1
Introduction

1.1 Why Should We Care About Voice Quality?

Whenever we speak, our voices convey information about us as individuals. Speakers may sound young, or tired, or elated, or distracted. They may sound as if they are drunk, or lying, or ill, or bearing secret, exciting news. By their voices, adult speakers usually reveal whether they are male or female, and in addition, they may signal that they come from Texas, or Wisconsin, or France. Over the telephone or radio we may recognize the speaker as someone we know, or we may form a distinct impression of the physical appearance of someone we have never seen. The impressions listeners gain from voices are not necessarily accurate; for example, everyone has known the surprise of meeting a telephone acquaintance who does not match the mental picture we have previously formed of them. Despite such occasional mismatches, voice quality is one of the primary means by which speakers project their identity – their "physical, psychological, and social characteristics" (Laver, 1980, p. 2) or their "auditory face" (Belin, Fecteau, and Bedard, 2004) – to the world.

Table 1.1 non-exhaustively summarizes some of the kinds of judgments that listeners make when listening to voices. These human abilities arise from a long evolutionary process, and many animal species, including primates (Cheney and Seyfarth, 1980), wolves (Goldman, Phillips, and Fentress, 1995), penguins (Jouventin and Aubin, 2002), frogs (Bee, 2004), and bats (Balcombe and McCracken, 1992) use vocal quality to signal or perceive size, threat, and kin relationships. Human infants' ability to recognize their mothers' voices is in place at birth (DeCasper and Fifer, 1980), and responses to maternal voices can be measured *in utero*, suggesting such abilities develop even before birth (Hepper, Scott, and Shahidullah, 1993; Kisilevsky *et al.*, 2003). Voice conveys much of the emotion and attitude communicated by speech (Williams and Stevens, 1972; Banse and Scherer, 1996; Ellgring and Scherer, 1996; Van Lancker and Pachana, 1998; Breitenstein, Van Lancker, and Daum, 2001). Alterations in voice quality relative to the speaker's normal vocal delivery may signal

Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception,
First Edition. Jody Kreiman and Diana Sidtis.

Table 1.1 Some kinds of judgments listeners make from voice.

Spoken message
Physical characteristics of the speaker
Age
Appearance (height, weight, attractiveness)
Dental/oral/nasal status
Health status, fatigue
Identity
Intoxication
Race, ethnicity
Sex
Sexual orientation
Smoker/non-smoker
Psychological characteristics of the speaker
Arousal (relaxed, hurried)
Competence
Emotional status/mood
Intelligence
Personality
Psychiatric status
Stress
Truthfulness
Social characteristics of the speaker
Education
Occupation
Regional origin
Role in conversational setting
Social status

irony or sarcasm (Van Lancker, Canter, and Terbeek, 1981). Changes in rate and fundamental frequency affect the perceived "competence" (Brown, Strong, and Rencher, 1974) or credibility (Geiselman and Bellezza, 1977) of a speaker. Voice quality provides cues that indicate order of turn-taking in conversation (Schegloff, 1998; Wells and Macfarlane, 1998) and helps resolve sentential ambiguities (Kjelgaard, Titone, and Wingfield, 1999; Schafer, Speer, Warren, and White, 2000). Listeners may also judge the speaker's sexual preference (Linville, 1998; Munson and Babel, 2007), status as native or nonnative speaker (Piske, MacKay, and Flege, 2001), and a myriad of personality factors (Scherer, 1979) based on voice quality cues.

This book describes the manner in which these kinds of information are conveyed to listeners, and how listeners draw conclusions – correctly or incorrectly – about speakers from their voices. Many of the points described are illustrated by recorded examples provided on the accompanying web site.

For example, consider the voice in audio sample 1.1. As you listen to this brief speech sample, you will probably automatically gather information about the speaker. Listeners agree that the speaker is female. Although opinions differ, listeners are likely to think that the speaker is adult but not elderly, cheerful, confident,

alert, and in good health. She is American, but does not have a pronounced regional, social, or ethnic accent. She sounds average or slightly above average in height and weight. She seems educated and is speaking carefully. She does not sound like a smoker. You probably do not recognize the voice, but it may remind you of someone you know.

Compare this talker to the voice in audio sample 1.2. This speaker is also female, but the voice sounds like a much older person. She has a New England accent, and the rhythm of her speech is unusual, making her sound rather upper-class or snobby (or merely self-conscious) to some listeners. She is not tired, depressed, or angry, but she is not obviously happy, either, and may be bored. Her voice is somewhat hoarse, suggesting that she is or has been a smoker, but she does not seem ill. Listeners disagree somewhat about her height and weight, but generally estimate that she is average or slightly below average in height, and slightly above average in weight.

The voice of a speaker with a vocal pathology is presented in audio sample 1.3. Even this short sample may produce complex impressions of old age, illness, and unattractiveness, along with a sense of the speaker's emotions or mood, intelligence, and competence. Patients who develop a voice disorder often complain that the disordered voice is not really their voice, and does not convey who they are. In some cases, patients dislike the image they portray so much that they avoid speaking, resulting in significant social and work-related difficulties. Severe voice quality problems may also interfere with speech intelligibility, creating a handicap in the communication of verbal information (Kempler and Van Lancker, 2002).

The strong impressions conveyed by voice quality are often manipulated by the media for multiple purposes. For example, in the classic film *Singin' in the Rain* (Freed, Kelly, and Donen, 1952), the shrill, loud voice of the character Lina Lamont (played by actress Jean Hagen) surprises and amuses because it does not fit her appearance (a beautiful, smiling blonde) or the elegant, poised, sophisticated personality she visually projects. This contrast – a prototypically silly voice in an elegant physique – forms a running joke throughout the film, playing off such lines as, "What do you think I am, dumb or something?" spoken in the abrasive voice stereotypically associated with a vulgar, uneducated, shrewish female. More often, voices are selected to fit the intended message. Documentary films enhance credibility through the use of a male narrator whose voice carries the stereotype of an authoritative figure who is solid, mature, calm, highly intelligent, and dignified. In the field of advertising, impressions conveyed by voice quality are integral to establishing a product image. Consider the characteristics projected by the voices typically used in advertisements for luxury automobiles. Low pitch, breathy quality, and a fairly rapid speaking rate produce the image of an intimate message from a mature but energetic male who possesses authority, sex appeal, social status, and "coolness." These vocal attributes are appropriate to the economic niche for the product and imply that its owners are powerful, sexy, and affluent.

Given the wide range of information listeners derive from voices, it is not surprising that scholars from many different disciplines have studied the production and perception of voice. Table 1.2 lists some of these disciplines, along with a sampling of typical research questions. These research questions encompass much of human existence, and indicate how central voice quality is to human life.

Table 1.2 Disciplines incorporating the study of voice and voice quality.

Discipline	*Some typical research questions*
Acoustics	Deriving reliable and meaningful acoustic measures of voices
Animal behavior	Vocal recognition of kin and social information by nonhuman animals
Biology	Biological and evolutionary significance of vocalization
Computer science, signal processing, information	Transmission, measurement, and synthesis of voice
Forensic science, law enforcement	Reliability and verification of "earwitness" testimony; assessment of truthfulness from voice
Linguistics, phonetics	Meanings of vocal quality in speech
Medicine:	
Developmental biology	Infant voice recognition
Gerontology	Voice quality changes in aging
Neurology	Brain function underlying vocal behaviors
Obstetrics	Prenatal voice perception
Otolaryngology	Voice disorders
Pediatrics	Childrens' processing of vocal information
Physiology	Control of phonation
Respiration	Role of breathing in vocalization
Surgery	Effects of surgical interventions in the vocal tract on voice; cosmetic changes for transgendered voices
Music:	
Singing	The singing voice: many questions
Vocal coaching	The effects of training on the voice
Physics	Vibrating laryngeal tissues; relation of vibration to sound; patterns of airflow through the glottis
Psychology:	
Cognitive psychology	Speaker recognition and its causes; interaction between speech recognition and voice quality
Clinical psychology	Detecting depression, psychopathology, and personality in the human voice
Social psychology	Voices as signals of social relationships including conversational turn taking, sarcasm, and successful con-artistry
Neuropsychology	Brain mechanisms underlying the perception and production of voice cuing personal identity as well as mood and motivation
Psychophysics	Relevant acoustic voice features for perception
Psycholinguistics	Voice information in meaning comprehension for grammatical structure and nonliteral meanings
Sociology	Voice types associated with social groups and their development
Speech science	Normal voice and speech production
Speech pathology	Effects of vocal pathologies on voice quality
Theater arts	Voice as artistic instrument

1.2 What is Voice? What is Voice *Quality?* The Definitional Dilemma

The terms "voice" and "voice quality" are variously used, sometimes apparently interchangeably, and deriving consistent definitions has not proven easy. Adding to the confusion, authors also discuss a range of specific voice qualities (a creaky voice, a breathy voice), qualities associated with a speaker's internal or physical state (a sad voice, a tired voice; a sexy voice), and so on, without benefit of a theoretical framework linking all these usages. We attempt to distinguish these meanings usefully by discussing the terms here.

Although a clear definition of *voice* is a prerequisite to its study, the broad range of functions subserved by voice has made it difficult to provide a single, all-purpose definition that is valid and useful across disciplines, scholarly traditions, and research applications. As voice scientist Johann Sundberg has noted (1987), everyone knows what voice is until they try to pin it down, and several senses of the term are in common use. In scientific usage (and throughout this book), the term "voice" has a physical and physiological base that refers to the acoustic signal (as generated by the voice production system), while "voice quality" refers to the perceptual impression that occurs as a result of that signal, analogous to the distinction between "frequency" (a physical property of vibration) and "pitch" (a listener's sensation). Definitions of voice fall into two general classes. In the first, voice can be defined very narrowly in physiological terms as "sound produced by vibration of the vocal folds." Were this definition applied, voice would include only those aspects of the signal that are attributable to the action of the vocal folds, and would exclude the acoustic effects of vocal tract resonances, vocal tract excitation from turbulent noise, or anything else that occurs during speech production other than the action of the vocal folds. (Chapter 2 describes the voice production process in detail.) This definition corresponds approximately to the linguistic voicing feature that phonetically distinguishes voiced from voiceless sounds (for example, /s/ from /z/) in many languages. Authors who use the term "voice" in this sense (for example, Brackett, 1971) typically distinguish voice from speech. Voice in this sense is also synonymous with the term "laryngeal source," which emphasizes the fact that vocal fold vibrations are the acoustic energy source for much of speech.

Anatomical constraints make it difficult to study voice as narrowly defined. The larynx is located fairly low in the neck (see Chapter 2), and vocal fold function is difficult to observe directly for long periods of time. Short sequences of open vowel phonation can be inspected through the use of a laryngeal mirror (see Sidebar in Chapter 2). Direct views of some aspects of laryngeal vibrations are available using endoscopic imaging technology and either stroboscopy[1] (for example, Hertegard and Gauffin, 1995) or high-speed imaging (for example, Koike and Hirano, 1973; Berry, Montequin, and Tayama, 2001; Deliyski *et al.*, 2008). Laryngeal vibrations can also be studied experimentally using excised larynx preparations (for example, van den Berg, 1968; Berry, 2001). Some authors have used the output of devices like the laryngograph

[1] A technique by which rapid vocal fold vibrations are apparently "slowed" through use of a strobe light so that they can be easily viewed.

(for example, Abberton and Fourcin, 1978) or electroglottograph (Kitzing, 1982) to measure the action of the laryngeal source. However, to study voice in its narrow sense, most researchers adopt the more practical expedient of controlling for all non-laryngeal contributions to the sounds a speaker makes by restricting voice samples to steady state vowels (usually /a/). This practice does not fully eliminate the contributions of non-laryngeal factors (such as vocal tract resonances) to the voice signal, but it does hold such factors relatively constant across utterances and talkers. This approach is the most common implementation of narrow definitions of voice.

Voice as a physiological and physical phenomenon can also be defined very broadly as essentially synonymous with speech. Besides details of vocal fold motions, voice in this sense includes the acoustic results of the coordinated action of the respiratory system, tongue, jaw, lips, and soft palate, both with respect to their average values and to the amount and pattern of variability in values over time.

The term *voice quality* belongs properly to the realm of perception, and refers to how the voice sounds to a listener. Both "voice" and "voice quality" can be defined (and approached) narrowly or broadly, and each is best considered as analogous to a two-sided coin, melding the production characteristics of one side to the perceptual characteristics of the other side. Like the term "voice," "voice quality" can be used very narrowly, to specify a single aspect of the phonatory process such as the perceived amount of unmodulated airflow present in the voice signal; less narrowly, to mean the perceived result of the process of phonation; or broadly, to mean a listener's response to the overall sound of speech. Because these terms appear in various contexts, their specific use also depends on purpose and perspective, so that providing a precise definition of either term is difficult. Definitions of voice quality abound, depending on interest and focus in each particular discipline. Listeners collate a very large amount of material when they gather information from the ongoing speech of individual talkers. Articulatory details, laryngeal settings, F0 and amplitude variations, and temporal patterning all contribute to how a speaker sounds (Banse and Scherer, 1996; cf. Sapir's (1927) notion of "speech as a personality trait"). Broad definitions of voice quality aim to reflect this fact, and generally portray quality as the end result of a complex sequence of cognitive, physiological, and acoustic events, the familiar "speech chain" (Figure 1.1; Denes and Pinson, 1993).

According to the speech chain, sound is produced by the actions of the speech production mechanism. The acoustic signal then travels to the ears of the listener and back to the speaker in the form of feedback. The auditory percept (a stretch of speech) is first processed peripherally within the mechanisms of the ear, followed by neurological activation of the 8th cranial nerve and the auditory pathway to the receiving areas in the brain (as described in Chapter 3). As increasingly complex cognitive processes are invoked, the stretch of speech under analysis may be described in terms of a number of complex messages (Table 1.1). As briefly reviewed above, voice patterns convey information (more or less successfully) about affect, attitude, psychological state, pragmatics, grammatical function, sociological status, and many aspects of personal identity, all of which emerges from this complex enfolding of phonatory, phonetic, and temporal detail.

Precisely which stage in this chain of events receives focus depends on the interest of the practitioner or experimenter, or on the task faced by the listener, and individual definitions of voice quality may vary according to intellectual tradition (Table 1.2).

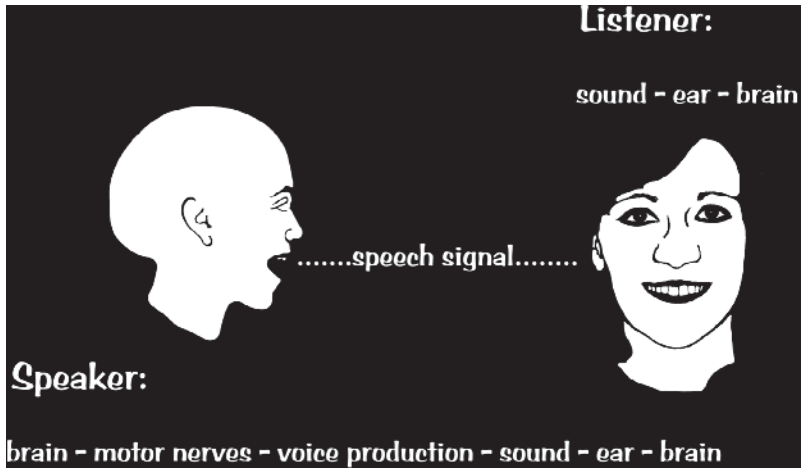


Figure 1.1 The speech chain, showing the transmission of information from a speaker to a listener. Voice production engages systems for respiration, phonation, and articulation.

For example, when surgeons use the term voice quality, they typically think in terms of physiological function, with secondary concern for the exact perceived quality that results from phonation. A typical physiologically-oriented definition characterizes voice quality as "sounds generated by the voice organ … by means of an air stream from the lungs, modified first by the vibrating vocal folds, and then by the rest of the larynx, and the pharynx, the mouth, and sometimes also the nasal cavities" (Sundberg, 1987: 3). Engineers are often interested in the acoustic waveform that correlates with vocal sound, and therefore define voice quality in terms of acoustic attributes that are (presumptively) perceptually important, without particular regard for the mechanisms that produced the sound. In contrast, psychologists are not especially interested in how the voice is physically produced or in the acoustic features of each utterance, but instead approach voice quality solely in terms of higher-level perceptual attributes.

Given that voice quality is by definition a perceptual response to an acoustic signal, one approach to providing a definition is to specify the nature of the interaction between a sound and a listener that results in quality. This is the approach taken by the American National Standards Institute (ANSI) Standard definition, which defines the quality (or timbre) of a sound as "that attribute of auditory sensation in terms of which a listener can judge that two sounds similarly presented and having the same loudness and pitch are dissimilar" (ANSI Standard S1.1.12.9, p. 45, 1960; cf. Helmholtz, 1885). This definition introduces a number of complications that are not apparent in simpler, narrower, physiologically-based definitions. By the ANSI Standard definition, quality is acoustically multidimensional, including the shape and/or peaks of the spectral envelope,[2] the amplitude of the signal and its fundamental frequency, the extent to which the signal is periodic or aperiodic, and the extent and pattern of changes over time in all these attributes (Plomp, 1976). This large number of degrees

[2] The spectral envelope refers to the way in which acoustic energy is distributed across the different frequencies in the voice.

of freedom makes it difficult to operationalize the concept of quality, particularly across listening tasks. The perceived quality of a single voice sample may also vary from occasion to occasion as listeners focus on different aspects of a sound in different contexts, or as different listeners attend to different aspects of the same sound; and what a given listener attends to when judging voices varies from voice to voice, according to task demands. According to the ANSI Standard definition, quality is a perceptual response in a specific psychophysical task (determining that two sounds are dissimilar), and it is unclear how this definition might generalize to other common, seemingly-related tasks like speaker recognition or evaluation of a single stimulus. Excluding pitch and loudness from what we call "vocal quality" is troublesome, because pitch and loudness are consistently found to be highly important characteristics of voice, on which listeners depend heavily for many kinds of judgments (as described in subsequent chapters). Evidence also suggests that quality may not be independent of frequency and amplitude (Melara and Marks, 1990; Krumhansl and Iverson, 1992), as the ANSI definition seemingly requires. Finally, this definition is essentially negative: It states that quality is not pitch and loudness, but does not indicate what it does include (Plomp, 1976). Such complications have led to frequent criticism of the ANSI definition, which some claim amounts to no definition at all (see, for example, Bregman, 1994, for review).

Dissatisfaction with this situation has led some voice researchers to adopt definitions of quality that simply echo the narrow or broad definitions of voice described above, so that voice quality is characterized in physiological, not perceptual, terms. Consistent with narrow definitions of voice, vocal quality may be defined as the perceptual impression created by the vibration of the vocal folds. More broadly, and parallel to broad definitions of voice, voice quality may be defined as the perceived result of coordinated action of the respiratory system, vocal folds, tongue, jaw, lips, and soft palate. For example, Abercrombie viewed voice quality as "those characteristics which are present more or less all the time that a person is talking: It is a quasi-permanent quality running through all the sound that issues from his mouth" (1967: 91). Similarly, Laver referred to voice quality as "a cumulative abstraction over a period of time of a speaker-characterizing quality, which is gathered from the momentary and spasmodic fluctuations of short-term articulations used by the speaker for linguistic and paralinguistic communication" (1980: 1). Such definitions do very little to specify listeners' contributions to quality, which are essential to defining what is after all a perceptual phenomenon. For example, the perceptual importance of different aspects of a voice depends on context, attention, a listener's background, and other factors (Kreiman, Gerratt, Precoda, and Berke, 1992; Gerratt, Kreiman, Antoñanzas-Barroso, and Berke, 1993; Kreiman, Gerratt, and Khan, 2010), and is affected by the listening task (Gerratt *et al.*, 1993; Gerratt and Kreiman, 2001a; Kreiman, Gerratt, and Antoñanzas-Barroso, 2007). Thus, the measured response to a given voice signal is not necessarily constant across listeners or occasions.

Some of the difficulty that arises when contemplating the nature of quality may be due to the fact that quality is often treated as analogous to pitch and loudness, the two other perceptual attributes of sound specified in the ANSI Standard definition. Authors often discuss *the* pitch or *the* loudness of a signal, presumably because these factors can be scaled unidimensionally, from low to high or faint to strong (Plomp,

1976), and because the anatomy of the auditory system is fairly consistent across individuals, so that responses to fundamental frequency and intensity are reasonably consistent, at least in the auditory periphery (but see Krishnan, Gandour, and Bidelman, 2010, for evidence of listener differences even at this level of processing). In fact, some authors even treat pitch and fundamental frequency, or loudness and intensity, as synonymous in informal writing. This creates the expectation that the acoustic correlates of quality should be fairly consistent from listener to listener, and that the same cues should operate across all voices, as fundamental frequency is the major cue to pitch, but not the only cue (see, for example, Thomas, 1969), and intensity is the primary cue to loudness (Fletcher and Munson, 1933). However, quality is multidimensional. It cannot be successfully scaled unidimensionally; and because more than one possible cue to quality exists, the possibility of listener differences is always present, so that quality can never have fixed acoustic determinants (Kreiman, Gerratt, and Berke, 1994). Given this fact, the perceptual response evoked by a voice signal will always depend on factors like task demands, and listener attention will vary across the multiple facets of the signal, so that some are more important than others from occasion to occasion (although experimental controls can minimize the effects of these factors, as discussed below). For this reason, a single perceived quality may not consistently result from a given signal, relative to the listener. In contrast, pitch and loudness do not ordinarily vary in this way, because of their more-or-less unidimensional nature.

The strength of the ANSI Standard definition is that it incorporates the inherently multivariate nature of voice quality by treating sound quality as the result of a perceptual process rather than as a fixed quantity, and highlights the importance of both listeners and signals in determining quality. Listeners usually listen to voices in order to gather information about the environment, and the information they attend to varies with their purpose and with the information available from a particular utterance. Considered in this light, the ANSI Standard definition has distinct advantages; in fact, its limitations can be reduced by broadening the definition to include different tasks, rather than narrowing its focus to include only a small set of specific acoustic variables. Voice quality may best be thought of as an interaction between a listener and a signal, such that the listener takes advantage of whatever acoustic information is available to achieve a particular perceptual goal. Which aspects of the signal are important will depend on the task, the characteristics of the stimuli, the listener's background, perceptual habits, and so on. Given the many kinds of information listeners extract from voice signals, it is not surprising that these characteristics vary from task to task, voice to voice, and listener to listener.

Studies of familiar voice recognition (van Dommelen, 1990; Remez, Fellowes, and Rubin, 1997) highlight the importance of signal/listener interactions in voice perception. Specific phonatory and articulatory information is key to identifying some individual voices, but not relevant to others (Van Lancker, Kreiman, and Wickens, 1985a), such that three conditions of signal alteration (backwards, rate changed to slower or faster speech) affected the recognizability of individual voices differently. Perceptual processing of voice quality differs qualitatively depending on whether the listener is familiar or unfamiliar with the voice (see Kreiman, 1997, and Chapter 6 for review). Listeners' perceptual strategies can thus be expected to vary depending on the differential familiarity of the voices. Listeners' attention to different cues to voice

identity also depends on the total voice pattern in which the cue operates (Van Lancker *et al.*, 1985a; Van Lancker, Kreiman, and Emmorey, 1985b), so that the importance of a single cue varies across voices as well as listeners. Definitions of quality that depend exclusively on aspects of production or on the signal cannot account for such effects. Voice quality is the result of perceptual processes, and must be defined in terms of both signals and listeners.

Although the great majority of studies of voice maintain a firm distinction between production and perceptual aspects of voice, some recent work in dialogic[3] linguistics abandons this distinction in favor of a view of voice as inextricable from a communicative context, so that production and perception are inseparably linked (Bertau, 2008). In this view, a human voice is a concrete, perceivable event that is inseparable from (and thus indexes) the body that produced it, which in turn shapes the sound of the voice. At the same time, voice manifests the speaker's abstract, unobservable consciousness, thus representing the whole person and "underscoring the physicality of psychological self" (p. 97). Further, the speaking person exists in a communicative context that necessarily includes a listener (somewhat reminiscent of the sound made by a tree falling in the woods), and the voice that is produced cannot be separated from the act of listening that provides the context for production. As Bertau writes,

> So, "voice" is a vocal-auditory event, and it is a concept belonging to a certain socioculturally constructed way of expression. The uttered voice is absolutely individual, coming from a unique body, but this body is located in specific sociocultural contexts and has a history of action, movements, labels, etc. So, the voice, too. As for every human expression, the voice is individual and societal, both aspects being the facets of a wholeness … (pp. 101–2)

As we will see in subsequent chapters, this viewpoint is helpful when considering the neuropsychology of voice production and perception, both of which suggest that voice reflects the whole physical and social self and is shaped in part by communicative context. This view is also consistent with studies showing how speakers subtly adjust their accents to mutually match elements of the speech of the co-participant. Thus, while the voice pattern is uniquely expressing personal characteristics, it is also capable of adjusting to the voice pattern of the other (Pardo, 2006). Interestingly, this broad view of voice as a perceived manifestation of the total self (or consciousness), cast in another format, is seen in self-help books describing ways to "find and use your natural voice" that "represents us well" (Boone, 1991, p. 6). Such popularized notions are fully consonant with the view that voice expresses who we are, both in isolation and with respect to other individuals.

1.3 Measuring Voice Quality

Given the difficulties inherent in defining voice and voice quality, it is not surprising that considerable confusion also surrounds quality measurement. By its nature, quality is essentially psychoacoustic: It is the psychological impression created by a

[3] As the name suggests, this branch of linguistics studies interactional language use, including negotiation, mediation, social identity in partnered communication, identification with and influence of interlocutors, and the like.

physical stimulus, and thus depends on both the listener and the voice, as discussed above. However, the psychoacoustic study of complex multidimensional auditory signals is in its infancy (for example, Melara and Marks, 1990; see Yost *et al.*, 1989, for review), and little research has examined the perceptual processes listeners apply to voice signals. Research has focused instead on deriving and defining static descriptive labels for voices. In this approach, vocal quality is treated as if it can be decomposed into a set of specific features or elements, whose presence or absence characterize a speaker's voice.

The most common approach to the problem of specifying voice quality is simply to create a long list of terms to describe listeners' impressions, essentially decomposing overall "quality" into a set of component "qualities." Listeners then assess quality by rating the extent to which a voice possesses each feature. (Alternatively, listeners may simply mark as present the features they hear in the voice in question.) It can be difficult to determine the basis on which terms in such lists have been selected, and labels like these for quality tend to be rather mixed in their level of description. They may describe voices visually (for example, brilliant, dark), kinesthetically (strained, tight), physically (heavy, thin, pointed), aesthetically (pleasing, faulty), with reference to anatomy (pectoral, nasal), and so on (for example, Orlikoff, 1999).

Such dimensional approaches to measuring voice quality depend on descriptive traditions rather than theory, and have changed only superficially in nearly 2000 years. Table 1.3 includes three lists of descriptive features for voices, one venerable (Julius Pollux, 2nd century AD; cited by Austin, 1806) and two modern (Moore, 1964, cited by Pannbacker, 1984; Gelfer, 1988). A few differences exist among these lists. For example, the oldest list includes terms related to the personality and emotional state of the speaker (confused, doleful), and terms related to articulation and rhetorical ability (articulate, distinct), reflecting the importance of rhetoric in Roman culture (see Gray, 1943, or Laver, 1981, for review). More modern compendia include terms like "breathy" and "nasal" that are commonly used in the study of vocal pathology. However, similarities among the lists are striking. Although alignment of terms across lists is approximate, only eight of forty terms lack at least one close counterpart in the other lists, mostly due to the loss of terms for enunciation, emotion, or rhetorical style in the modern vocabulary for voice, as noted above.

The bomb threat form shown in Figure 1.2 is a modern forensic application of this "list of features" approach to quality assessment and speaker recognition. In completing this form, the listener is asked to judge an eclectic array of vocal descriptors, including the speaker's physical characteristics (age, sex, race), their emotional state (angry, calm), and their identity (familiar, disguised voice), and to describe the dynamics of the utterance (rate, loudness) and the quality of the voice (whispered, nasal, raspy, ragged). As an applied tool, this questionnaire includes commonly known articulatory disorders (lisp, stutter) and nonverbal modes (crying, laughter). This large and heterogeneous set of descriptors has a specific purpose in narrowing down a field of suspects and coordinating voice identity information with other evidence obtained in an investigation. For example, if a listener can correctly judge a caller's sex when completing the form, this alone would eliminate half the population as suspects. However, these descriptors alone would not be useful across the general population in specifying voice quality or uniquely identifying a set of suspects with any degree of confidence.

Table 1.3 Venerable and modern labels for voice quality.

After Julius Pollux, 2nd century AD[a]	*Moore, 1964*[b]	*Gelfer, 1988*
High (altam)	–	High
Powerful (excelsam)	Ringing	Strong, intense, loud
Clear (claram)	Clear, light, white	Clear
Extensive (latam)	Rich	Full
Deep (gravam)	Deep	Resonant, low
Brilliant (splendidam)	Bright, brilliant	Bright, vibrant
Pure (mundatam)	–	–
Smooth (suavam)	Cool, smooth, velvety	Smooth
Sweet (dulcem)	–	–
Attractive (illecebrosam)	Pleasing	Pleasant
Melodious, cultivated (exquisitam)	Mellow	Mellow, musical
Persuasive (persuasibilem)	–	–
Engaging, tractable (pellacem, tractabilem)	Open, warm	Easy, relaxed
Flexible (flexilem)	–	Well-modulated
Executive (volubilem)	–	Efficient
Sonorous, harmonious (stridulam)	Chesty, golden, harmonious, orotund, round, pectoral	Balanced, open
Distinct (manifestam)	–	–
Perspicuous, articulate (perspicuam)	–	–
Obscure (nigram)	Dark, guttural, throaty	Husky, guttural, throaty
Dull (fuscam)	Dead, dull, heavy	Dull, heavy, thick
Unpleasing (injucundam)	–	Unpleasant
Small, feeble (exilem, pusillam)	Breathy	Breathy, soft, babyish
Thin (angustam)	Constricted, heady, pinched, reedy, shallow, thin	Thin
Faint (difficilem auditu, molestam)	Whispery	Weak
Hollow, indistinct (subsurdam, obscuram)	Covered, hollow	Muffled
Confused (confusam)	–	–
Discordant (absonam)	Blatany, whiney	Strident, whining
Unharmonious, uncultivated (inconcinnam, neglectam)	Coarse, crude	Coarse, gruff
Unattractive, unmanageable (intractabilem)	–	Shaky
Uninteresting (inpersuasibilem)	Blanched, flat	–
Rigid (rigidam)	Hard, tight	Monotonous, constricted, flat
Harsh (asperam)	Harsh, strident, twangy	Harsh, gravelly
Cracked (distractam)	Pingy, raspy	Strained, raspy, grating, creaky

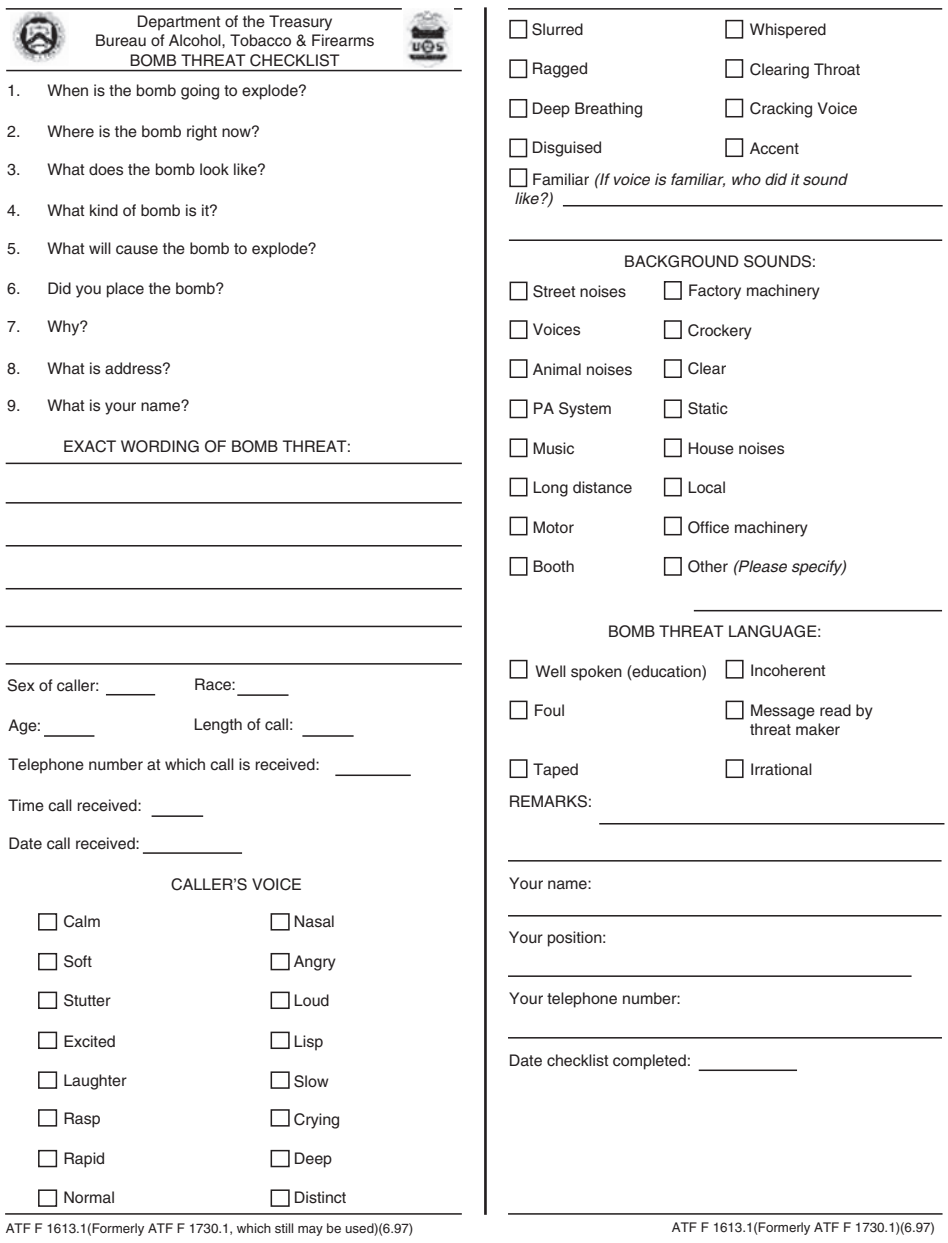
Table 1.3 (*Cont'd*).

After Julius Pollux, 2nd century AD[a]	*Moore, 1964*[b]	*Gelfer, 1988*
Doleful (tristem)	–	–
Unsound, hoarse (infirmam, raucam)	Faulty, hoarse, poor, raucous, rough	Hoarse, rough, labored, noisy
Brassy (aeneam)	Buzzy, clangy, metallic	Metallic
Shrill, sharp (acutam)	Cutting, hooty, piercing, pointed, sharp, shrill	Shrill, sharp
–	Nasal	Nasal
–	Denasal	Denasal
–	Toothy	–

Notes:
[a] Cited in Austin (1806).
[b] Cited in Pannbacker (1984).

Redundancies and ambiguities are common in lists of terms, which tend to be exhaustive rather than efficient. To address the problem of which terms to include in a voice quality assessment protocol, some researchers have applied factor analysis, a statistical procedure that reduces large lists of overlapping features to small non-redundant sets. In such studies, listeners evaluate each of a set of voices on a large number of rating scales like those in Table 1.3. Two general approaches have been used in voice quality research. In the first (Holmgren, 1967), voice samples (spoken passages of text) are rated on a relatively small set of scales that have been selected to represent an a priori underlying set of factors. Because no standard factors or dimensions have been established for voice quality, such studies have adopted previously-proposed dimensions (for example, potency, evaluation, activity) and scales (for example, sweet/sour, strong/weak, hot/cold; Osgood, Suci, and Tannenbaum, 1957) that are not necessarily applicable to voice quality. Alternatively, investigators have asked listeners to rate voice samples (again, spoken sentences or passages of text) on large sets of voice quality scales that do not derive from an a priori factor structure (Voiers, 1964; Fagel, van Herpt, and Boves, 1983). Such exploratory studies attempt to ensure that all possible perceptual factors are represented in the derived factors by oversampling the semantic space for voice quality.

In either case, statistical analysis of listeners' ratings produces a small number of orthogonal factors that capture as much of the variance in the underlying ratings as possible. Each original scale is given a weight on each factor, so that scales that are strongly related to the factor receive large weights, and scales that are weakly related to the factor receive low weights. Factors are then given summary labels based on the scales that they comprise. For example, a factor with large weights on scales like "fast," "agitated," "tense," "busy," and "exciting" might be labeled "animation" (Voiers, 1964), while one with large weights on scales like "vivacious," "expressive," "melodious," "cheerful," "beautiful," "rich," and "active" might be labeled "melodiousness" (Fagel, van Herpt, and Boves, 1983).



Department of the Treasury
Bureau of Alcohol, Tobacco & Firearms
BOMB THREAT CHECKLIST

1. When is the bomb going to explode?
2. Where is the bomb right now?
3. What does the bomb look like?
4. What kind of bomb is it?
5. What will cause the bomb to explode?
6. Did you place the bomb?
7. Why?
8. What is address?
9. What is your name?

EXACT WORDING OF BOMB THREAT:

Sex of caller: _____ Race: _____

Age: _____ Length of call: _____

Telephone number at which call is received: _____

Time call received: _____

Date call received: _____

CALLER'S VOICE

☐ Calm ☐ Nasal
☐ Soft ☐ Angry
☐ Stutter ☐ Loud
☐ Excited ☐ Lisp
☐ Laughter ☐ Slow
☐ Rasp ☐ Crying
☐ Rapid ☐ Deep
☐ Normal ☐ Distinct

ATF F 1613.1(Formerly ATF F 1730.1, which still may be used)(6.97)

☐ Slurred ☐ Whispered
☐ Ragged ☐ Clearing Throat
☐ Deep Breathing ☐ Cracking Voice
☐ Disguised ☐ Accent
☐ Familiar *(If voice is familiar, who did it sound like?)* _____

BACKGROUND SOUNDS:

☐ Street noises ☐ Factory machinery
☐ Voices ☐ Crockery
☐ Animal noises ☐ Clear
☐ PA System ☐ Static
☐ Music ☐ House noises
☐ Long distance ☐ Local
☐ Motor ☐ Office machinery
☐ Booth ☐ Other *(Please specify)* _____

BOMB THREAT LANGUAGE:

☐ Well spoken (education) ☐ Incoherent
☐ Foul ☐ Message read by threat maker
☐ Taped ☐ Irrational

REMARKS: _____

Your name: _____

Your position: _____

Your telephone number: _____

Date checklist completed: _____

ATF F 1613.1(Formerly ATF F 1730.1)(6.97)

Figure 1.2 A typical form for reporting a bomb threat. Listeners completing this form must judge the speaker's personal characteristics (sex, age, ethnicity), and rate the voice on an eclectic selection of characteristics, including terms related to emotional state, articulation, message content, and voice quality.

Voice feature schemes derived from factor analysis do have obvious advantages over large lists of terms. Such protocols typically include between three and six factors (Table 1.4), and thus are manageable for listeners and investigators alike. In theory, factors are independent of one another, reducing concerns about redundancies or

Table 1.4 Factor analytic studies of normal voice quality.

Speakers	*Stimuli*	*Listeners*	*Input scales*	*Derived factors*	*Reference*
5 male, 5 female	spoken passage	235	35 7-point bipolar	5 factors: Melodiousness Articulation quality Voice quality Pitch Tempo	Fagel *et al.* (1983)
16 male	sentences	32	49 7-point bipolar	4 factors: Clarity Roughness Magnitude Animation	Voiers (1964)
10 male	spoken passage	20	12 scales representing 4 underlying factors	2 factors: (1) Slow/fast, resting/busy, intense/mild, simple/complex; (2) Clean/dirty, beautiful/ugly	Holmgren (1967)

overlap across scales, while at the same time they capture much of the information in the scalar ratings, so economy is achieved with minimal loss of information. Finally, this approach preserves the descriptive tradition of quality assessment, because factors are defined in terms of the underlying scales. Thus, factor analytic approaches bring the impression of scientific rigor to the familiar descriptive approach to quality assessment.

Certain limitations to such approaches are also apparent. First, results of factor analytic studies depend on the input scales and stimuli. That is, a factor will not emerge unless that factor is represented in the set of rating scales and is also perceptually relevant for the specific voices and utterances studied. Studies often employ restricted populations of speakers, small sets of voices, and short stimuli. For example, the well-known GRBAS[4] protocol was developed from the results of factor analyses that used five steady-state vowels produced by only 16 speakers (Isshiki, Okamura, Tanabe, and Morimoto, 1969; see Hirano, 1981, for review). Such restrictions significantly limit the extent to which results can legitimately be generalized to the full spectrum of vocal qualities. Further, as can be seen in Table 1.4, results of factor analyses have varied substantially from study to study. The validity of the factors as perceptual features also depends on the validity of the underlying scales, which has never been established. Thus, even a large-scale factor analysis (or multiple analyses) will not necessarily result in a valid or reliable rating instrument for voice quality. Idiosyncrasies in labeling the factors may also obscure differences among studies. For example, in studies of pathological voice quality Isshiki *et al.* (1969) found a "breathiness" factor that loaded highly on the scales dry, hard, excited, pointed, cold, choked, rough, cloudy, sharp, poor, and bad, while a "breathiness" factor reported by Hammarberg,

[4] **G**rade (i.e., severity of deviation), **R**oughness, **B**reathiness, **A**sthenicity (or weakness), and **S**train.

Fritzell, Gauffin, Sundberg, and Wedin (1980) corresponded to the scales breathy, wheezing, lack of timbre, moments of aphonia, husky, and not creaky. Finally, Voiers (1964) reported perceptual factors related to statistically reliable constant listener biases and interactions between specific voices and listeners, in addition to factors related only to the target voices. Emergence of such factors suggests that an adequate perceptual model cannot be framed solely in terms of the stimuli, but must also account separately for differences among listeners. Overall, it thus appears that factor analysis has not convincingly identified scales for vocal quality that are independent and valid.

Dependence on underlying descriptive terminology can be avoided by deriving perceptual features for voices through multidimensional scaling (MDS), rather than factor analysis. In MDS listeners assess the similarity of the experimental voice stimuli directly (usually by listening to pairs of voices and rating their similarity), without reference to scales for specific qualities. The analysis produces a perceptual space from these similarity ratings, such that distances between voices in the space are proportional to the rated similarities (more similar = closer together). Dimensions in this space are then interpreted, usually by examining correlations between rated and/or measured characteristics of the input stimuli and stimulus coordinates or clustering of stimuli in the space. Through this process, exploratory MDS can reveal how overall vocal quality (as it determines similarities between voices) relates to scales for particular qualities. Discovery of a dimension that is highly associated with some specific quality provides evidence for the "psychological reality" of that particular quality as an important vocal feature.

Studies applying MDS to normal vocal qualities are listed in Table 1.5. As with factor analysis, results have varied substantially from study to study, with the exception that dimensions related to pitch (F0) emerge consistently across tasks and stimulus types. Some of these differences can be attributed to differences in study design. Note that three of these 11 studies used vowels as stimuli, while the rest used longer, more complex speech samples, which yield additional information and address questions about the broader definition of voice quality. For example, dimensions associated with stimulus duration or F0 variability typically emerge when sentence stimuli are employed, rather than steady-state vowels. Differences have also been reported in the perceptual features derived for male and female voices (Singh and Murry, 1978; Murry and Singh, 1980). However, variability in solutions has emerged due to factors other than stimulus characteristics. In particular, variability in the perceptual dimensions that emerge from studies of fixed sets of stimuli indicates that listeners differ both as individuals and as groups in the perceptual strategies they apply to voices (Gelfer, 1993; cf. Kreiman, Gerratt, and Precoda, 1990, or Kreiman *et al.*, 1992, who studied pathological voice quality). Thus, it does not appear that any specific features, other than F0, are always important for characterizing the quality of all voices under all circumstances.

Multidimensional scaling solutions may also leave large amounts of variance unaccounted for, and published reports may explain less than half of the variance in the underlying similarity judgments, even for simple vowel stimuli (Murry, Singh, and Sargent, 1977; Murry and Singh, 1980). This may occur because of the limited resolution of MDS: The number of extractable dimensions depends on the number of

Table 1.5 Multidimensional scaling studies of normal voice quality.

Speakers	*Stimuli*	*Listeners*	*Derived dimensions*	*Reference*
8 male	vowels	6	4 dimensions: F0 Glottal source spectrum Jitter Formant frequencies	Matsumoto *et al.* (1973)
9 male	phrase	15	3 dimensions: F0 Intensity Intonation pattern	Carterette and Barnebey (1975)
20 male	word	11	4 dimensions: F0 Utterance duration Speaker's age "Superior" vs. "inferior" voice quality	Walden *et al.* (1978)
10 male, 10 female	sentence	10	3 dimensions: Speaker sex Pitch (male voices only) Utterance duration (female voices only)	Singh and Murry (1978)
20 male	vowel	10	4 dimensions: Pitch Formant frequencies (2 dimensions) Perceived nasality	Murry and Singh (1980)
20 female	vowel	10	4 dimensions: Pitch Perceived breathiness Formant frequencies Perceived effort	Murry and Singh (1980)
20 male	passage	10	4 dimensions: Pitch and effort Perceived hoarseness Formant frequencies 1 uninterpreted dimension	Murry and Singh, 1980
20 female	passage	10	4 dimensions: Perceived effort and nasality Pitch Utterance duration 1 uninterpreted dimension	Murry and Singh (1980)
10 male	sentence	24	4 dimensions: Perceived masculinity Perceived creakiness Perceived variability Perceived mood	Kreiman and Papcun (1991)

Table 1.5 (*Cont'd*).

Speakers	*Stimuli*	*Listeners*	*Derived dimensions*	*Reference*
20 female	sentence	20 speech-language pathologists	5 dimensions: Pitch Loudness Perceived age Perceived variability Voice quality	Gelfer (1993)
20 female	sentence	20 untrained	2 dimensions: Pitch and resonant quality Variability, age, and rate	Gelfer (1993)

stimuli studied, which has been limited to twenty or less (although additional perceptual features may also be derived from clustering of stimuli in the space). It is possible that more dimensions (providing more explanatory power) exist in the data than can be extracted due to the small numbers of voices involved. Alternatively, large amounts of variance may remain unexplained because the dimensional model of quality implied by MDS and factor analytic studies does not provide a good description of how quality is perceived.

A study of pathological voice quality (Kreiman and Gerratt, 1996) supports the latter explanation. In that study, listeners judged the similarity of all possible pairs of vowel productions obtained from very large sets of speakers (80 males and 80 females) representing a variety of diagnoses and ranging in quality from nearly normal to severely disordered. In this study, use of vowel stimuli limited the information available to listeners, consistent with the narrow definition of voice, so that the perceptual task was somewhat simpler than with connected speech stimuli. Despite this simplification, multidimensional scaling solutions for male and female voices each accounted for less than half of the variance in the underlying data, and revealed two-dimensional solutions in which the most severely pathological voices were separated from voices with milder pathology. Separate analyses of the data from individual listeners accounted for more variance (56–83%). However, stimuli did not disperse in these perceptual spaces along continuous scale-like linear dimensions, but instead clustered together in groups that lacked subjective unifying percepts. Different voices clustered together for each listener; in fact, no two voices ever occurred in the same cluster for all listeners, suggesting that listeners lacked a common notion of what constitutes similarity with respect to voice quality, even when quality is narrowly defined. If listeners lack a common perceptual space for voice quality in its most restricted sense, then a single set of perceptual features for voice quality more broadly defined is not likely to be discoverable.

In the absence of empirical evidence for the validity of particular descriptors or dimensions, it is unclear why some should be included, and others excluded, in a descriptive framework for vocal quality. Further, each traditional descriptive label is holistic and independent, and labels do not combine to form a permutable set. This makes it diffi-

cult to understand precisely how qualities differ from one another, or how seemingly similar qualities are related. Finally, in this tradition it is often unclear how quality relates to other parts of the speech chain. In particular, there is no formal theoretical linkage between a given quality and the physiological configuration that produced it (although terms like "nasal" may imply in some cases that such a linkage exists).

The phonetic/articulatory features for voice quality proposed by Laver (1980, 2000; Ball, Esling, and Dickson, 2000) were designed in response to these limitations. In this approach, voice quality is characterized as "quasi-permanent" and derived cumulatively throughout an individual's vocal sound production (Abercrombie, 1967). It is then described in terms of the global long-term physiological configuration that (hypothetically) underlies the overall sound of a speaker's voice. Laryngeal and supralaryngeal aspects of voice are both specified, and are assumed to be auditorily separable. The specific features are derived from phonetics, and include laryngeal raising and lowering, lip rounding and spreading, jaw position (open, closed), tongue tip and body position (raised, lowered, advanced, retracted), pharyngeal constriction or expansion, velum position, and glottal state (modal voice, falsetto, whisper, creak, breathiness, harshness) (see Laver, 1980, 2000, for more details). This model of voice quality was originally developed to describe normal voices, but has been adapted as a clinical voice evaluation protocol called "vocal profile analysis" that is widely used in the United Kingdom and elsewhere (Laver, Wirz, Mackenzie, and Hiller, 1981; Wirz and Mackenzie Beck, 1995).

Vocal profile analysis is analytic, consistent with phonetic models of speech production, and nearly exhaustive in the physiological domain. Because quasi-independent features (or "settings") can combine in different ways, the system can be used to describe a broad range of voice qualities in a single framework, rather than applying vague terms whose relationships to each other are unclear. Thus, for example, "hoarse" voice might appear in this system as "deep, (loud), harsh/ventricular, whispery voice," or "gruff" voice might become "deep, harsh, whispery, creaky voice" (Laver, 1968). The primary limitation of this system is the fact that it models perception in terms of speech production processes without established or documented reference to a listener. That is, by describing voice quality in detailed terms of the supposed underlying physiological configuration, profile analysis indicates where perceptual information about quality *might* be found. However, it does not indicate which of the many aspects specified are meaningful, or, indeed, perceptible to listeners, how listeners actually use different features to assess quality, whether (or when, or why) some features might be more important than others, or how dimensions interact perceptually. The assumption that listeners are able to separate different features auditorily is also questionable, particularly given recent evidence that listeners have difficulty isolating individual dimensions of complex voice patterns (Kreiman and Gerratt, 2000a; Kreiman *et al.*, 2007; cf. Fry, 1968).

The results reviewed above indicate that the validity of dimensional and featural protocols for assessing overall voice quality remains questionable, although clinical applications of such featural systems are common. Despite the proliferation of rating systems, convergence to a general theory of voice perception remains elusive. These protocols model voice quality solely in terms of the voice signal itself, although couching many of the descriptive labels in perceptual terms. Most of these approaches imply that voice quality can reasonably be represented as a list or grouping of descriptors or

dimensions – that there is a list of attributes that listeners can and do attend to, and that the same set adequately describes all voices. Whether quality is broadly or narrowly construed, such frameworks imply a well-defined perceptual space for voice quality, applicable to all voices and true for all listeners, which listeners all exploit in essentially the same way. However, substantial evidence and theoretical considerations, many of which have been touched upon in this chapter, contradict these requirements. A well-defined, theoretically motivated set of features for voice has not emerged, despite many years' research; and listeners apparently exploit vocal signals in unique ways. Data thus suggest that efforts to specify a perceptually valid set of scales for voice quality are unlikely to succeed.

A further difficulty with dimensional protocols is their unreliability as measurement tools. Most studies of listener reliability have focused on pathological voices, due to the importance of scalar ratings in clinical assessments of voice quality (for example, Gerratt, Till, Rosenbek, Wertz, and Boysen, 1991). Across studies, scales, and statistics, average interrater reliability has ranged from extremely low ($r^2 = .04$) to extremely high (100% of ratings within +/– one scale value) (see Kreiman, Gerratt, Kempster, Erman, and Berke, 1993, for review). Analyses of the reliability with which listeners judge individual voices indicate that listeners almost never agree in their ratings of a single voice. Even using the simplest of phonated stimuli, the likelihood that two raters would agree in their ratings of moderately pathological voices on various seven-point scales averaged 0.21 (where chance is 0.14); further, more than 60% (and as much as 78%) of the variance in voice quality ratings was attributable to factors other than differences among voices in the quality being rated (Kreiman and Gerratt, 1998). The voice profile analysis system developed by Laver also falls short of desired levels of reliability. Wirz and Mackenzie Beck (1995) reported that 242 listeners who completed a three-day training course in the system's use rated voices within one scale value (of a possible 6) of a target score for 52%–65% of items in a post-test. Studies of rating reliability for normal voices are less common, but not more encouraging. For example, Gelfer (1988) asked trained and untrained listeners to rate 20 normal female voices (speaking sentences) on 16 different quality scales. Kendall's coefficient of concordance revealed only "modest to slight agreement for both groups" (0.14 to 0.69 across scales, with values averaging 0.33 overall) (Gelfer, 1988, p. 325).

In summary, despite a long history of research, significant difficulties continue to plague traditional approaches to vocal quality measurement. Such approaches suffer from possibly irresolvable issues of rating reliability and validity. It is not clear what if any features characterize quality, or how traditional descriptors or dimensions relate to overall quality (broadly or narrowly construed) or to each other. More modern articulatory distinctive-feature approaches are analytical and motivated by phonetic theory, but while they usefully enumerate articulatory possibilities, they do not predict listeners' behavior. Featural systems in general suffer from this limitation, because they model quality as if it inheres in voices, without also accounting for such listener-dependent factors as attention, experience, and response bias.

Given the difficulties, both theoretical and operational, inherent in measuring voice quality, some authors (particularly those studying pathological voices) have argued that perceptual measures of voice quality should be replaced with instrumental measures (see, for example, Orlikoff, 1999, for review). A variety of measures of voice and

Table 1.6 Measurement techniques for voice and vocal function.

Technique	*What it does*	*Sample references*
Acoustic measurements	Quantify F0, amplitude, resonance, and temporal parameters in the speech signal	Buder (2000)
Aerodynamic measures	Specify respiratory driving pressure and airflow through glottis and oronasal cavity	Warren (1996)
Anemometry	Measurement of oral and subglottal air flow velocities	Baken (1987), Tropea (1995)
Electroglottography (EGG)	Reflects vocal fold closure and separation cycles	Childers *et al.* (1990)
Functional MRI or PET	Graphic representation of activity during performance of behavioral tasks, for example, of tongue, brain	J. Sidtis (2007)
High-speed imaging	Provides photographic images of rapid movements in speech: tongue, vocal folds	Luchsinger and Arnold (1965)
Magnetic resonance imaging (MRI)/Computerized tomography (CT scans)	Graphic representation of vocal structures: vocal tract, brain	Kertesz (1994)
Movement transduction techniques	Gauges for tracking movements of velum, tongue, jaw, and lips	Baken (1987)
Palatography	Records tongue contacts with alveolar ridge and hard palate	Palmer (1973)
Ultrasound imaging	Use of high frequency sound signals to delineate boundaries between specific structures at rest or during movement	Kent (1997)
Videostroboscopy	Moving pictures of larynx taken in synchrony with a flashing light to reveal phonatory cycles	Baken (1987)

vocal function are available for use in living subjects (Table 1.6). In particular, a spectacular array of acoustic analysis techniques has been developed over the years, largely by speech scientists, enabling researchers to visualize and quantify the properties of the acoustic signals that are the carriers of voice quality information. Of course, these techniques must eventually interface with knowledge about psychophysical processes.

In contrast to perceptual measures, instrumental measures of aerodynamic, acoustic, or physiological events promise precision, reliability, and replicability. Considerations like these have motivated researchers to create several measurement systems for voice, including the Dysphonia Severity Index (Wuyts *et al.*, 2000) and the Hoarseness Diagram (Frohlich, Michaelis, Strube, and Kruse, 2000), whose purpose is to "establish an objective and quantitative correlate of the perceived vocal quality" (Wuyts *et al.*, 2000: 796). A popular software approach to quantitative

assessment of voice quality is the Multi-Dimensional Voice Program (MDVP; Kay Elemetrics Corp.). Using sustained vocalization as input, the program calculates 22 acoustic parameters, displaying these values in colorful snapshot form. However, because vocal quality is the perceptual response to a stimulus, development of instrumental protocols for measuring quality ultimately depends on our ability to define quality in a way that accounts for perceptual factors that introduce variability in listeners' judgments. Although it might be possible to devise objective methods to quantify specific quality dimensions, it is more difficult to set up general rules specifying which dimensions are selected and how they combine to produce a final evaluative judgment (Bodden, 1997). Further, no comprehensive theory exists describing the relationships between physiology, aerodynamics, acoustics, and vocal quality, so it is difficult to establish which instrumental measures ought to correspond to perceptually meaningful differences in vocal quality, or why such associations should exist. Existing research has been limited largely to correlational studies, which have produced highly variable results that are difficult to interpret. (See Kreiman and Gerratt, 2000b, for extended discussion.)

1.4 Alternatives to Dimensional and Featural Measurement Systems for Voice Quality

Finding valid and reliable alternatives to traditional voice quality scaling methods requires knowledge of the sources of listener disagreements. Previous studies of pathological voices (Kreiman, Gerratt, and Ito, 2007; see also Gerratt *et al.*, 1993; Kreiman and Gerratt, 2000a) have shown that traditional methods for rating voice quality can be modeled as a kind of matching task, in which external stimuli (the voices) are compared to stored mental representations that serve as internal standards for the quality of interest. Variability in ratings can be predicted with good accuracy (over 84% variance accounted for) by four factors: Instability of internal standards for different qualities; difficulties isolating individual attributes in complex acoustic voice patterns; measurement scale resolution; and the magnitude of the attribute being measured (Kreiman *et al.*, 2007). A protocol that does not rely on internal standards, and that makes it easier for listeners to focus their attention appropriately and consistently, would eliminate these sources of listener disagreement. One such approach (Gerratt and Kreiman, 2001a; Kreiman and Gerratt, 2005) applies speech synthesis in a method-of-adjustment task. In this task, listeners vary the control parameters of a voice synthesizer by moving sliding cursors, until the synthetic token they control represents an acceptable auditory match to a voice stimulus. When a listener chooses a match to a test stimulus, the synthesizer settings parametrically represent the listener's perception of voice quality. Because listeners directly compare each synthetic token they create to the target voice, they need not refer to internal standards for particular voice qualities, which may be varying and incomplete. Further, listeners can manipulate acoustic parameters and hear the result of their manipulations immediately. Such manipulations bring the particular acoustic dimension to the foreground, helping listeners focus their attention consistently, which is the most important factor governing reliability.

Data indicate that this method improves agreement among listeners (to over 95%) in their assessments of voice quality relative to traditional rating scale techniques (Kreiman *et al.*, 2007).

This method of quality measurement also provides other practical advantages. First, quality is measured in acoustic terms, so that the relationship between acoustic parameters and what a listener hears is established directly, rather than correlationally. Thus, measuring quality with synthesis can experimentally establish the perceptual relevance of different acoustic attributes of voice. Mappings between acoustics and quality also mean that hypotheses can be tested about the perceptual relationships between different signals, because quality is measured parametrically. The perceptual importance of different parameters can also be evaluated in naturally occurring complex multivariate contexts. (See Kreiman and Gerratt, 2005, for an example of this kind of application.)

Finally, note that this approach to quality measurement follows directly from the ANSI Standard definition of sound quality, in that it measures quality psychophysically as those aspects of the signal that allow a listener to determine that two sounds of equal pitch and loudness (the synthetic and natural voice samples) are different. In this method, listeners need not focus on single quality dimensions, eliminating concerns about scale validity; and they create a direct mapping between the acoustic signal and a perceptual response, thus modeling quality psychoacoustically.

In summary, voice quality is psychoacoustic in nature, and can most appropriately be measured by developing methods that can assess interactions between listeners and signals, rather than by treating quality solely as a function of the acoustic voice signals themselves. Reductionistic approaches like those reviewed above have not led to a satisfactory model of voice quality assessment by humans; and the study of voice quality has not received benefit of classic psychophysical research methods. Pitch and loudness can often be treated as if they were functions of the signal, because measures of frequency and intensity are fairly well correlated with listeners' perceptual judgments. However, this simplification is inappropriate in the case of quality, because quality is multidimensional and listeners are flexible and variable. This is the case even when the definition of voice is constrained to refer only to laryngeal aspects of sound production. The complexities multiply with broader definitions of voice and voice quality.

Issues of quality measurement have implications beyond the study of quality itself. Once the relationship between a signal and a percept is understood, it may be possible to determine which physiological parameters create perceptually meaningful changes in phonation. At present, it is not possible to determine which aspects of vocal physiology are perceptually important, in part because the relationship between perception and acoustics (which links production to perception in the "speech chain") is poorly understood. In the absence of theories linking physiology to acoustics to perception, observed correlations between acoustic measures of voice and listeners' judgments of a speaker's characteristics remain hard to interpret (Scherer, 1986). Some progress has been made in understanding the acoustic determinants of perceived affect, voice identity, and perception of physiological characteristics like age and sex, as described in the chapters that follow, but much remains to be discovered. Better methods of quality

assessment have important implications for understanding aspects of normal voice perception (age, gender, identity, etc.) that are based in physiology, extending them to the impact of habitual speech patterns on listeners' perceptions. An improved understanding of the issues surrounding measurement of vocal quality is a first step toward these broader goals.

1.5 Organization of the Book

This book comprises ten chapters, with Chapter 2 presenting an overview of vocal physiology and acoustics, and describing the technical details of how listeners produce voice and vary their pitch, loudness, and vocal quality. Chapter 3 reviews elementary facts about the neural substrates for voice production, including consideration of the extent to which the elements of vocalization, frequency, amplitude and timing are independently modulated and how different forms of vocalization in humans are related to those of nonhuman animals. We touch upon this question because scientists disagree on neurological correspondences between nonhuman and human vocalizations, and whether the animal substrates, in evolutionary history, are pertinent to explaining the development of human speech and/or language.[5] Chapter 3 also describes the fundamentals of audition as it pertains to voice perception, including information about the auditory periphery, the basic neurophysiology of the auditory system, and the resolution of elementary acoustic elements into complex patterns.

Chapter 4 describes listeners' abilities to judge a speaker's physical and personal characteristics – age, sex, race, and so on – from voice. In Chapter 5, we review behavioral studies of the perception of familiar and unfamiliar voices, beginning with nonhuman animals and ending with a psychological model of human voice recognition. Chapter 6 reviews neuropsychological approaches to voice perception. Drawing on studies of the neurological substrates and processes relevant to perceiving and recognizing vocal patterns, this chapter develops a model of brain function underlying human voice recognition. Chapter 7 returns to the perception of unfamiliar voices, this time in forensic contexts. Chapter 8 examines linguistic prosody, including phonological, grammatical, semantic, and pragmatic uses of voice quality. Chapter 9 reviews the extent to which listeners can judge a speaker's personality, character, attitude and emotional state from voice, along with the influence of voice characteristics on comprehension of affective and attitudinal meanings. Chapter 10, which we call "Miscellany," describes some remaining manifestations of the voice in singing, voice printing and lie detection, advertising, speech synthesis, and problems encountered in film dubbing. As the title of the book implies, our goal is to provide the most current perspectives on voice perception as gained from many sources and disciplines.

[5] The terms "speech" and "language" are used by linguists to refer to different entities. Speech is performance, the motoric, physical realization of language ability; language is competence or the abstract mental ability.