost the same, but not exactly. For instance, observed agreement for *close jaw* was 0.9730 and expected agreement 0.9734 (even though in Table A2 they are rounded to 0.97). This implied that the Fleiss kappa yielded a negative κ, meaning that agreement was worse than chance. Given the unlikely event of the above and considering the odd effect of invariant values on the calculation of this statistic, we treat these cases as undefined values. Apart from *close jaw*, this only happened in: *lip rounding*, *labiodentalization*, *extensive mandibular range*, *retracted tongue tip* and *backed tongue body.*

ranged between 73% (*fronted tongue body*) and 87% (*creaky* and *nasal*). The only other setting that occurs very frequently in this corpus is *breathy*, with an average intra-rater agreement of 83% across the three raters.

3.3 Correlation results

Table 2 below shows positive and negative correlations found between pairs of settings. We only show those correlations which are higher than |.35|, that is, higher than weak–moderate correlations. Using the guide that Evans (1996) suggests for the interpretation of the absolute value of *r*, none of the correlations between settings is very strong ($r = .80–1.0$). However, a strong positive correlation is found between *raised larynx* and *tense larynx* ($r = .62$) and a strong negative correlation between *lax vocal tract* and *tense vocal tract* ($r = –.65$). The rest of the correlations are moderate ($r = .40–.59$), with a weak to moderate correlation ($r = .20–.39$) found between *creaky* and *lowered larynx*. Most correlations were highly significant, even after Bonferroni correction (see Section 2.2.3.2).

Table 2 Correlated VPA settings using Contingency Coefficients (*C*) and Spearman's *r*, ordered from highest to lowest following the *C* ranking. Italics: negative correlations (Spearman's *r*).

Correlated VPA settings		*C*	Spearman's *r*	*p*-value
Raised larynx	Tense larynx	.58	.62***	5.38×10^{-12}
Harsh	Tense larynx	.57	.36*	.0003
Lax larynx	Lowered larynx	.52	.57***	7.92×10^{-10}
Creaky	Lax larynx	.45	.46***	2.09×10^{-6}
Advanced tongue tip	Fronted tongue body	.41	.38**	8.84×10^{-5}
Creaky	Lowered larynx	.35	.35*	.0006
Creaky	Whispery	.37	*–.36**	.0003
Lowered larynx	Tense larynx	.46	*–.47****	7.35×10^{-7}
Creaky	Raised larynx	.44	*–.43****	1.11×10^{-5}
Lax larynx	Raised larynx	.47	*–.51****	8.62×10^{-8}
Lowered larynx	Raised larynx	.51	*–.55****	3.01×10^{-9}
Lax larynx	Tense larynx	.57	*–.66****	1.33×10^{-13}
Lax vocal tract	Tense vocal tract	.61	*–.73****	7.51×10^{-18}

* $p < .0006$, ** $p < .0001$, *** $p < .000012$ (with Bonferroni correction)

4 Discussion

4.1 Reliability of ratings

The first aim of this study was to assess the reliability of VPA ratings across three different analysts. For that purpose, we tested several measures of inter-rater agreement. The results showed that overall agreement – considering absolute values – is very good for most settings, with a considerable number reaching 100% agreement. However, seven settings out of 32 improve considerably if percentage agreement is measured within one scalar degree: *advanced tongue tip*, *fronted tongue body*, *nasal*, *tense vocal tract*, *lax vocal tract*, *creaky* and *breathy*. Interestingly, in the discussion which followed the calibration meetings, all these settings occupied much of the raters' discussion, suggesting that there were clearly some issues involved in their definition, labelling or perceptual salience. It seems logical, therefore, that the agreement reached for them was not as high as with other settings.

The use of percentage agreement presented some problems that were discussed in relation to the perfect agreement reached in settings that were simply not present in this population.

When we calculated Fleiss' kappa – typically considered a more reliable inter-rater measure – the issue of invariant values was revealed. It seems that this chance-corrected measure, which relies on the calculation of expected and observed agreement, is not the most appropriate for this type of variables, where raters often agree completely because they only use one value of the variable: most typically neutral. While other inter-rater measures could be investigated for those specific cases in future studies, the kappa analysis proves useful overall, yielding very good agreement results for most settings.

While most κ values lie between fair and moderate agreement, these results are considerably better than those reported by Webb et al. (2004), also using the VPA protocol. Although it is not advisable to compare kappa values across different studies (Uebersax 1987), the highest κ in Webb et al. (2004) is 0.32, obtained for *whispery*, followed by 0.24 for *harshness*. Other settings achieve much lower agreement (e.g. *lip rounding*/*spreading* 0.03 or *pharyngeal constriction* 0.07). Indeed, Sellars et al. (2009) found that many studies focusing on the GRBAS scheme report the highest κ as no better than 'moderate', for overall grade. As for a previous study reporting inter-rater agreement on VPA settings, Beck (1988, as cited in Beck 2005) only provides percentage agreement results for a normophonic adult population, with only two raters. For a number of settings, inter-rater agreement was no higher than 50%: tongue tip (36%), tongue body (28%), laryngeal tension (48%), or whisperiness (48%). Inter-rater results for pathological voices using the VPA scheme were also modest, in view of the results presented in Wirz & Mackenzie Beck (1995). Inter-rater reliability fluctuates widely across studies, regardless of the characteristics of the speaker (i.e. normophonic or pathological), which confirms the need for a theoretical framework to explain such variation (Kreiman et al. 1993).

The pairwise results for this study show that there were no striking differences between the agreements reached by any pair of analysts versus the other pairs. This is especially revealing, taking into account both native language differences and the different experience reported by each rater. In turn this provides strong evidence that calibration sessions are necessary and useful. As a case in point, *murmur* achieves the highest agreement of all settings. This was not included in the guidelines used in Beck (2007), so raters agreed on an ad hoc definition of this setting, and in view of the results they were mostly coincident in when and how to use it.

All things considered, it seems clear that there are idiosyncrasies and biases in each individual analyst, i.e. strengths and weaknesses in their assessment of specific settings. Indeed, the analysts were aware of such issues while conducting the analyses. Some biases reflect personal interests, prior assumptions, and possibly difficulty in perceptual categorization or separation of some dimensions. We think this is true for any sort of phonetic analysis, but a team approach seems to resolve most of the errors and strongest individual biases.[7] Future studies will, nevertheless, include more raters. The small sample considered in this study – although in line with previous studies – may have played a role in the results obtained.

4.2 Correlation of settings

The second aim of this investigation was to evaluate to what extent VPA settings are independent. All the correlations that were perceived during the meetings held by the raters (see Section 3.1) were confirmed using statistical tests, although with varying strength. The strongest positive correlation was found between *raised larynx* and *tense larynx* ($r = .62$)

[7] Following the recommendation of an anonymous reviewer, in order to reiterate the point that a team approach is optimal in forensics, where it is now mandatory to have some sort of checking, we refer the reader to the Forensic Regulator's Codes of Practice and Conduct for forensic science providers and practitioners in the Criminal Justice System (Forensic Science Regulator 2016a) and the accompanying appendix for Speech and Audio Services (Forensic Science Regulator 2016b).

and the strongest negative correlation between *lax vocal tract* and *tense vocal tract*. (r = –.73). While the former points to articulatory configurations that co-occur – hence they are also correctly assessed perceptually – the latter suggest that the raters managed to define two opposite settings in a satisfactory, albeit predictable, way: they do not co-occur in perception, and while they can in principle both be used by the same speaker at different times, most did not in this corpus.

Interestingly, two of the most strongly correlated setting pairs include *tense larynx*, which tends to co-occur with *raised larynx* and *harsh* phonation. The combination of laryngeal tension and raising the larynx, or a harsh voice and tensing the larynx was expected.[8] As a matter of fact, the cluster analysis carried out by San Segundo et al. (2018) revealed that speakers with *tense larynx*, *raised larynx* and *harsh* phonation cluster together and are clearly separated from a second cluster including speakers with *lax larynx*, *lowered larynx* and other types of phonation. Likewise, *advanced tongue tip* and *fronted tongue body* show a weak–moderate correlation. Most of these correlations seem to be physiologically motivated. However, Gold & Hughes (2015) point out that correlations in FSC may be due to sociolinguistic reasons too (e.g. belonging to the same speech community and adopting the same voice patterns).

All in all, we consider that none of these correlations means that any particular pair of settings should be collapsed. None of the correlations appear strong enough to suggest that the settings involved should be collapsed to a single dimension. On the contrary, it seems that raters may still find a speaker's larynx raised but not perceptibly tense. Therefore, it seems justified to keep both labels, at least in order to characterize this population.

4.3 VPA for forensic speaker characterization: Challenging settings and issues for future consideration

Finally, we had a general objective, which was to provide a methodological example of how the VPA can be adapted for different purposes, such as forensic speaker characterization. For that aim, we provided an account of practical issues derived from the different calibration meetings held by the raters (Section 3.1). While the discussion of these results was in part intertwined with the detailed account of methodological issues, the main points can be summarized as follows.

First, some settings appear less perceptually salient than others, at least to these raters and in respect of this particular corpus. Such settings present some difficulties in terms of agreeing on a definition. As expected, these difficulties ran parallel with a slight-fair agreement achieved by the three raters, although not in all cases. For instance, *vocal tract tension* or *creaky phonation* required extensive discussion by raters, and the inter-rater results were indeed fair-moderate. *Nasal* or *fronted tongue body*, on the other hand, did not provoke much discussion in the meetings but both settings yielded similar degrees of agreement. The reason for this is that both settings were so frequent in this population that to some extent they can be considered accent features. For that reason, agreement increased considerably when measured within one scalar degree. From a forensic perspective, however, rarity of a setting in the relevant population is very important for the strength of the evidence. Therefore, the question arises as to whether it is much more important that experts agree on the presence of

[8] We noted before that there are different types of *creak* and hence more acoustic investigations are necessary on this issue. We thank an anonymous reviewer for highlighting that more research is also needed to explore the correlation between *creak* and *lax larynx*. According to Ladefoged & Maddieson (1996), *creaky voice* is at the closed end of a continuum of laryngeal closure, and *breathy* at the other, with *creaky voice* associated with 'a great deal of tension in the intrinsic laryngeal musculature' (p. 53). In our perceptual ratings *creaky* sometimes co-occurs with *lax larynx* and sometimes with *tense larynx*, which could suggest that there are lax and tense kinds of *creak* or *creaky voice*.

very unusual settings (e.g. *falsetto* or *tremor*) than they slightly disagree on settings that are undoubtedly present in the speaker.

Finally, we briefly discuss here some settings that could be useful to include in the VPA protocol for other investigations and maybe when analysing other speaker populations. For instance, we had removed 'audible nasal escape' from our protocol. Sometimes there is also something close to audible oral escape, which could be marked as 'inadequate breath control' (a label for which is included in the supplementary page of the VPA protocol in Beck 2007). Both were apparent for some speakers, e.g. #028 (both oral and nasal) and #063 (mainly oral). However, Beck (2007) defines audible nasal escape as a (presumably pathological) feature that particularly affects obstruents. In this investigation, it was originally interpreted as something being produced outside of the actual speech (extralinguistic or paralinguistic), mostly at turn ends or hesitation junctures. Probably 'breath support' covers this effect, but it may be preferred to separate oral and nasal breath control.

Furthermore, there is no label for overall rapidity or any rhythmic aspects in the VPA, although some features related to temporal organization or prosodic variables do appear on the supplementary page of Beck (2007). These features were salient in many of the assessed speakers (e.g. fast rate for #020; slow rate for #030). Articulation rate is commonly commented upon in FSC, and a number of rhythmic measures have also been proposed (Leemann, Kolly & Dellwo 2014, Dellwo, Leemann & Kolly 2015). However, these aspects are typically considered separately from VQ. Similarly, some of the speakers in this population were impressionistically described as 'lively' or 'active'. This reflected both rapidity and also extensive pitch movement (e.g. #022, #072). The opposite is also perceived in other speakers described as 'monotonous'.

Sibilance is a setting that appears in the version of the VPA protocol modified by JP French Associates for forensic casework. It was not included in the scheme used here because it was deemed to be covered by the setting *advanced tongue tip*. However, if high agreement between raters is not achieved using only that setting and its three possible scalar degrees, raters might consider adding 'sibilance'.

In terms of phonation variation, several speakers (e.g. #009 and #046) were very distinctive, largely because most phonation settings, including intermittent *falsetto*, could be found in their recordings. While some raters had initially marked most phonation settings as intermittent – additionally using *tremor* to capture the variation/inconsistency – the ad hoc use of this label does not comply with the definition in Beck (2007). Therefore, the question arises as to whether the raters are satisfied with using a wide range of labels individually or whether there should be some sort of global label for these variables. For instance, the GRBAS protocol reserves the 'G' for overall dysphonia.

Diplophonia also appears in the VPA supplementary page (Beck 2007). We looked at some speech fragments acoustically and found clear examples of this, e.g. speaker #032. We discussed whether it should be added as a new label but it was eventually not added for two main reasons. On the one hand, it seems that this is a feature which benefits from – and perhaps even requires – acoustic examination in order to confirm its presence. While Beck (2007) suggests that the acoustic signal can be consulted if something is not clear, we were not doing this with the rest of the settings. On the other hand, we felt that the perceptual effect of diplophonia in this corpus could be covered by a combination of the settings *constricted pharynx*, *tense larynx* or *harsh voice*. Diplophonia requires the simultaneous presence of two separate tones or pitches, typically associated with hoarseness (Dejonckere & Lebacq 1983). See Moisik (2013) for a description of the role of the epilarynx (epilaryngeal vibration) in the diplophonic voice as well as in other voice qualities such as *constricted pharynx* or *tense larynx*.

Apart from the inter-rater and correlation results, we have discussed in detail a method for the perceptual assessment of voices. In view of the good inter-rater results, we consider the different stages of the method described in Table 1 to be a workable methodology. When a considerably large number of voices need to be assessed by two or more raters, the analysis

of a small set of speakers, followed by different calibration sessions, seems a good two-stage methodology to follow in future studies.

5 Conclusions

In this study we have examined the agreement reached by three phoneticians in the perceptual assessment of VQ using an adapted version of the VPA scheme. Undertaking this investigation is of forensic relevance, as in the context of FSC experts can be required by the court to provide a reliability measure of the methods used. Despite the fact that inter-rater agreement seems to depend greatly on the specific VPA setting, we have shown that it is possible to reach a high degree of inter-rater agreement, provided several calibration and training sessions are conducted. This has been the case here, even though raters' experience and even mother tongue differed.

The methodology described here represents a step towards formalizing VQ for wider use. Furthermore, the reliability of VPA settings has been seldom explored, especially with spontaneous speech samples in a non-pathological population, which makes this investigation relevant for a wider research community with an interest in the use of the VPA. For instance, the search for procedures which can compensate or alleviate disagreement is still a relevant issue in a range of phonetic-perceptual areas of research, as we highlighted in the introduction.

In a second step, we examined how independent the VPA settings are. This is also of forensic importance, as FSC analyses should rely on non-correlated features in order to avoid over-weighting of evidence. We suggested a methodology for agreeing on a single version of VPA scores, which served to calculate correlations between settings. These correlations were not higher than .62 (moderate correlation), which we interpreted as meaning that none of the correlated settings should be collapsed into a single one. Although it could be safer to state in forensic reports that they are not completely independent, each of them still provides specific information for the characterization of a speaker.

As for directions for future work, more investigations are needed into the acoustic correlates of certain settings. For instance, definitions for vocal tract tension seem much less precise than for the rest of the settings, or made in impressionistic terms, like 'muffled' voice for *lax vocal tract*. Although those terms can be very useful for calibration purposes, empirical tests of whether tension relates to any prosodic aspect, such as speech rate (as suggested by Beck 2007), would be necessary. Along this line, one should consider how to align labels on perceptual profiles with acoustic measures and automatic tools.

If the ultimate goal of perceptual VPA assessment is forensic use, other corpora or speaking tasks should be used for training. DyViS Task 2 seemed a good starting point for future assessments of telephone-degraded recordings. However, this Task 2 is an information exchange with an accomplice, which might entail problems such as accommodation towards the accomplice, or acting behaviour (i.e. not naturalness) exhibited in certain speakers. Finally, different perceptual methods could be envisaged to minimize any serial effect in speaker rating, whereby the speaker that has just been rated might influence the ratings for the next. For instance, if one speaker is extremely creaky and the next one only slightly so, the contrast might lead the analyst to score the second speaker lower on this setting than might otherwise have been the case.

All in all, we conclude that the perceptual assessment of VQ through the use of the VPA is a valuable tool in fields such as forensic phonetics but, foremost, that it can be adapted and modified to a range of research areas, and not necessarily limited to the evaluation of pathological voices in clinical studies. In the same way that 'perceptual ratings of voice quality are commonly used as one aspect of a battery of voice outcomes in the voice efficacy literature' (Carding et al. 2001: 128), the VPA protocol has to be understood as one tool in the (forensic) phonetician's toolbox.

Acknowledgements

This research was conducted as part of the Arts & Humanities Research Council (AHRC) funded grant *Voice and Identity – Source, Filter, Biometric* (AH/M003396/1). We thank three anonymous reviewers for their useful comments as well as the DyViS team for their support which has allowed us to conduct this research. Thanks also to the audiences at IAFPA conferences for their comments on earlier versions of parts of this material.

Appendix. Modified Vocal Profile Analysis (VPA) protocol and inter-rater agreement results

	FIRST PASS		SECOND PASS				Notes
				Slight	Mrkd.	Extrm.	
	Neutral	Non-Neutral	SETTING	1	2	3	
A. VOCAL TRACT FEATURES							
Labial			Lip rounding/protrusion				
			Lip spreading				
			Labiodentalisation				
			Extensive labial range				
			Minimised labial range				
Mandibular			Close jaw				
			Open jaw				
			Extensive mandibular range				
			Minimised mandibular range				
Lingual tip/blade			Advanced tongue tip/blade				
			Retracted tongue tip/blade				
Lingual body			Fronted/raised tongue body				
			Backed/lowered tongue body				
			Extensive lingual range				
			Minimised lingual range				
Pharynx			Pharyngeal constriction				
			Pharyngeal expansion				
Velopharyngeal			Nasal				
			Denasal				
Larynx height			Raised larynx				
			Lowered larynx				
B. OVERALL MUSCULAR TENSION							
Vocal tract tension			Tense vocal tract				
			Lax vocal tract				
Laryngeal tension			Tense larynx				
			Lax larynx				
C. PHONATION FEATURES							
			Present		Scalar Degree		
					Slight	Mrkd.	Extrm.
	SETTING		Neutral	Non-neutral	1	2	3
Voicing type	Falsetto						
	Creaky						
	Whispery						
	Breathy						
	Murmur						
	Harsh						
	Tremor						

Figure A1 Modified Vocal Profile Analysis (VPA) protocol.

Table A1 Inter-rater agreement.

	Absolute agreement (%)				Agreement within 1 scalar degree (%)			
Setting (% occurence)	ES/PF	ES/JPF	JPF/PF	Mean	ES/PF	ES/JPF	JPF/PF	Mean
Lip rounding (< 10)	96	96	100	97	96	96	100	97
Lip spreading (< 10)	94	95	95	95	94	95	95	95
Labiodentalization (< 10)	98	100	98	99	98	100	98	99
Extensive labial range (< 10)	100	100	100	100	100	100	100	100
Minimized labial range (< 10)	100	100	100	100	100	100	100	100
Close jaw (< 10)	96	96	100	97	96	96	100	97
Open jaw (< 10)	100	100	100	100	100	100	100	100
Ext. mandibular range (< 10)	99	99	100	99	99	99	100	99
Min. mand. range (< 10)	96	96	98	97	98	98	98	98
Advanced tongue tip (56)	55	56	66	**59**	69	73	78	**73**
Retracted tongue tip (< 10)	92	99	92	94	93	99	92	95
Fronted tongue body (98)	33	43	31	**36**	51	69	62	**60**
Backed tongue body (< 10)	97	97	100	98	97	97	100	98
Ext. ling. range (< 10)	98	99	99	99	100	100	100	100
Min. ling. range (< 10)	98	98	100	99	99	99	100	99
Pharyngeal constriction (< 10)	97	95	98	97	98	97	99	98
Pharyngeal expansion (< 10)	97	98	97	97	99	100	99	99
Nasal (92)	43	36	49	**43**	66	75	75	**72**
Denasal (< 10)	90	87	92	90	91	88	93	91
Raised larynx (34)	78	73	71	74	85	84	79	82
Lowered larynx (43)	62	70	71	67	72	79	79	76
Tense vocal tract (51)	53	55	59	**55**	75	65	66	**68**
Lax vocal tract (43)	66	55	58	**59**	76	65	71	**70**
Tense larynx (37)	69	66	68	67	74	80	74	76
Lax larynx (47)	66	69	51	62	71	85	58	71
Falsetto (< 10)	100	100	100	100	100	100	100	100
Creaky (83)	42	37	59	**46**	80	79	85	**81**
Whispery (10)	90	94	88	91	95	98	95	96
Breathy (73)	49	42	64	**52**	72	77	85	**78**
Murmur (< 10)	99	100	99	99	100	100	100	100
Harsh (31)	75	74	76	75	84	80	84	82
Tremor (< 10)	100	100	100	100	100	100	100	100
Overall rate	82.13	82.03	83.72	82.59	88.38	89.78	89.53	89.06

Notes: Raters' initials: ES, PF and JPF. **Bold**: percentage where the use of 'within 1 scalar degree' agreement improves the absolute agreement considerably (> 10 points). Setting frequency of occurrence (first column) based on the cross-rater agreed version (figures presented in San Segundo et al. 2016).

Table A2 Inter-rater agreement: Fleiss' kappa (FK).

Setting	FK observed agreement	FK expected agreement	Fleiss' kappa
Lip rounding	0.97	0.97	Undefined
Lip spreading	0.95	0.93	0.18
Labiodentalization	0.99	0.99	Undefined
Extensive labial range	1	1	Undefined
Minimized labial range	1	1	Undefined
Close jaw	0.97	0.97	Undefined
Open jaw	1	1	Undefined
Ext. mandibular range	0.99	0.99	Undefined
Min. mandibular range	0.97	0.93	0.53
Advanced tongue tip	0.59	0.36	0.35
Retracted tongue tip	0.94	0.94	Undefined
Fronted tongue body	0.36	0.35	0.01
Backed tongue body	0.98	0.98	Undefined
Ext. lingual range	0.99	0.94	0.77
Min. lingual range	0.99	0.97	0.49
Pharyngeal constriction	0.97	0.93	0.49
Pharyngeal expansion	0.97	0.93	0.59
Nasal	0.43	0.35	0.13
Denasal	0.90	0.87	0.22
Raised larynx	0.74	0.51	0.46
Lowered larynx	0.67	0.45	0.41
Tense vocal tract	0.55	0.43	0.22
Lax vocal tract	0.59	0.43	0.29
Tense larynx	0.67	0.51	0.34
Lax larynx	0.62	0.45	0.31
Falsetto	1	1	Undefined
Creaky	0.52	0.31	0.31
Whispery	0.91	0.80	0.53
Breathy	0.52	0.31	0.31
Murmur	0.99	0.96	0.83
Harsh	0.75	0.55	0.43
Tremor	1	1	Undefined

Note: Undefined κ is due to invariant values, i.e. expected and observed agreement is the same.

References

Abberton, Evelyn & Adrian Fourcin. 1978. Electrolaryngography. In Martin J. Ball & Chris Code (eds.), *Instrumental clinical phonetics*, 119–148. London: Whurr.

Abercombie, David. 1967. *Elements of general phonetics*. Edinburgh: Edinburgh University Press.

Ball, Martin J., John H. Esling & B. Craig Dickson. 1995. The VoQS system for the transcription of voice quality. *Journal of the International Phonetic Association* 25(2), 71–80.

Beck, Janet Mackenzie. 1988. *Organic variation and voice quality*. Ph.D. dissertation, University of Edinburgh.

Beck, Janet Mackenzie. 2005. Perceptual analysis of voice quality: the place of Vocal Profile Analysis. In Hardcastle & Beck (eds.), 285–322.

Beck, Janet Mackenzie. 2007. *Vocal profile analysis scheme: A user's manual*. Edinburgh: Queen Margaret University College–QMUC, Speech Science Research Centre.

Bele, Irene V. 2007. Dimensionality in Voice Quality. *Journal of Voice* 21(3), 257–272.
Bell, Alexander M. 1867. *Visible speech: The science of universal alphabetics*. London: Simpkin, Marshall & Company.
Carding, Paul, Eva Carlson, Ruth Epstein, Lesley Mathieson & Christina Shewell. 2001. Re: Evaluation of voice quality; Letters to Editor. *International Journal of Language and Communication Disorders* 36, 127–143.
Catford, J. C. 1977. *Fundamental problems in phonetics*. Edinburgh: Edinburgh University Press.
Coadou, Marion & Abderrazak Rougab. 2007. Voice quality and variation in English. *Proceedings of 16th International Congress of Phonetic Sciences* (ICPhS XVI), Saarbrücken, Germany, pp. 2077–2080.
Cohen, Jacob. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20(1), 37–46.
Curtin, François & Pierre Schulz. 1998. Multiple correlations and Bonferroni's correction. *Biological Psychiatry* 44(8), 775–777.
Dejonckere, Philippe H. & Jean Lebacq. 1983. An analysis of the diplophonia phenomenon. *Speech Communication* 2(1), 47–56.
Dellwo, Volker, Adrian Leemann & Marie-José Kolly. 2015. Rhythmic variability between speakers: Articulatory, prosodic, and linguistic factors. *The Journal of the Acoustical Society of America* 137(3), 1513–1528.
Eckert, Hartwig & John Laver. 1994. *Menschen und Ihre Stimmen*. Weinheim: Beltz Psychologie Verlags Union.
Eskenazi, Laurent, Donald G. Childers & Douglas M. Hicks. 1990. Acoustic correlates of vocal quality. *Journal of Speech, Language, and Hearing Research* 33(2), 298–306.
Esling, John H. 1978. *Voice quality in Edinburgh: A sociolinguistic and phonetic study*. Ph.D. dissertation, University of Edinburgh.
Esling, John H. & Jimmy G. Harris. 2005. States of the glottis: An articulatory phonetic model based on laryngoscopic observations. In Hardcastle & Beck (eds.), 347–383.
Evans, James D. 1996. *Straightforward statistics for the behavioral sciences*. Pacific Grove, CA: Brooks-Cole.
Fleiss, Joseph L. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5), 378–382.
Forensic Science Regulator. 2016a. *Codes of practice and conduct for forensic science providers and practitioners in the Criminal Justice System*, Issue 3. Birmingham: Forensic Science Regulator Publications.
Forensic Science Regulator. 2016b. *Codes of practice and conduct for forensic science providers and practitioners in the Criminal Justice System*, Appendix: Speech and Audio Forensic Services, FSR-C-134, Issue 1. Birmingham: Forensic Science Regulator Publications.
Foulkes, Paul & Gerard J. Docherty. 2000. Another chapter in the story of /r/: 'Labiodental' variants in British English. *Journal of Sociolinguistics* 4, 30–59.
Foulkes, Paul & Peter French. 2012. Forensic speaker comparison: A linguistic-acoustic perspective. In Lawrence Solan & Peter Tiersma (eds.), *The Oxford handbook of language and law*, 557–572. Oxford: Oxford University Press.
Freelon, Deen. 2010. ReCal: Intercoder reliability calculation as a web service. *International Journal of Internet Science* 5(1), 20–33.
French, Peter, Paul Foulkes, Philip Harrison, Vincent Hughes, Eugenia San Segundo & Louisa Stevens. 2015. The vocal tract as a biometric: Output measures, interrelationships, and efficacy. *Proceedings of the 18th International Congress of Phonetic Sciences* (ICPhS XVIII), Glasgow.
French, Peter & Louisa Stevens. 2013. Forensic speech science. In Mark J. Jones & Rachael-Anne Knight (eds.), *Bloomsbury companion to phonetics*, 183–197. London: Continuum.
Gil, Juana & Eugenia San Segundo. 2013. La cualidad de voz en fonética judicial [Voice quality in Forensic Phonetics]. In Elena Garayzábal & Mercedes Reigosa (eds.), *Lingüística forense. La lingüística en el ámbito legal y policial* [Forensic Linguistics: Linguistics in legal and criminal contexts], 154–199, Madrid: Euphonía Ediciones.

Hammarberg, Britta. 2000. Voice research and clinical needs. *Folia Phoniatrica et Logopaedica* 52, 93–102.

Hanson, Helen M. 1997. Glottal characteristics of female speakers: Acoustic correlates. *The Journal of the Acoustical Society of America*. 101(1), 466–481.

Hardcastle, William J. & Janet Mackenzie Beck (eds.). 2005. *A figure of speech: A Festschrift for John Laver*. London & Mahwah, NJ: Laurence Erlbaum.

Harmegnies, Bernard & Albert Landercy. 1988. Intra-speaker variability of the long term speech spectrum. *Speech Communication* 7(1), 81–86.

Hirano, Minoru. 1981. *Clinical examination of voice*. Vienna & New York: Springer.

Hirano, Minoru, Yasuo Koike & Hans Von Leden. 1968. Maximum phonation time and air usage during phonation. *Folia Phoniatrica et Logopaedica* 20(4), 185–201.

Honda, Kiyoshi, Tatsuya Kitamura, Hironori Takemoto, Seiji Adachi, Parham Mokhtari, Sayoko Takano, Yukiko Nota, Hiroyuki Hirata, Ichiro Fujimoto, Yasuhiro Shimada, Shinobu Masaki, Satoru Fujita & Jianwu Dang. 2010. Visualisation of hypopharyngeal cavities and vocal-tract acoustic modelling. *Computer Methods in Biomechanics and Biomedical Engineering* 13(4), 443–453.

Honikman, Beatrice. 1964. Articulatory settings. In David Abercrombie, Dennis Butler Fry, Peter MacCarthy, Norman Carson Scott & John L. M. Trimm (eds.), *In honour of Daniel Jones*. London: Longmans.

Hughes, Vincent, Philip Harrison, Paul Foulkes, Peter French, Colleen Kavanagh & Eugenia San Segundo. 2017. Mapping across feature spaces in forensic voice comparison: The contribution of auditory-based voice quality to (semi-)automatic system testing. *Interspeech*, Stockholm, 3892–3896.

Keating, Patricia & Marc Garellek. 2015. Acoustic analysis of creaky voice. Presented at LSA Annual Meeting, Portland, OR.

Keller, Eric. 2004. The analysis of voice quality in speech processing. Tutorial at Research Workshop Nonlinear Speech Processing: Algorithms and Analysis, 12–17 September, Vietri, Italia.

Kempster, Gail B., Bruce R. Gerratt, Katherine V. Abbott, Julie Barkmeier-Kraemer & Robert E. Hillman. 2009. Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology* 18(2), 124–132.

Kent, Ray D. 1996. Hearing and believing: Some limits to the auditory-perceptual assessment of speech and voice disorders. *American Journal of Speech-Language Pathology* 5(3), 7–23.

Kitzing, Peter. 1982. Photo- and electroglottographical recording of the laryngeal vibratory pattern during different registers. *Folia Phoniatrica* 34, 234–241.

Klatt, Dennis H. & Laura C. Klatt. 1990. Analysis, synthesis, and perception of voice quality variations among female and male talkers. *The Journal of the Acoustical Society of America* 87(2), 820–857.

Klug, Katharina. 2017. Refining the Vocal Profile Analysis scheme (VPA) for forensic purposes. Poster presented at the International Association of Forensic Phonetics and Acoustics (IAFPA) Conference, Split, Croatia, 9–12 July 2017.

Kreiman, Jody & Bruce R. Gerratt. 1998. Validity of rating scale measures of voice quality. *The Journal of the Acoustical Society of America* 104(3), 1598–1608.

Kreiman, Jody & Bruce R. Gerratt. 2010. Perceptual assessment of voice quality: Past, present, and future. *SIG 3 Perspectives on Voice and Voice Disorders* 20(2), 62–67.

Kreiman, Jody, Bruce R. Gerratt & Mika Ito. 2007. When and why listeners disagree in voice quality assessment tasks. *The Journal of the Acoustical Society of America* 122(4), 2354–2364.

Kreiman, Jody, Bruce R. Gerratt, Gail B. Kempster, Andrew Erman & Gerald S. Berke. 1993. Perceptual evaluation of voice quality: Review, tutorial, and a framework for future research. *Journal of Speech, Language, and Hearing Research* 36(1), 21–40.

Kreiman, Jody & Diana Sidtis, D. 2011. *Foundations of voice studies: Interdisciplinary approaches to voice production and perception*. Boston, MA: Wiley-Blackwell.

Ladefoged, Peter. 1971. *Preliminaries to linguistic phonetics*. Chicago, IL: University of Chicago Press.

Ladefoged, Peter & Ian Maddieson. 1996. *The sounds of the world's languages*. Malden, MA: Blackwell.

Landis, J. Richard & Gary G. Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics* 33(2), 363–374.

Laver, John. 1975. *Individual features in voice quality*. Ph.D. dissertation, University of Edinburgh.

Laver, John. 1980. *The phonetic description of voice quality*. Cambridge: Cambridge University Press.
Laver, John. 1991. *The gift of speech: Papers in the analysis of speech and voice*. Edinburgh: Edinburgh University Press.
Laver, John. 1994. *Principles of phonetics*. Cambridge: Cambridge University Press.
Laver, John. 2000. Phonetic evaluation of voice quality. In Raymond D. Kent & Martin J. Ball (eds.), *Voice quality measurement*, 37–48. San Diego, CA: Singular Publications.
Laver, John & Robert Hanson. 1981. Describing the normal voice. In John K. Darby (ed.), *Speech evaluation in psychiatry*, 51–78. New York: Grune and Stratton.
Laver, John, Sheila Wirz, Janet Mackenzie Beck & Steven Hiller. 1981. A perceptual protocol for the analysis of vocal profiles. *Edinburgh University Department of Linguistics Work in Progress* 14, 139–155.
Leemann, Adrian, Marie-José Kolly & Volker Dellwo. 2014. Speaker-individuality in suprasegmental temporal features: Implications for forensic voice comparison, *Forensic Science International* 238, 59–67.
Löfqvist, Anders. 1986. The long-time-average spectrum as a tool in voice research. *Journal of Phonetics* 14, 471–475.
Ma, Estella P. M. & Edwin M. L. Yu. 2005. Multiparametric evaluation of dysphonic severit. *Journal of Voice* 20(3), 380–390.
Mackenzie, Kenneth, Ian J. Deary, Cameron Sellars & Janet A. Wilson. 1998. Patient reported benefit of the efficacy of speech therapy in dysphonia. *Clinical Otolaryngology & Allied Sciences* 23(3), 284–285.
Maryn, Youri, Nelson Roy, Marc De Bodt, Paul Van Cauwenberge & Paul Corthals. 2009. Acoustic measurement of overall voice quality: A meta-analysis. *The Journal of the Acoustical Society of America* 126(5), 2619–2634.
Michel, John F. 1968. Fundamental frequency investigation of vocal fry and harshness. *Journal of Speech and Hearing Research* 11(3), 590–594.
Moisik, Scott. 2013. *The epilarynx in speech*. Ph.D. dissertation, University of Victoria, Canada.
Multon, Karen D. 2010. Interrater reliability. In Neil J. Salking (ed.), *Encyclopedia of research design*, vol. 2, 626–628. Thousand Oaks, CA: SAGE.
Nolan, Francis. 2005. Forensic speaker identification and the phonetic description of voice quality. In Hardcastle & Beck (eds.), 385–411.
Nolan, Francis, Kirsty McDougall, Gea De Jong & Toby Hudson. 2009. The DyViS database: Style-controlled recordings of 100 homogeneous speakers for forensic phonetic research. *International Journal of Speech, Language & the Law* 16(1), 31–57.
Powers, David M. 2012. The problem with kappa. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 345–355. Association for Computational Linguistics.
Rose, Phil. 2006. Accounting for correlation in linguistic-acoustic likelihood ratio-based forensic speaker discrimination. *Proceedings of Odyssey Speaker and Language Recognition Workshop*, Puerto Rico, 1–18.
San Segundo, Eugenia, Paul Foulkes, Peter French, Philip Harrison & Vincent Hughes. 2016. Voice quality analysis in forensic voice comparison: Developing the Vocal Profile Analysis scheme. Communication presented at the International Association of Forensic Phonetics and Acoustics (IAFPA) conference, York, 24–27 July 2016.
San Segundo, Eugenia, Paul Foulkes, Peter French, Philip Harrison, Vincent Hughes & Colleen Kavanagh. 2018. Cluster analysis of voice quality ratings: Identifying groups of perceptually similar speakers. *Proceedings of Phonetics and Phonology in the German-speaking Countries* (P&P13), 173–176. Berlin: AG Elektronisches Publizieren.
San Segundo, Eugenia, Paul Foulkes & Vincent Hughes. 2016. Holistic perception of voice quality matters more than L1 when judging speaker similarity in short stimuli. *Proceedings of the 16th Australasian Conference on Speech Science and Technology* (ASSTA), University of Western Sydney, Australia, 309–312.

San Segundo, Eugenia & José A. Mompeán. 2017. A simplified Vocal Profile Analysis Protocol for the assessment of voice quality and speaker similarity. *Journal of Voice* 31(5), 644.e11–644.e27.

San Segundo, Eugenia, Athanasios Tsanas & Pedro Gómez-Vilda. 2017. Euclidean distances as measures of speaker dissimilarity including identical twin pairs: A forensic investigation using source and filter voice characteristics. *Forensic Science International* 270, 25–38.

Scherer, Klaus R. 1979. Personality markers in speech. In Klaus Scherer & Howard Giles. (eds.), *Social markers in speech*, 147–209. Cambridge: Cambridge University Press.

Scott, Sheila & Fi Caird. 1983. Speech therapy for Parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry* 46(2), 140–144.

Sellars, Cameron, A. E. Stanton, Alex McConnachie, Catherine P. Dunnet, L. M. Chapman, C. E. Bucknall & Kenneth Mackenzie. 2009. Reliability of perceptions of voice quality: Evidence from a problem asthma clinic population. *The Journal of Laryngology & Otology* 123(7), 755–763.

Stemler, Steven E. 2004. A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation* 9(4), 1–19.

Stevens, Louisa & Peter French. 2012. Voice quality in studio quality and telephone transmitted recordings. Presented at British Association of Academic Phoneticians (BAAP) Conference, Leeds, March 2012.

Stuart-Smith, Jane. 1999. Glasgow: Accent and voice quality. In Paul Foulkes & Gerard J. Docherty (eds.), *Urban voices: Accent studies in the British Isles*, 203–222. London: Arnold.

Trask, Robert L. 1996. *A dictionary of phonetics and phonology*. London: Routledge.

Uebersax, John S. 1987. Diversity of decision-making models and the measurement of inter-rater agreement. *Psychological Bulletin* 101(1), 140–146.

Webb, A. L., Paul Carding, Ian J. Deary, Kenneth Mackenzie, Nick Steen N & Janet A. Wilson. 2004. The reliability of three perceptual evaluation scales for dysphonia. *European Archives of Otorhinolaryngology* 261, 429–434.

Wells, John C. 1982. *Accents of English*. Cambridge: Cambridge University Press.

Wilson, D. Kenneth. 1987. *Voice problems of children*. Baltimore, MA: Williams & Wilkins.

Wirz, Sheila & Janet Mackenzie Beck. 1995. Assessment of voice quality: The vocal profile analysis scheme. In Sheila Wirz (ed.), *Perceptual approaches to communication disorders*, 39–55. London: Whurr.

Wormald, Jessica. 2016. *Regional variation in Panjabi English*. Ph.D. dissertation, University of York.