The Journal of the Acoustical Society of America

FEBRUARY 12 2024

Information conveyed by voice quality[a)] FREE

Jody Kreiman

Check for updates

J. Acoust. Soc. Am. 155, 1264–1271 (2024)
https://doi.org/10.1121/10.0024609

View Online
Export Citation
CrossMark

20 February 2024 03:40:19

ASA
LEARN MORE
Advance your science and career as a member of the
Acoustical Society of America

ASA
ACOUSTICAL SOCIETY OF AMERICA

JASA ARTICLE

Information conveyed by voice quality[a)]

Jody Kreiman[b)]
Departments of Head and Neck Surgery and Linguistics, University of California, Los Angeles, Los Angeles, California 90095-1794, USA

ABSTRACT:
The problem of characterizing voice quality has long caused debate and frustration. The richness of the available descriptive vocabulary is overwhelming, but the density and complexity of the information voices convey lead some to conclude that language can never adequately specify what we hear. Others argue that terminology lacks an empirical basis, so that language-based scales are inadequate *a priori*. Efforts to provide meaningful instrumental characterizations have also had limited success. Such measures may capture sound patterns but cannot at present explain what characteristics, intentions, or identity listeners attribute to the speaker based on those patterns. However, some terms continually reappear across studies. These terms align with acoustic dimensions accounting for variance across speakers and languages and correlate with size and arousal across species. This suggests that labels for quality rest on a bedrock of biology: We have evolved to perceive voices in terms of size/arousal, and these factors structure both voice acoustics and descriptive language. Such linkages could help integrate studies of signals and their meaning, producing a truly interdisciplinary approach to the study of voice.
© *2024 Acoustical Society of America*. https://doi.org/10.1121/10.0024609

(Received 21 September 2023; revised 8 January 2024; accepted 9 January 2024; published online 12 February 2024)

[Editor: James F. Lynch] Pages: 1264–1271

20 February 2024 03:40:19

I. INTRODUCTION

This project began with a simple question: Why do we insist on describing voices with terms like "breathy" and "rough?" Ample evidence has long indicated that listeners do not agree when asked to rate voices on these scales (e.g., Kreiman and Gerratt, 1998; Webb *et al.*, 2003). In fact, listeners do not even agree whether a voice is or is not breathy or rough (Kreiman and Gerratt, 2000). Lack of agreement appears to compromise meaning to the point that some have argued that descriptive terms in general are useless as measures of voice quality (Bregman, 1990; Kreiman *et al.*, 1994; Kreiman and Gerratt, 1998). Nevertheless, these terms remain in constant use in clinical, scientific, and informal contexts and have been in such use for centuries [Austin, 1806; Lichte, 1941, Hirano, 1981; Kempster *et al.*, 2009; Malawey, 2020; see Kreiman and Sidtis (2011) for review]. Given their well-established weaknesses as measurement tools and apparent lack of stable indexical meaning across listeners, how do we justify our ongoing practice of labeling voices in this way?

This question points to the general problem of creating a comprehensive model of voice that relates perceived quality to physical voice signals, and vice versa. An extensive literature describes attempts to quantify what listeners hear in terms of instrumental measures of signal characteristics, and a similarly large literature addresses listeners' experiences in qualitative terms; but attempts to associate measurements with descriptions of voice have not been fruitful to date, and we have not solved the problem of connecting voice signals to the meaning they convey to listeners, despite these long-standing efforts.

In this paper, I describe a possible way to resolve this issue in the study of voice. I begin by briefly reviewing the literature from different disciplines on descriptive terms for voice. I will argue that a small number of dimensions emerge rather consistently from these studies, along with many other terms that vary widely from study to study. Next, I review work showing that acoustic variability within and across speakers can be characterized by a few dimensions that are very widely shared, regardless of the speaker's gender, the language spoken, or the kind of speech sample produced, again accompanied by a larger number of other parameters whose salience depends on linguistic factors and on the idiosyncrasies of the particular voice in question. Finally, I argue that frequently emerging descriptive terms align well with these shared acoustic dimensions and that both are associated with physical size, reproductive fitness, and arousal across many species and are thus biologically significant. I conclude that the most commonly applied terms for voice remain useful—and continue to be used—because our use of language is partially structured by biology. That is, we perceive voice in terms of these factors, and our terminology reflects this structure without conscious design, because these aspects of the meaning of a signal are part of our evolutionary heritage. This provides a link between qualitative and quantitative approaches to measuring voice that is independent of both speaker and listener, thus potentially forming a common foundation for both

[a)]This paper is part of a special issue on Iconicity and Sound Symbolism. A preliminary version of this work was presented in "Labels for voices," International Congress of Phonetic Sciences, Prague, Czech Republic, August 2023.
[b)]Email: jkreiman@ucla.edu

 © 2024 Acoustical Society of America

https://doi.org/10.1121/10.0024609

kinds of study and a potential basis for truly interdisciplinary approaches to voice.

II. THE "DUAL NATURE" OF VOICE QUALITY

Two primary approaches to the study of voice quality—descriptive and quantitative—are apparent in the literature (e.g., Barsties and De Bodt, 2015; van Elferen, 2018; Vélez, 2018). Both have the same general goal of explaining the relationship between voice production and how the resulting signal is perceived, but they differ in the starting point of their inquiries. Descriptive studies examine quality from the perspective of the listener. Such studies start with perceptual measures or descriptions of a heard voice and then seek the source of these impressions in the vocal signal or production system. In comparison, quantitative studies start by targeting some aspect of phonation or of the acoustic signal and then attempt to identify the perceptual correlates of that measured attribute. I will review each approach briefly in turn.

A. Descriptive and qualitative studies

Whether focused on specific terms like "breathy" or on more elaborate or nuanced descriptions, descriptive studies of voice begin with a listener's percept—with the meaning conveyed by voice quality—and seek to explain why that particular message was received. Because such studies focus on what is heard, which is taken as given, they necessarily assume that voice quality is a function of the listener and not of the speaker (e.g., Fales, 2002; Eidsheim, 2019). Rather like a tree falling in the woods, which generates unheard (and thus meaningless) vibrations in the absence of a hearer, a speaker produces something audible, but precisely what is heard (quality) depends not just on the physical signals and the speaker's intentions, but also on the listener's affect and memory, attention, the conversational setting, cultural structures, and a multitude of other factors (Hajda *et al.*, 1997; Kreiman and Sidtis, 2011; Heidemann, 2016). This perspective is nicely illustrated by the writing of American author Raymond Chandler (1888–1959):

> Then she laughed. It was almost a racking laugh. It shook her as the wind shakes a tree. I thought there was puzzlement in it, not exactly surprise, but as if a new idea had been added to something already known and it didn't fit. Then I thought that was too much to get out of a laugh (Chandler, 1939, p.196).

The meaning the protagonist (private investigator Philip Marlowe) derives from this voice is clear, but it is difficult for readers to imagine, let alone agree upon, what the voice actually sounds like; and other listeners might well draw a different impression from this laugh than Marlowe does. Acoustic analysis would be uninformative in this case, because the description reflects Marlowe's personal evaluation of what he hears, but not anyone else's impression, and not the actual sounds themselves.

Additional examples from Chandler's writing (Table I) further illustrate the denseness and variety of the information conveyed by voice, descriptions of which are seemingly limited only by listeners' imaginations, attention to different aspects of what they hear, and taste for flights of fancy. The apparent contradiction between an essentially infinite range of possible interpretations derived from a single discrete signal has long been an issue in voice studies (e.g., Fales, 2002). Some scholars have argued that descriptive terminology derives from whimsy, analogy, metaphor, and historical tradition and lacks an empirical basis, so that language-based measurement systems are *a priori* theoretically inappropriate measurement tools (Kreiman and Sidtis, 2011). Others have concluded that even given the vast range and power of descriptive language, it remains inadequate to specify what we hear (Kendall and Carterette, 1993; Malawey, 2020).

This gap between physical signals and meaning has also impeded efforts to create a broadly applicable standardized set of terms for different voice qualities, because reliable associations between signals and descriptors have not emerged even from experimental studies [see Kreiman and Sidtis (2011) for review, or Dolan and Rehding (2018) for review of studies of musical timbre that ask similar questions and use similar methods]. Experimental approaches to descriptive terminology seek to constrain the descriptions applicable to voice quality by identifying terminological redundancies and overlap, thus reducing the set of possible descriptors to a more manageable, empirically derived set of orthogonal semantic dimensions that underlie the profusion of possible descriptors found in the literature. A variety of techniques have been applied to isolate these superordinate dimensions. For example, investigators may ask listeners to select the adjectives they feel apply to voice from a long list (Gelfer, 1988; Carron *et al.*, 2017) or to perform a free description task (e.g., Paz *et al.*, 2022). In another approach, listeners rate voices (or other sounds) on large sets of descriptive (e.g., breathy, strained) or semantic-differential scales (hot-cold, loud-soft, big-small), which may be general purpose (Osgood *et al.*, 1957) or specific to voice quality (Fagel *et al.*, 1983). In a third method, listeners assess the similarity of voices heard in pairs (e.g., Murry and Singh, 1980; Kreiman and Gerratt, 1996; Baumann and Belin, 2010). A data reduction step generally follows such

TABLE I. Detective Philip Marlowe's impressions of some of the voices he hears in the course of his investigations, from the novels of Raymond Chandler. Chandler's writing includes many such vivid descriptions of voice.

"Please don't get up," she said in a voice like the stuff they use to line summer clouds with. —*The Long Goodbye* (Chandler, 1953, p. 95)
The voice I heard was an abrupt voice, but thick and clogged, as if it was being strained through a curtain or somebody's long white beard. —*The Little Sister* (Chandler, 1949, p. 41)
He sounded like a man who had slept well and didn't owe too much money. —*The Big Sleep* (Chandler, 1939, p. 43)

20 February 2024 03:40:19

perceptual assessments. Studies using semantic-differential scaling typically apply factor analysis to reveal the primary dimensions underlying the ratings (e.g., Osgood *et al.*, 1957; Voiers, 1964; Hirano, 1981). Similarity judgments may be analyzed using multidimensional scaling, which outputs an *n*-dimensional space in which the distances between pairs of voices reflect their rated similarity (e.g., Kreiman and Gerratt, 1996; Baumann and Belin, 2010). In either case, experimenters interpret the derived scales or dimensions of the perceptual space *post hoc* in terms of their statistical association with instrumental measures (a "rate and correlate" approach). Finally, in a non-experimental approach with the same goals, clinicians and others have devised "consensus models" for describing voice. In this case, experts meet in groups and agree on a list of terms that they feel are important descriptors (Hammarberg and Gauffin, 1995; Porcello, 2004; Kempster *et al.*, 2009; Lechien *et al.*, 2023). Results again vary across these studies, and no agreed-upon descriptive protocol exists, although a number have been proposed (Hirano, 1981; Wirz and MacKenzie Beck, 1995; Bele, 2007; Kempster *et al.*, 2009).

In summary, descriptive approaches to voice quality assessment can provide significant insight into the detailed meaning listeners derive when they hear a voice. However, what is heard and how it is understood do not depend solely on the physical voice signals, so it is unclear how meaning can be fully explained or predicted by acoustic or other instrumental measures. Research has not produced a standardized vocabulary, either within or across disciplines, and it is not clear how even commonly used terms relate to the information listeners glean from voices. Thus, although such studies are an essential part of a comprehensive approach to voice, they are not sufficient on their own to document either what listeners hear or why they hear what they do.

Nevertheless, certain terms recur across studies, despite our inability to explain their origin or meaning. A re-examination of the literature indicates that across papers, a small set of dimensions has, in fact, rather consistently emerged, despite differences in the methods, voices, listeners, and descriptive scales under examination. These dimensions include something analogous to brightness/brilliance/sharpness/clarity, which is associated with the distribution of spectral energy in the voice; breathiness and/or roughness, associated with noise or spectral irregularity; and fullness/richness, associated with the location of the spectral centroid [e.g., Lichte, 1941; Voiers, 1964; Plomp (1970), cited in Hermes (2023); von Bismarck, 1974; Pratt and Doak, 1976; Wallmark and Kendall, 2021]. Reviews and meta-analyses (Maryn *et al.*, 2009; Barsties and De Bodt, 2015; Barsties *et al.*, 2019) confirm this pattern. For example, Barsties *et al.* (2019) reviewed many protocols for clinical quality assessment and concluded that only scales for breathiness, roughness, and strain are widely accepted. Interestingly, similar sets of scales have emerged across cultures and languages (Alluri and Toiviainen, 2012; Zacharakis *et al.*, 2014) and from studies of instrumental timbre and animal vocalization (Lichte, 1941; McAdams *et al.*, 2006; Wallmark and Kendall, 2021).

B. Quantitative studies

A second approach to assessing voice quality involves identifying salient or potentially meaningful physical attributes of the voice or voice production system and measuring them instrumentally. This approach assumes that quality inheres in the voice signal: The production mechanism creates the acoustic voice signal, which is causally linked to the body that produced the sounds and necessarily reflects the physical properties of the speaker (Heidemann, 2016; Malawey, 2020). This signal in turn serves as input to the perceptual mechanism. Because measures of acoustic signals quantify the input listeners have to work with, such measures should shed light on how listeners perceive voices.

Hundreds of studies have examined the relationship between voice production, acoustic (or other instrumental) measures of voice, and/or voice quality [see, e.g., Buder (2000) for review]. Such studies have somewhat varied goals. For example, many studies have examined the acoustic attributes associated with a speaker's identity or emotional state (e.g., Becker *et al.*, 2022; Gobl and Ní Chasaide, 2003; López *et al.*, 2013). Additional recent work uses physical and computational models of production to examine the physiological control of specific acoustic measures that are associated with quality [Table II; see, e.g., Sundberg (1987) for an introduction to voice production or Zhang (2016a) for a more thorough discussion]. Studies using this approach have begun to explain the ways in which speakers consciously or unconsciously manipulate specific psychoacoustically motivated acoustic parameters (e.g., Zhang, 2016b) but often do not measure perception directly. In contrast, experimenters in typical older studies (e.g., Ryan and Burk, 1974; Streeter *et al.*, 1983; Södersten and Lindestad, 1990) use regression or correlation to demonstrate the relationship between instrumental measures and ratings of individual

TABLE II. Acoustic variables used to assess acoustic variability within and between speakers. From Lee *et al.* (2019). H1* − H2* = the difference in the amplitudes of the first and second harmonics, corrected for the influence of formants on harmonic amplitudes; H2* − H4* = the difference in the amplitudes of the second and fourth harmonics, corrected as above; H4* − H2kHz* = the difference in the amplitudes of the fourth harmonic and the harmonic closest in frequency to 2 kHz, corrected as above; H2kHz* − H5kHz = the difference in the amplitudes of the harmonic closest in frequency to 2 kHz, corrected as above, and that closest in frequency to 5 kHz; CPP = cepstral peak prominence (Hillenbrand *et al.*, 1994); energy = the root mean square energy calculated over five pitch pulses; SHR = the amplitude ratio between subharmonics and harmonics (Sun, 2002).

Variable categories	Acoustic variables
Voice source	F0
Formant frequencies	F1, F2, F3, F4, formant dispersion
Harmonic source spectral shape	H1* − H2*, H2* − H4*, H4* − H2kHz*, H2kHz* − H5kHz
Spectral noise	CPP, energy, SHR
Variability	Coefficients of variation for all measures

20 February 2024 03:40:19

https://doi.org/10.1121/10.0024609

qualities or characteristics like breathiness, age, or emotional state (a "rate and correlate" approach), with the goal of finding reliable instrumental substitutes for unstable perceptual judgments. Such studies have not identified consistent correlates for familiar descriptive terms, much less for the complex nuances of meaning found in everyday use [see Kreiman and Sidtis (2011) for review]. Results vary widely across studies, due in part to the lack of theoretical motivation for the particular instrumental parameters studied; differences in the methods, voices, and qualities studied; and differences in how individuals define and assess the target voices and qualities. For example, there are multiple kinds of creaky voice that listeners can easily distinguish, yet which are all labeled creaky (Gerratt and Kreiman, 2001; Keating *et al.*, 2015). Similarly, a voice with a steeply falling source spectrum can be perceived as breathy even in the absence of turbulent noise (e.g., Garellek *et al.*, 2013).

Neither approach addresses the superordinate problem of explaining how voice production engenders meaning. Individuals who are familiar with acoustic measures may be able to derive a rudimentary sense of what a voice might sound like from an acoustic profile, but this fuzzy approximation gives no sense of what the sound might mean. For example, imagine a voice produced with relatively high formant frequencies, high and moderately variable F0, a relatively flat source spectrum, and a rapid speech rate, possibly signaling someone who was smiling and energetic ("…a man who had slept well and didn't owe too much money"; Table I). However, this same set of measures could also be produced by someone who is anxious because they are lying, or someone who is angry, or young, or small in stature. At the same time, our well-rested, well-heeled acquaintance could speak with a slow rate, a relatively steep source spectrum, and a soft volume, conveying the calm that comes with money and rest. Or that same slow, soft voice could instead represent distraction or a lack of engagement. The possibilities are virtually limitless, so no immutable meaning can ever attach to a specific profile of instrumental measures.

It is, however, possible to quantify quality in the limited sense of that which makes it possible to say that two signals are the same or different (ANSI, 1960). Using analysis-by-synthesis, Kreiman and colleagues (2021) identified a set of acoustic measures that combine to specify the sound of a voice with enough precision that voice samples synthesized using these parameters are indistinguishable from natural target voice samples. By specifying the acoustic parameters that are both necessary and sufficient to quantify quality in this limited sense, this psychoacoustic model provides a partial bridge between production and perception. However, this model does not connect voice production to meaning in any broader sense, and, equally importantly, it does not explain why these particular acoustic parameters emerge as perceptually important.

C. The timbral abyss

To summarize, quantitative methods have increased our understanding of the relationship between the voice production system and the sounds it produces and shed light on the causal links between the acoustic voice signal and the body that produced it. Such methods (which are common in clinical studies of voice disorders) also generalize well across studies and voices, unlike descriptive studies whose results are specific to the listener, the set of voices studied, and the terminology used to describe them. However, acoustic measures can only predict which signals will sound the same or different (Kreiman *et al.*, 2021) and have not consistently explained even the most common meanings a listener attributes to a signal, much less what it sounds like for a voice to be "strained through a long white beard" (Chandler, 1949). At the same time, descriptive studies can provide important insights into specific voices in specific contexts, but results typically do not generalize well to other contexts, other listeners, or other voices (Heidemann, 2016). In this sense, qualitative descriptions do not actually measure, or even specify, voice quality in any useful way. This is inevitable, given the density and complexity of the kinds of meaning conveyed by voice, but greatly limits the use of qualitative approaches for uncovering general truths about quality or voice perception.

It thus seems that no one analysis approach on its own is sufficient to assess quality. There is no apparent way to use acoustics to assess signal meaning, and there is no way to explain what listeners hear in acoustic terms. Descriptions of voices convey meaning but not the sound of the voice, and instrumental measures can characterize a sound, but not its meaning. A gap thus exists between physical sounds and perceived quality in our models of voice, the so-called "timbral abyss" (van Elferen, 2017; Wallmark, 2022). This abyss represents an apparently uncrossable structural "conceptual and methodological barrier that prevents the reconciliation and integration of perspectives" (Wallmark, 2022, p. 12). It is also reflected in the siloing of research efforts into studies of production, perception, or acoustics, with relatively few studies attempting to relate one domain to the others.

III. UNITING FORM AND MEANING IN THE STUDY OF VOICE QUALITY

On further consideration, however, the existence of the timbral abyss does not make sense. The arguments just put forth imply that there is no way to create a single theoretical framework describing voice production and perception together as one system. However, voice production and perception clearly function as a unified system of communication (Sidtis and Kreiman, 2012; Pisanski *et al.*, 2022). The ability to recognize familiar individuals or to assess another's emotional or physical state (even when the other is from another species or even another class of animals; Filippi *et al.*, 2017; Congdon *et al.*, 2019; Thévenet *et al.*, 2023) is widely distributed across animals, and similar cues are associated with arousal and reproductive fitness across species as well [see Andics and Faragó (2019) for review]. Such abilities require that voice production and perception coevolve, so that animals can both produce the required

20 February 2024 03:40:19

messages and understand them as listeners (Darwin, 1871; Pisanski *et al.*, 2022). The wide distribution of these abilities further points to an evolutionary origin (e.g., Darwin, 1872; Sidtis and Kreiman, 2012; Elemans *et al.*, 2015). It is, thus, unreasonable to conclude that production and perception cannot be explained in a single theoretical framework, given that they have seemingly evolved to work together in so many ways. How, then, and to what extent, can we combine these different facets of quality in a single multidisciplinary theoretical framework? What is the source of the timbral abyss, and how can we bridge it to connect perception and production in a single model of voice?

One impediment to a unified theory of voice seems to be the assumption that quality inheres solely in the listener. If quality is not consistently associable with acoustic signals, then there is no way to bridge the abyss; but is *all* meaning *always* necessarily a function of the listener, or are there attributes that carry consistent meaning, regardless of who is speaking or listening? Some terms for quality seemingly emerge fairly reliably from qualitative studies of voice. Are there also acoustic attributes that consistently vary across speakers in parallel with these terms?

To investigate this question, Lee and colleagues (Lee *et al.*, 2019; Lee and Kreiman, 2022a,b) assessed acoustic variability within and across speakers by measuring acoustic variables drawn from the psychoacoustic model of voice quality (Kreiman *et al.*, 2021; Table II) for large sets of speakers. Variables in this model have been shown to be perceptually important, so applying the psychoacoustic model ensures that the set of measurements accurately and adequately characterizes the quality (in the ANSI sense) of the voices studied. Thus, the structure of a voice space derived from these measures should define a perceptual space for quality (again, in the ANSI sense).

The variables in this model were measured from recordings of read speech produced by 158 female and male speakers of American English, Seoul Korean, Hmong, and Thai [see Lee *et al.* (2019) for analysis details]. Principal component analysis (PCA) was applied to determine which acoustic variables accounted for perceptually important differences among speakers. Two sets of parameters consistently emerged from analyses of speaker groups and of individual speakers: the balance between high-frequency harmonic and inharmonic energy in the voice and formant dispersion. All the speakers we studied, regardless of sex and age and across the native languages studied to date, seemingly shared this pattern of acoustic variability [see also Johnson and Babel (2023) for related studies of Cantonese]. These components can be thought of as defining a low-dimensional "voice space" that represents the primary ways in which voices differ from each other acoustically. The fact that the same components emerged for every speaker examined to date implies that this simple voice space may be universal and provides a single basis for quick-and-dirty assessment of voice similarity across speakers and languages [see Baumann and Belin (2010), who found a similar space in a study of 32 speakers producing French vowels].

The finding that certain patterns of variability recur so very consistently across speakers, regardless of sex or language spoken, suggests that these aspects are also evolutionarily derived. There is substantial evidence consistent with this hypothesis. The first principal component to emerge from Lee and colleagues' PCA analyses always reflected variations in the balance of harmonic and inharmonic energy in the voice source. This combination of parameters is often associated with a quality continuum from "strained" (or "pressed") to "breathy" (e.g., Gordon and Ladefoged, 2001; Kreiman *et al.*, 2012; Zacharakis *et al.*, 2014; Anikin, 2020), which signals arousal across many species (e.g., Anikin, 2020; Congdon *et al.*, 2019; Pisanski *et al.*, 2022). This same information is employed by many species [including pigs, seals, dolphins, bats, many primates, and others; see Briefer (2012) for review] for assessing the hostile or friendly intent of another animal and for communicating one's own intent, thus providing a survival benefit by potentially altering the behavior of another animal (Owren and Rendall, 1997). Formant dispersion, which also emerges consistently from early principal components, is related to the frequency of the spectral centroid and the balance of spectral energy and varies with the size of the speaker's vocal tract. This parameter serves to signal both dominance and reproductive fitness across many species (Fitch, 1997; Pisanski *et al.*, 2014), again providing a potential survival benefit. The biological importance of these parameters and their wide distribution across taxa argue strongly that they are central to the function and meaning of phonation.

Of course, the results just described derive from read speech, which may have limited the acoustic variables that emerged as shared. For example, F0 is important for voice perception in humans and other animals (e.g., Teichroeb *et al.*, 2012; Puts *et al.*, 2006), but its variability across speakers is limited in read speech, which tends to be highly stylized. Similarly, nonlinear acoustic phenomena have been repeatedly associated with arousal and negative affect in many animals (e.g., Briefer, 2012; Anikin *et al.*, 2020), but they do not occur too commonly in read utterances of "A pot of tea helps to pass the evening." While further investigations using a wider range of speech samples will clarify the range of acoustic information that is shared across speakers, we hypothesize that any measures that consistently account for variance across a range of talkers will be related to arousal, reproductive fitness, and/or dominance, consistent with the pattern just described.

Computational and physiological studies of phonation [reviewed in Zhang (2023)] provide additional evidence consistent with an evolutionary basis for the wide distribution of these particular voice qualities. These studies indicate that vocal fold shape governs the harmonic spectrum and spectral noise levels via changes in thickness: Thicker folds are associated with more high-frequency energy in the voice, and thinner folds are associated with more inharmonic excitation (e.g., Zhang, 2016a,b, 2023). Vocal fold thickness is also an important part of airway protection, which is the primary function of the vocal folds and larynx.

20 February 2024 03:40:19

This function is shared across species, and the larynx has been highly conserved evolutionarily, with notable structural similarities across animals ranging from crocodiles to virtually every species of mammal (Negus, 1949). Thus, those aspects of phonation that are oldest—dating at least to the split from reptiles—are also associated with control of the acoustic parameters that are associated with the quality dimensions that appear to characterize voice across species.

IV. THE TIMBRAL ABYSS IS NOT A BOTTOMLESS PIT

Empirical evidence, thus, points to a few acoustic factors that define a simple two-dimensional voice space that applies to all voices studied to date. The dimensions of this space (variability in the balance of harmonic and inharmonic energy and formant dispersion) align well with those that commonly emerge from descriptive studies of voice (brightness, breathiness, roughness, and richness). This correspondence between the most common terms for voice and parameters that define the human acoustic voice space suggests that the meaning voices carry rests on a bedrock of biology. Voice perception and production are structured as they are because these particular dimensions, evolved over time, provide a survival benefit, possibly apply to every voice, *and thus carry a consistent meaning regardless of who is talking or who is listening*. That is, we (and other animals) have evolved to produce and to perceive voices in ways that reflect size and state of arousal, and these factors form a basis for both voice acoustics and the language we most commonly use to describe voices. These dimensions comprise part of the biological purpose of phonation and, hence, its meaning: "Breathy" and "rough" remain useful because they directly link bodies and signals to perceived voices and to biological meaning, seemingly without the need to consider the specific perceptual, cognitive, social, or emotional context surrounding the act of hearing, and because they reliably carry (seemingly) universal meanings.

Wallmark and Kendall (2018) have suggested a similar association between physical and perceptual measures of quality, pointing out that some commonly applied descriptors reflect the fact that voices come from bodies and that this may account in part for the fact that these terms in particular tend to reappear across studies, cultures, and languages. The arguments in this paper take this account further, by examining not just why some descriptors link bodies to perceived voices, but also why this specific set of descriptors serves this purpose.

Of course, there is much more to voice acoustics than just these few shared dimensions. In fact, nearly half of acoustic variance for individual voices is idiosyncratic, presumably representing individual anatomy, habits, stylistic flourishes, and other such factors that help to link signals to particular speakers (Lee *et al.*, 2019). The small set of acoustic measures discussed in this paper is also inadequate to completely specify the sound of a voice in the ANSI sense, although they are a subset of the parameters in a psychoacoustic model that does specify why specific voice samples sound the same or different (Kreiman *et al.*, 2021). The qualitative dimensions discussed in this paper are similarly only a small subset of the labels or phrases that can be used to describe voice quality (e.g., Pedersen, 2008). Finally, finding empirical support for a small set of qualitative descriptors of voice does not mean such descriptors are good tools for quantifying quality. There is ample evidence that ratings on scales like "breathiness" and "roughness" are unreliable and subject to many kinds of measurement error (Kreiman and Gerratt, 2000).

Given the abundance of possible measures of voice and of ways of describing what is heard, it is unlikely that a model that completely maps from one domain to another will emerge, either in theory or in practice (Hermes, 2023). Nevertheless, the points of coincidence identified in this paper between qualitative and quantitative analyses of voice, although only a small part of "voice" as a whole, do suggest that certain fundamental kinds of information link signals and specific aspects of the meaning they convey, without reference to external variables or context. In other words, although the parameters described here in no way comprise a comprehensive model of voice quality, these results do suggest that there exists a bedrock of meaning, derived from the biological functions subserved by voice, that seemingly underlies and unites qualitative and instrumental approaches to voice. This foundation explains some aspects of the meaning of voice in terms of specific aspects of production and acoustics, and vice versa, thus spanning, at least in part, the timbral abyss.

The task remaining for humanists and empiricists alike is to consider the extent to which a shared theoretical foundation can inform and advance our work. Understanding the meanings that inhere in voices, versus those that derive from listener-based factors, could inform humanistic discussions of voice quality, and a focus on those acoustic aspects of voice that are inherently and consistently meaningful could guide and structure the development of better acoustic and biomechanical models of voice. Exploiting this common foundation shared by humanistic and empirical approaches to voice could help integrate studies of physical signals and their meaning, leading eventually to a truly interdisciplinary approach to voice.

ACKNOWLEDGMENTS

The author thanks Zhaoyan Zhang and Nina Eidsheim for helpful discussions and comments on earlier versions of this paper, a preliminary version of which appears in the *Proceedings of the 2023 International Congress of Phonetic Sciences* (Kreiman, 2023).

AUTHOR DECLARATIONS

Conflict of Interest

The author has no conflicts to disclose.

DATA AVAILABILITY

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

20 February 2024 03:40:19

Alluri, V., and Toiviainen, P. (**2012**). "Effect of enculturation on the semantic and acoustic correlates of polyphonic timbre," Music Percept. **20**, 297–310.
Andics, A., and Faragó, T. (**2019**). "Voice perception across species," in *The Oxford Handbook of Voice Perception*, edited by S. Früholz and P. Belin (Oxford University, Oxford), pp. 363–392.
Anikin, A. (**2020**). "A moan of pleasure should be breathy: The effect of voice quality on the meaning of human nonverbal vocalizations," Phonetica **77**, 327–349.
Anikin, A., Pisanski, K., and Reby, D. (**2020**). "Do nonlinear vocal phenomena signal negative valence or high emotion intensity?," R. Soc. Open Sci. **7**, 201306.
ANSI (**1960**). ANSI S1.1-1960, *Acoustical terminology* (American National Standards Institute, New York).
Austin, G. (**1806**). *Chironomia* (Cadell and Davies, London), pp. 553–554 (reprinted by Southern University Press, Carbondale, IL, 1966).
Barsties, B., and De Bodt, M. (**2015**). "Assessment of voice quality: Current state-of-the-art," Auris Nasus Larynx **42**, 183–188.
Barsties, B., Ulozaitė–Staniėn, N., Petrauskas, T., Uloza, V., and Maryn, Y. (**2019**). "Diagnostic accuracy of dysphonia classification of DSI and AVQI," Laryngoscope **129**, 692–698.
Baumann, O., and Belin, P. (**2010**). "Perceptual scaling of voice identity: Common dimensions for different vowels and speakers," Psychol. Res. **74**, 110–120.
Becker, K., Khan, S. D., and Zimman, L. (**2022**). "Beyond binary gender: Creaky voice, gender, and the variationist enterprise," Lang. Var. Change **34**, 215–238.
Bele, I. (**2007**). "Dimensionality in voice quality," J. Voice **21**, 257–272.
Bregman, A. S. (**1990**). *Auditory Scene Analysis: The Perceptual Organization of Sound* (MIT, Cambridge, MA).
Briefer, E. F. (**2012**). "Vocal expression of emotions in mammals: Mechanisms of production and evidence," J. Zool. **288**, 1–20.
Buder, E. H. (**2000**). "Acoustic analysis of voice quality: A tabulation of algorithms 1902–1990," in *Voice Quality Measurement*, edited by R. D. Kent and M. J. Ball (Singular, San Diego, CA), pp. 119–244.
Carron, M., Rotureau, T., Dubois, F., Misdariis, N., and Susini, P. (**2017**). "Speaking about sounds: A tool for communication on sound features," J. Des. Res. **15**, 85–109.
Chandler, R. (**1939**). *The Big Sleep* (Vintage Crime, New York). Page references are to the 1992 edition.
Chandler, R. (**1949**). *The Little Sister* (Vintage Crime, New York). Page references are to the 1992 edition.
Chandler, R. (**1953**). *The Long Goodbye* (Vintage Crime, New York). Page references are to the 1992 edition.
Congdon, J. V., Hahn, A. H., Filippi, P., Campbell, K. A., Hoanget, J., Scully, E. N., Bowling, D. L., Reber, S. A., and Sturdy, C. B. (**2019**). "Hear them roar: A comparison of black-capped chickadee (*Poecile atricapillus*) and human (*Homo sapiens*) perception of arousal in vocalizations across all classes of terrestrial vertebrates," J. Comp. Psychol. **133**, 520–541.
Darwin, C. (**1871**). *The Descent of Man, and Selection in Relation to Sex* (Princeton University, Princeton, NJ).
Darwin, C. (**1872**). *The Expression of the Emotions in Man and Animals* (John Murray, London).
Dolan, E. I., and Rehding, A. (eds.) (**2018**). *The Oxford Handbook of Timbre* (Oxford University, Oxford).
Eidsheim, N. (**2019**). *The Race of Sound* (Duke University, Durham, NC).
Elemans, C., Rasmussen, J., Herbst, C., Düring, D., Zollinger, S., Brumm, H., Srivastava, K., Svane, N., Ding, M., Larsen, O. N., Sober, S. J., and Švec, J. G. (**2015**). "Universal mechanisms of sound production and control in birds and mammals," Nat. Commun. **6**, 8978.
Fagel, W. P. F., van Herpt, I. W. A., and Boves, L. (**1983**). "Analysis of the perceptual qualities of Dutch speakers' voice and pronunciation," Speech Commun. **2**, 315–326.
Fales, C. (**2002**). "The paradox of timbre," Ethnomusicology **46**, 56–95.
Filippi, P., Congdon, J. V., Hoang, J., Bowling, D. L., Reber, S. A., Pašukonis, A., Hoeschele, M., Ocklenburg, S., de Boer, B., Sturdy, C. B., Newen, A., and Güntürkün, O. (**2017**). "Humans recognize emotional arousal in vocalizations across all classes of terrestrial vertebrates: Evidence for acoustic universals," Proc. R. Soc. B **284**, 20170990.
Fitch, W. T. (**1997**). "Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques," J. Acoust. Soc. Am. **102**, 1213–1222.
Garellek, M., Keating, P., Esposito, C. M., and Kreiman, J. (**2013**). "Voice quality and tone identification in White Hmong," J. Acoust. Soc. Am. **133**, 1078–1089.
Gelfer, M. P. (**1988**). "Perceptual attributes of voice: Development and use of rating scales," J. Voice **2**, 320–326.
Gerratt, B. R., and Kreiman, J. (**2001**). "Toward a taxonomy of nonmodal phonation," J. Phon. **29**, 365–381.
Gobl, C., and Ní Chasaide, A. (**2003**). "The role of voice quality in communicating emotion, mood, and attitude," Speech Commun. **40**, 189–212.
Gordon, M., and Ladefoged, P. (**2001**). "Phonation types: A cross-linguistic overview," J. Phon. **29**, 383–406.
Hajda, J. M., Kendall, R. A., Carterette, E. C., and Harshberger, M. L. (**1997**). "Methodological issues in timbre research," in *Perception and Cognition of Music*, edited by I. Deliege and J. Sloboda (Psychology, Hove, UK), pp. 253–306.
Hammarberg, B., and Gauffin, J. (**1995**). "Perceptual and acoustic characteristics of quality differences in pathological voices as related to physiological aspects," in *Vocal Fold Physiology: Voice Quality Control*, edited by O. Fujimura and M. Hirano (Singular, San Diego), pp. 283–300.
Heidemann, K. (**2016**). "A system for describing vocal timbre in popular song," Music Theory Online **22**, 1–17.
Hermes, D. J. (**2023**). *The Perceptual Structure of Sound* (Springer, Eindhoven, Netherlands).
Hillenbrand, J., Cleveland, R., and Erickson, R. (**1994**). "Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech," J. Speech Lang. Hear. Res. **37**, 769–778.
Hirano, M. (**1981**). *Clinical Examination of Voice* (Springer, New York).
Johnson, K. A., and Babel, M. (**2023**). "The structure of acoustic voice variation in bilingual speech," J. Acoust. Soc. Am. **153**, 3221–3238.
Keating, P. A., Garellek, M., and Kreiman, J. (**2015**). "Acoustic properties of different kinds of creaky voice," in *Proceedings of the 18th International Congress of Phonetic Sciences*, August 10–14, Glasgow, UK (International Phonetic Association, London).
Kempster, G. B., Gerratt, B. R., Verdolini Abbott, K., Barkmeier-Kraemer, J., and Hillman, R. E. (**2009**). "Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol," Am. J. Speech Lang. Pathol. **18**, 124–132.
Kendall, R. A., and Carterette, E. C. (**1993**). "Verbal attributes of simultaneous wind instrument timbres: I. von Bismarck's adjectives," Music Percept. **10**, 445–468.
Kreiman, J. (**2023**). "Labels for voices," in *Proceedings of the 2023 International Congress of Phonetic Sciences*, August 7–11, Prague, Czech Republic (GUARANT, Prague, Czech Republic).
Kreiman, J., and Gerratt, B. R. (**1996**). "The perceptual structure of pathologic voice quality," J. Acoust. Soc. Am. **100**, 1787–1795.
Kreiman, J., and Gerratt, B. R. (**1998**). "Validity of rating scale measures of voice quality," J. Acoust. Soc. Am. **104**, 1598–1608.
Kreiman, J., and Gerratt, B. R. (**2000**). "Sources of listener disagreement in voice quality assessment," J. Acoust. Soc. Am. **108**, 1867–1876.
Kreiman, J., Gerratt, B. R., and Berke, G. S. (**1994**). "The multidimensional nature of pathologic vocal quality," J. Acoust. Soc. Am. **96**, 1291–1302.
Kreiman, J., Lee, Y., Garellek, M., Samlan, R., and Gerratt, B. R. (**2021**). "Validating a psychoacoustic model of voice quality," J. Acoust. Soc. Am. **149**, 457–465.
Kreiman, J., Shue, Y.-L., Chen, G., Iseli, M., Gerratt, B. R., Neubauer, J., and Alwan, A. (**2012**). "Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation," J. Acoust. Soc. Am. **132**, 2625–2632.
Kreiman, J., and Sidtis, D. (**2011**). *Foundations of Voice Studies* (Wiley, Malden, MA).
Lechien, J. R., Geneid, A., Bohlender, J. E., Cantarella, G., Avellaneda, J. C., Desuter, G., Sjogren, E. V., Finck, C., Hans, S., Hess, M., Oguz, H., Remacle, M. J., Schneider-Stickler, B., Tedla, M., Schindler, A., Vilaseca, I., Zabrodsky, M., Dikkers, F. G., and Crevier-Buchman, L. (**2023**). "Consensus for voice quality assessment in clinical practice: Guidelines of the European Laryngological Society and Union of the European Phoniatricians," Eur. Arch. Otorhinolaryngol. **280**, 5459–5473.
Lee, Y., Keating, P., and Kreiman, J. (**2019**). "Acoustic voice variation within and between speakers," J. Acoust. Soc. Am. **146**, 1568–1579.
Lee, Y., and Kreiman, J. (**2022a**). "Acoustic voice variation in spontaneous speech," J. Acoust. Soc. Am. **151**, 3462–3472.

20 February 2024 03:40:19

Lee, Y., and Kreiman, J. (**2022b**). "Linguistic versus biological factors governing acoustic voice variation," in *Proceedings of INTERSPEECH 2022*, September 18–22, Incheon, South Korea (International Speech Communication Association, Baixas, France), pp. 640–643.
Lichte, W. H. (**1941**). "Attributes of complex tones," J. Exp. Psychol. **28**, 455–480.
López, S., Riera, P., Assaneo, M. F., Eguía, M., Sigman, M., and Trevisan, M. (**2013**). "Vocal caricatures reveal signatures of speaker identity," Sci. Rep. **3**, 3407.
Malawey, V. (**2020**). *A Blaze of Light in Every Word: Analyzing the Popular Singing Voice* (Oxford University, New York).
Maryn, Y., Roy, N., De Bodt, M., Van Cauwenberge, P., and Corthals, P. (**2009**). "Acoustic measurement of overall voice quality: A meta-analysis," J. Acoust. Soc. Am. **126**, 2619–2634.
McAdams, S., Giordano, B., Susini, P., Peeters, G., and Rioux, V. (**2006**). "A meta-analysis of acoustic correlates of timbre dimensions," J. Acoust. Soc. Am. **120**, 3275–3276.
Murry, T., and Singh, S. (**1980**). "Multidimensional analysis of male and female voices," J. Acoust. Soc. Am. **68**, 1294–1300.
Negus, V. E. (**1949**). *The Comparative Anatomy and Physiology of the Larynx* (Grune and Stratton, New York).
Osgood, C. E., Succi, G. J., and Tannenbaum, P. H. (**1957**). *The Measurement of Meaning* (University of Illinois, Champaign, IL).
Owren, M. J., and Rendall, D. (**1997**). "An affect-conditioning model of nonhuman primate vocal signaling," in *Perspectives in Ethology*, edited by D. H. Owings, M. D. Beecher, and N. S. Thompson (Plenum, New York), Vol. 12, pp. 299–346.
Paz, K. E. d. S., Almeida, A. A., Behlau, M., and Lopes, L. W. (**2022**). "Descriptors of breathy, rough, and healthy voice quality in common sense," Audiol. Commun. Res. **27**, e2602.
Pedersen, T. H. (**2008**). *The Semantic Space of Sounds* (Delta, Hørsholm, Denmark).
Pisanski, K., Bryant, G. A., Cornec, C., Anikin, A., and Reby, D. (**2022**). "Form follows function in human nonverbal vocalisations," Ethol. Ecol. Evol. **34**, 303–321.
Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J. M., Röder, S., Andrews, P. W., Fink, B., DeBruine, L. M., Jones, B. C., and Feinberg, D. R. (**2014**). "Vocal indicators of body size in men and women: A meta-analysis," Anim. Behav. **95**, 89–99.
Plomp, R. (**1970**). "Timbre as a multidimensional attribute of complex tones," in *Frequency Analysis and Periodicity Detection in Hearing*, edited by R. Plomp and G. Smoorenburg (Seithoff, Leiden, Netherlands), pp. 397–414.
Porcello, T. (**2004**). "Speaking of sound: Language and the professionalization of sound-recording engineers," Soc. Stud. Sci. **34**, 733–758.
Pratt, R., and Doak, P. (**1976**). "A subjective rating scale for timbre," J. Sound Vib. **45**, 317–328.
Puts, D. A., Gaulin, S. J. C., and Verdolini, K. (**2006**). "Dominance and the evolution of sexual dimorphism in human voice pitch," Evol. Hum. Behav. **27**, 283–296.
Ryan, W. J., and Burk, K. W. (**1974**). "Perceptual and acoustic correlates of aging in the speech of males," J. Commun. Disord. **7**, 181–192.
Sidtis, D., and Kreiman, J. (**2012**). "In the beginning was the familiar voice: Personally familiar voices in the evolutionary and contemporary biology of communication," Integr. Psychol. Behav. Sci. **46**, 146–159.
Södersten, M., and Lindestad, P. A. (**1990**). "Glottal closure and perceived breathiness during phonation in normally speaking subjects," J. Speech Lang. Hear. Res. **33**, 601–611.
Streeter, L. A., MacDonald, N. H., Apple, W., Krauss, R. M., and Galotti, K. M. (**1983**). "Acoustic and perceptual indicators of emotional stress," J. Acoust. Soc. Am. **73**, 1354–1360.
Sun, X. (**2002**). "Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, May 13–17, Orlando, FL (IEEE, New York), pp. 333–336.
Sundberg, J. (**1987**). *The Science of the Singing Voice* (Northern Illinois University, de Kalb, IL).
Teichroeb, L. J., Riede, T., Kotrba, R., and Lingle, S. (**2012**). "Fundamental frequency is key to response of female deer to juvenile distress calls," Behav. Process. **92**, 15–23.
Thévenet, J., Papet, L., Coureaud, G., Boyer, N., Levréro, F., Grimault, N., and Mathevon, N. (**2023**). "Crocodile perception of distress in hominid baby cries," Proc. R. Soc. B **290**, 20230201.
van Elferen, I. (**2017**). "Drastic allure: Timbre between the sublime and the grain," Contemp. Music Rev. **36**, 614–632.
van Elferen, I. (**2018**). "Timbrality: The vibrant aesthetics of tone color," in *The Oxford Handbook of Timbre*, edited by E. I. Dolan and A. Rehding (Oxford University, Oxford), pp. 68–91.
Vélez, D. V. (**2018**). "The matter of timbre: Listening, genealogy, sound," in *The Oxford Handbook of Timbre*, edited by E. I. Dolan and A. Rehding (Oxford University, Oxford), pp. 22–51.
Voiers, W. D. (**1964**). "Perceptual bases of speaker identity," J. Acoust. Soc. Am. **36**, 1065–1073.
von Bismarck, G. (**1974**). "Timbre of steady tones: A factorial investigation of its verbal attributes," Acta Acust. united Acust. **30**, 146–159.
Wallmark, Z. (**2022**). *Nothing but Noise: Timbre and Musical Meaning at the Edge* (Oxford University, Oxford).
Wallmark, Z., and Kendall, R. A. (**2018**). "Describing sound: The cognitive linguistics of timbre," in *Oxford Handbook of Timbre*, edited by E. I. Dolan and A. Rehding (Oxford University, Oxford), pp. 578–608.
Wallmark, Z., and Kendall, R. A. (**2021**). "Describing sound: The cognitive linguistics of timbre," in *The Oxford Handbook of Timbre*, edited by E. I. Dolan and A. Rehding (Oxford University Press, Oxford), pp. 579–608.
Webb, A. L., Carding, P. N., Deary, I. J., MacKenzie, K., Steen, N., and Wilson, J. A. (**2003**). "The reliability of three perceptual evaluation scales for dysphonia," Eur. Arch. Otorhinolaryngol. **261**, 429–434.
Wirz, S., and MacKenzie Beck, J. (**1995**). "Assessment of voice quality: The vocal profiles analysis scheme," in *Perceptual Approaches to Communication Disorders*, edited by S. Wirz (Whurr, London), pp. 39–55.
Zacharakis, A., Pastiadis, K., and Reiss, J. D. (**2014**). "An interlanguage study of musical timbre semantic dimensions and their acoustic correlates," Music Percept. **31**, 339–358.
Zhang, Z. (**2016a**). "Mechanics of human voice production and control," J. Acoust. Soc. Am. **140**, 2614–2635.
Zhang, Z. (**2016b**). "Cause-effect relationship between vocal fold physiology and voice production in a three-dimensional phonation model," J. Acoust. Soc. Am. **139**, 1493–1507.
Zhang, Z. (**2023**). "Vocal fold vertical thickness in human voice production and control: A review," J. Voice (published online).

20 February 2024 03:40:19