LABELS FOR VOICES

Jody Kreiman

Departments of Head and Neck Surgery and Linguistics, UCLA
jkreiman@ucla.edu

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 leads 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 acoustic characterizations have also had limited success: acoustics 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 voice 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 voice.

Keywords: voice quality, terminology

1. INTRODUCTION

The problem of how to characterize voice quality is an endless source of debate and frustration across disciplines. The richness of the vocabulary available to describe voice is overwhelming, but a shared vocabulary (for example, for clinical or pedagogical use) has not been validated, despite many efforts [e.g., 1, 2]. Although there is some agreement that a few specific terms (especially breathiness and roughness) are important dimensions of quality, listeners do not agree about the extent to which a given voice demonstrates these attributes, or even whether or not they are present at all [3]. Further, the relationship between such terms and the information listeners glean from voice is not clear, so that description does not precisely predict what listeners hear or the way in which voices convey that information. This weak linkage has led some scholars to conclude that language can never adequately specify what we hear [4, 5]. Others have argued that terminology derives from whimsy, analogy, metaphor, and historical tradition and lacks an empirical basis, so that language-based scales are theoretically inappropriate measurement tools *a priori* [6]. However, efforts to provide meaningful acoustic characterizations of the messages conveyed by voice have also had limited success (see [6] for review), and it remains difficult to predict even basic attributes of the speaker (for example, sex, age, emotional state, identity) from instrumental measures. Acoustic parameters may precisely quantify physical sound patterns [7], but beyond that they cannot explain the meaning a listener attributes to a signal, because such meaning derives not only from the sound, but also from the emotional, situational, and cultural context in which the utterance takes place and from the listener's attitudes and background, among many other factors. Listener performance on relatively simple tasks like determining whether two voice samples represent the same or two different speakers (or even the same or two different tokens) cannot be explained by the properties of the signal without reference to higher-level cognitive variables reflecting the listener's contribution to the perceptual process [8, 9].

Thus, a gap emerges between physical sounds and perceived quality in our models of voice, the so-called "timbral abyss" [10]. Acoustic studies quantify signals and link them causally to the bodies that produced them, but do not predict or provide access to any broader meaning conveyed, while studies of the meanings carried by voice are not presently able to explain how such meaning emerges from discrete acoustic signals. How, and to what extent, can we combine these different facets of quality in a single multidisciplinary theoretical framework?

In this paper, I describe a possible partial solution to this long-standing issue in the study of voice quality. I begin by briefly reviewing the literature on descriptive terms for voice from different disciplines. I will argue that a small number of dimensions consistently emerge from these studies (along with a multitude of other terms that vary widely from study to study), and that not only lay individuals, but scholars and clinicians persist in using these terms despite the fact that they cannot be applied reliably to describe or quantify signals. Next, I review work

showing that acoustic variability within and across virtually all speakers can be characterized by a few dimensions that are shared by everyone, regardless of the speaker's gender, language spoken, or the kind of speech sample produced, again accompanied by a large number of other parameters whose relevance depends on the idiosyncrasies of the particular voice in question. Finally, I argue that frequently-emerging descriptive terms align well with these shared acoustic dimensions, which are associated with physical size 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 voice signal are part of our evolutionary heritage. This provides a link between qualitative and quantitative approaches to measuring voice, forming a common foundation for both kinds of study and a potential basis for truly interdisciplinary approaches to voice studies.

2. THE "DUAL NATURE" OF VOICE QUALITY

Critical theorists (e.g., [11]) propose that the timbral abyss arises from the apparent dual nature of voice quality. From an empirical perspective, quality ultimately depends on speech production (vocal fold movements, vocal tract configurations, and so on (e.g., [12]). This system creates the acoustic voice signal, which can be quantified and serves as input to the voice processing (perceptual) mechanism. From a humanistic perspective, however, voice quality resides in the listener, and not in the speaker. Rather like a tree falling in the woods, which generates meaningless vibrations in the absence of a hearer, a speaker produces something to hear, but what is heard (quality) depends not just on the physical signals, but also on the listener's affect and memory, the conversational setting, cultural structures, and a multitude of other factors [6, 12, 13]. Descriptive terminology remains the most common approach to capturing these aspects of perception, which cannot be assessed by acoustic measures alone because they are not solely functions of the voice signal. Such terminology is relied upon in clinical settings, in humanistic research, and in common discourse.

Thus a gap emerges between acoustic and qualitative approaches to quality. Acoustic studies can provide measures that can reasonably be applied to other voices in other studies, and can precisely quantify quality in the ANSI sense [14] as what makes one voice sound the same or different from another [7]. Qualitative studies can provide very detailed descriptions of specific voices in specific contexts, but results typically do not generalize well to other contexts or other voices [12]. 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 in general. Acoustic measures also cannot fully characterize quality, because they cannot tap into listeners' contributions to what they hear. Thus, no one kind of analysis appears to be adequate on its own to assess quality, and there is no obvious way to reconcile the different approaches, so that the timbral abyss appears uncrossable.

3. QUALITATIVE STUDIES

Qualitative studies of voice have a long history (see [6] for review), and are supplemented by extensive studies of musical timbre which ask similar questions and use similar methods (see [15] for review). A variety of techniques have been applied to explore the semantic dimensions of quality. Typically authors ask listeners to rate voices (or other sounds) on sets of semantic-differential scales [16] and then apply factor analysis to reveal the structure underlying the ratings [e.g., 16, 17, 18]. For example, Bele [2] asked listeners to rate normal voices on 15 visual analog scales. Four underlying dimensions emerged from factor analyses of these data: sonority, irregularity, noise, and phonatory effort. Similar dimensions (severity of pathology, roughness, breathiness, weakness, and strain) emerged from a study of hoarse voices [17]; and the popular CAPE-V protocol for clinical voice assessment includes scales for severity of pathology, roughness, breathiness, and strain [1]. Across studies, despite differences in the voices, listeners, and scales under examination (see [6] for review), a small set of dimensions consistently emerges across papers. 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., 18-21]. Interestingly, similar sets of scales have emerged across cultures and languages [22, 23], and from studies of instrumental timbre and animal vocalization.

4. ACOUSTIC VARIATIONS WITHIN AND AMONG SPEAKERS

The question thus arises: Why these terms, and not others? Recent studies of within- and between-speaker acoustic variability suggest a possible answer. Analyses of the speech of large groups of female and male speakers of English, Seoul Korean, Hmong, and Thai [24-26] show remarkably consistent patterns of acoustic variability that are seemingly shared by all speakers, regardless of sex, age, or native language. These factors can be thought of as a low-dimensional "voice space" that represents the ways in which voices differ from each other acoustically. The first dimension of this space always reflects 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" (or "bright") to "breathy" [27, 28], which signals arousal across many species ([27, 29]). The second dimension is associated with formant dispersion, which is sexually dimorphic and varies with the size of the vocal tract. Formant dispersion serves to signal both dominance and reproductive fitness across many species [30]. A second set of dimensions describes variability that is shared by speakers with a common native language, but not by speakers of other languages. These dimensions reflect the phonological structure of the language being spoken [25]: For example, a factor representing H1-H2 (the difference in the amplitudes of the first and second harmonics) appears in the space for speakers of Hmong, which has a phonemic contrast between breathy and modal phonation, but not for English speakers, given that English does not have such a contrast. A final set of dimensions accounts for idiosyncratic variance components that vary from speaker to speaker. Thus, acoustic voice spaces appear to be structured first by biologically driven factors that describe acoustic variability for virtually every voice, secondly by characteristics of the language spoken, and lastly by other miscellaneous attributes of an individual's voice.

5. CONNECTING QUALITATIVE AND ACOUSTIC ASPECTS OF VOICE

The results in Section 3 strongly parallel those in Section 4. The primary dimensions that emerge from qualitative voice research—brightness, breathiness, roughness, and richness, which are associated with the distribution of spectral energy, irregularity/noise, and the location of the spectral centroid—are in essence the same dimensions that emerge from quantitative studies of voice acoustics (variability in the balance of harmonic and inharmonic energy and formant dispersion). These characteristics have empirical associations with biologically important voice characteristics, and are meaningful attributes of vocalization across species. This correspondence between the most common terms for voice and parameters that define the (seemingly universal) human acoustic voice space suggests that the meaning voices carry rests on a bedrock of biology. That is, we have evolved to phonate and to perceive voices in ways that reflect size and state of arousal, and these factors are the basis for both voice acoustics and the language we most commonly use to describe voices.

Wallmark and Kendall [31] have previously 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 re-appear 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, 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 [32]. This small set of acoustic measures is also inadequate to completely specify the sound of a voice in the ANSI sense, although they are an important part of a psychoacoustic model that does specify why specific voice samples sound the same or different [7]. The qualitative and acoustic dimensions discussed in this paper are also only a tiny subset of the measures and labels that can be used to assess or describe voice quality. Recall that both factor analytic and acoustic analyses produced a mass of idiosyncratic detail along with a few widely applicable dimensions. Further, language, as noted above, is a virtually limitless tool for evoking the meaning of a voice. It may be (in fact, it seems likely) that a model that completely maps from one domain to another is both theoretically and empirically impossible. This conclusion is consistent with the observation that many descriptions of voice convey its meaning, but not necessarily the way it sounds. Consider, for example, Raymond Chandler's description of private investigator Philip Marlowe's first encounter with a potential client:

> 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. [33]

Acoustic analysis would be uninformative in this case, which reflects Marlowe's evaluation of what he hears, but not the actual sounds themselves, which are rather hard to imagine.

Nevertheless, the small point of coincidence between qualitative and quantitative analyses of voice suggests that there is at least a foundation of information that links signals and specific aspects of the meaning they convey—in other words, that dependably links speakers to listeners without reference to external variables or context. 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 both qualitative and acoustic approaches to voice. This foundation provides a way of explaining 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 foundation can inform and advance their 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 the proposed 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.

6. REFERENCES

[1] 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.

[2] Bele, I. 2007. Dimensionality in voice quality, *J. Voice* 21, 257 – 272.

[3] Kreiman, J., Gerratt, B.R., and Berke, G.S. 1994. The multidimensional nature of pathologic vocal quality. *J. Acoust. Soc. Am.* 96, 1291 – 1302.

[4] Kendall, R.A., and Carterette, E.C. 1993. Verbal attributes of simultaneous wind instrument timbres: I. von Bismarck's adjectives. *Music Percep.* 10, 445 – 468.

[5] Malawey, V. 2020. *A Blaze of Light in Every Word: Analyzing the Popular Singing Voice*. Oxford.

[6] Kreiman, J., Sidtis, D. 2011. *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception.* Wiley-Blackwell.

[7] Kreiman, J., Lee, Y., Garellek, M., Samlan, R., Gerratt, B.R. 2021. Validating a psychoacoustic model of voice quality. *J. Acoust. Soc. Am.* 149, 457 – 465.

[8] Kreiman, J., Gerratt, B.R., and Khan, S. D.(2010. Effects of native language on perception of voice quality. *J. Phonetics* 38, 588 – 593.

[9] Lavan, N., Burston, L.F.K., and Garrido, L. 2018. How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British J. Psychol.* 110, 576 – 593.

[10] Wallmark, Z. 2022. *Nothing but Noise: Timbre and Musical Meaning at the Edge.* Oxford.

[11] Fales, C. 2002. The paradox of timbre. *Ethnomusicology* 46, 56 – 95.

[12] Heidemann, K. 2016. A system for describing vocal timbre in popular song. *Music Theory Online* 22, 1 – 17.

[13] Hajda, J.M., Kendall, R.A., Carterette, E.C., Harshberger, M.L. 1997. Methodological issues in timbre research. In: Deliege, I., Sloboda, J. (eds), *Perception and Cognition of Music*. Psychology Press, 253 – 306.

[14] ANSI. 1960/1994. *Psychoacoustic Terminology: Timbre.* American National Standards Institute.

[15] Dolan, E.I., Rehding, A. (eds) 2018. *The Oxford Handbook of Timbre* (Oxford).

[16] Osgood, C.E., Succi, G.J., Tannenbaum, P.H. 1957. *The Measurement of Meaning*. University of Illinois Press.

[17] Hirano, M. 1981. *Clinical Examination of Voice.* Springer.

[18] Voiers, W.D. 1964. Perceptual bases of speaker identity. *J. Acoust. Soc. Am*. 36, 1065 – 1073.

[19] Lichte, W.H. 1941. Attributes of complex tones. *J. Exp. Psychol.* 28, 455 – 480.

[20] Pratt, R., Doak, P. 1976. A subjective rating scale for timbre. *Journal of Sound and Vibration* 45, 317 – 328.

[21] von Bismarck, G. 1974. Timbre of steady tones: A factorial investigation of its verbal attributes. *Acta Acustica united with Acustica* 30, 146 – 159.

[22] Zacharakis. A., Pastiadis, K., Reiss, J.D. 2014. An interlanguage study of musical timbre semantic dimensions and their acoustic correlates. *Music Percep.* 31, 339 – 358.

[23] Alluri, V., Toiviainen, P. 2012. Effect of enculturation on the semantic and acoustic correlates of polyphonic timbre. *Music Percep.* 20, 297 – 310.

[24] Lee, Y., Keating, P., Kreiman, J. 2019. Acoustic voice variation within and between speakers. *J. Acoust. Soc. Am* 146, 1568 – 1579.

[25] Lee, Y., Kreiman, J. 2022. Acoustic voice variation in spontaneous speech. *J. Acoust. Soc. Am.* 151, 3462 – 3472.

[26] Lee, Y., and Kreiman, J. 2023. Within- versus between-speaker acoustic variability in Thai. Presented at the 184th Meeting of the Acoustical Society of America.

[27] 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.
[28] Kreiman, J., Shue, Y.-L., Chen, G., Iseli, M., Gerratt, B.R., Neubauer, J., 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.
[29] Congdon, J.V., Hahn, A.H., et al. 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.
[30] 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.
[31] Wallmark, Z., Kendall, R.A. 2018. Describing sound: The cognitive linguistics of timbre. In: Dolan, E.I., Rehding, A. (eds), *The Oxford Handbook of Timbre.* Oxford, 578 – 608.
[32] Kreiman, J., Gerratt, B.R. 2000. Sources of listener disagreement in voice quality assessment. *J. Acoust. Soc. Am*. 108, 1867 – 1876.
[33] Chandler, R. 1949. *The Little Sister*. Vintage, p. 41.