A GENERAL THEORY OF SINGING VOICE PERCEPTION: AN INTRODUCTION

Ian Howell, DMA
New England Conservatory of Music
24 September 2018
Introduction

Singing is, at its core, an aural art form. Like our other senses, hearing limits and colors our experience of reality. Yet many of the models we use to explain the acoustic behavior of the singing voice rely on computer measurements, predominantly of speech. These measurements tell us something true about the pressure wave as measured in the air, but miss qualitative aspects of sound dependably imparted by the process of hearing. By coming to understand how the ear and hearing brain behave, we can ask how one might possibly perceive a singing voice. This is another way of asking what sounds the human voice is capable of making, which cuts directly to the heart of what it is to train a voice.

The human voice can transmit complex—even abstract—ideas through language, create purely musical or emotionally motivated sounds, or do both at the same time. Linguistics and language perception and cognition are deep and fruitful fields of study, and the models that voice pedagogy borrows (e.g. the International Phonetics Alphabet, vowel formant maps, spectrographic analysis) clearly bring powerful tools to bear on the study of the singing voice. As helpful as these tools may be to characterize spoken language, they frequently fall short of describing the qualities prized in singing voices. Such qualities are vital to characterizing musical sounds but may have little to no impact on language cognition. These include distinctions along the continuums of bright and dark, loud and quiet, high and low, rough or pure, tone or noise, and sound or silence—the qualitative oppositions that transmit meaning in musical sounds, gestures, and phrases. Singers share a common, aurally transmitted understanding of what is meant, felt, and heard when we discuss the sound of a singer, be it a clear or muddy vowel at a certain pitch, a specific registration choice, the difference between sufficient or insufficient
glottal closure, or an overall color to define a genre. These are aspects of sound quite separate
from any linguistic meaning the singer may convey.

Phonemes in spoken language are groupings of subtly different sounds. These differences
are meaningful, but not technically consequential. In singing, the physical adjustments associated
with such subtly different sounds may mean the difference between success and strain. Anyone
who sings in a style that requires deviations from a speech shape vocal tract to effectively
resonate higher pitches (broadly called vowel modification) understands that singers leave the
strict definitions of IPA symbols as pitch rises. Any treble singer who can sing above the staff
similarly understands that high pitches are qualitatively and experientially different from low
pitches. Additionally, phonemes in language are interdependent; it is as often phonetic context
that imbues a sound with linguistic meaning as it is the objective timbral character.¹ However,
singers frequently make sounds out of linguistic context; what else is a melisma or vocalise but
an exploration of the objective timbral character of the singer’s sound? It is perhaps more
profitable to think of such changes in a singing voice as one would in a violin or piano;
instrument with definite tonal characteristics based on pitch range and emotional affect.

Were one to begin the study of the timbral qualities of such a musical instrument with a
clean slate, it is doubtful that the obvious first step would be to remove all the sound below and
above the analog telephone bandwidth (roughly 300Hz-3400Hz).² However, with rare exception
that is exactly what our singing voice community has done. Our models for understanding the

¹ See Mark H. Ashcraft, Cognition (Upper Saddle River, NJ: Pearson Prentice Hall, 2006), 382-3, for a
discussion of coarticulation. See also Terrance M. Nearley, “Static, dynamic, and relational properties in vowel
between the context effect and inherent quality in current speech research.
² See www.telecomabc.com/b/bandwidth.html.
sound of a singing voice fall back on either the frequency bandwidth of the first two spectral peaks ($F_1$ and $F_2$), or those spectral peaks plus the Singer’s Formant Cluster. At best, this approach falls short of effectively describing the singing voice. At worst, this actually points our ears and imaginations away from what could be pedagogically helpful ideas. It is understandable and explainable, but it is a limitation we may benefit from removing. Into this ambiguous space—a space that voice teachers explore every day—I would like to introduce a general theory of singing voice perception. This is a way to discuss the sound of the singing voice in practical and objective perceptual terms that transcend the ambiguities built into the study of spoken language.

The actual phenomenon of a singing voice is perhaps too complex to grasp all at once. Instead, we use models to teach important, generalizable concepts. It is the nature of a model to exclude information, and the art of creating a good model rests on knowing what information may be left out. Psychoacoustics, the field of study concerned with the perception of sound, offers important tools to flesh out our models. Rather than simply observe objectively measurable aspects of a singer’s spectrum, we can begin to ask questions like: What does that aspect of the singer’s spectrum sound like? What does it contribute to the whole? Does it stimulate my ear in a way that the act of perception dependably introduces qualitative elements of timbre? Can I train my ear to listen for those qualities and link them to successful or unsuccessful attempts at a technical goal? Can I hear everything I see on a spectrogram?

It is important to mention that everything that follows builds on what we currently know and teach. I aim to add information and detail, not to arbitrarily tear down the models that we currently use, and perhaps most importantly not to call into question the logic or motivations that led to their adoption. If anything, I hope to ask how we might understand the singing voice if our
models simply include more knowledge from psychoacoustics. This is a process that has already taken place with the incorporation of anatomy & physiology, linguistics, and acoustics into voice pedagogy. In these cases, the interdisciplinary mingling often fleshes out what works well, generates new frameworks for teaching, and in some cases reforms what have been questionable assumptions. Delightfully, the specific needs and questions of the singing voice community also feed back into these bodies of research, at times challenging basic assumptions for the better. What follows similarly enlivens what works well in our current models, and seeks to fill in the unsatisfying exceptions our current models force us to accommodate.

**It is About Time: Making Vocal Timbre a More Manageable Problem**

Much ink has been spilled exploring how complex and problematic the definition of timbre is: “Timbre is 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.”\(^3\) Patterson (2010) writes,

> Informally, the standard definition of timbre is regarded with considerable amusement. You might expect the definition of timbre to tell you something about what timbre is, but all the definition tells you is that there are a few things that timbre is not. It is not pitch, it is not loudness, and it is not duration. It is everything else."\(^4\)


\(^4\) Patterson, 223.
Cogan (1969) similarly laments, “Timbre, of all the parameters of music, is the one least considered. It lacks not only an adequate theory, but even an inadequate one.”⁵ Timbre is a catch-all term to characterize the differences between two sounds as they unfold over time.

Slawson (1985) encourages, “limiting the scope,” of inquiry into timbre by separating its characteristics into those that can exist in the abstract and those that only exist in the passage of time.⁶ This may sound hideously complicated for someone studying voice pedagogy, but stick with it. There is something valuable on the other side of this idea. His analogy using visible light is a helpful place to start:

A light may be described as… slowly changing in color, but… changing shade is not itself regarded as a color. When we say that a sound color has no temporal aspect, this rules out of consideration all changes in sounds. That is, a sound may be heard to be changing from one color to another, but the change itself is not a sound color.⁷

Slawson’s definition of the term sound color is fairly narrow in context; however, it does point to similar aspects of a singer’s sound that exist as abstract perceptual qualities. One may discuss brightness in a singing voice, for example. Exactly how bright that voice is, and whether that brightness changes in quantity or frequency range over time is powerfully relevant to the timbre of the voice. However, the qualitative characteristic of brightness exists as an abstract idea, despite the fact that one may discuss how its presence or absence unfolds over time. By contrast, qualities like ease of production, stability, vocal efficiency, or even something as ubiquitous as a crescendo only exist through the passage of time. They arise because of accumulated information

---

in the mind of the listener, not because of an inherent quality present in any given moment. It is helpful then to think of other qualities of a voice that exist outside of the passage of time. They are temporally invariant; their defining qualities do not vary with time.

The limited scope approach put forth by Slawson points to a set of perceptual characteristics that are powerfully relevant to understanding the singing voice, complement our current acoustic models, and have a strong basis of support within the psychoacoustics literature. These temporally invariant characteristics of a singer’s sound are: pitch, auditory roughness, and tone color. Taken together, they offer a framework for discussing the sound of a singer in objective terms separate from the context of the sung language. My hope is that this framework may serve as a neutral home base for understanding the sound of a singer. This is not a simple trick to address one particular aspect of the voice, nor is it only relevant to one style of singing. There is not just one single practical application. It is as close to a fundamental shift in the way one might think about and listen to a singing voice as I have experienced in my life. I encourage you to give this material time to sink in, attempt to experience it as aurally as you can, and consider how it has the potential to inform nearly everything about the way that you listen to a singer.

**New Models for Understanding the Singing Voice**

Most scientific models for thinking about the singing voice center on either the physiological means of production—what muscles engage to position and move the singing body, and how do the vocal folds behave in the flow of air—or on acoustical measurements—how a computer mathematically divides the sound of the singer into a visible spectrum.
However, these are moments in the process of creating and perceiving a voice, not immutable distillations. It is worth asking whether these models offer a comprehensive view of the voice. In particular, it is worth asking whether the visible spectrum created by a fast Fourier transform—the process that takes a complex compression wave and divides it in real time into harmonics of differing intensities, and the primary way we visually display a voice in the voice pedagogy literature—represents the way that the voice is either produced or perceived.

It may be helpful for the reader to consider that voicing is a process with many steps, and that each step transforms a physical phenomenon. The pressurizing of subglottal air in many ways represents the potential to voice, and one cannot layer complex aesthetic qualities onto it. The train of air puffs that escape the rapidly opening and closing glottis while sustaining a pitch creates a repeating pattern of pressure waves. All the information we understand as the spectrum of this source sound is contained within the rate and intensity of the rise and fall of this pressure wave. This phenomenon is perhaps best modeled by a simple line graph charting transglottal airflow against time. When the glottis closes, and the vocal tract becomes an effective resonator, the vocal tract reshapes this simple pressure wave. What was a repeating rise and fall becomes a much more complex pattern with multiple rises and falls of varying power. This pattern continues to repeat at the same frequency of the source puff train, preserving pitch while changing timbre. We may think of the vocal tract as a grouping of individual resonances, and indeed one can dependably effect change in just a portion of the radiated spectrum by specific physiological manipulations, but the entire vocal tract appears to act on the entire source pressure

---

wave as a whole.\footnote{Worth mentioning here is an unpublished presentation given by Dr. Brad Story at the 2016 Pan American Vocology Association’s national conference. In it he demonstrated a model that disrupts the idea that the vocal tract functions as a set of isolated resonators. Instead, various adjacent cross sections of the vocal tract cause opposite rises and falls in the frequency centers of the various standing resonances.} In other words, it seems credible to think of the transformation of the source signal by the vocal tract as a transformation of a simple pressure wave into a complex pressure wave, rather than the interaction of independent voice source harmonics and vocal tract resonances. Note that this is despite the fact that every step along the way can be understood through the lens of a fast Fourier transform, which appears to imply the opposite. This reshaped pressure wave is further changed by the acoustical properties of the room (signal chain if amplified singing) and ear canal. Once the energy of the wave is transferred from the last line of air molecules to the ear drum, the physical limitations of the middle and inner ear come into play. Again, at any point one could impose a fast Fourier transform on the signal and understand the sound according to its spectral components. However, this does not mean that a spectral analysis captures the physical nature of the phenomenon at that moment.

The last step in this process is the one most frequently overlooked in voice pedagogy. The act of hearing, or more specifically, auditory transduction, does not just pass through the physical properties of the pressure wave from the air to the brain. It actually introduces qualitative aspects of sound relevant to the percept of the singing voice. This is the process by which the vibrations striking the ear drum are turned into the nerve impulses received and interpreted by the brain. Every step in the process of auditory transduction—be it the reshaping of the pressure wave by the ear canal, the flexibility and mobility of the eardrum and bones of the middle ear, or the spatially organized frequency division and resynthesis of the vibrations by
the inner ear—changes aspects of the sound before it reaches the brain. We must keep in mind that the results of a fast Fourier transform represent a single moment in this process, perhaps most accurately mapping to the response of a structure within the cochlea called the basilar membrane. In a computer analysis of sound within spectrographs like VoceVista, Overtone Analyzer, Praat, and similar, we see a frozen moment in the process of sound generation and perception. In a real human, this frequency division is quickly followed by a further transformation of the sound, which introduces qualitative aspects of timbre not reflected in a spectrogram.

This idea is worth spinning out a little further. Despite important non-linear feedback loops, broadly speaking the act of singing is a linear process. The feedback of the sound in the room may change the next sound the singer makes, but the singer cannot change the pressure wave once it leaves their body. Similarly, once the pressure wave enters the ear, that percept can no longer be affected by the acoustics of the room. Applying a fast Fourier transform—inserting the cochlea—to the pressure wave at any step of this complex process absolutely tells us something valuable. It is helpful to understand the spectrum created by the vocal folds as they thicken or thin, for example. The high frequency energy present when the former (and not the latter) is perceived is perceptually relevant and appears in a spectrum much more intuitively than a flow glottogram or waveform. However, we must be cautious not to arbitrarily impose the cochlea at this earlier step in the process. The cochlea is not, in reality, stimulated by the vocal fold source sound prior to the filtering of the vocal tract. We do not hear the vocal fold source sound. We hear the interaction of the source pressure wave, the vocal tract, the room and/or signal chain, and the hearing process as a functional whole.
Dividing the complex wave into its harmonic components—a translation rather than transformation—may make it more difficult to conceptualize how the physical pressure wave transforms as it moves from the glottis through the vocal tract, and again how that transformed pressure wave is further changed by the acoustics of the room. We should absolutely use fast Fourier transform technology to convert counterintuitive graphs into more apprehensible graphs to explain the steps of voicing. However, we must be cautious to switch back to a more accurate model before we can imagine how the pressure wave changes as it moves on to the next step. In short, a spectrogram always needs to be qualified. If it seeks to represent some aspect of the pressure wave prior to the transformation by the inner ear, we must recognize this risks misrepresenting the physical nature of the phenomenon itself. If it seeks to represent the sound as perceived, we must recognize that the inner ear introduces qualitative aspects of the sound after the frequency division displayed on the spectrogram. A spectrogram or power spectrum (the two most common images derived from a fast Fourier transform) is a frozen moment in the process of producing and perceiving sound, and the moment in that process that a spectrogram captures does not represent what came before or after particularly well.

We are then forced to consider what a spectrogram or power spectrum leaves out. In 2017, I proposed a framework for thinking about the objective aspects of timbre relevant to a voice. The model explores the voice through the simplified intersection of three perceptual qualities: pitch, auditory roughness, and tone color. These qualities are quickly identifiable, easily differentiated, and do not exist prior to the perceptual processes of the cochlea. My hope is

that this model allows voice teachers and students to begin to attach pedagogically relevant qualities to the spectrogram and power spectrum images that they see in their textbooks and journal articles. However, the more meaningful practical application may lay in ear training and functional listening. One should be able to predict and discuss the sound (not just the phoneme) captured in those images with reasonably high accuracy. One should also be able to understand, label, and predict obligate changes in a voice as the singer changes pitch, vowel, intensity, and registration. What will follow is broadly based on the principle at the core of the preceding pages: The ear coequally determines the sound of a singing voice. The act of sound perception takes place in a physical world by means of physical systems that follow rules. By understanding these rules, we may peer behind the curtain of that most ephemeral aspect of the singing voice, 

the sound.