VOCAL PEDAGOGY AND PEDAGOGICAL VOICES

Gregory H. Wakefield

Electrical Engineering and Computer Science, College of Engineering
Otorhinololaryngology, School of Medicine
Performing Arts Technology, School of Music
The University of Michigan
ghw@umich.edu

ABSTRACT

Singers learn from their teachers lessons that to the outsider are not transparently obvious. Some of these lessons are discussed in the paper, and their application to problems in sound quality, music information retrieval, and the modeling of the singing voice are presented.

1. INTRODUCTION

Bel Canto is Italian for "beautiful singing". The present paper considers vocal training in the Bel Canto tradition, in part, because it has, for over 200 years, been the dominant form of vocal training in Western classical music[1][1]. The Bel Canto tradition has also been the subject, in recent years, of textbooks that attempt to codify, or at least document, the teaching techniques which have been handed down from master to student (e.g., [2-6]. Finally, Bel Canto, more than other singing tradition, is receiving the greatest attention among voice scientists and otolaryngologists with respect to the training and care of the professional voice (e.g., [7-10].

The story is told about a young singer who was taken under the wing by a famous Italian voice teacher. Every day, the student would practice his vocal exercises, following exactly the instructions of his teacher. Every week, the student would appear at the appointed time for his lesson with the teacher. During the first few minutes of the lesson, the voice would be warmed up after which the teacher would determine how well the exercises had been mastered. By the end of each lesson, some of the past week's exercises would have been dropped, others modified, and new exercises would have been added. This process continued for five years. Then, the day came when the student appeared at the appointed time for his lesson, but, the teacher, rather than going to the piano to begin the warm-up exercises, approached the student with an open hand. "Our work is now completed," he told the student, and with the firm shake of his hand, he said, "now, go out and sing."

If you include the time it takes to grow the tree, building a voice is a bargain compared with building a violin. However, one cannot go out and purchase a Caruso or a Callas, as one can, at least in principle, purchase a Stradivarius or a Guaneri. By definition, a great voice lasts, at most, a lifetime; in practice, a lifetime is a gross exaggeration. Indeed, the parallels between the elite athlete and the elite singer run deeper than those between other elite instrumentalists and those who choose the voice as their instrument. Such parallels begin with the typical span of one's career, one to two decades, and end with the observation that with proper training, exercise, and general care of the body, it is possible to add another decade or two [3, 4].

A luthier works the wood, initially at a rapid pace in very coarse fashion, and then at a slower pace using a variety of sanding materials. The builder's goal is to shape the wood so that its complete range of resonances is realized. Simple formulas, such as the shapes of the face plate and f-holes, guide the early stages of string building. The work slows in the final stages as no acoustic theory exists to guide the sanding required to realize the resonances a given piece of wood supports [11, 12].

An expert luthier is likely to create a dozen or more instruments per year. An expert voice teacher is likely to require four years or more in creating a single instrument, although the time spent per week (one hour) is significantly less for the voice "builder" than for the luthier. As is the case with violin making, simple rules dictate the coarse adjustments of the many mechanisms involved in sound production, although only in the past twenty years has the scientific basis of a few of these rules been established [9, 10]. Similarly, finer adjustments are made by the voice teacher based on experience and on their familiarity with the "grains" of the specific voice.

In speaking with voice teachers and luthiers about their craft, both emphasize the importance of listening, closely and carefully, to the instrument's properties as they build it. Both have standard ways for querying the instrument. For the violinist, it is the tap response at various locations along the face plate as he or she sands. For the voice teacher, it is a set of vocalise, exercises for the voice. For both, in the final stages of shaping the instrument, the process is one closer to the scientific method than artistic interpretation. Hypotheses are generated, tested, and discarded as the instrument is successively refined. What they

1. Others will rightly point out that the German, French, and English traditions stand alongside the Italian as methods for training the classical voice. For our purposes, we consider the differences among these traditions small when any of them are compared with the types of vocal production found in musical theater, Chinese opera, jazz, or country-western. Furthermore, the dominance of the Italianate tradition in contemporary music conservatories justifies the emphasis on Bel Canto in our discussions.

listen for as they iterate through the process remains an open question.

Our own work on voice began in 1999 with the vocal analogue to the perennial question, "What makes a Stradivarius, a Stradivarius"? What is it that allows us to recognize Caruso, Schwarzkopf, Fischer-Dieskau, or Hotter as different from other tenors, sopranos, mezzo's baritones, or bass-baritones? In building the voice, what is it that the "biological Stradivarius's" of the world of vocal pedagogues listen for? Like all great questions, the answer four years later remains unknown, but we have answered several questions of smaller scope, and have raised several more.

2. SINGING AS A SEQUENCE OF PITCHES

Singing is certainly about words, but first and foremost, it is about the proper sequencing of pitches over time. As such, singing can be used to query other listeners about the name of a piece of music. The written phrase "da-da-da daaaah" doesn't convey all that the sung version of the opening motif to the first movement of Beethoven's 5th Symphony does. Upon hearing the sung version, listeners familiar with the music will recognize the sequence of pitches as referring to Beethoven's 5th. Neither the melody nor the rhythm need be exact; listeners will still associate the sung passage with the desired target despite a highly erroneous query.

Within the music library community, such queries "by humming", as they are known, are sufficiently common that a number of indexes have been built around them over the past 75 years. In recent years, computer-based systems for automatically searching a database of melodies have been proposed, e.g., [13, 14] in which query-by-humming serves as a standard user interface. A similar system has been under development at the University of Michigan [15], which is designed for web-based searches of large music databases.

There are several problems with the use of singing as a query. Singers, in general, do not keep a steady pitch. This problem is not limited to novice or amateur singers. On even a good day, a professional chamber choir may find itself slowly going flat, as an ensemble, over the course of an a capella performance. Anecdotal evidence suggests that descending intervals, particularly half-steps, tend to be stretched, so that a choir, intent on keeping in tune, will have their pitch pulled down by the stretched descending intervals of any one vocal part. Such transpositional errors where the key signature changes over time are common among novice or amateur singers.

In addition, singers may not always sing the correct pitch. The vocal folds bear little resemblance to the action of a piano, for example, so that it isn't a matter of one particular pitch being repeatedly out of tune. Rather, singing is more like playing a single-string violin where the performer cannot rely on cross-string relationships as landmarks to help keep pitch intervals in tune. Worst yet, there is no vocal equivalent to an "open string" so there is no chance to return to proper concert pitch and re-tune.

Besides errors of transposition and pitch interval, singers commonly produce rhythmic errors and insertion/deletion errors, where entire musical notes are either added or eliminated from the sequence. These errors must be considered when designing the algorithms that extract the pitch-duration sequence as well as the algorithms that score the pitch-duration sequence against items stored in the database. Finally, these errors suggest the need for a means whereby the errors might be corrected by the singer. Imagine a search engine for a database of textual material that does not allow the user to correct their typing errors. How this need can be met, particularly when the user is often illiterate to musical notation, remains a challenge.

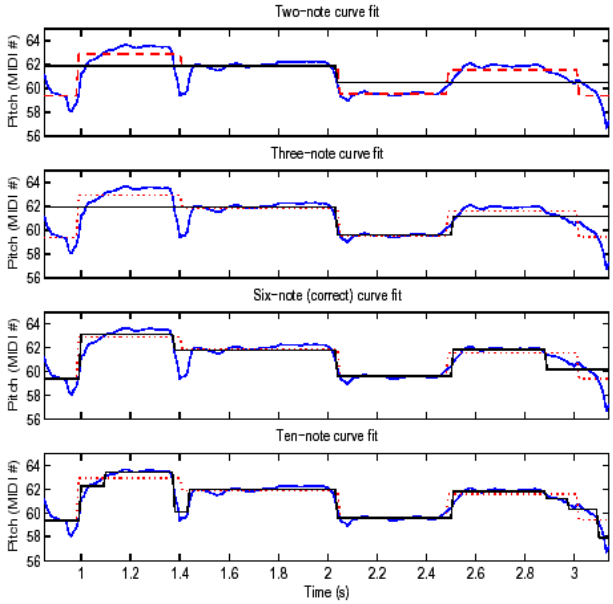

Figure 1. *Extracted pitch contours for a sample passage from "Rock-a-Bye Baby" with manual (dashed) and automatic segmentation (solid). The automatic segmentation algorithm was based on a dynamic programming solution for finding the best fitting N piecewise constant segments to the data (from [16]).*

Even in the absence of user error, there are features of singing that affect the accuracy by which the acoustic signal can be segmented into individual pitch-duration events. Fig. 1 shows various segmentation outcomes for a pitch track from three seconds of the folk song, "Rock-a-Bye Baby". The pitch track, which was obtained using a standard algorithm [17], shows properties that are typical of singers - the pitch is rarely steady, transitions between pitches are often in the form of glides, and these transitions often exhibit overshoot and undershoot behavior. The dashed lines in each panel show the manual segmentation of the pitch sequence, while the solid lines show the best-fitting piecewise constant functions using dynamic programming [16] where the number of notes is varied across panels.

As the example shows, the transformation of pitch contour into a sequence of constant pitches is not obvious. The "rules" used by the listener in manually segmenting the pitch contour are not readily deduced, and differ across listeners. Because of the hard-threshold nature of segmentation, insertion and deletion errors are likely, as are errors in rhythm, due to improper placement of onset or offset boundaries.

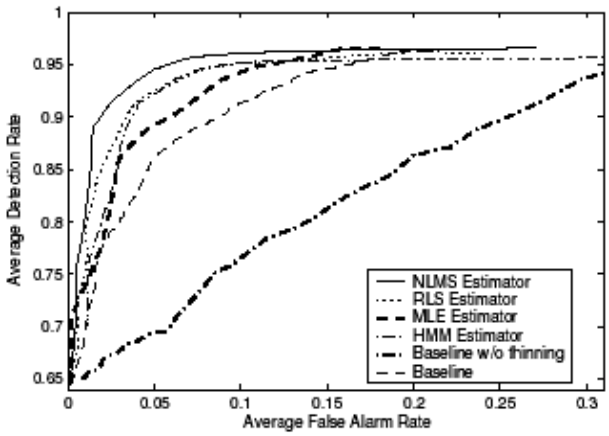

Figure 2. *ROC curves for different segmentation algorithms (from [18]).*

The situation is only made worse when a sample is drawn from a well-trained singer. The vibrato that is often present in the trained voice is typically a semitone in frequency excursion at a rate of 6 Hz. This feature further exacerbates precise determination of transitions between notes as well as accurate estimates of the pitch of each note.

Because of the degraded nature of the pitch contour, many music information retrieval (MIR) systems have used a coarse quantization of the pitch contour to represent the pitch sequence for database search. The idea is to replace an error-corrupted sequence of 12 possible pitches/octave with an errorless sequence of substantially fewer pitches/octave so that the queries are always error-free.

In our own studies, we have shown that this strategy does not necessarily improve the accuracy of the MIR system. Fig. 2 shows receiver operating characteristics for several candidate segmentation algorithms measured over a small corpus of sung melodies. These algorithms include as baseline one drawn from the MELDEX system[19] [14] based on thresholding the derivative of the pitch contour. The remaining estimators are drawn from the standard signal processing literature for their application to "change detection": a nonlinear LMS (perceptron), a Recursive Least Squares estimator, a Hidden Markov Model estimator, and a Maximum Likelihood Estimator.

As can be seen, the parameters of each segmentation algorithm can be adjusted to balance hits (proper detection of a segment boundary) against false alarms (insertion of a segment boundary when one is not present). In addition, by using an estimator structure that is based on statistical models of the signal and "corrupting noise", segmentation performance is substantially improved over the baseline technique.

Classification results are shown in Fig. 3 when the segmented queries are compared with the database of melodies. In this case, both the queries and the melodies in the database are quantized to a fixed number of pitches per octave. An "edit distance" (see [20]) is then created between the quantized query and each member of the database. This edit distance is designed to produce a low-complexity approximation to the type of distance metric produced by dynamic time warping.

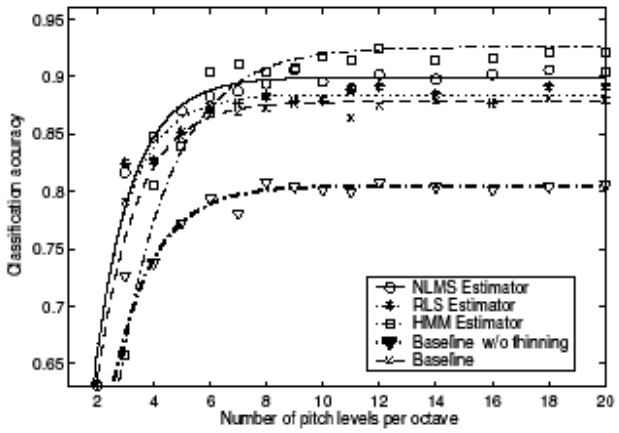

Figure 3. Classification accuracy for four note segmenters as a function of the resolution of the uniform pitch quantizer (from [18]).

From these results, it can be seen that performance differences exist among the different methods and the number of quantization levels. In particular, the baseline method performs the poorest, but also reaches its maximum performance with relatively coarse quantization (6 notes/octave). In contrast, the HMM segmenter is the best performer, even though it isn't the best in the ROC analysis. It also appears to improve relatively little for finer quantization than 6 notes/octave. Nevertheless, the differences among the signal-processing techniques are small when compared with the baseline performance. Furthermore, the ability to systematically trade false-alarm and misses on the front end is impacting our design of classifiers on the back end.

3. VOCAL RANGE

If the baseball-park soloist is lucky, the "Star Spangled Banner" has been pitched appropriately for their vocal range. For those in the stadium trying to sing along, few are likely to find they can sing without abruptly lowering or raising their pitch by an octave. Even the octave leap in "Happy Birthday" is enough of a challenge to prove a useful screening device when auditioning younger singers for choir. Thus, while we sing sequences of pitches, not everyone can sing all the pitch sequences that can be sung.

Each person has a range of pitch over which they are most comfortable singing. From the standpoint of audition, a two octave range is very small when compared to hearing over a frequency range of 20 to 20 kHz. Yet, from the standpoint of singing, two octaves may very well put you on stage. Of course, it is a long way from generating fundamentals over a two octave range to becoming an elite singer, but most professional classical singers perform music within a two-octave range of pitch. For these singers, the vocal range is likely to extend to three octaves or more, at least with respect to phonation. Such singers choose to concentrate their training over the smaller, two-octave range, but numerous anecdotes exist, including the one of Caruso, in which they demonstrate their broader range. During a performance at the Metropolitan Opera, the baritone was having

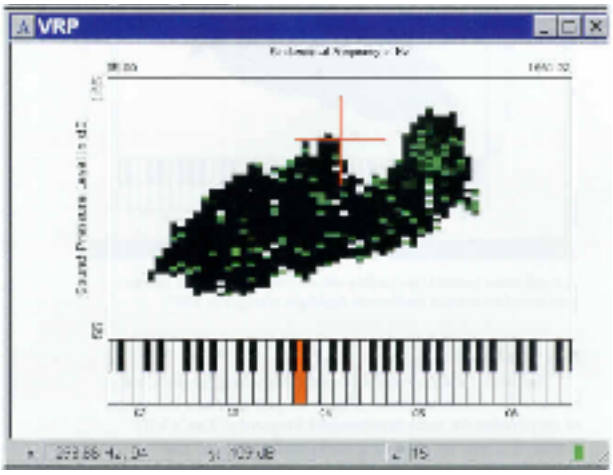

Figure 4. *A sample of the Voice Range Profile. or phonetogram, for a singer using a commerical software package. The shaded region indicates operating points in the space of all fundamental frequencies and sound pressure levels for the singer.*

a particularly difficult day. When it came time for the baritone's aria, Caruso sang it instead.

The dynamic range of the voice also varies across individuals. Within a given singer, the acoustic power radiated by the voice increases with increasing fundamental frequency for constant input gain because of the natural relationship between the generation mechanism and frequency [10]. Across singers, maximum acoustic output differs tremendously and may exceed 105 dB SPL in some cases.

The acoustic power of the voice also depends upon the efficiency of the vocal tract as a resonator. Because of the forward-placement of the vowels, Italian and German are better suited than English, French, or Russian, for example, for creating an oral cavity that approximates an exponential horn [10]. It shouldn't be surprising, therefore, that the experimental art and craft of vocal pedagogy, in which the demand for a singer to be heard in large acoustic spaces, has flourished for the past three centuries in Italy and Germany. For the native-English speaker, shaping the oral cavity "as if you had an egg in your mouth" is a very awkward posture from which to sing; for native-Italian or German speakers, such a shape describes the natural placement for speech.

Early attempts [21] to objectively quantify a singer's vocal range gave rise to the phonetogram or Voice Range Profile (VRP) [22, 23]. The VRP plots the region of the fundamental-frequency/ intensity space over which a speaker or singer can phonate. Fig. 4 shows an example of the VRP from a commercially-available software package [24]. The bounded region represents the singer's operating range, and various statistics have been proposed for quantifying the region and comparing across singers. From a pedagogical perspective, the VRP is reminiscent of a classic Italian vocalise known as *messa di voce,*[1] which is intended to aid the singer in mastering the nuances of their instrument [26, 27].

Over the past thirty years, the VRP has become a well-used tool, both for clinicians and voice scientists, in characterizing the

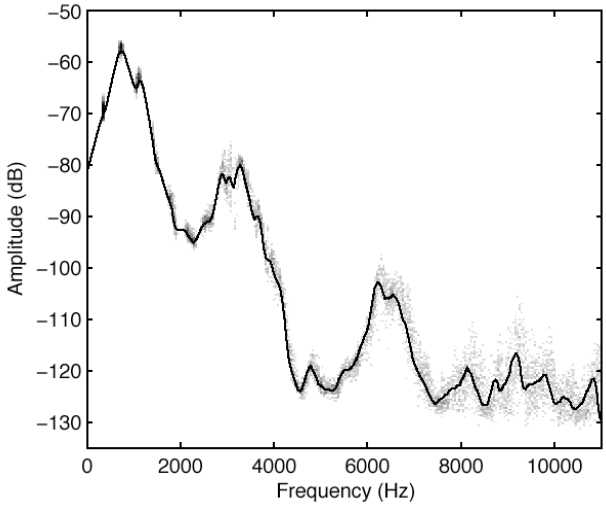

Figure 5. *The Composite Transfer Function (CTF) for a 1.5 second sample of the sung vowel /a/. The CTF (solid line) lumps the acoustics of the vocal tract with the spectral envelope of the glottal source. The CTF is obtained from the instantaneous amplitudes and frequencies (dots) of the waveform by minimizing a regularized cost function (from [36]).*

normal, abnormal, and highly-trained voice [28-32]. It has been used to document male/female and untrained/trained differences among singers, and has become one of the standard battery of measurement tools for the clinician.

4. SINGER'S FORMANT

Intensity alone does not guarantee that a voice can be heard over an orchestra, despite the dynamic range suggested by the VRP. In his 1977 paper, Sundberg compared the long-term average spectrum (LTAS) of an orchestra and a tenor [33]. The spectrum of the tenor essentially follows the characteristic rolloff of the orchestra's for frequencies up to 2000 Hz and beyond 3500 Hz. However, between 2000 and 3500 Hz, the tenor's spectrum lies significantly above that of the orchestra's. Part of this difference can be accounted for by the second formant of the voice, which can be as high as 2300 Hz for [i] vowels. The remaining source of difference is what is known as the *singer's formant*.

While formants are a well-defined property of the vocal tract [34], estimating their center frequencies and bandwidths is a standard problem in system identification for which "sufficiently

1. In *messa di voce*, the singer slowly crescendos from a very soft to a very loud level and then decrescendos back to a very soft level. Throughout the exercise, which is intended to last between 10 and 20 seconds, the singer must hold pitch and vowel constant. *Messa di voce* has a long and illustrious history, with one early pedagogue advocating that no further exercise be considered until the *messa di voce* is mastered, that is, until there are no shifts in pitch or breaks in the voice over the entire pitch range for the singer [25].

rich" driving functions are required. In particular, the harmonic glottal sources found in voiced speech and singing become increasingly *less rich* with increasing fundamental frequency. This is a well-known problem in automatic speech recognition systems where higher-pitched voices are difficult to transcribe. Given that the typical singer speaks in the lower 1/3 of their pitch range, the problem of formant identification is only exacerbated when going from speech to singing. For example, the tenor's high "A" generates a "picket-fence" spectrum with samples every 440 Hz. Using such evidence limits the accuracy by which one can measure the width and location of resonances. However, 440 Hz is well within the middle of the pitch range for female voices. The soprano's A at 880 Hz degrades the accuracy even further. At the opposite end, decreasing the fundamental increases the density of the harmonic samples and thereby improves accuracy.

Because of the role that fundamental frequency plays in estimating properties of formants, the singer's formant has been studied most extensively in male voices, where its center frequency is found to lie in the range of 2500 to 3200 Hz. Evidence suggests that center frequency of the formant is related to vocal category with basses having the lowest center frequency, on average, and tenors having the highest. Fant noted that the first two formants vary as a function of vowel, and that the third through fifth formants are more specific to the individual [34]. The location of the formant and its emergence as an outcome of vocal training suggests that it is an additional resonance to those normally found in the speaking voice, rather than a reinforcement of the third formant.

Our work has concentrated on characterizing sing-er-specific features of the formants of the female voice [35]. In [36], a regularized estimate of the composite transfer function (CTF) (see [37]) is proposed based on measures of the instantaneous frequency and amplitude of the partials obtained from the modal time-frequency distribution [38]. We take specific advantage of vibrato to track the local variations in instantaneous amplitude and frequency of each partial. In general, as these partials vary in frequency over time, their change in amplitude reflects, in part, the resonance properties of the vocal tract. In addition to the effects of vocal tract transmission, there are short-time changes in amplitude due to variations in the gain of the glottal source which must be regressed out of the vocal tract estimates. Finally, as we are interested in characterizing the types of adjustments singers make in the vocal tract as a function of pitch, e.g., vowel adjustment and neutralization, for each isolated pitch, there is likely to be bands of frequency for which we have no amplitude estimates. A smoothness constraint is therefore imposed on the shape of the CTF to allow us to interpolate over regions of missing data.

An example of the CTF for a sung /a/ is shown in Fig. 5. Each dot in the figure is the instantaneous amplitude and frequency of one of the partials from the 1.5 second sample. Regions where no data exist are smoothed according to the regularizer for frequency variation in the CTF. In general, no adjustment is made for transmission loss at the lips, so the overall gain of the CTF rolls off more rapidly than is usually reported for speech spectra.

Average CTFs for six sopranos are shown in Fig. 6. As can be seen, there is evidence of a broad, weak resonance in the soprano

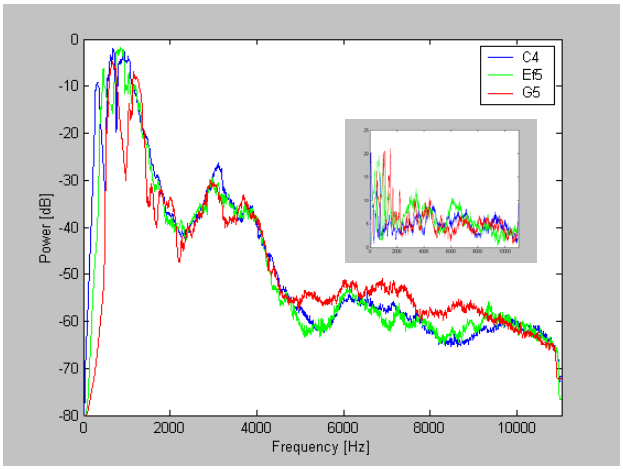

Figure 6. *Average CTFs for six sopranos are shown. The parameter is the pitch at which each CTF was obtained. The inset shows the standard deviation of the data. Each partition is 10 dB.*

data between 3000 and 4000 Hz. From these results, along with those collected by other researchers using either LTAS [39] or a partial-tracking methodology similar to ours [40], we see that the strong narrowband resonance associated with the singer's formant in males appears to be absent in females. In its place appears to be a diffuse resonance which many attribute to a coalescing of formants F3-F5 to create something akin to a singer's formant.

Before we conclude that a singer's formant is absent in the female voice, however, it is important to consider the data shown in the inset of Fig. 6. The inset shows the standard deviation of the data as a function of frequency. As can be seen (with the help of Acrobat's zoom function), the deviations across the soprano CTFs within the 3000-4000 Hz range is on the same order as the magnitude of the resonance. This would be the case if each individual soprano showed a narrower resonance which varied in location as a function of singer.

Fig. 7 shows a set of CTFs as a function of pitch for the vowel /o/ sung by one of the sopranos in our study. The very narrow ridge in the neighborhood of 4 kHz is similar in structure to what is observed in the male voice at much lower frequencies. Similar ridges are observed for all of the singers in the study, but at different frequencies. In addition, the location of the ridge is observed to vary up to several hundred Hz as a function of pitch and also changes slightly with respect to vowel. These three sources of variation are responsible for the smooth, weak resonance observed in the averaged data.

Whether the mechanism of the singer's formant is the same for males as for females cannot be addressed by the present data. However, the fact that female voices also exhibit narrowband resonances is consistent with the perceptual correlate of the singer's formant known as "ring". Vocal pedagogues listen for ring, a quality of vocal brilliance, as they work with the singer on shaping their instrument.

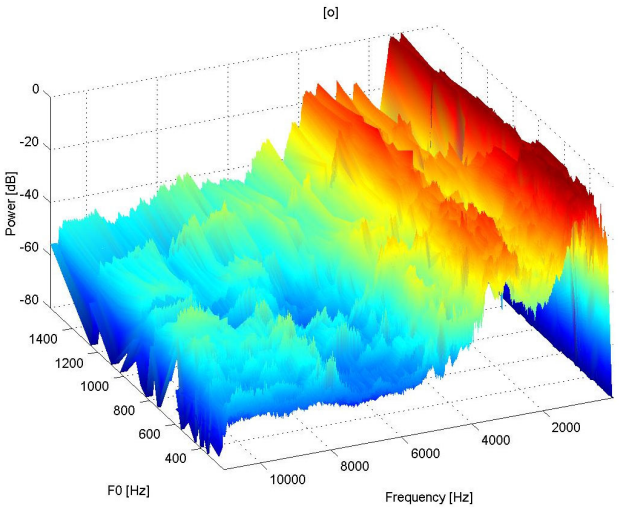

Figure 7. *Composite Transfer Functions are shown for the vowel /o/ as a function of pitch for a particular soprano. Note the very narrow resonant band in the neighborhood of 4 kHz, which in location over several hundred Hz as a function of pitch. This display is termed the singerScape, to distinguish it from other surface visualizations, such as the spectrogram, which represents the spectrum of the signal.*

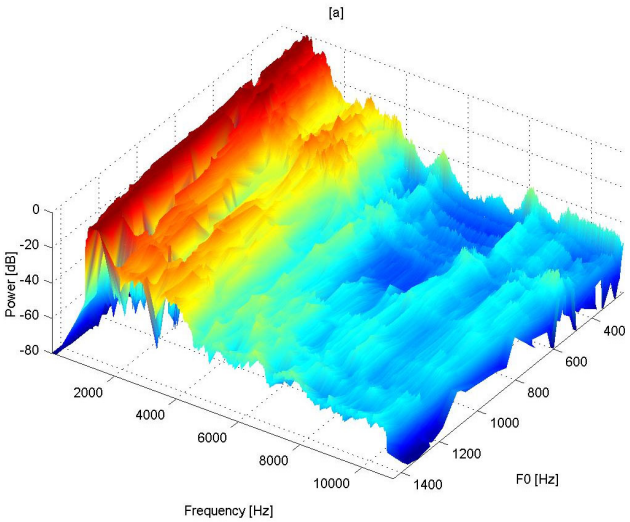

Figure 8. *A soprano's singerScape for the vowel /a/ over a 2.5 octave range. The narrow ridge in the neighborhood of 4 kHz is broken for fundamentals between 600 and 800 Hz, which corresponds to this singer's passaggio.*

Our longitudinal evidence suggests that such training builds up the ridges, where otherwise there are very weak resonances. An example is shown in Fig. 8 for a soprano. The formant ridge in the neighborhood of 4 kHz is broken between 600 and 800 Hz when vocalizing on the vowel /a/. This region is one of the *passaggi* in her voice, where there are relatively abrupt shifts in the vocal mechanisms used to control and support phonation. Examining the surfaces for other vowels from this singer shows a smoothly varying ridge throughout the passaggio region (see Fig. 7, for example). At the time the data were gathered, the /a/ was the last remaining vowel to be shaped.

5. VIBRATO

Ask someone to identify one feature about a classical singer's voice, and they will likely focus on the singer's vibrato. Yet, it is an interesting fact that vibrato is not directly trained into the voice. There are no vibrato vocalise, as there are exercises for flexibility, dynamic control, and pitch range. Rather, many voice teachers will indicate that vibrato is a natural outcome of "opening up the voice". It is also an interesting fact that the mechanism responsible for vibrato has yet to be identified. A likely candidate appears to be entrainment of a local neural circuit that controls vocal fold tension, but this has yet to be fully investigated.

The presence and type of vibrato is used by the teacher as an indicator of whether the vocal apparatus is properly engaged. A vibrato that is too slow (significantly less than 5 Hz) or too fast (significantly greater than 6 Hz) is often indicative of improper breath support. The so-called warblers (slow vibrato) and bleaters (fast vibrato) are encouraged to sing on a straight tone for a period of time while appropriate adjustments in breath support are made. As vocal range is stretched, the initial sessions will exhibit little vibrato in the singer's voice. As the singer learns to balance the various mechanisms, vibrato emerges as part of the process. This emerges guides the balancing process.

We have investigated the extent to which vibrato varies as a function of pitch in a sample of female voices [37]. Fig. 9 shows a typical instantaneous fundamental frequency for one of the singers in our study. The horizontal line shows the nominal frequency that corresponds to the perceived pitch. The vibrato signal is shown by the function with + symbols at local maxima. Finally, the solid line shows the best-fitting constant-amplitude sinusoid to the data.

As is evident from the data, the regularity of the vibrato rate is very high. This is typical of vibrato production observed in our singers. Over short periods of time, vibrato rate behaves as a well-tuned oscillator. As time extends, small changes in vibrato rate are observed to occur. In contrast, the regularity of the vibrato excursion is low. The maxima deviate significantly from a constant. This behavior does not become more extreme with time, but appears to be a part of the ongoing production of vibrato.

Besides the gross features of the vibrato signal, there are fine details that appear in the multi-phase structures found particularly around the local maxima. These do not appear to be entirely due to artifacts of the estimator. Though a feature of the signal, they do not appear to be a regular "signature" within or across singers. In our initial analysis, we determined that signals synthesized with and without these features were not perceptually discriminable, so, we adopted two pairs of measures, vibrato rate and its variability and vibrato excursion and its variability, as a means to summarize the data.

Fig. 10 shows results from two of these measures, the average vibrato rate and vibrato excursion as a function of pitch,

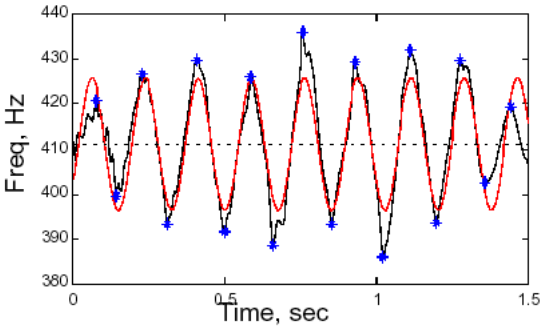

Figure 9. *A sample of the fundamental frequency as a function of time for a particular sung vowel along with a best-fitting constant-amplitude sinusoid. Local maxima (+) are used to determine the relative stabililty of the vibrato excursions.*

for the twelve singers in our study. As can be seen, no single trend holds for all singers. S07 exhibits a downward trend in both variables with increasing pitch, in contrast to M03, who exhibits an upward trend.

Subsequent observations suggest that these data are more reflective of the general variability found in vibrato, rather than a strong functional dependence of vibrato rate and excursion on pitch. There is some argument in the literature that a stable vibrato is indicative of a well-trained classical singer. What stable means, however, in another matter. In our own studies, we observe that stability, in all likelihood, is characterized by a range of values and that listeners are able to discriminate among vibrato conditions over this range.

6. SINGER IDENTITY[1]

The previous sections have considered a variety of properties of the voice, from its ability to accurately sing a sequence of pitches, to more individual features, such as the range of vocal production, the perceptual attribute of ring, or vibrato. Can Caruso be defined by a combination of these features, or is the singer's identity more than this?

In a series of studies, we have looked more closely at the question of singer identity. Our latest work utilizes the CTF representation described in Section 4. A standard quadratic classifier was used for classification with each of the twelve singers treated as a separate class. The classifier was trained by computing the sample mean μ_k and sample covariance Σ_k of the training instances that belong to the kth singer for all k. The classifier assigns a class label c_x to an instance with feature vector x using the formula

$$c_x = \underset{x}{\operatorname{argmin}} (x-\mu_k)^t \Sigma_k^{-1} (x-\mu_k) \qquad (1)$$

Principal component analysis is performed on the estimates of the CTF (in dB) to reduce the dimensionality of the feature space and the first P principal components are selected as features for classification.

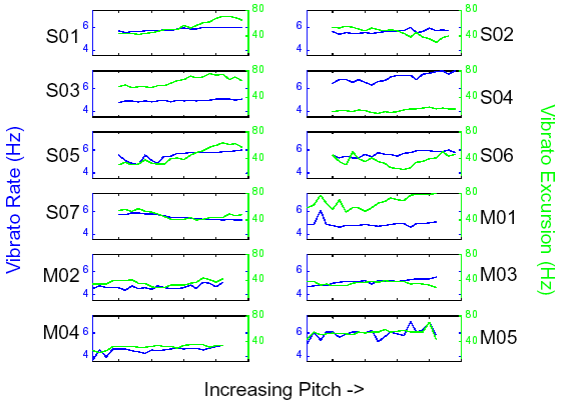

Figure 10. *The dependence of vibrato rate (left scale) and vibrato excursion (right scale) is shown as a function of pitch for twelve female singers (7 sopranos and 5 mezzos) in the Michigan study.*

Because of the well-known dependence of the spectral shape on the vowel being sung, we divide the data into five subsets, one for each vowel, to determine the extent to which classifier performance depends on vowel. The training-set proportion was set to 0.5. Fig. 11 shows the fraction of correctly classified elements in the test set as a function of the dimensionality of the feature space, P. From this figure, we note that the performance is nonmonotonic in P with a broad maximum between 10 and 20. For $P = 16$, which roughly corresponds to the maximum for this data set, the classifier achieves test set accuracies of 94% to 97% depending upon the vowel. The training set accuracy for this value of P is greater than 99.5% for all vowels. Certain vowels in this data set appear to be more easily distinguished than others. Specifically, /a/ systematically yields the highest accuracy, while /u/ yields the lowest accuracy[2].

A number of characterizations of the classifier are reported in [36]. In general, picking a fixed dimensionality for all vowels and singers yields classification scores of 95% for 32 features based on training with as little as 25% of the data. Fig. 12 shows the practical case when the classifier trained on the entire data from the five-note exercises is used to classify on the basis of the singer's arias. In this case, each aria was segmented for "stable" notes, e.g., pitches which were sustained for a sufficiently long period to time to yield stable estimates of the CTF. The CTFs were subsequently projected onto the feature space and classification performed. The columns of confusion matrix shown in Fig. 12 indicate the correct classification while the rows indicate the classifier's selection. Overall test-set accuracy in this task is 71%. Of the twelve singers, m04 and s06 are correctly classified the most often with accuracies of 100% and 96.4%,

1. Portions of the text in this section also appeared in [41].

2. The drop in accuracy for high values of P appears to be a result of classification using ill-conditioned class covariance matrices. We note that this occurs at a much lower dimensionality for /u/ than for other vowels.

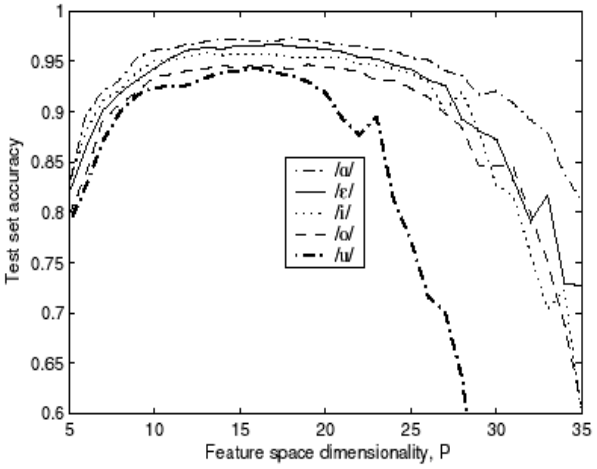

Figure 11. *Test set accuracy for a classifier as a function*

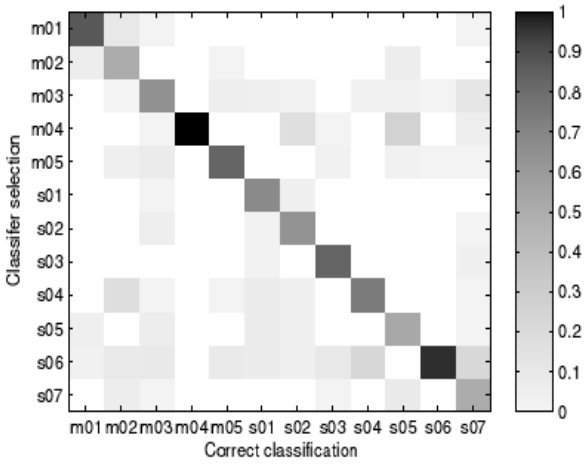

Figure 12. *The confusion matrix for a classifier trained on the vocal exercise data and tested on arias.*

respectively. The most misclassified singers are m02, s05 and s07 with accuracies of 51%, 53%, and 51% respectively. While many of the singers are misclassified quite often, we note that none of the singers are misclassified more than half of the time. Thus, for this small selection of singers, all twelve arias would be properly classified by a classifier which selects the majority of each individual note classification.

These classification results suggest that there are consistent differences among singers in a 32-dimensional parameter space such that a classifier trained on a singer's vocal exercise may correctly identify the singer of an aria. Further experiments shed light on the nature of "identity" in this 32-dimensional space. For example, training on a reduced band of frequencies, say 1-5 kHz, degrades the performance of the classifier. Thus, formants well above the fifth are useful in creating a statistical description of the singer. The same is true if one trains on a subset of the vowels. More surprisingly, perhaps, is that the same is also true if one trains on a smaller set of pitches. Evidently, from a statistical standpoint, how a singer shapes their vocal tract to handle their middle range of pitches is not redundant with how they handle their upper range.

Such observations clearly fly in the face of our everyday experience. Caruso, for example, sounds like Caruso whatever pitch or vowel he happens to be singing. Few would ever mistake Caruso for Domingo or Melchior for Gedda, regardless of what aria they happen to be singing. However, everyday experience may also be misleading.

In [42], we reported on the results of a perceptual version of the classification task above. Assuming that "Caruso is Caruso", each of the twelve singers was represented by one of their 5-note vocalise. The listener's task was to classify each of the remaining 5-note vocalise samples with the appropriate singer. Of four subjects, only one scored substantially above chance in the experiment. Three of the four subjects had never heard the singers before. The fourth subject, who was able to classify the singers, was also the person who had collected and processed each singer's data. When the experiment was repeated by substituting an aria from each singer as their identifier, each of the remaining three subjects improved in their ability to classify the singer.

7. VOCAL PEDAGOGY AND PEDAGOGICAL VOICES

7.1. Core

The failure to perceptually generalize a singer beyond a half octave of their vocal range was a surprise to us, as it was for Erickson, who arrived at similar conclusions using a different experimental paradigm [43]. From a personal perspective, it raised a question about the practice of vocal pedagogy. After all, if listeners can't generalize beyond a half-octave, how does the voice teacher recognize the "right" voice as he or she coaches the tenor student, for example, in the development of their head voice?

I raised the question with George Shirley during a voice lesson shortly after we had completed the human classification studies. "Core," he replied. "I listen for your core and then make sure you don't deviate from that wherever and whatever you sing." This response prompted a series of studies, using samples of my own voice obtained during voice lessons, to find the acoustic attributes of core.

In [44], I confessed that the acoustical differences between instances of an open-throat G4 deemed "precisely so" by my teacher and those not worthy of producing again were on the order of measurement/modeling error, despite the fact that as I listened to the exercise, I could hear what I took to be more "core" in the better instances. A year of study later, I came to a better understanding about "core". The hint was that my production of G4 was never produced in isolation, as if this were a key on a piano or position on a cello fingerboard. Vocalise exercises are never about pitches in isolation, but instead, the exercises are always designed in the context of other pitches and/or dynamic levels. The admonition to the singer is always "sing it legato", that is, sing it so that the transitions are as smooth as possible.

The recognition that G4 was never judged in isolation led to thinking of each vocal exercise not as a set of isolated composite transfer functions but as a surface. In [45], we introduced the singerScape as such a visualization and observed that "precisely so" surfaces were those with smooth variations as pitch or dynamic level varied. Fulfilling the goal laid out in our talk the preceding year, we demonstrated how the "present sound of one's future voice" could be created by altering local discontinuities in the CTF surface structure and synthesizing the result. In the case of the "passaggio gap" shown in the singerScape of Fig. 8, not only did appropriate surface interpolation yield a "nicer looking picture", but it created an acoustic target for that soprano which, to the ears of a number of voice teachers, preserved what they would label as her core.

As impersonators clearly demonstrate, the human voice is remarkably flexible. For the singer and their teacher, "legato" appears to be the practice by which "core" is preserved across pitch, vowel, and dynamic level. In my own experience, what appears to be "core" has changed as different parts of my voice have been engaged. As choices are made for handling the mixture of head and chest voice, breath support, and resonance, the student and their teacher re-visit the entire range of vocal production and use new found possibilities to guide the re-shaping of older, more time-worn habits. My "bass-baritone" naive approach to singing, with its darker and thicker edges, has been replaced by a more brilliant focused tenor approach which creates a smoothly varying surface over a much wider operating range. This is likely to be what is lacking in grosser measures, such as the Vocal Range Profile, in coming to understand what changes occur under vocal training.

7.1. Your job is to sing, my job is to listen

Voice lessons are a scientific process of hypothesis testing by which you and your teacher successively approximate a well-blended instrument. The role of feedback is crucial to this process as is the role of repetition. The phrase "precisely so" has become to me more rewarding than getting an "A" on a quiz, but the follow-up "now, do it again" reminds me that singing is not about outcome as much as it is about the actions that lead to the outcome. For someone trained to listen, breaking this particular feedback loop has been hard. You would think that listening when you sing is everything; it has little to do with singing at all.

Before I develop this point, however, let me clarify that listening obviously plays an important role in singing. If you don't listen to the musicians around you, you may be off-pitch, late in your entry, or delayed in your cut-offs. You may fail to be a part of the ensemble, although in many cases as a soloist, you are to stand-out from the ensemble in ways that the rest of the musicians should not.

What you don't do in singing is spend much time *listening to yourself*. It has taken me several years to learn that listening to myself to gauge whether I'm really "doing it again" gets in the way of doing it. Part of this is the nature of the task itself: if you're hearing it, it is way too late in the process to do anything about it. Singing is more like the task of the baseball pitcher than the outfielder. Once the ball is hit, the outfielder can rely on closed-loop control to monitor the position of the ball, their position on the field, and reduce the difference between the two by feedback control. Once the ball is thrown, in contrast, there is nothing the pitcher can do to take the throw back. Pitching is all in the set-up; it is open-loop control and so is singing. This is so, surprisingly enough, even as one traverses a sustained passage of notes: if you fail to launch the sequence with the proper trajectory, it is very difficult to error-correct along the way. I don't play golf, but George Shirley tells me that it is all in the swing, and so is the production of sound.

The de-emphasis of auditory feedback in vocal production is underscored by two standard practices in the teaching process. Before George Shirley says "do it again" and after he says "precisely so", he says, "what did you feel". Introspection appears to be extremely important in the process of learning to sing better, but the self-monitoring deliberately ignores auditory sensation and focuses attention on a variety of proprioceptive cues instead. You feel a particular shape in your mouth, a tingling on the sides of your nose; yes, you even feel as if your forehead is a stove pipe and the sound, for certain pitches and vowels, is shooting straight through it. To verbally coach a student in these proprioceptive cues ("sing as though the sound appears to float about two inches outside your mouth") may not be as productive as encouraging the student to develop their own language for the cues and to learn how to properly set-up the sound for their own bodies.

The development of proprioceptive monitoring/target-matching not only makes sense with respect to what a singer can directly control, but it also makes sense with respect to the daily vagaries of the instrument. String players will adjust their instrument to the needs of the day by applying different types of rosin to the bow. Wind instrumentalists keep their reed-making kits handy and brass players understand when they have no more "lip". For the singer, however, how much sleep you got the night before, what you had for breakfast, whether you've properly hydrated, etc., make a real difference in how well the mechanisms responsible for singing will work. Some days are easier than others; you simply learn to make the proprioceptive adjustments necessary to achieve the proper physiological launching of the sound.

The second practice in teaching is the relative lack of acoustic modeling by the teacher. Rarely will George Shirley ever demonstrate the target I am trying to achieve. Perhaps in the case of a bass coaching a soprano, this makes sense, but even when the teacher and student are in the same general voice category, acoustic modeling is rarely practiced. Part of this is likely to be the fact that we, as singers or speakers, cannot hear what we sound like to the listener. Thus, matching an acoustic target requires going through an intermediate model that compensates for the differences.

A second, and I suspect a much more powerful reason acoustic modeling is not productive is that listening to one's voice is a very unreliable cue. Indeed, one would expect that, in time, a good singer/listener would be able to learn the proper compensations necessary to match an acoustic target. However, in my experience, I find auditory feedback to vary dramatically, depending on the acoustics of the room. Acoustic self-monitoring, even for such gross features as loudness level, is a very poor cue as some rooms and stages support the voice well

and others simply "suck it all up" and provide little "kick-back" as the acousticians call it. Acoustic self-monitoring for adjusting finer features of the voice is likely to be impossible. The reliance on proprioceptive targets, in this case, makes sense. Rooms don't change the experience of setting-up and launching the sound. The cues are always there, you count on them when you need them, and you let the sound take care of itself.

7.1. The Secret Code

The concept of core and the role of proprioceptive feedback were important lessons to learn, and they ran counter to my thinking as an outsider looking in. The third lesson that the study of voice has taught was, for me, the most surprising of all. Well before I began voice lessons, I was curious about what makes a great musical interpretation. To come to a better understanding of the question, I acquired several recordings of a Schubert song cycle and compared what I heard with what was notated in the score. I came to suspect that there must be a secret code, or notational system, that singers use to fill in all that is missing in the score. Each performer, I believed, created valid and breath-taking renditions of the song cycle, by carefully notating each small note-by-note nuance and then practicing that interpretation over and over again.

I carried this belief into my first experience in studio class. Like most students, 60 minutes per week are spent in voice lessons and another 120 minutes are spent in studio class, where all of your teacher's students gather to perform for each other. These performances are typically not just matters of getting up, singing, and sitting down. Rather, each student's performance may last from 15-45 minutes, during which the teacher coaches on matters of technique and interpretation and other students provide feedback.

I was very excited to learn that one of the senior students was going to perform in studio one of the pieces from the Schubert song cycle I had analyzed. I hoped that his score would be passed around for each of us to see and appreciate the secret notational code he used to flesh out what Schubert had written. With all this in mind, I was a little disappointed, therefore, when his performance didn't have that immediate presence I had come to expect from the recordings, and discounted this as an issue of maturity and experience. I suspected there were also flaws in his secret coding of the piece and was relieved when George Shirley, after complimenting the student on their work, said "let's go over some things." Finally, I would get to see the notational system that no one writes about, but, clearly, everyone must use.

"How old are you?" asked George of the student. I was rather surprised by this line of questioning, as it would appear to have little bearing on the notational system that needed improvement. His questions continued. "Do you really love the girl you are singing about? When was the first time you kissed? What did she say after that? Where were you born? Are your parents still living? Do you have brothers and sisters? Can you remember the first time you ever kissed anyone?" The questions were not addressed to the student, but to the character who was singing the song. After answering these questions and several others, George had the student speak the English translation of the German while the pianist played the accompaniment. He asked the student to deliver the lines as someone who was born in a small town outside Munich, who had first kissed a girl on a crisp fall day near the church graveyard after Sunday mass, and who was beginning to doubt that his new love would turn out to be the girl of his dreams. Listening to the English rendition, I had to admit, created a moving performance. At the end of the performance, we were all very quiet. "Now," George said after a moment of silence, "you are ready to sing the lied."

Like Dorothy when she opened the black-and-white door and walked into the color of Oz, the performance that followed had all the presence, legitimacy, and nuance of any of the recordings I owned when compared with the black-and-white version we had heard earlier. Indeed, in many respects, it was better because it was live. And, like Dorothy, the black-and-white secret codes and notational systems fell away and I came to understand that the process works at a much more basic level, using far more subtle and powerful paints.

8. ACKNOWLEDGMENTS

This work was funded, in part, by an NSF ITR and by the Office for the Vice President of Research at the University of Michigan.

My collaborators have been many, including colleagues from the University of Michigan's School of Music: Profs. Shirley, Herseth, and Cheek from the voice faculty, Profs. Blackstone, Morrison, and Snow from the choral conducting faculty, Profs. Simoni and Rush from performing arts technology, Prof. Andre from musicology, as well as the contributions of new genre artists in the School of Art and Design, Michael Rodemer and Stephanie Rowden. In addition, the MUSART research effort in music information retrieval has helped shape much of the more recent work. This team includes Profs. Birmingham and Jagadish (EECS), and Dannenberg (CS from Carnegie Mellon University). Dr. Hogikyan (Otolaryngology) and Prof. Gross (Kinesiology) have been instrumental in moving our work toward the biology of voice, much as Ann Arbor's luthier, Joseph Curtin, has moved us toward thinking of vocal training as the crafting of an instrument where the dividing lines between performer, instrument, sound, and hearing must be deliberately blurred.

Within my own research group, Maureen Mellody, Mark Bartsch, and Norman Adams have shared in the exploration. Rebekah Nye played an important role in recruiting the seven soprano and five mezzo vocal performance majors in the School of Music from whom several hours of recording have led to years of research. Juliet Petrus and Michael Turnblom also contributed substantially not only to our database, but also to our growing list of questions. Finally, a special thanks to George Shirley, who realized long before I did the necessity of studying voice "the old fashion way" if one really wants to study voice scientifically, and took me under his wing.

9. REFERENCES

[1] Stark, J.A., *Bel Canto: A History of Vocal Pedagogy.* 2000, Toronto: University of Toronto Press.

[2] Doscher, B.M., *The Functional Unity of the Singing Voice*. 2 ed. 1994: Scarecrow Press.

[3] Miller, R., *The Structure of Singing: System and Art in Vocal Technique*. 1986, Oxford: Oxford University Press.
[4] Miller, R., *The Art of Singing*. 1996, Oxford: Oxford University Press.
[5] Coffin, B., *Coffin's Overtones of Bel Canto Accompanying Chart*. 1980: Scarecrow Press.
[6] Brown, O.L., *Discover Your Voice: How to Develop Healthy Voice Habits*. 1996: Singular Publishing.
[7] Bunch, M., *Dynamics of the Singing Voice*. 4 ed. 1997, New York: Springer Verlag.
[8] Sataloff, R.T., *Professional Voice: The Science and Art of Clinical Care*. 2nd ed. 1997: Singular Publishing.
[9] Sundberg, J., *The Science of the singing voice*. 1987, DeKalb, IL: Northern Illinois University Press.
[10] Titze, I.R., *Principles of voice production*. 1994, Englewood Cliffs: Prentice Hall.
[11] Curtin, J. personal communication. 1998: Ann Arbor.
[12] Cremer, L., *Physics of the Violin*. 1984, Cambridge: MIT Press.
[13] Ghias, A., et al., *Query by humming: Musical information retrieval in an audio database*. ACM Multimedia, 1995: p. 231-236.
[14] McNab, R.J., et al., *The New Zealand Digital Library MELody inDEX*. 1997, D-Lib Magazine.
[15] Birmingham, W.P., et al. *MUSART: Music Retrieval Via Aural Queries*. in *ISMIR2001: International Symposium on Music Information Retrieval*. 2001. Bloomington, IN.
[16] Adams, N.H., *Automatic Segmentation of Sung Melodies*. 2002, The University of Michigan. p. 1-30.
[17] Boersma, P., *Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound*. Proceedings of the Institute of Phonetic Sciences of the University of Amsterdam, 1993. **17**: p. 97-110.
[18] Adams, N.H., M. Bartsch, and G.H. Wakefield. *Coding of sung queries for music information retrieval*. in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*. 2003. Mohonk, New Paltz, NY.
[19] McNab, R.J., L.A. Smith, and I.H. Witten. *Towards the Digital Music Library: Tune Retrieval from Acoustic Input*. in *ACM Digital Libraries Conference*. 1996. Bethesda, MD.
[20] Durbin, R., et al., *Biological Sequence Analysis*. 2001, Cambridge, UK: Cambridge University Press.
[21] Wolf, S., D. Stanley, and W. Sette, *Quantitative studies on the singing voice*. Journal of the Acoustical Society of America, 1935. **6**: p. 255-66.
[22] Damste, P., *The phonetogram*. Practica Otorhinolaryngol., 1970. **32**: p. 185-187.
[23] Schutte, H., *Over het fonetogram*. Logopedie en Foniatrie, 1975. **47**: p. 82-92.
[24] *Voice Range Profile, Model 4326*, Kay Elemetrics.
[25] Rossini, G., *Gorgheggi e Solfeggi*. 1825, Milan: Ricordi.
[26] Concone, J., *Thirty Daily Exercises*. 1894, New York: Schirmer.
[27] Titze, I.R., et al., *Messa di voce: An investigation of the symmetry of crescendo and decrescendo in a singing exercise*. Journal of the Acoustical Society of America, 1999. **105**: p. 2933-2940.
[28] Gramming, P. and J. Sundberg, *Spectrum factors relevant to phonetogram measurement*. Journal of the Acoustical Society of America, 1988. **83**: p. 2352-60.
[29] Pabon, J. and R. Plomp, *Automatic phonetogram recording supplemented with acoustical voice-quality parameters*. Journal of Speech and Hearing Research, 1988. **31**: p. 710-22.
[30] Sulter, A.M., et al., *A structured approach to voice range profile (phonetogram) analysis*. Journal of Speech and Hearing Research, 1994. **37**: p. 1076-1085.
[31] Sulter, A., H. Schutte, and D. Miller, *Differences in phonetogram features between amela nd female subjects with and without vocal training*. Journal of Voice, 1995. **9**: p. 363-77.
[32] Titze, I.R., *Acoustic interpretation of the voice range profile (phonetogram)*. Journal of Speech and Hearing Research, 1992. **35**: p. 21-34.
[33] Sundberg, J., *The Acoustics of the singing voice*. Scientific American, 1977. **236**: p. 82-91.
[34] Fant, G., *Acoustic Theory of Speech Production*. 1960, The Hague: Mouton.
[35] Wakefield, G.H., M.A. Bartsch, and F. Herseth. *Finding formants, fixed or otherwise*. in *Voice Foundation Symposium*. 2002. Philadelphia.
[36] Bartsch, M.A. and G.H. Wakefield, *Singing voice identification using spectral envelope estimation*. IEEE Transactions on Speech and Audio Processing, 2003: p. submitted for publication.
[37] Mellody, M., F. Herseth, and G.H. Wakefield, *Modal distribution analysis and synthesis of a soprano's sung vowels*. Journal of Voice, 2001. **15**(4): p. 469-482.
[38] Pielemeier, W.J. and G.H. Wakefield, *A high resolution time-frequency representation for musical instrument signals*. Journal of the Acoustical Society of America, 1996. **99**(4): p. 2382-2396.
[39] Sundberg, J., *The singer's formant revisited*. 1995, KTH: Stockholm. p. 83-96.
[40] Weiss, R., J. Brown, and J. Morris, *Singer's formant in sopranos: Fact or fiction?* Journal of Voice, 2001. **15**(4): p. 457-468.
[41] Wakefield, G.H. and M.A. Bartsch. *Where's Caruso? Singer identification by listener and machine*. in *Cambridge University Music Processing Colloquium*. 2003. Cambridge.
[42] Mellody, M. and G.H. Wakefield. *Application of stimulus sample discrimination to the perceptual evaluation of synthesized sounds*. in *International Conference on Auditory Displays*. 2001. Helsinki.
[43] Erickson, M., *Discrimination Functions: Can They Be Used to Classify Singing Voices*, Journal of Voice, 2001. **15**(4): 492-502.
[44] Wakefield, G.H., M. Bartsch, and G.I. Shirley. *On the present sound of one's future voice: Lessons from a case study of the tenor voice*. in *Voice Foundation Symposium*. 2001. Philadelphia.
[45] Wakefield, G.H., et al. *Surveying the SingerScape: Visualization tools for characterizing the singing voice*. in *Voice Foundation Symposium*. 2002. Philadelphia.