Loquens 1(1)
January 2014, e009
eISSN 2386-2637
doi: http://dx.doi.org/10.3989/loquens.2014.009

Toward a unified theory of voice production and perception

Jody Kreiman[1], Bruce R. Gerratt[1], Marc Garellek[2], Robin Samlan[3] and Zhaoyan Zhang[1]

[1] Bureau of Glottal Affairs, Department of Head and Neck Surgery, UCLA School of Medicine, Los Angeles, CA USA
[2] Department of Linguistics, UC San Diego, San Diego, CA USA
[3] Department of Speech, Language, & Hearing Sciences, University of Arizona, Tucson, AZ USA
e-mail: jkreiman@ucla.edu, bgerratt@ucla.edu, mgarellek@ucsd.edu, rsamlan@email.arizona.edu, zyzhang@ucla.edu

Citation / Cómo citar este artículo: Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., and Zhang, Z. (2014). Toward a unified theory of voice production and perception. *Loquens, 1*(1), e009. doi: http://dx.doi.org/10.3989/loquens.2014.009

ABSTRACT: At present, two important questions about voice remain unanswered: When voice quality changes, what physiological alteration caused this change, and if a change to the voice production system occurs, what change in perceived quality can be expected? We argue that these questions can only be answered by an integrated model of voice linking production and perception, and we describe steps towards the development of such a model. Preliminary evidence in support of this approach is also presented. We conclude that development of such a model should be a priority for scientists interested in voice, to explain what physical condition(s) might underlie a given voice quality, or what voice quality might result from a specific physical configuration.

KEYWORDS: voice quality; voice production; modeling; synthesis; acoustics

RESUMEN: *Hacia una teoría unificada de la producción y la percepción de la voz*.- En la actualidad quedan por contestar dos cuestiones importantes relacionadas con la voz, a saber: (1) cuando la cualidad de la voz cambia, ¿qué alteración en el mecanismo vocal es la responsable?; y (2) si se produce un cambio en el sistema de producción de la voz, ¿qué cambio puede esperarse en la cualidad de voz percibida auditivamente? Sostenemos que la única respuesta posible a estas preguntas reside en un modelo de voz integrado que una producción y percepción, y describimos pasos hacia el desarrollo de tal modelo. Presentamos evidencias preliminares para respaldar esta propuesta. Concluimos que el desarrollo de semejante modelo debería ser una prioridad para los científicos interesados en la voz con el fin de explicar qué condición o condiciones físicas podrían subyacer a una cualidad de voz determinada, o qué cualidad de voz podría derivar de una configuración física específica.

PALABRAS CLAVE: cualidad de voz; producción de voz; modelo; síntesis; acústica

1. INTRODUCTION: WHAT IS A UNIFIED THEORY OF VOICE, AND WHY DO WE NEED ONE?

In general, speakers phonate in order to convey information (linguistic or paralinguistic; intentionally or unintentionally) to a listener. The stages of transmitting information in this way can be described by the well-known "speech chain" (Figure 1; Denes & Pinson, 1993). We presently know a good deal about the individual steps along the chain, including motor planning, laryngeal innervation, tissue properties, the biomechanics of laryngeal vibrations, aeroacoustics, acoustics and resonance, and voice perception. However, very few studies address the manner in which information is transmitted from one stage to the next, much less from one end of this chain to the other. As a result, two important questions about voice remain unanswered: 1) When voice quality changes in some way, what caused the change? and 2) If a change occurs in voice production, what will be the resulting perceived change in quality? In this paper, we motivate a model of voice that is designed to answer these questions, and describe our preliminary steps towards generating this model.

In our view, these two questions define the primary goals of the study of voice. Because voice production, acoustics, and perception are all parts of the same communicative process, understanding the communicative function of any of these aspects of voice—laryngeal/physiologic, acoustic, or perceptual—requires knowledge of how each stage interacts with the others in the transmission of vocal information. Details of voice production, acoustics and quality may be misinterpreted without considering the other domains. For example, dozens of different measures of acoustic jitter, shimmer, and harmonics-to-noise ratios (HNRs) have been proposed (see Buder, 2000, for review), presumably because the authors assumed that jitter and shimmer were important vocal characteristics. Hundreds of research papers have examined the correlations between ratings of voice quality and these acoustic measures (see e.g. Maryn, Roy, De Bodt, Van Cauwenberge, & Corthals, 2009, for review), and many more examined correlations between measured perturbation and voice physiology or vocal diagnosis (see e.g. Roy et al., 2013, for review). However, acoustic perturbation measures are not individually informative about voice quality, because listeners cannot hear even large differences in jitter or shimmer (although they are sensitive to changes in the overall level of harmonic vs. inharmonic energy in the voice source; Kreiman & Gerratt, 2005). Further, jitter, shimmer, and noise tell us little about voice production, because they have multiple neurological, biomechanical, aerodynamic, and acoustic causes (see Titze, 1994, for review). Thus, these studies have not resulted in any significant insight into voice production or perception, because questions about causation are difficult to answer without a model explicitly linking production to perception. In another example, clinicians applying stroboscopy or high-speed video imaging often interpret asymmetric vocal fold motion as evidence of vocal pathology. However, although asymmetries sometimes co-occur with abnormal voice qualities, asymmetrical vibration can also occur without any negative effect on the sound of the voice. High-speed video and audio recordings demonstrating such an asymmetry in a normal speaker are presented in the supplemental material [S1] accompanying this paper (see also Zhang, Kreiman, Gerratt, & Garellek, 2013). Again, no theoretical model exists to predict which asymmetries have perceptual consequences, and which do not.

Thus, apart from its basic science interest, a theory describing the links between voice production and perception would also have substantial clinical importance, because the clinical process used to diagnose and treat voice disorders involves a search for cause and effect from one system to another. The primary measure of treatment outcome in voice therapy is perceived voice quality—a patient is not well until their voice sounds better, no matter what the values of instrumental measures may be. Thus, identifying and treating the cause of a deviation in voice quality requires knowledge of which physiological change is responsible for the quality deviation, and predicting treatment outcome requires knowledge of the links between changes in laryngeal physiology and the resulting perceived changes in quality.

Because the acoustic signal links production to perception, our approach to understanding how speakers and listeners produce and perceive communicative changes in voice quality begins with these three steps:

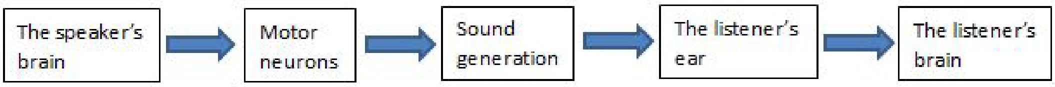
1. Link perception to acoustics by explaining quality in terms of perceptually valid acoustic measures that combine to fully determine voice quality.
2. Link voice production to acoustics and perception by determining which changes in the physiological voice source produce perceptible changes in the acoustic signal.
3. Iterate until the two sets of acoustic parameters align.

We discuss our progress towards each of these goals in what follows. Note that in this approach, quality—the speaker's ultimate concern—"drives" the model. Important acoustic changes are identified by assessing their perceptual salience, after which the acoustic changes that account for what listeners hear can be used to generate hypotheses about what physical changes have important perceptual consequences. By identifying perceptually-important vocal attributes and then examining the glottal pulse shapes associated with these attributes, we will be able to highlight the physical attributes that are important in communication, thus potentially providing data to focus physical modeling efforts towards the physiologic aspects of greatest perceptual importance to speakers and listeners.

2. WHAT IS QUALITY AND HOW SHOULD IT BE MEASURED?

Like pitch and loudness, quality results from an interaction between a listener and a signal. A significant body of behavioral and neuropsychological data (e.g., Andics et al., 2010; Kreiman, Gerratt, & Ito, 2007b; Kreiman & Sidtis, 2011; Latinus, McAleer, Bestelmeyer, & Belin, 2013; Lavner, Rosenhouse, and Gath, 2001; Li & Pastore, 1995; Melara & Marks, 1990) shows that listeners perceive voice quality as an integral pattern, rather than as the sum of a number of separate features (the view implied by use of rating scales). For example, studies of voice recognition from synthetically-altered stimuli indicate that the perceptual importance of a given feature depends on the values of the other attributes of the pattern, and not solely on the value of the feature itself (Van Lancker, Kreiman, & Emmorey, 1985; Van Lancker, Kreiman, & Wickens, 1985). Similarly, in priming experiments, reaction times to famous voices were

Figure 1: The speech chain, describing the transmission of information from a speaker to a listener. The speaker's brain generates an intent to phonate and a set of commands to the relevant muscles; sound is generated when the articulators modulate airflow through the glottis and vocal tract; this sound is transduced by the listener's ear and transformed into neural messages, which are perceived and interpreted by the listener's brain. Adapted from Denes and Pinson (1993).



significantly faster when listeners had previously heard a different exemplar of the voice. Because the priming effect was produced by different samples of each voice, it appears that the benefit derives from the complete voice pattern, not from the specific details of a given sample, again consistent with the view that voices are processed as patterns, and not as bundles of features (Schweinberger, Herholz, & Stief, 1997). In the same manner, listeners appear largely unable to isolate single dimensions in a voice pattern (Kreiman et al., 2007b). Data also demonstrate that harmonic and inharmonic (noise) components of the voice source interact perceptually (Kreiman & Gerratt, 2012), so that listeners' sensitivity to either acoustic attribute depends on the levels of energy in both; and sensitivity to tremor rates depends on tremor amplitude (Kreiman, Gabelman, & Gerratt, 2003). Thus, neither the perceptual meaning of a given quality dimension nor the perceptual significance of an acoustic measure can be assessed without knowledge of the context provided by the complete voice pattern in which the feature or measure functions. It follows that partitioning the overall quality of a voice into separate factors like "breathiness" or "roughness" and asking listeners to isolate and rate qualities is unlikely to tell us enough about how a listener actually perceives either the specific quality or overall quality, so that the sum of a set of individual rating scale responses is not informative enough about how a voice sounds or how it compares to other voices.

If quality is integral, as these studies indicate, then valid measurement requires quantifying the entire voice pattern. To achieve this goal, we apply analysis-by-synthesis to copy each voice sample with a speech synthesizer (Kreiman, Antoñanzas-Barroso, & Gerratt, 2010). Because the acoustic synthesizer parameters combine to completely re-create the perceived voice pattern, they can be considered a psychoacoustic model of voice quality that parametrically represents an integral voice pattern and objectively quantifies a subjective percept.

3. LINKING VOICE QUALITY TO ACOUSTICS

The next step in model development is the selection of parameters to map between acoustics and perception. An adequate voice source model should 1) include enough parameters that it can model any voice quality; and 2) should only include parameters to which listeners are sensitive. In other words, the parameters in the set should be both necessary and sufficient to model voice quality. Development of our psychoacoustic model began with the assumptions that listeners are more likely to pay attention to those acoustic parameters that actually vary across voices (so that they meet the "necessary" test), and that parameters that are constant across voices are less likely to be perceptually important. (For example, if every speaker spoke with exactly the same range of f_0 values, f_0 would not be useful for distinguishing among speakers.) To determine the parameters that actually do vary across speakers—and thus may be perceptually salient—we performed a principal components analysis of the spectra of 70 voices (Kreiman, Gerratt, & Antoñanzas-Barroso, 2007a). FFT spectra for these voices were calculated and normalized to the amplitude of the first harmonic. Spectral envelopes were estimated by connecting the harmonic peaks, and seventy equally-spaced points were chosen along each envelope. Amplitude values for these points served as input to the principal components analysis. Results indicated that four factors accounted for most of the variance in source spectral shape across voices: the source spectral slope above 4 kHz, the slope below 450 Hz, and the slope from 1.5 kHz to 4 kHz (two factors). Similar analyses of a large set of acoustic measures showed significant variability across voices in the relative amplitudes of the first and second harmonics (H1-H2), the relative amplitudes of the second and fourth harmonics (H2-H4),[1] overall spectral slope, and high frequency noise excitation. Our initial perceptual studies therefore focused on these factors.[2]

To assess model sufficiency throughout the course of model development, we used the UCLA voice synthesizer to copy-synthesize several hundred voices

[1] Two measures of the difference in the amplitudes of the first two harmonics are in current use. The first, H1-H2, is measured directly from the source spectrum of the voice, usually as estimated via inverse filtering. This is the measure used in our research. The second, designated H1*-H2*, is estimated from the complete voice signal as recorded at the mouth, but with corrections for the influence of the formants on harmonic amplitudes (a kind of virtual inverse filtering; Hanson, 1997; Hanson & Chuang, 1999).

[2] Note that, although this model describes the spectrum of the voice source (to facilitate mapping to perception, which is usually easier to describe in the spectral domain), most other source models (for example, the Liljencrants-Fant [LF] model [Fant, Liljencrants, & Lin, 1985] or the Fujisaki-Ljungqvist model [Fujisaki & Ljungqvist, 1986]) describe changes in source pulses over time. We return to this issue in the concluding section of this paper.

over a period of several years. The software and procedures used are fully described in Kreiman et al. (2010). Briefly, speakers with and without vocal pathology were selected at random from a large library of voices recorded with a Brüel & Kjær ½" microphone during clinical evaluation. Voices ranged from normal to severely disordered in quality, and a very wide range of diagnoses were represented, including reflux, mass lesions, and functional and neurogenic disorders. The harmonic part of the voice source was estimated by inverse filtering a representative cycle of phonation, and source spectra were fitted with the model (Figure 2). The inharmonic part of the source spectrum was estimated using a cepstral-domain analysis (de Krom, 1993), and f_0 and amplitude contours were tracked on the original voice sample. Finally, the voice was resynthesized by combining these parameters with a model of the vocal tract (estimated by LPC), and all parameters were adjusted until the synthetic copy formed an acceptable match to the natural token. Examples of natural and modeled tokens are included in the supplemental material [S2] accompanying this paper.

We then asked listeners to compare the synthesized tokens to the natural voice samples in a series of "same/different" (AX) tasks. Examination of cases in which the synthetic tokens were distinguished from the natural target stimuli at greater than chance levels suggested that more detail was needed in our modeling of the source spectrum above H4 (e.g., Kreiman, Garellek, & Esposito, 2011; Kreiman & Gerratt, 2011). As a result, we removed the parameter H4-5 kHz from the model and replaced it with two new parameters: the spectral slope from the fourth harmonic to the harmonic nearest 2 kHz in frequency (H4-2 kHz) and the spectral slope from that harmonic to the harmonic nearest 5 kHz in frequency (2 kHz-5 kHz). We then repeated the same/different task, with the result that listeners were unable to consistently distinguish synthetic from natural tokens (d′ < 1). Although evaluation is ongoing, we conclude for the present that the current model (Table 1) provides enough detail to describe the majority of normal and pathological voice qualities.

Establishing the necessity of each parameter as part of the model requires a series of experiments to determine how sensitive listeners are to changes in that parameter. To that end, we began by defining sensitivity as the ratio of the smallest difference in a parameter that listeners can consistently detect (the just-noticeable difference, or JND) to the overall variability of that parameter across speakers (Kreiman & Gerratt, 2010). We reasoned that the smaller the JND was relative to variability, the more information that parameter potentially carried to listeners. To calculate these ratios, we first estimated the range of each model parameter across natural voices by modeling 144 voice samples (79 female, 65 male) via analysis-by-synthesis, and then measuring each of the source model parameters from the modeled source spectra. Samples ranged from normal to severely disordered in quality, and were unselected with respect to diagnosis and the specific voice quality. H1-H2 and H2-H4 values generally ranged from 0-20 dB, while spectral slopes for H4-2 kHz and 2 kHz-5 kHz ranged more widely, from 0 dB-40 dB (see Kreiman, Garellek, Samlan, & Gerratt, 2014, for detailed results).

Table 1: Components of the psychoacoustic model of voice quality and associated voice synthesis parameters.

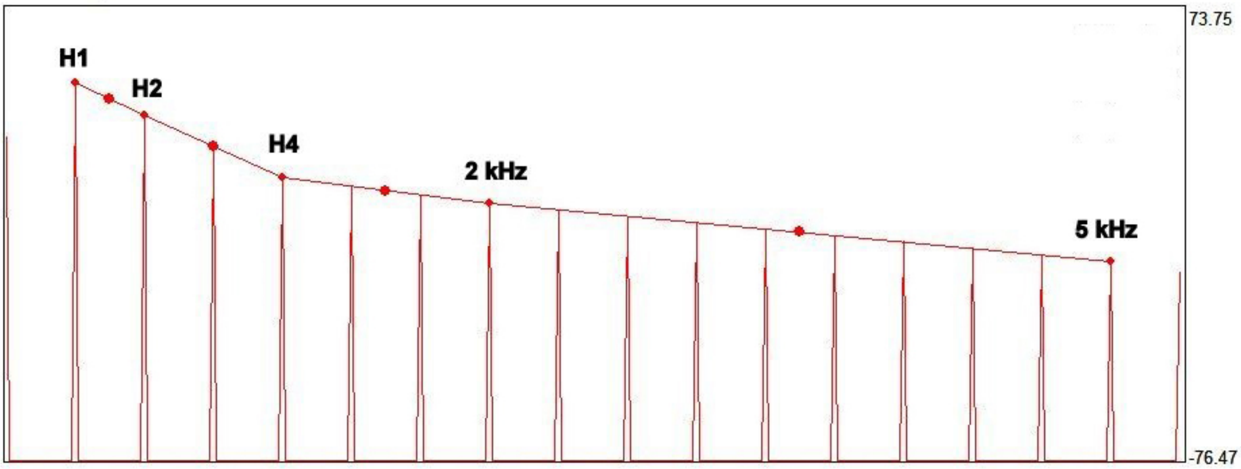
Model Component	Parameters
Harmonic source spectral shape	H1-H2
	H2-H4
	H4-2 kHz
	2 kHz-5 kHz
Inharmonic source excitation	Spectrally-shaped noise-to-harmonics ratio
Time-varying source characteristics	f_0 mean and standard deviation (or f_0 track)
	Amplitude mean and standard deviation (or amplitude track)
Vocal tract transfer function	Formant frequencies/bandwidths
	Spectral zeroes/bandwidths

We next conducted a series of experiments using a one up, two down protocol (Levitt, 1971) to determine the smallest change in each parameter that listeners can reliably detect (e.g., Garellek, Samlan, Kreiman, & Gerratt, 2013; Kreiman & Gerratt, 2012). We synthesized series of stimuli in which a single source spectral parameter was varied in very small steps, and then played pairs of these stimuli to listeners in a same/different (AX) task. When listeners correctly perceived a difference between the stimuli, the difference between stimuli in the next pair decreased; when listeners incorrectly judged the stimuli to be the same, the difference was increased, with the pattern of trials iterating until results began to oscillate around a single difference value which was defined as the JND. (See Kreiman et al., 2014, for details of methods and analyses.) Results are summarized in Table 2. Because the amount of change listeners can hear is small relative to the variability of the parameters across speakers, we tentatively conclude that these parameters are potentially informative to listeners, and that the set of parameters that constitutes the psychoacoustic source model meets the "necessary" test.

Table 2: The ratio of listener sensitivity (JND) to parameter variability across speakers, for the four source model parameters. Data from Kreiman et al. (in preparation).

	Female speakers	Male speakers
H1-H2	0.17	0.24
H2-H4	0.09	0.13
H4-2 kHz	0.09	0.09
2 kHz-5 kHz	0.26	0.29

Figure 2: The four-parameter source spectral model, fitted to the spectrum of a natural voice. The voice source was estimated via inverse filtering, and its spectrum was then calculated via fast Fourier transform. Differences in the amplitudes of individual harmonics are altered so that they conform to the slope of the appropriate model segment.



4. ADDITIONAL EVIDENCE FOR THE PSYCHOACOUSTIC MODEL

This psychoacoustic model makes implicit claims about voice production. First, if voice quality is described by a specific set of acoustic parameters, then speakers must be able to control these parameters or their physiological precursors in order to convey information to listeners. Conversely, aspects of voice production that speakers can easily manipulate should produce perceptible changes in voice quality, which should be measurable with the parameters in the psychoacoustic model.

Some evidence from studies of linguistic uses of voice quality is consistent with the first of these claims, particularly with respect to H1-H2 (or H1*-H2*). In languages with phonemic contrasts in voice quality, speakers must change source characteristics to distinguish meanings, and evidence that they do this in consistent ways supports the notion that they are able to control specific source spectral attributes. For example, in White Hmong (a language in which changes in voice quality accompany some tones), increases in both H1-H2 and H2-H4 (especially in combination) increased the likelihood of perceiving phonemic breathiness, consistent with the view that the percept of breathiness is influenced by a steep drop in harmonic energy in the lower frequencies (Garellek et al., 2013). Speakers of a number of other languages, including Gujarati, Mazatec, Chong, and Green Mong, also distinguish word meanings via differences in H1-H2 (e.g., Andruski & Ratliff, 2000; Blankenship, 2002; Fischer-Jørgensen, 1967; see DiCanio, 2009, and Garellek & Keating, 2011, for review). More directly, Esposito (2012) combined electroglottographic (EGG) measures of laryngeal closing speed and closed quotient with simultaneously-gathered acoustic measures of the source spectrum to examine the physiological and acoustic determinates of the phonation contrast in White Hmong, which has tones characterized by differences in both f_0 and phonation type (breathy, modal, and creaky). Closed quotient was a good predictor of H1*-H2* (r = -0.6, p < .05), which in turn reliably distinguished breathy voice from modal and creaky voice.

Additional evidence comes from a high-speed imaging study of changes in glottal configuration with changes in voice quality along a continuum from breathy to pressed (Kreiman et al., 2012). In this study, six speakers produced steady-state vowels while varying f_0 and voice quality. Measures of the glottal open quotient (OQ) and the asymmetry quotient were made from the high-speed images, and H1*-H2* was measured synchronously from audio recordings of the same utterances. Across speakers and voice qualities, OQ, the asymmetry coefficient, and fundamental frequency accounted for an average of 74% of the variance in H1*-H2*. However, individual speakers used several strategies for varying voice quality, including manipulating glottal gap size, changing OQ, varying f_0, and altering the skewness of glottal pulses. Thus, H1*-H2* can be predicted from glottal configuration with good overall accuracy, although its relationship to phonatory characteristics is complex and speaker dependent.

It is not surprising that speakers would have a variety of phonatory strategies available to them for manipulating H1-H2 in speech. Listeners are highly sensitive to the relative amplitudes of the lowest harmonics (Kreiman & Gerratt, 2010), which convey both paralinguistic information about a variety of personal and interpersonal attributes (see Kreiman & Sidtis, 2011, for review) and linguistic information, as just described. The ability to use different movements to produce the same speech sound has been described for the oral articulators (e.g., Guenther, 1994), and a similar facility for phonation may arise from attempts to produce a particular quality, whether for linguistic or paralinguistic reasons, in the context of different combinations of simultaneous pitch and/or loudness goals.

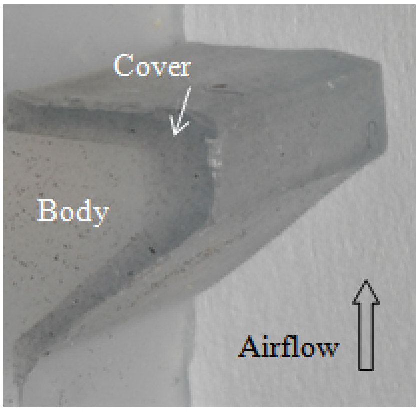
5. ADDITIONAL EVIDENCE FOR THE PSYCHOACOUSTIC MODEL

The second claim implicit in our psychoacoustic model of voice quality is that aspects of voice production that speakers can easily manipulate should produce perceptible changes in voice quality (which also should be quantifiable via the parameters in the psychoacoustic model). This in turn implies that examining the perceptual consequence of changes in physiology will allow us to identify perceptually-relevant mechanical or behavioral manipulations that may be attempted in the clinic. Unfortunately, studies manipulating vocal physiology cannot be conducted in humans, who lack the ability to consciously control individual laryngeal muscles, vocal fold stiffness, glottal gap size and location, and so on. However, we can apply various physical, computational, and ex vivo models of phonation to study the cause-effect relationship between voice production and voice quality by varying parameters of voice production (e.g., vocal fold geometry, stiffness, muscle stimulation, subglottal pressure, etc.) one at a time and observing the consequence on vocal fold vibration, voice acoustics, and voice quality.

Laryngeal modeling has a long history (e.g., Ishizaka and Flanagan, 1972; Titze & Talkin, 1979; Berry, Herzel, Titze, & Krischer, 1994; Steinecke & Herzel, 1995; Story & Titze, 1995; Zhang, Neubauer, & Berry, 2006, 2007; Mendelsohn & Zhang, 2011; Xue, Mittal, Zheng, & Bielamowicz, 2012), but most studies assess only the physical and/or acoustic results of model permutations, without evaluation of any perceptual consequences. One exception to this rule is Zhang et al. (2013), who investigated the acoustic and perceptual consequences of left-right stiffness mismatches in a mechanical self-oscillating vocal fold model. It is generally assumed that left-right stiffness mismatches like those occurring in unilateral vocal fold paralysis or paresis lead to left-right asymmetry in vocal fold vibration, which is often an indication for surgical intervention. However, it is unclear whether left-right stiffness mismatches and the resulting left-right vibrational asymmetry are always perceptually significant. In other words, the consequences of variability in the material properties and geometry of vocal folds on voice quality are not well understood, so we do not know if vibrational asymmetry (or other deviations from normal vocal fold movement) leads to acoustic changes that people can hear.

To address this question, a body-cover two-layer mechanical vocal fold model was used (Figure 3). A series of left-right asymmetric conditions with varying left-right mismatches in body stiffness were created by varying the body-layer stiffness of the left vocal fold model while the right vocal fold remained unchanged. All vocal fold models had identical vocal fold geometry and cover-layer stiffness. For each asymmetric vocal fold model, phonation tests were performed using a flow-ramp procedure in which the flow rate was increased in steps from zero to a value above onset of vibration. The outside acoustic signals recorded at a subglottal pressure 10% above onset were used in subsequent acoustic analysis and perceptual tests. Measures of source spectral slope were extracted (as discussed above) for each asymmetric condition. In addition, the number of harmonics below 8 kHz in the sound spectrum was also measured. For perceptual tests, listeners were asked to evaluate the voice samples in a sort-and-rate task (Figure 4), in which they sorted the voice samples along a straight line so that tokens that sounded similar were placed close together on the line (Granqvist, 2003; Zhang et al., 2013).

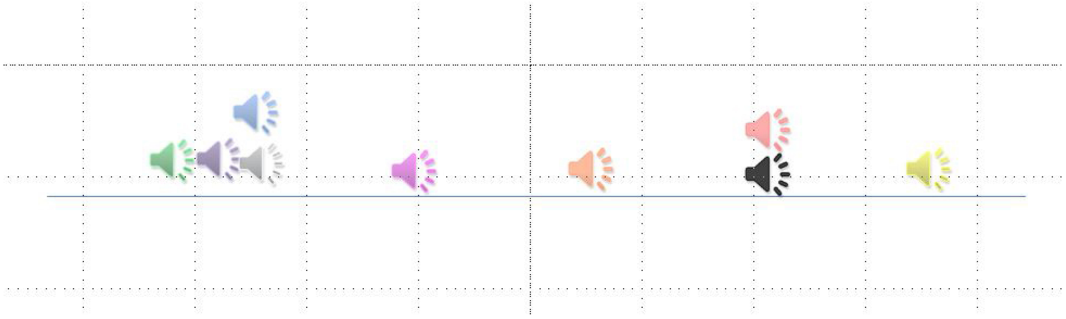
Figure 3: The two-layer cover-body vocal fold model used in Zhang et al. (2013).



This study revealed two regimes of distinct vibratory patterns with varying left-right stiffness mismatch. For conditions with a large left-right stiffness mismatch, only the soft-body fold was excited while the stiff-body fold barely moved, which led to weak excitation of high-order harmonics. For small left-right stiffness mismatches, both folds were strongly excited but the stiff fold always led in phase in their motion. The outside sound in this regime had strong excitation of high-order harmonics. Perceptual tests also demonstrated two clusters, each corresponding to one of the two vibratory regimes. There was no significant difference between voice samples within the same perceptual regimes.

This study showed that changes to the degree of left-right stiffness mismatch and the resulting left-right vibratory asymmetry did not produce perceptually significant differences in quality unless the stiffness mismatch was large enough to cause a qualitative change in vibratory mode (a bifurcation). This suggests that a vibration pattern with left-right asymmetry does not necessarily result in a salient deviation in voice quality, and thus may not always be of clinical significance. Perceptual changes were explicable with reference to the psychoa-

Figure 4: The user interface from the sort-and-rate perceptual task. Listeners click on an icon to play a voice sample, then drag the icons so that those that sound similar are placed close together on the line, and those that sound different are farther apart.



coustic model parameters, including spectral slopes and the noise-to-harmonics ratio, consistent with the general framework being developed here.

A similar approach has also been used recently by Samlan and colleagues (Samlan & Story, 2011; Samlan, Story, & Bunton, 2013), who studied the relationship between kinematic, acoustic, and perceptual measures using voice samples generated with a computational vocal fold model coupled to a model of the vocal tract. For example, Samlan and Story (2011) manipulated vocal process separation, vocal fold bulging, the "nodal point ratio" (the ratio of the point at which mucosal fold motion begins to overall vocal fold thickness), and epilaryngeal area, and measured the effects on H1-H2 and on the cepstral peak prominence (CPP; Hillenbrand & Houde, 1996), a measure of the relative levels of harmonic and inharmonic energy in the voice. Samlan et al. (2013) added measures of spectral slope and ratings of perceived breathiness to the mix. They found a clear relationship between CPP, separation of the vocal processes, and ratings of breathiness (presumably related to increases in turbulent noise with increasing glottal gaps), with additional variance explained by nodal point ratio, vocal fold bulging, and spectral slope. The relationship between measures of spectral slope and model parameters depended on severity of rated breathiness: H1-H2 was a better predictor of mild breathiness of the kind often associated with "vocal weakness," while overall spectral slope was a better predictor when significant high-frequency noise was present in the voice. This finding reflects both the complexity of causation in vocal physiology and the perceptual multidimensionality of breathiness (Kreiman, Gerratt, & Berke, 1994).

Modeling studies like these are attractive because they allow simultaneous direct manipulation of many parameters in a well-controlled laboratory setting. The limitations of this approach lie in the vocal fold model used, or specifically, how realistically these models (the mechanical or computational model in the examples above) reproduce the physiology and physics of human phonation. Ideally, we would like to model phonation in a living human being, but direct manipulation and measurement of muscle activities and vocal fold properties (geometry and stiffness) are currently impossible in living human subjects, due to the great sensitivity and relative inaccessibility of the larynx. To overcome this problem, an ex-vivo perfused living model of human phonation has been developed (Berke, Mendelsohn, Howard, & Zhang, 2013).[3] In this model, a human larynx and trachea are harvested from an organ donor post mortem and perfused with oxygenated blood. The tissue remains viable for several hours, and because the laryngeal nerves and muscles are still living, they can be directly stimulated in a well-controlled laboratory setting, as opposed to mechanical manipulations in ex vivo models in which the material properties of the muscles and other tissues change post-mortem. This model makes it possible to study the effects of known levels of actual human laryngeal muscle activation on vocal fold stiffness and geometry. It also allows us to study interactions among muscles (for example, the thyroarytenoid and cricothyroid) in investigations of the control of pitch, loudness, and voice quality. Although use of this model is only beginning, when combined with perceptual testing and acoustical analysis, it promises to provide new data about the precursors and correlates of changes in voice quality.

6. CONCLUSIONS, FUTURE WORK, AND IMPLICATIONS FOR CLINICAL PRACTICE

The studies reviewed in this paper suggest that phonation is best viewed as part of a communicative process, the pieces of which are difficult to understand out of the context of the entire process. Thus, understanding and ultimately predicting how speakers produce the intended voice quality (and how disorders disturb this process) requires a unified model of voice that links production to perception.

[3] Video examples can be found at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3562273/figure/v1/ and at http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3562273/figure/v2/.

Many issues await resolution as we work towards this goal. Because phonation takes place in the time domain while perception depends largely on spectral information, understanding the relationship between perception and production requires mapping between time and spectral domain representations, which has proven difficult (e.g., Fant, 1995; Ni Chasaide & Gobl, 1997). More than one physical configuration may produce the same voice quality; conversely, large changes in configuration may not result in changes in quality. Variables in the current voice source model certainly interact: for example, we know that the perceptual salience of changes in high-frequency harmonics depends on the signal-to-noise ratio and on the shape of the noise spectrum (Kreiman & Gerratt, 2012). Finally, the extreme complexity of the phonatory system (and of human communication in general) and the difficulty inherent in observing and measuring many aspects of phonation make it hard both to gather all the needed data regarding interactions among factors, and to model those data once they are gathered. Despite these complications and complexities, we argue that the systematic approach described in this paper will eventually make it possible to understand how features of the voice production system combine with attributes of the perceptual system to transmit voice information from speakers to listeners, but only if the research community considers this a primary goal for voice research.

ACKNOWLEDGEMENTS

This research was funded by NIH/NIDCD grants DC01797, DC009229, and DC011299. A preliminary version of this paper was presented at the 2013 ASHA Convention in Chicago, IL. Software described in this paper is available free of charge from the authors.

REFERENCES

Andics, A., McQueen, J. M., Petersson, K. M., Gál, V., Rudas, G., & Vidnyánszky, Z. (2010). Neural mechanisms for voice recognition. *NeuroImage, 52*, 1528–1540. http://dx.doi.org/10.1016/j.neuroimage.2010.05.048

Andruski, J., & Ratliff, M. (2000). Phonation types in production of phonological tone: The case of Green Mong. *Journal of the International Phonetic Association, 30*, 37–61. http://dx.doi.org/10.1017/S0025100300006654

Berke, G., Mendelsohn, A. H., Howard, N. S., & Zhang, Z. (2013). Neuromuscular induced phonation in a human ex vivo perfused larynx preparation. *Journal of the Acoustical Society of America, 133*, EL114–EL117. http://dx.doi.org/10.1121/1.4776776

Berry, D. A., Herzel, H., Titze, I. R., & Krischer, K. (1994). Interpretation of biomechanical simulations of normal and chaotic vocal fold oscillations with empirical eigenfunctions. *Journal of the Acoustical Society of America, 95*, 3595–3604. http://dx.doi.org/10.1121/1.409875

Blankenship, B. (2002). The timing of nonmodal phonation in vowels. *Journal of Phonetics, 30*, 163–191. http://dx.doi.org/10.1006/jpho.2001.0155

Buder, E.H. (2000). Acoustic analysis of voice quality: A tabulation of algorithms 1902-1990. In R.D. Kent & M.J. Ball (Eds.), *Voice quality measurement* (pp. 119–244). San Diego, CA: Singular.

de Krom, G. (1993). A cepstrum-based technique for determining a harmonics-to-noise ratio in speech signals. *Journal of Speech and Hearing Research, 36*, 254–266.

Denes, P. B., and Pinson, E. N. (1993). *The speech chain* (2nd ed.). New York, NY: WH Freeman.

DiCanio, C. T. (2009). The phonetics of register in Takhian Thong Chong. *Journal of the International Phonetic Association, 39*, 162–188. http://dx.doi.org/10.1017/S0025100309003879

Esposito, C. M. (2012). An acoustic and electroglottographic study of White Hmong phonation. *Journal of Phonetics, 40*, 466–476. http://dx.doi.org/10.1016/j.wocn.2012.02.007

Fant, G. (1995). The LF model revisited. Transformations and frequency domain analysis. *STL-QPSR, 36*(2–3), 119–156.

Fant, G., Liljencrants, J., & Lin, Q. (1985). A four-parameter model of glottal flow. *STL-QPSR, 26*(4), 1–13.

Fischer-Jørgensen, E. (1967). Phonetic analysis of breathy (murmured) vowels in Gujarati. *Indian Linguistics, 28*, 71–139.

Fujisaki, H., & Ljungqvist, M. (1986). Proposal and evaluation of models for the glottal source waveform. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 11), 1605–1608. http://dx.doi.org/10.1109/ICASSP.1986.1169239

Garellek, M., & Keating, P. (2011). The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. *Journal of the International Phonetic Association, 41*, 185–205. http://dx.doi.org/10.1017/S0025100311000193

Garellek, M., Keating, P., Esposito, C., & Kreiman, J. (2013). Voice quality and tone identification in White Hmong. *Journal of the Acoustical Society of America, 133*, 1087–1089.

Granqvist, S. (2003). The visual sort and rate method for perceptual evaluation in listening tests. *Logopedics Phoniatrics Vocology, 28*, 109–116. http://dx.doi.org/10.1080/14015430310015255

Guenther, F. H. (1994). A neural network model of speech acquisition and motor equivalent speech production. *Biological Cybernetics, 72*, 43–53. http://dx.doi.org/10.1007/BF00206237

Hanson, H. M. (1997). Glottal characteristics of female speakers: Acoustic correlates. *Journal of the Acoustical Society of America, 101*, 466–481. http://dx.doi.org/10.1121/1.417991

Hanson, H. M., & Chuang, E. S. (1999). Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *Journal of the Acoustical Society of America, 106*, 1064–1077. http://dx.doi.org/10.1121/1.427116

Hillenbrand, J., & Houde, R. A. (1996). Acoustic correlates of breathy vocal quality: Dysphonic voices and continuous speech. *Journal of Speech and Hearing Research, 39*, 311–321.

Ishizaka, K., & Flanagan, J. L. (1972). Synthesis of voiced sounds from a two-mass model of the vocal cords. *Bell System Technical Journal, 51*, 1233–1268. http://dx.doi.org/10.1002/j.1538-7305.1972.tb02651.x

Kreiman, J., Antoñanzas-Barroso, N., & Gerratt, B. R. (2010). Integrated software for analysis and synthesis of voice quality. *Behavior Research Methods, 42*, 1030–1041. http://dx.doi.org/10.3758/BRM.42.4.1030

Kreiman, J., Gabelman, B., & Gerratt, B. R. (2003). Perception of vocal tremor. *Journal of Speech, Language, & Hearing Research, 46*, 203–214. http://dx.doi.org/10.1044/1092-4388(2003/016)

Kreiman, J., Garellek, M., & Esposito, C. (2011). Perceptual importance of the voice source spectrum from H2 to 2 kHz. *Journal of the Acoustical Society of America, 130*, 2570. http://dx.doi.org/10.1121/1.3655295

Kreiman, J., Garellek, M., Samlan, R. A., and Gerratt, B. R. (2014). Perceptual sensitivity to a model of the voice source spectrum. Manuscript in preparation.

Kreiman, J., & Gerratt, B. R. (2005). Perception of aperiodicity in pathological voice. *Journal of the Acoustical Society of America, 117*, 2201–2211. http://dx.doi.org/10.1121/1.1858351

Kreiman, J., & Gerratt, B. R. (2010). Perceptual sensitivity to first harmonic amplitude in the voice source. *Journal of the Acoustical Society of America, 128*, 2085–2089. http://dx.doi.org/10.1121/1.3478784

Kreiman, J., & Gerratt, B. R. (2011). Modeling overall voice quality with a small set of acoustic parameters. *Journal of the Acoustical*

Society of America, 129, 2529. http://dx.doi.org/10.1121/1.3588381

Kreiman, J., & Gerratt, B. R. (2012). Perceptual interaction of the harmonic source and noise in voice. *Journal of the Acoustical Society of America, 131*, 492–500. http://dx.doi.org/10.1121/1.3665997

Kreiman, J., Gerratt, B. R., & Antoñanzas-Barroso, N. (2007a). Measures of glottal source spectrum. *Journal of Speech and Hearing Research, 50*, 595–610. http://dx.doi.org/10.1044/1092-4388(2007/042)

Kreiman, J., Gerratt, B. R., & Berke, G. S. (1994). The multidimensional nature of pathologic vocal quality. *Journal of the Acoustical Society of America, 96*, 1291–1302. http://dx.doi.org/10.1121/1.410277

Kreiman, J., Gerratt, B. R., & Ito, M. (2007b). When and why listeners disagree in voice quality assessment tasks. *Journal of the Acoustic Society of America, 122*, 2354–2364. http://dx.doi.org/10.1121/1.2770547

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. *Journal of the Acoustical Society of America, 132*, 2625–2632. http://dx.doi.org/10.1121/1.4747007

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies. An interdisciplinary approach to voice production and perception*. Malden, MA: Wiley-Blackwell. http://dx.doi.org/10.1002/9781444395068

Latinus, M., McAleer, P., Bestelmeyer, P. E. G., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology, 23*, 1075–1080. http://dx.doi.org/10.1016/j.cub.2013.04.055

Lavner, Y., Rosenhouse, J., & Gath, I. (2001). The prototype model in speaker identification by human listeners. *International Journal of Speech Technology, 4*, 63–74. http://dx.doi.org/10.1023/A:1009656816383

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America, 49*, 467–478. http://dx.doi.org/10.1121/1.1912375

Li, X., & Pastore, R. E. (1995). Perceptual constancy of a global spectral property: Spectral slope discrimination. *Journal of the Acoustical Society of America, 98*, 1956–68. http://dx.doi.org/10.1121/1.413315

Maryn, Y., Roy, N., De Bodt, M., Van Cauwenberge, P., & Corthals, P. (2009). Acoustic measurement of overall voice quality: A meta-analysis. *Journal of the Acoustical Society of America, 126*, 2619–2634. http://dx.doi.org/10.1121/1.3224706

Melara, R. D., & Marks, L. E. (1990). Interaction among auditory dimensions: Timbre, pitch, and loudness. *Perception and Psychophysics, 48*, 169–178. http://dx.doi.org/10.3758/BF03207084

Mendelsohn, A. H., & Zhang, Z. (2011). Phonation threshold pressure and onset frequency in a two-layer physical model of the vocal folds. *Journal of the Acoustical Society of America, 130*, 2961–2968. http://dx.doi.org/10.1121/1.3644913

Ni Chasaide, A., & Gobl, C. (1997). Voice source variation. In W. J. Hardcastle & J. Laver (Eds.), *The handbook of phonetic sciences* (pp. 427–461). Oxford, UK: Blackwell.

Roy, N., Barkmeier-Kraemer, J., Eadie, T., Sivasankar, M. P., Mehta, D., Paul, D., and Hillman, R. (2013). Evidence□based clinical voice assessment: A systematic review. *American Journal of Speech-Language Pathology, 22*, 212–226. http://dx.doi.org/10.1044/1058-0360(2012/12-0014)

Samlan, R. A., & Story, B. H. (2011). Relation of structural and vibratory kinematics of the vocal folds to two acoustic measures of breathy voice based on computational modeling. *Journal of Speech, Language, & Hearing Research, 54*, 1267–1283. http://dx.doi.org/10.1044/1092-4388(2011/10-0195)

Samlan, R. A., Story, B. H., & Bunton, K. (2013). Relation of perceived breathiness to laryngeal kinematics and acoustic measures based on computational modeling. *Journal of Speech, Language, & Hearing Research, 56*, 1209–1223. http://dx.doi.org/10.1044/1092-4388(2012/12-0194)

Schweinberger, S. R., Herholz, A., & Stief, V. (1997). Auditory long-term memory: Repetition priming of voice recognition. *Quarterly Journal of Experimental Psychology Section A: Human Experimental Psychology, 50*, 498–517. http://dx.doi.org/10.1080/713755724

Steinecke, I., & Herzel, H. (1995). Bifurcations in an asymmetric vocal fold model. *Journal of the Acoustical Society of America, 97*, 1874–1884. http://dx.doi.org/10.1121/1.412061

Story, B. H., & Titze, I. R. (1995). Voice simulation with a body-cover model of the vocal folds. *Journal of the Acoustical Society of America, 97*, 1249–1260. http://dx.doi.org/10.1121/1.412234

Titze, I. R. (1994). *Principles of voice production*. Englewood Cliffs, NJ: Prentice Hall.

Titze, I. R., & Talkin, D. T. (1979). A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation. *Journal of the Acoustical Society of America, 66*, 60–74. http://dx.doi.org/10.1121/1.382973

Van Lancker, D., Kreiman, J., & Emmorey, K. (1985). Familiar voice recognition: Patterns and parameters. Part I: Recognition of backward voices. *Journal of Phonetics, 13*, 19–38.

Van Lancker, D., Kreiman, J., & Wickens, T. D. (1985). Familiar voice recognition: Patterns and parameters. Part II: Recognition of rate-altered voices. *Journal of Phonetics, 13*, 39–52.

Xue, Q., Mittal, R., Zheng, X., & Bielamowicz, S. (2012). Computational modeling of phonatory dynamics in a tubular three dimensional model of the human larynx. *Journal of the Acoustical Society of America, 132*, 1602–1613. http://dx.doi.org/10.1121/1.4740485

Zhang, Z., Kreiman, J., Gerratt, B. R., & Garellek, M. (2013). Acoustic and perceptual effects of changes in body layer stiffness in symmetric and asymmetric vocal fold models. *Journal of the Acoustical Society of America, 133*, 453–462. http://dx.doi.org/10.1121/1.4770235

Zhang, Z., Neubauer, J., & Berry, D. A. (2006). The influence of subglottal acoustics in laboratory models of phonation. *Journal of the Acoustical Society of America, 120*, 1558–1569. http://dx.doi.org/10.1121/1.2225682

Zhang, Z., Neubauer, J., & Berry, D. A. (2007). Physical mechanisms of phonation onset: A linear stability analysis of an aeroelastic continuum model of phonation. *Journal of the Acoustical Society of America, 122*, 2279–2295. http://dx.doi.org/10.1121/1.2773949

SUPPLEMENTAL MATERIAL

S1. High-speed video showing asymmetrical vocal fold motion with normal voice quality. The accompanying audio file was synchronously recorded with this video.

Video file:	asymm_male.mp4
Audio file:	asymm_male.mp3

S2. Examples of natural voice samples and copies synthesized using the psychoacoustic voice source model and the UCLA voice synthesizer.

Example 1:	female1_natural.mp3 female1_synthetic.mp3
Example 2:	female2_natural.mp3 female2_synthetic.mp3
Example 3:	male1_natural.mp3 male1_synthetic.mp3
Example 4:	male2_natural.mp3 male2_synthetic.mp3

All these files are accessible from the html version of the paper (click here).