Eric J. Humphrey, Sravana Reddy, Prem Seetharaman, Aparna Kumar, Rachel M. Bittner, Andrew Demetriou, Sankalp Gulati, Andreas Jansson, Tristan Jehan, Bernhard Lehner, Anna Kruspe, and Luwei Yang

An Introduction to Signal Processing for Singing-Voice Analysis

High notes in the effort to automate the understanding of vocals in music

Digital Object Identifier 10.1109/MSP.2018.2875133
Date of publication: 24 December 2018

Humans have devised a vast array of musical instruments, but the most prevalent instrument remains the human voice. Thus, techniques for applying audio signal processing methods to the singing voice are receiving much attention as the world continues to move toward music-streaming services and as researchers seek to unlock the deep content understanding necessary to enable personalized listening experiences on a large scale. This article provides an introduction to the topic of singing-voice analysis. It surveys the foundations and state of the art in computational modeling across three main categories of singing: general vocalizations, the musical function of voice, and the singing of lyrics. We aim to establish a starting point for practitioners new to this field and frame near-field opportunities and challenges on the horizon.

Power of the human voice

The human voice dominates nearly all music cultures. The voice, through singing, can function as a musical instrument and at the same time convey semantic meaning. Theory from the field of psychology suggests that people generally find the human voice especially salient and powerful and that the human voice is a meaningful factor, perhaps the most meaningful factor, in affecting our music-listening behavior. Research has suggested that music exists because of the complex system that enables humans to communicate, interpret, and feel emotions via vocal sounds [1]. Given such strong anthropological links between music and voice, it is unsurprising that singing plays a prominent role in modern music culture; karaoke, for example, is a billion-dollar worldwide industry.

Thus, digital signal processing research has long focused on methods and techniques for modeling the human voice. Early progress in efforts to encode and transmit speech for telecommunication systems [2] paved the way for singing-information processing, the study of signal processing techniques on the human voice in musical contexts [3]. Singing information processing can be represented as a cyclic system where, under ideal conditions, an audio signal is transformed, via analysis, into high-level descriptors or symbols, such as pitch or lyrics; rich symbolic information can then be transformed, via synthesis, into audio signals of singing; and, falling between analysis

and synthesis, effects can be applied to either audio or symbolic information by manipulating intermediary representations between the two domains. A popular vocal effect, for example, is that of pitch correction ("autotune"), where a vocal audio signal is analyzed, the estimated pitch over time is quantized to a given key, and the voice signal is resynthesized.

Starting around the end of the 20th century, the field of music information retrieval (MIR) has developed techniques and methods for various applications of singing-information processing. While many researchers have made contributions to this field, the work of two groups in particular stands out: the Music Technology Group (MTG) at Universitat Pompeu Fabra in Spain, under the direction of Xavier Serra, and the National Institute of Advanced Industrial Science and Technology (AIST) of Japan, under the direction of Masataka Goto. Researchers at the MTG have a long history of advancing the state of the art in singing-voice synthesis, resulting in both commercial products and published studies [4]. Meanwhile, the efforts of AIST are noteworthy for their novelty and breadth, spanning use cases in music production, education, and consumption [5]. One of the more comprehensive reviews of singing-information processing research to date appeared as a tutorial at the 16th International Society for Music Information Retrieval Conference in Málaga, Spain, in 2015 [43]. This tutorial provided an exhaustive list of methods, data sets, tools, and applications, including real-world examples of different singing styles.

Given the pervasiveness of voice in music, demand is keen for improvements in singing-information processing. Now that music-streaming services are the de facto way for people across the world to not only listen to music but also to discover new songs, personalized recommendation is a very promising application. A recent study confirms that music-streaming listeners are especially attuned to the perception of singing [6]. Of several hundred users surveyed (1.2% response rate), listeners indicated that vocals (29.7%), lyrics (55.6%), or both (16.1%) are among the salient attributes they notice in music. Additionally, the four most important "broad" content categories were found to be emotion/mood, voice, lyrics, and beat/rhythm. Meanwhile, listeners said and the seven most important vocal semantic categories are skill, "vocal fit" (to the music), lyricism, the meaning of lyrics, authenticity, uniqueness, and vocal emotion. High-level content attributes like these can be combined with traditional recommendation approaches (e.g., collaborative filtering, factorization machines, or deep networks) to reach a level of nuance that would be difficult to achieve with user-interaction signals alone (e.g., explicit feedback or curated playlists). Furthermore, content-informed methods are necessary for cold-start recommendation (i.e., discovery), an inherent problem for algorithms that rely solely on user signals. Though expert-backed approaches, like the one taken by the Music Genome Project (https://www.pandora.com/about/mgp), have made considerable progress over the last decade, the demand for further improvements is rising along with the seemingly limitless growth in the amount of digital music content and in the number of listeners. Only through automation of music-content description will it be possible to match so much content to so many listeners.

In this article, we focus specifically on the challenge of automatically characterizing attributes of the voice in music as a self-contained and independently testable problem. A holistic view of singing analysis is diagramed in Figure 1, which provides the basic structure of this article. We first outline the fundamentals of the human voice and singing, provide notation to represent singing in recorded music, and introduce common computational models of the voice. Different applications of singing analysis are then grouped by their relationships to music and natural language: vocalized sound in general, voice in musical contexts, and the singing of lyrics. Having outlined approaches to automatically characterizing the voice, we offer

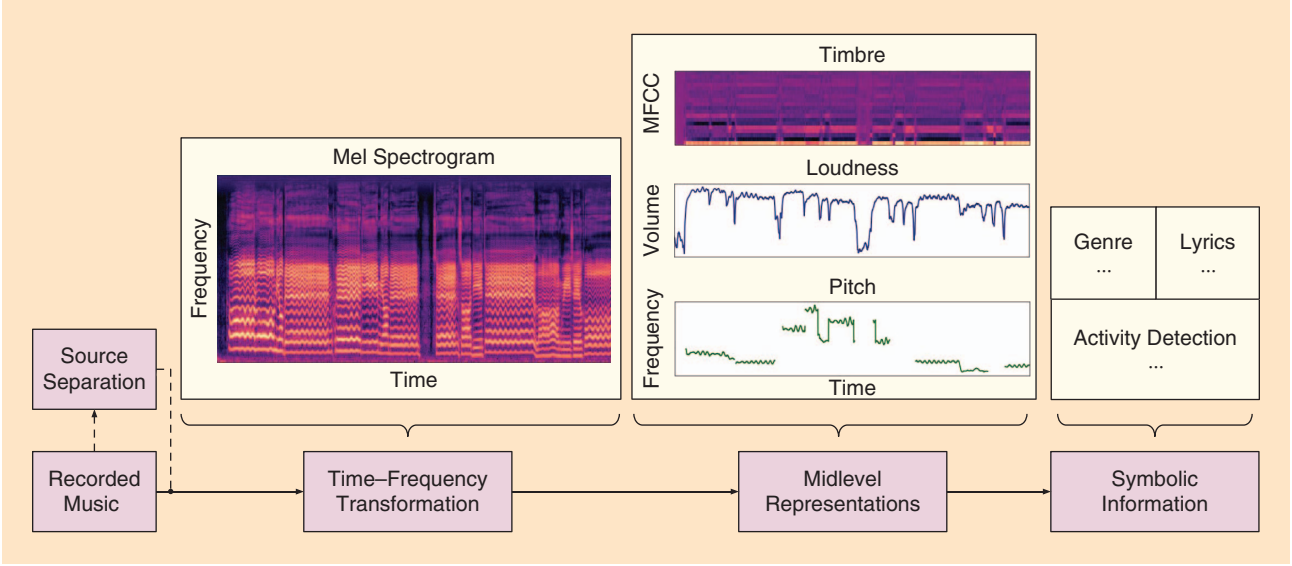

FIGURE 1. A high-level overview of singing-analysis systems: recorded music is optionally preprocessed by a source-separation algorithm before undergoing feature transformations to extract descriptors or symbolic information. Depending on the task, machine learning may be applied between these operations or in an "end-to-end" fashion.

some concrete next steps in this research lineage, and we conclude with an assessment of potential challenges and opportunities facing singing analysis research.

Fundamentals of singing

The compression and expansion, or rarefaction, of air molecules causes a propagation of oscillations known as an *acoustic wave*. These fluctuations can be expressed as a combination of pure sinusoids such that frequencies in the range of 20 to 20,000 Hz are perceived by humans as sound. Classified as an aerophone in the Hornbostel–Sachs taxonomy, the human voice produces sound by moving air, forced from the diaphragm, across the vocal cords, causing them to resonate. This harmonic sound is then shaped via the mouth, with varieties of sibilance added from the teeth, lips, and tongue. The physiological formation of different sounds in the vocal cords and glottis is known as *phonation*, which is how humans convey different phonemes in speech and different voicing styles in singing.

Computational approaches to modeling the human voice fall into either physical or spectral categories [4]. Much is understood about the human vocal organs, and so physical models can be used to demonstrate how the voice produces sound. Source–filter theory, an approach that applies to a variety of string and wind instruments as well, represents sound production as a two-stage process, where a source signal is convolved with the impulse response of a filter. The source can be either voiced (e.g., periodic vowels like *[a]*) or unvoiced (e.g., aperiodic fricatives like *[f]*). In the case of a voiced source signal, the vocal folds vibrate and generate a signal similar to that of a vibrating string. The pitch or fundamental frequency (f_0) of a voiced sound is determined by the rate at which the vocal folds vibrate, and subsequent peaks created at multiples of f_0 are called *harmonics*. Higher frequencies are damped, sloping downwards at approximately −12 dB per octave. In the case of an unvoiced source signal, turbulent noise is created with the teeth, lips, tongue, and, in case of whispering, the glottis. The vocal tract, a tube-shaped acoustic resonator that acts as a filter, is assumed to be independent of the source signal. The resonance frequencies are the direct consequence of the vocal tract, causing what are known as *formants*. They are the main contributor to the spectral envelope of the voice (i.e., the relative amplitudes of the harmonic series) and change along with the length and shape of the vocal tract. Compared to the vibrating vocal folds (source), the vocal tract (filter) can only exhibit relatively slow alternations. Formants allow for the articulation of different vowels and a wealth of different timbres.

Due to the independence of source and filter, it is possible by estimating one component to reconstruct the second. Thus in vocal-signal analysis, the spectral envelope is of specific interest, since it determines the timbre—everything that is not pitch or loudness—to a large degree. One prominent method to estimate the filter/spectral envelope is linear prediction, and its results are the linear predictive coefficients (LPCs) [2]. The basic idea is that the current amplitude of a time-varying digital signal is predictable (approximately) from a linear combination of its past values. The error of this linear model equals the source signal relating to vocal fold characteristics, thus making the source and filter separable.

In contrast, spectral approaches measure the relative contributions of sinusoidal components in signals, often through short-time analysis under assumptions of local stationarity. One of the earliest approaches used sinusoidal modeling, which fits the frequencies and amplitudes of a number of time-varying oscillators to a signal. This method was later extended to model the residual signal as either noise alone or both noise and transients [4]. Though it has the properties of being both compact and complete, sinusoidal modeling can be computationally expensive and quite sensitive to the presence of other signals. As a result, it is more common to model vocal-tract characteristics via mel-frequency cepstral coefficients (MFCCs). MFCCs have been used specifically for music analysis since being introduced by [7] and, until the recent popularization of deep learning, served as one of the standard features in speech and music timbre analysis.

MFCCs are computed through a two-stage process. First, a mel filter bank is applied to the audio signals, typically via the fast Fourier transform for efficiency, such that frequency components are collapsed into 30–120 half-overlapping triangular-shaped filters along a frequency scale grounded in psychoacoustics. Next, the signals are transformed into the cepstral domain by computing and applying a discrete cosine transform (DCT) to the log-magnitude spectra, thus decorrelating the mel-filter bank coefficients. Discarding some of the higher-order coefficients of the DCT results in the representation of a low-pass-filtered spectral envelope, which can be reconstructed by applying the inverse DCT. More recently, the "fluctogram" has been proposed as an alternative time–frequency representation specific to the singing voice. Designed to encode the temporal evolution of the fundamental frequency and its harmonics [8], the fluctogram is computed for several frequency bands based on the cross-correlation of a log-scaled spectrum to the succeeding spectrum, exploiting the characteristic of the voice as a continuous pitched source.

Importantly, the motivation for these models is based on the assumption that the signal of interest contains only a single voice recorded in isolation. However, most recordings in consumer music settings are the result of professional sound production, also referred to as "mixing," an artistic process that combines a number of audio signals arranged in time, subject to any number of complex effects processors (e.g., compression, equalization, reverb, and distortion). For clarity, this process can be expressed as the summation of N digital audio signals, notated as $x[t] = \sum_{n=0}^{N} \alpha_n[t] * f(x_n[t] \,|\, \phi_n[t])$, where α defines a time-varying gain and f an arbitrary, often nonlinear, effects chain with its composite parameters $\phi_n[t]$. In this article, we use "recorded music" to mean the resulting signal $x[t]$, and "voice" as all K signals, x_k, $K \leq N$, that were produced by human voices (note, however, that the true number of voice signals, K, in a recording will not necessarily correspond to the number of distinct voices a listener perceives).

Often in music, one or more of these voice signals will emerge as the "lead" voice, whereby a typical listener perceives a single voice as being particularly salient. Robust, human-level

understanding of singing in recorded music therefore presents the additional complex task of first identifying the voice amid multiple sounds before extracting some desired high-level information.

When creating the architecture for vocal analysis systems to operate on recorded music, any one of three basic approaches can be taken. First, a system could be designed to only consider parts of the music signal where the voice is naturally isolated (i.e., points at which all nonvocal signals are silent). This approach is conceptually straightforward, but has three major drawbacks. The system is limited by its ability to discriminate a solo voice from all other conditions, and any errors will propagate through the system. There are no guarantees that isolated vocals will occur with sufficient frequency in a recording to perform some task. Even so, occasional views of the signal will be inadequate for applications that require comprehensive information regardless of interference (e.g., transcription of melody or lyrics).

Another approach, described in a large body of work in source separation of music, attempts to isolate a sound source of interest given a mix of other signals [9]. Source-separation algorithms generally fall into one of two categories: those that exploit domain knowledge of music in the application of signal decomposition algorithms (e.g., independent components analysis, nonnegative matrix factorization, robust principal components analysis) or those that use data-driven methods that act as filters to directly produce the voice signal in isolation. To the former, the singing voice is often sparse and nonrepetitive in a musical mixture, and algorithms can exploit these properties to perform singing-voice separation [10]. Accompaniment is often considered "low rank," in that it consists of instruments (e.g., drums or guitars playing repetitive patterns), whereas the voice is monophonic and irregular. In a complementary fashion, audio decomposition techniques can be applied in a cascaded fashion to disassemble the music recording into a set of midlevel components that are fine enough to model various characteristics of the singing voice, while coarse enough to keep an explicit semantic meaning of the components [11]. More recently, deep neural networks have emerged in singing-voice separation as powerful nonlinear filters. These algorithms are trained on existing pairs of aligned mixture and isolated voice signals, with the objective of minimizing the error between the true and estimated vocal signals. Modern deep-learning approaches show particular promise, and various works continue to explore different architectures, objective functions, and data sources [12]. To chart progress in this area, the Signal Separation Evaluation Campaign is an annual community-led event organized to systematically and reproducibly compare source-separation algorithms [13].

The third, and most direct, approach is to develop models or features that can characterize the voice despite the presence of interfering signals. In practice, MFCCs or LPCs have proven to be reasonably useful as a consequence of standard practice in sound production; typically, though by no means always, lead vocals are the predominant signal in the mix, and thus vocal information also tends to dominate these representations. For some tasks, feature engineering has proven rather effective, but there are obvious limitations to this approach. More generally, given advances in machine learning, and particularly deep learning, generic time–frequency representations (e.g., MFCCs or spectrograms) or raw time-domain waveforms may be used as inputs to deep neural networks. Data-driven methods enable the system to tease apart signal attributes relevant to voice given an objective, but present their own challenges with respect to data collection, training, and computation. We will see how these three approaches are applied as a function of the task, model, and data.

Singing analysis applications

From the perspective of web-scale music listening, singing-voice analysis aims to extract high-level information from audio signals to enable systems to address some user need (e.g., find instrumental music or songs without expletives). This application space is broad, given the range of sounds the human voice can produce, and so it is helpful to distinguish between the different categories of sound within this space. Musicality and natural language can be represented as two partially overlapping subsets (Figure 2), whose union lies within a larger space of vocalization: for example, one can sing without adhering to the rules of any natural language (e.g., humming or scat), communicate via speech amusically, or produce a variety of sounds that qualify as neither. The ability of humans to comprehend information in musical or linguistic contexts is achieved through high-level cognition built upon lower-level perceptual faculties.

Noting that significant time and attention has been paid to the computational analysis of speech [2], we focus our attention here on three types of singing, each with an eye toward the corresponding musical applications:

- *Vocalization*: acoustic primitives of voice that are common to both musical and linguistic contexts, contributing to such tasks as vocal activity, technique classification, and vocalist identification
- *Vocal music*: singing in musical contexts, which give rise to intonation, melody, and genre by establishing or reinforcing the elements of harmony, rhythm, and timbre

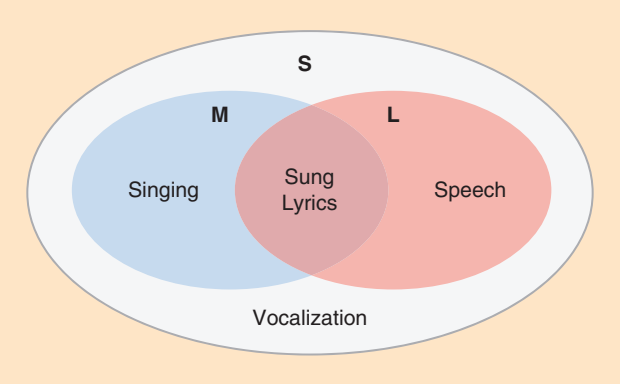

FIGURE 2. An illustration of the set relationship between musicality across the space of sound produced by the human voice (S) such that "singing" comprises vocalizations in a musical grammar (M), "speech" as vocalizations in a linguistic grammar (L), and "sung lyrics" as the intersection of the two, $M \cap L$.

- *Sung lyrics*: the intersection of musicality and language, with applications similar to those in speech recognition, such as language identification, audio–text alignment, and transcription.

Before proceeding, we offer a few notes for consideration. First, these domains are ordered by level of abstraction, which serves as an approximate guide of computational difficulty (e.g., vocal activity is simpler than melody estimation, and both are simpler than lyrics transcription; this is not to say, however, that any of these tasks are trivial, as all are open research areas). Related tasks typically employ similar approaches, and lower-level tasks or representations are often reused in higher-level ones. Finally, the applications presented here are connected to salient dimensions reported by listeners when relevant, both to motivate and identify opportunities for future work.

Vocalization

As described previously, vocalization encompasses the superset of sounds produced by the human voice. Given that listeners are particularly sensitive to the presence of voice generally, the first stage in singing analysis aims to characterize the acoustic primitives of voice. These systems focus on the human voice as a sound source and thus share the common properties that they are not inherently constrained to musical applications. As a result, these systems find additional application in higher-level voice-analysis systems (e.g., only apply lyrics transcription when the voice is present to reduce errors).

Activity detection

The automatic detection of singing voice in recorded music finds immediate use in recommendation contexts (e.g., identifying "focus" music). Referred to as *vocal activity detection* (*VAD*), such systems typically predict the likelihood of vocal activity on short time-scales (i.e., 1 s to dozens of seconds) and can be applied convolutionally over longer signals to produce time-varying estimates; others aim to make predictions over a complete recording. Continuous-valued likelihoods may be simply thresholded at some bias point to produce binary decisions between vocal or instrumental states. Alternatively, in time-varying estimates, postprocessing [e.g., hidden Markov models (HMMs) or median filtering] may be used to prevent spurious or brief detection intervals.

At a high level, two basic approaches may be taken to detect the presence of a singing voice from an observation. The traditional approach involves feature engineering in combination with such classifiers as random forests, support vector machines (SVMs), or neural networks. The current state of the art with this approach uses fluctogram and delta-MFCC features (i.e., first-order difference) that are fed to a long short-term memory recurrent neural network [8]. Alternative approaches use deep neural networks in an end-to-end fashion. The current state of the art with this approach produces results similar to those of its feature-engineered counterpart when trained without data augmentation [14]. With data augmentation, the results seem to be superior, but it is still not clear how previous approaches would also benefit from data augmentation.

One particular challenge faced in VAD systems is a heightened sensitivity to data-set composition and domain transfer for training and evaluation. Both prior discussed approaches yield models that appear to distinguish even highly harmonic instruments producing voice-like pitch trajectories from actual singing voices, as demonstrated by extremely low false-positive rates on specifically curated tests. However, it is especially important to make use of instrumental music to better assess performance [8]. Training with instrumental music helps decrease false-positive rates, while evaluating on instrumental music can reveal certain weaknesses in a given model. Algorithms insensitive to variations of the level of loudness may allow for meaningful comparison. Otherwise, a performance gap between two methods—one loudness-invariant, the other not—could possibly be caused by a convenient level of loudness for the loudness-sensitive method. To give an example, for a loudness-sensitive method the number of false positives will often decrease along with the level of loudness, contrary to the output of a loudness-invariant method, where the number of false positives stays constant.

Technique classification

Machine perception of vocal technique, a burgeoning area of research in singing-voice analysis, relates to a listener's affinity or aversion to a music recording. Phonation modes are important building blocks of more advanced vocal techniques and corresponding analysis systems, such as genre recognition or lyrics transcription. Technique modeling can be seen as a more granular form of general vocal-activity detection, where short-time observations are classified into the kind of vocal activity present. To these ends, the Phonation Modes data set consists of sung vowels in one of the four main phonation modes: breathy, pressed, flow, and neutral [15]. By using a model of singing voice that simulates airflow and pressure through the vocal folds, the authors of the data set achieve an accuracy of 65% with a four-way classifier.

VocalSet is a singing voice data set that consists of these more advanced vocal techniques [16]. These vocal techniques include vibrato, straight, breathy, vocal fry, lip trill, trill, trillo, inhaled singing, belting, and spoken. Some of these techniques are found in a basic vocal repertoire, such as vibrato or trill, while others, like inhaled singing or vocal fry, are found in more advanced repertoires. Figure 3 shows spectrograms of each of these techniques for a male singer in the data set. The spectrograms of each technique are visually different, despite coming from the same singer with the same musical intention (e.g., singing scales, arpeggios, and long tones). VocalSet was collected by recruiting professional singers to sing examples of each of these techniques. The data set consists of 20 singers (11 female), each singing these ten techniques on scales, arpeggios, and long tones. VocalSet contains 10.1 h of recordings. Using deep convolutional neural networks, the authors of the data set achieved a precision of 0.676 and a recall of 0.619 in a ten-way classification setup.

Notably, the role of phonation in performance varies across musical cultures. Computational and quantitative techniques have been used to study variations of singing technique in the

Beijing Opera as a result of educational influence [17]; founded by different instructors, the students of different schools inherit the corresponding vocal production characteristics. Going beyond the subjective description of singing style (e.g., sweet, clear, fragile), the authors take into account a diverse set of audio features common in music-signal analysis, and experimental results support previous findings in the musicology literature.

Singer identification

The automatic identification of vocalists in music audio can help address metadata errors and identify collaborations in

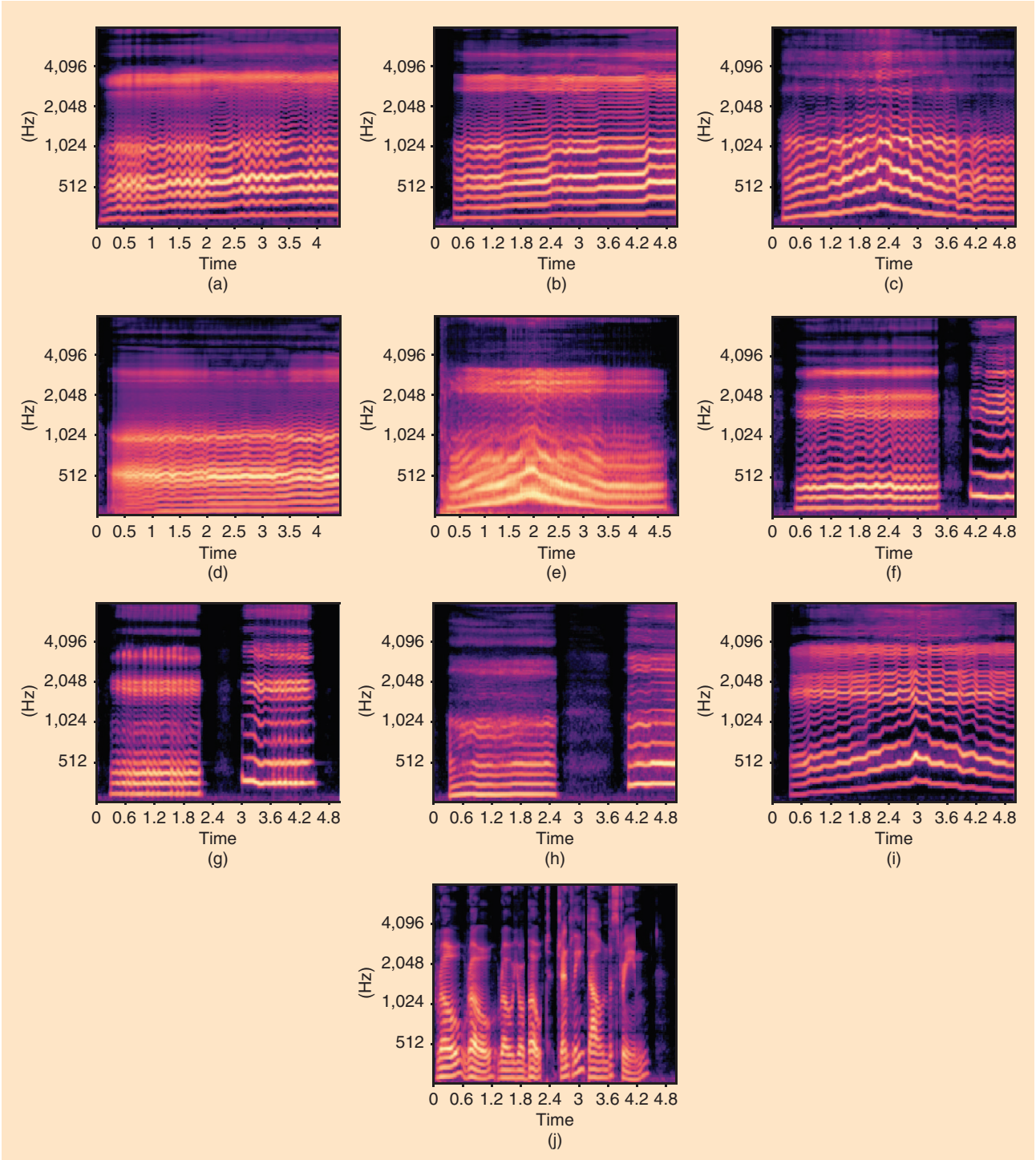

FIGURE 3. Mel spectrograms of the ten vocal techniques contained in the VocalSet data set: (a) vibrato, (b) straight, (c) breathy, (d) vocal fry, (e) lip trill, (f) trill, (g) trillo, (h) inhaled singing, (i) belting, and (j) speaking. Each is a performance of a specific vocal technique by the same male singer. Different vocal techniques produce characteristic spectrograms.

recordings, two commonly recurring challenges. As yet another degree of specificity beyond technique modeling, the problem of vocalist identification is one that stands to benefit greatly from data-driven methods. While efforts in singer identification (singer-ID) have produced few results, one system of note proceeds by extracting vocal segments from songs, computing some engineered feature representation, and classifying with a machine-learning model of choice (e.g., SVMs or Gaussian mixture models) [18]. Singer-ID is distinct from the recognition of vocal technique alone in two ways: 1) longer time scales may be necessary to distinguish among different vocalists; and 2) it remains unclear what the preceptual or computational limits of singer-ID might be in terms of accuracy or performance. However, given that music collections typically provide artist labels on recordings, singer-ID presents an interesting opportunity due to the availability of data for supervised machine learning.

Vocal music

Building upon general vocalizations, we now focus on the analysis of the singing voice in musical contexts specifically. While singing may also convey natural language, "vocal music" is defined as the musical compositions or performances that feature one or more human voices. This entails an understanding that singing conforms to the basic dimensions of music: harmony (pitch), rhythm (timing), and timbre (source discrimination). However, while timbre encompasses the distinguishing traits of a particular sound source—here, the human voice—a singer is considered a monophonic instrument, i.e., of a single pitch. While the human voice is capable of producing multiple pitched sounds simultaneously, the practice is uncommon and not considered here. As a result of emphasis placed on harmony in traditional music theory practice, the analysis of singing often, though not exclusively, focuses on pitch.

Intonation

The harmonic basis on which a piece of music is built is known as *intonation*. In popular Western music, the common tuning system is known as *12-tone equal temperament* and has standardized by convention on A4 = 440 Hz. While some popular instruments produce sound in quantized pitch intervals (e.g., piano), the human voice is capable of producing arbitrary pitch. Some non-Western music traditions, such as Indian art music (IAM), take other approaches to intonation that complicate the design of signal processing systems, making intonation a relevant research topic. For context, IAM refers to two art music traditions of the Indian subcontinent, Hindustani music (also known as *North Indian music*) and Carnatic music (also known as *South Indian music*). Both Hindustani and Carnatic music are singing-centric traditions, and therefore the voice effectively dictates the intonation used in a piece. Rāga is defined as the melodic framework in IAM and serves as the core musical concept used in composition, performance, music organization, and pedagogy. Hindustani and Carnatic music is characterized by different melodic attributes, such as svaras (roughly speaking, notes), intonation of the svaras, and characteristic melodic phrases.

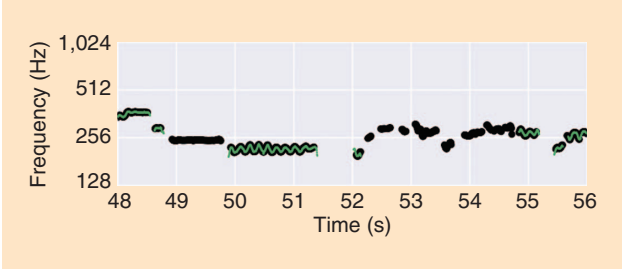

FIGURE 4. A vocal melody estimated by an algorithm (green) against the ground truth (human labeled) vocal melody (black).

Due to the importance and variation inherent to pitched singing, the lack of simplifying assumptions around tuning complicates the automatic analysis of these kinds of music. Carnatic music, for example, does not make use of an equal-tempered tuning schema, being closer to five-limit just intonation, whereas Hindustani music can be explained by a mixture of equal-tempered tuning and five-limit just intonation (a five-limit tuning system uses powers of two, three, and five to compute notes relative to a reference frequency). The intonation of svaras is an important characteristic of a rāga, and so detailed pitch distributions are informative as a result. It has been shown, for example, that the shape of the pitch histogram for different svaras can assist in automatic identification of rāgas [19]. Since there exists subtle intonation differences across rāgas, the frequency resolution chosen for intonation analysis in IAM is much higher than that for many other music traditions.

Melody estimation

The task of determining the pitch, or fundamental frequency, of the singing voice in music over time is generally referred to as *vocal melody estimation*. Estimated melodies are typically represented in the form of time series (time, pitch), where the interval between time steps is small (e.g., 10 ms), and pitch values are continuous (measured in hertz) values rather than as discrete note values. Figure 4 shows an example of a vocal melody estimated by an algorithm (green) plotted against the ground truth vocal melody (black) for a short excerpt. Note how by representing the pitch values on a continuous rather than discrete frequency grid, information, such as vibrato, is captured between 50 and 51 s in the figure. Additionally, note that part of the task is also to determine where no vocal melody is present.

There are three common types of approaches to vocal melody estimation [20]: salience, source separation, and machine-learning based. Salience based methods leverage the assumption that vocals exhibit a known harmonic series. To exploit this information, these approaches first estimate a vocal salience representation, a time–frequency representation derived from a short-time Fourier transform, realized by reweighting the amplitude of each time–frequency bin based on the presence or absence of related harmonics. The purpose of this is twofold: 1) to de-emphasize content that is not part of the vocal melody and 2) to emphasize content that is likely part of the vocal melody (i.e., content with many related harmonics). Salience representations are computed, for example, via harmonic summation, harmonic percussive

source separation, or filtering/equalization. After computing a salience representation, these methods often apply heuristics-based rules for selecting the most likely vocal melodies from the computed representation. Source separation-based methods first isolate the singing voice and subsequently apply a pitch tracker in order to compute the melody, or conversely they jointly estimate the singing-voice audio signal and the vocal melody. More recently, machine-learning methods have been used to turn the task into a classification problem by discretizing the frequency space with at least one class per semitone and predicting the most likely class over time [21], [22]. Alternatively, machine learning can be used to learn robust salience representations [23].

Vocal melody estimation has a number of applications in musical indexing and retrieval. A long-standing goal of MIR is known as *query-by-humming*, where a listener can search a collection of content by vocalizing a given melody. The ability to find specific recordings by melody would likely result in related results and similarity-based retrieval. Additionally, melody is a predominant feature of music and would further inform higher-level analysis, such as pattern discovery and structural segmentation (e.g., thumbnailing or chorus detection).

Estimation of the predominant melody is also at the core of singing-voice analysis in IAM [24]. In a typical performance, the main vocalist is accompanied by another melodic instrument, almost like a lagging imitation of the lead. There are approaches that exploit this convention by tracking the two melodic contours simultaneously, one of which being that of the lead vocalist. Attempts have been made to automate the selection of pitch contour corresponding to the lead artist by using temporal instability of the voice harmonics. Due to the subtle nuances in the temporal evolution of the melodies (specifically in the transitory regions between two svaras), the entire pitch contour is often used as a midlevel feature for singing-voice analysis. Often, steady-state regions and transitory regions in a melody are segmented for better characterization of the melodies.

Genre

Among the more abstract concepts in music, genre is used to describe the musical categories that emerge naturally from a culture's influence on itself. A genre is established through the use or reuse of certain musical aspects, such as structural form, instrumentation, or melodic patterns, which leads to shared understanding across groups of people. Various forms of rock prominently feature distorted guitars, for example, while blues is known for dominant chords and 12-bar phrasing.

While there are numerous, often inscrutable characteristics that may contribute to the boundaries of a genre, it is relevant here to consider those that place a specific emphasis on the singing voice. One instance is that of subgenres of metal music, which are characterized by extreme vocal effects [25]. One of the primary motivations behind singing-voice analysis in IAM is for automatic rāga identification. Recently, a technique called *time-delayed melody surfaces* has been shown to capture continuous tonal and temporal characteristics of these melodies, resulting in a significant improvement in rāga recognition accuracy [26]. Rap is another notable instance of a genre identified in large part by distinctive rhythmic voice delivery characteristics. It has been demonstrated that only 11 perception-inspired features lead to 91% classification accuracy between rapping and singing with only 3-s isolated vocal segments [27]. The most salient feature was found to be the ratio of voiced frames to nonsilent frames, confirming the prominent role of rhythm and lack of melodic characteristics of rapping, in contrast with the more melodic nature of traditional singing found in contemporary rhythm and blues music.

Genre can also serve as a suitable proxy for singing style, a musically appealing but difficult to define characterization of vocal performance (e.g., theatrical, aggressive, or powerful). Vocal-specific features, such as statistics computed over fundamental frequency (f_0) contours, are useful for discriminating between different singing styles in both supervised and unsupervised approaches [28]. Clustering these features has enabled the semantically meaningful organization of a collection of 50,000 excerpts of folk music from around the world, while large-scale embeddings for vocal style are also a promising avenue of research [29].

Sung lyrics

Viewed from the perspective of linguistics, human vocal communication with language has four dimensions [30]:

- *Phonemes*: the building blocks of vocalized language, representing discrete units of sound
- *Prosody*: the articulation of phonemes over time, including aspects of inflection, duration, rate, or intonation
- *Vocabulary*: the combination of phonemes into words as higher-level sound objects
- *Grammar*: the sequential, structural composition of words.

At the intersection of music and natural language, the singing of lyrics presents unique difficulties beyond those typically faced in speech processing alone [31]. Often the rules of grammar are bent or ignored for artistic reasons (e.g., rhyme). Prosodic elements are constrained by the melodic and rhythmic dimensions of a musical work and not necessarily by the language in which the lyrics are performed. For example, the typical fundamental frequency for female speech lies between 165 and 200 Hz, while in singing it can reach more than 1,000 Hz. This is further complicated in a tonal language like Chinese, where the inflection of pitch is also used to convey semantic meaning. As a result, traditional speech corpora are insufficient for building data-driven models for singing analysis, given the degree of domain transfer between spoken language and vocal music. Meanwhile, accompanying instrumentation complicates traditional assumptions regarding noise in speech processing, in that typically all signals in recorded music are both harmonically and temporally correlated. With that in mind, we now turn our attention to methods for language identification, the alignment of audio and lyrics, and lyrics transcription.

Language identification

Singing language identification (SLID) can be seen as a simplification of comprehensive lyrics transcription. In music services for global populations, the predominant language of performance

is a valuable attribute: It provides deeper insight into music catalogs in linguistically diverse settings, such as India or the Philippines; and, through greater comprehension of the content, enables a deeper understanding of a listener's language preferences. The latter is a complex issue facing recommender systems because of asymmetrical preferences toward music consumed in different origins (e.g., users in country *X* might listen to music from country *Y*, but not the inverse).

SLID systems conventionally approach the task by modeling the statistics of phonemes over long time scales, building different templates on a per-language basis. One modern effort of note is that of [32], which focuses on 25 languages drawn from 25,000 music videos. The authors explore a variety of feature representations, leveraging both acoustic and visual descriptors aggregated over the temporal context of the signal, fed into a number of binary SVM classifiers (one per language). Experimental results show that a mix of acoustic features—spectrograms, MFCCs, and stabilized auditory images—led to a performance on a test set of 44.7%; by adding visual features, the system achieved 47.8% accuracy. Interestingly, this system considers general-purpose feature representations, placing the burden of modeling on a powerful classifier, and calls into question the need to distinguish between vocal and nonvocal segments.

Audio–lyrics alignment

Time-alignment of lyrics with the corresponding audio is necessary for such popular applications as karaoke and subtitling of music videos. The availability of alignments also makes possible a host of applications, such as automatic radio edits, playback starting/ending at specified lines, and analyses of how words in music correspond to beats, melodies, and other musical structures [33]. Manual alignments do not scale to large collections of audio, raising the need for accurate automated alignment algorithms.

The goal of automated alignment, shown in Figure 5, is to take the audio and lyrics and produce a time alignment of the two inputs. Alignments are typically at the word level, but may also be at the level of lines or phonemes, depending on the downstream application. Line-level alignments may be sufficient for such products as subtitling or some karaoke interfaces. LyricAlly is one system of note that detects such structural elements as beats and rhythm, which are used to segment the audio into the introduction, verses, chorus, bridge, and coda [34]. The lines in the lyrics corresponding to these sections are then aligned to the segmented audio. Word-, syllable-, or phoneme-level alignments require greater precision. Some works rely on annotations, such as Musical Instrumental Digital Interface (MIDI) files or lead sheets; however, these cannot generalize to unannotated music.

The speech technology community uses a method called *forced alignment* to time-align audio and transcripts. Forced alignment involves finding the Viterbi path through HMMs that map phonemes to MFCCs or other features of the acoustics. These HMMs are trained from large corpora of transcribed speech. Several speech toolkits, such as CMU Sphinx (https://cmusphinx.github.io), the Hidden Markov Model Toolkit (http://htk.eng.cam.ac.uk), and Kaldi (http://kaldi-asr.org) implement forced alignment, including the ability to train the acoustic HMM models, with wrappers, such as the Montreal Forced Aligner (http://montreal-forced-aligner.readthedocs.io), providing interfaces to these programs. Forced alignment works best when line or phrase-level boundaries are specified, since alignment quality degrades with audio longer than a minute. Forced alignment forms the basis of most lyrics–audio alignment algorithms. However, some characteristics of singing make it challenging to apply alignment models developed for speech to music [35].

Introduced earlier, lyrics alignment is one area that makes use of vocal detection and separation as preprocessing steps before alignment to mitigate challenges posed by recorded music. In addition, it is possible to reduce the sound of

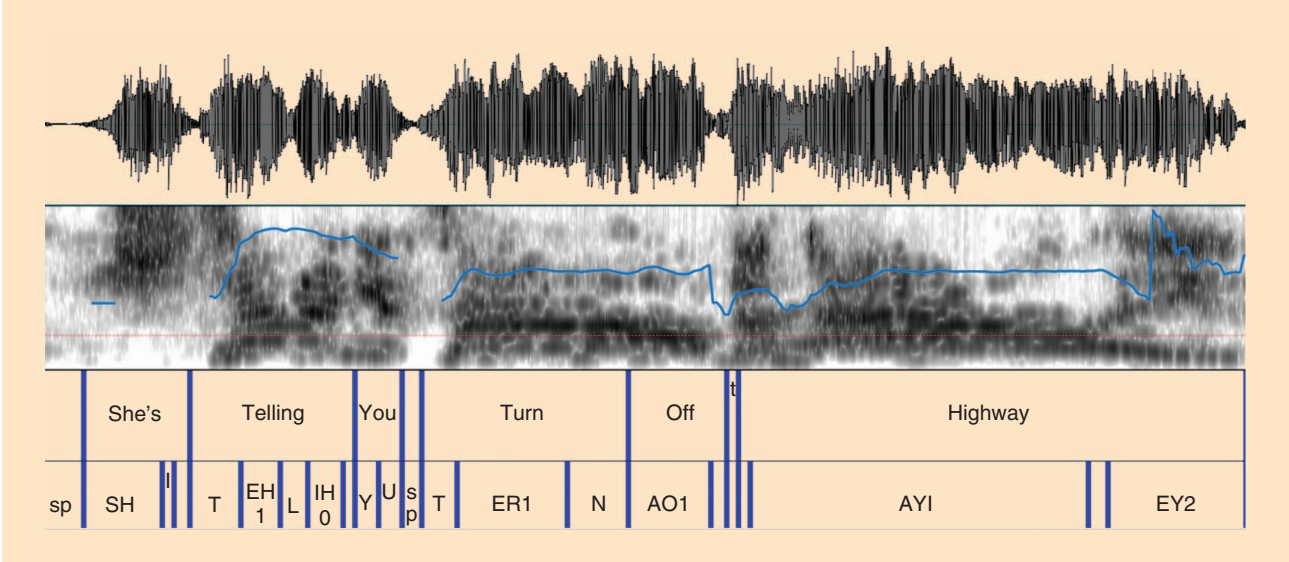

FIGURE 5. Visualization of automated word- and phoneme-level alignments from a segment of a song generated with the Praat software (http://www.fon.hum.uva.nl/praat).

accompanying instruments with f_0 estimation and resynthesis and to adapt the acoustic HMM models trained on speech to a small corpus of music [36]. Systems may also use placeholders in the HMM for such words as "yeah yeah" that may not be captured in the reference lyrics. Integrating musical information, such as chord sequences, is also helpful for improving lyrics-alignment performance [37].

Lyrics transcription

Lyrics transcription is generally performed in two steps: first, phoneme probabilities are recognized in the singing audio by using an acoustic model; then, the results are processed with a language model to obtain plausible word sequences. As in speech recognition, most early lyrics-transcription systems relied on HMMs for acoustic modeling. Due to the lack of lyrically transcribed singing data, many systems trained acoustic models on read speech, with language models built on actual texts of lyrics. For example, language modeling can be achieved with a finite-state automaton fitted to the lyrics of a collection of Japanese children's songs [38]. The system is tested on sung phrases consisting of five words, without accompaniment, achieving a word error rate of 36%. By training speaker-specific acoustic models, the word error rate is lowered to 27%.

Several improvements have been proposed that incorporate intuition about human perception to lyrics. It is noted that source separation can be used as a preprocessing technique to improve model accuracy. Repetition and structure in music, such as the chorus, may also be exploited to improve transcription accuracy [39]. Three different strategies are proposed for combining individual results: feature averaging, selection of the chorus instance with the highest likelihood, and combination using the Recognizer Output Voting Error Reduction (ROVER) algorithm. Twenty unaccompanied English-language songs from the Real World Computing (RWC) database were used for testing; chorus sections were selected manually. The best-instance selection and the ROVER strategies improve results significantly; with the ROVER approach and a general-purpose language model, the phoneme error rate is 74% (versus 76% in the baseline experiment), while the word error rate is improved from 97% to 90%. Interestingly, cases with a low baseline result benefit the most from exploiting repetition information.

To overcome the lack of realistic training data, forced-alignment algorithms may be used to fit a set of unaccompanied singing with unaligned lyrics [40]. For example, deep neural networks are trained on MFCCs of music signals to produce singing-specific acoustic models. These models produce better results compared to those trained on speech, with the phoneme error rate falling to 80%. Notably, both word and phoneme error rates are expected to be higher in lyrics transcription than in speech recognition. While the limits of human lyrics recognition are unknown, the phenomenon of "misheard" lyrics is common [41].

A simplified form of lyrics transcription is the ability to pinpoint specific words (e.g., expletives) in recordings. Many song lyrics contain expletives, and there are numerous scenarios in which it is necessary to know when these words occur (e.g., "family-friendly" listening sessions). In the case of airplay, expletives are commonly "bleeped" or acoustically removed. The task of finding such words is based on the alignment strategies described previously, taking advantage of the wide availability of textual lyrics. The system proceeds by automatically aligning text lyrics to audio, searching for predefined expletives in the result, and subsequently modifying the signal where any flagged instances occur (e.g., adding white noise as an obfuscation) [40]. The test data set consists of 80 popular songs, most of them hip-hop. Annotations indicated 711 instances with 48 expletives on these songs, and the matching textual, unaligned lyrics were manually retrieved from the Internet. Using the acoustic models described therein, 92% of the expletives were detected in their correct positions with a tolerance of 1 s.

Next steps

Getting started with singing analysis

As illustrated by the breadth of the previous section, singing-voice analysis is a diverse area of study with potential to enable a variety of large-scale applications. However, this rich array of possibilities may also make it difficult to decide where and how to first dive into this topic. To help direct new explorations in singing-voice analysis, there are three tasks we recommend as good entry points: vocal-activity detection, singer-ID, and SLID. Each can be framed as a straightforward classification problem with objective evaluation measures (i.e. precision, recall, f-score) and in each case the task of finding or collecting labeled data is relatively easy. To further facilitate this exploration, we also provide an open-source software tutorial for self-guided exploration (https://github.com/spotify/ieee-spm-vocals-tutorial).

Vocal-activity detection is a logical starting point for those new to music signal processing with an interest in singing analysis. Recognizing vocal activity as a low-level percept, computational systems can focus on short-time observations drawn from audio signals, simplifying both labeling and modeling as a binary classification task. Given the increasingly mature state of machine learning, the challenge of building a VAD system resides more in obtaining or curating data for training and evaluation. The two conventional data sets used in VAD research are the Jamendo collections, though newer collections like MedleyDB (http://medleydb.weebly.com/), OpenMIC-2018 (https://github.com/cosmir/openmic-2018), or AudioSet (http://research.google.com/audioset/) provide more data for training such models. A particular advantage of VAD as a task is that its simple framing allows one to study the effects of data-set composition on model performance. As mentioned previously, the inclusion of a cappella (solo voice) or instrumental music in a data set can help address false negatives or false positives, respectively, but it is also possible to synthesize more training data from multitrack recordings (e.g., MedleyDB).

Another attractive, near-field opportunity suitable for newcomers to the topic of singing-voice analysis is that of singer-ID. As discussed, methods for singer-ID are somewhat under-represented in the literature, leaving ample room to improve upon the state of the art. Additionally, there is often a 1:1 correspondence between recording artist (or group) and vocalist (i.e., a band features a single singer in all of its recordings), and it is

possible to collect large data sets for training machine-learning models without too much effort. This observation can be combined with modern source-separation algorithms to produce reasonable approximations of vocals in isolation, mitigating any confounding factors of instrumentation. This approach can be applied to the Free Music Archive (FMA) data set (https://github.com/mdeff/fma), which contains 100,000 recordings from more than 16,000 unique artists, with more than 1,000 artists having at least 20 recordings. Alternatively, Stanford's Digital Archive of Mobile Performances collection (https://ccrma.stanford.edu/damp/) features 35,000 solo voice recordings from roughly 350 amateur singers, which mostly bypasses the need for source-separation preprocessing. This data could be used to train a model as in the VAD scenario, with a classifier applied to short-time observations of audio signals. We emphasize that these artist–singer labels can be used to fit deep-learning models whose intermediary representations (e.g., the penultimate layer) can be used as an embedding model for similarity and retrieval.

A third accessible voice-analysis application is that for identifying the language of the song. While there is traditionally no mutually agreed upon data set for this problem, the FMA contains non-English-language tags for several hundred recordings, and global music services no doubt contain playlists or artists that consist of music performed in a given language. Similar to the formulation of singer-ID, language identification may benefit from the application of source separation as a preprocessing step, and there is considerable opportunity to advance the state of the art in the area of sung lyrics.

Challenges and opportunities

Singing-analysis research is rich with opportunities and challenges. We summarize a few. Subjective evaluation of singing-voice models, as in source separation and similarity, remains a challenge [42]. Objective metrics of source-separation quality are widely used (e.g., signal–noise ratio) but their ability to mirror perception is limited. Expert or crowdsourcing listening tests are often used, but researchers have yet to adopt a standard and well-controlled protocol. Singing-style models have mostly been evaluated using listening tests, and these have been small in scale due to the significant human effort involved. Larger models that cover diverse music require more quantitative methods. There is not yet a standard for benchmarking models of vocal style, for defining vocal similarity or style, or for quantifying listeners' perception of the singing voice. While there is some work that investigates the relationship of phonation modes with vocal styles, it is unclear how it relates to perception of the voice and remains an open area of research.

Machine-learning-based approaches are becoming ubiquitous to most aspects of computational analysis of vocals, but we have yet to see the kinds of dramatic improvements that have been achieved recently in related fields. On reflection, this is likely due to a lack of large, readily available collections for music signal processing research, like ImageNet for object recognition. Thus, while the newer data sets mentioned here, such as the FMA, may help address this shortcoming, more effort is needed to curate or mine large-scale data sets for other tasks in singing-voice research. For example, user-contributed lyrics are widely available on the Internet, and the ability to align these text documents with audio would transform the field.

Curating labeled music data sets for every task may prove cost prohibitive, given the skills required, as in the case of melody annotation. For these tasks, it may be more practical in the short term to artificially generate training data from symbolic signals, such as MIDI files and lead sheets, using realistic instrument synthesizers. This is not yet feasible for all tasks involving vocals, since modern voice synthesizers have yet to fully replicate natural singing. However, advances in melody estimation may provide realistic voice approximations, thereby producing more realistic data for training. Similarly, vocal source separation or an increase in the availability of multitrack recordings makes it possible to create mixes of arbitrary pairs of vocals and instrumentals. Importantly, unlabeled vocal music content is abundant. By mining large catalogs of music, we can build weakly labeled training sets or investigate multimodal approaches to data-set creation (e.g., music videos that feature lyrics).

Finally, most music informatics research is focused on analyzing commercially produced music content, which typically is created by professional musicians and follows basic tenets of music in accordance with the relevant genre or tradition. On the other hand, content produced by amateurs is not bound to follow these tenets and often poses a challenge to the existing singing information processing approaches. In recent years, the volume of such content and applications has risen significantly, often in the context of music education and gaming, (e.g., karaoke applications). The imprecision of amateur singing may be more pronounced than that for instrumental performances by amateurs since the frequencies produced by the voice are not naturally quantized, like they are, for example, for the flute, and have neither tangible nor visual feedback, as with a violin. Given that there are vastly more amateur than professional singers, the automatic analysis of the singing voice presents a considerable opportunity to enhance the human experience of music.

Authors

Eric J. Humphrey (ejhumphrey@spotify.com) received his B.S. degree in electrical engineering from Syracuse University, New York, his M.S. degree in music engineering technology from the University of Miami, Florida, and his Ph.D. degree in music technology from New York University, where he worked with Juan Pablo Bello in the Music and Audio Research Laboratory. He is a machine-learning engineering manager at Spotify in New York City, helping teams research and develop machine-learning algorithms to improve the experience of listeners around the world. Previously at Spotify, he was a senior researcher focusing on machine-learning approaches to understanding music audio signals. Beyond research, he is also a singer–songwriter and multi-instrumentalist.

Sravana Reddy (sravana@spotify.com) received her B.S. degree in computer science, mathematics, and creative writing from Brandeis University, Waltham, Massachusetts, and her Ph.D. degree in computer science from the University of

Chicago. She has spent time at the University of Southern California's Information Sciences Institute in Los Angeles, Dartmouth College Hanover, New Hampshire, and Wellesley College, Massachusetts. She is a machine-learning engineer at Spotify in Boston, where she works on projects related to natural language processing and machine learning. Her research spans natural language processing, speech, machine learning, and linguistics, with a particular emphasis on language variation, including both dealing with it in practical systems and analyzing it using large corpora. Her interests also include applications of computation to literature and writing.

Prem Seetharaman (prem@u.northwestern.edu) received his B.S. degree in computer science with a second major in music composition from Northwestern University Evanston, Illinois, where he is currently a Ph.D. candidate working with Bryan Pardo. He works on problems in creativity support tools, audio source separation, and machine learning. In addition to research, he is an active composer and musician in the Chicago, Illinois, area.

Aparna Kumar (aparna@spotify.com) received her B.S. degree in physics from Drexel University, Philadelphia, Pennsylvania, and her Ph.D. degree from the School of Computer Science at Carnegie Mellon University, Pittsburgh, Pennsylvania. She is a senior research scientist at Spotify in New York City, focusing on audio understanding, perceptual evaluation, user modeling, and data mining for business applications. Her research began in computational biology. Her prior work includes mining pathology images, experimental design, and data collection for oncology drug development.

Rachel M. Bittner (rachelbittner@spotify.com) received her B.S. degree in mathematics and her B.M. degree in music performance from the University of California, Irvine. She received her M.S. degree in mathematics from New York University's Courant Institute in 2013. She received her Ph.D. degree in music technology from New York University, working in the Music and Audio Research Laboratory with Juan Pablo Bello, with her dissertation focus on the application of machine learning to fundamental frequency estimation. Previously, she was a research assistant at NASA Ames Research Center working with Durand Begault in the Advanced Controls and Displays Laboratory. Her research interests are at the intersection of audio signal processing and machine learning, applied to musical audio.

Andrew Demetriou (andrew.m.demetriou@gmail.com) received his B.A. degree in political science and philosophy from Queens College, City University of New York, and his M.S. degree in social psychology from Vrije Universiteit, Amsterdam. He is currently a Ph.D. candidate in the Multimedia Computing Group at the Technical University at Delft, The Netherlands. His academic interests focus on the intersection of the psychological and biological sciences and the relevant data sciences. His academic interests also extend to furthering our understanding of love, relationships, and social bonding; optimal, ego-dissolutive, and meditative mental states; and people performing, rehearsing, and listening to music.

Sankalp Gulati (sankalp.gulati@gmail.com) received his B.Tech. degree in electrical and electronics engineering from the Indian Institute of Technology, Kanpur, India, and his M.S. degree in sound and music computing from the Universitat Pompeu Fabra, Barcelona, Spain. He received his Ph.D. degree from the University of Pompeu Fabra in Barcelona, Spain, where he worked the Music Technology Group with Xavier Serra on the CompMusic project. His research interests include signal processing, time series analysis, and machine learning applied to audio music signals. He has years of industrial experience working in the domain of audio and speech technologies, music content analysis, music education, and is currently working on machine learning and artificial intelligence in the area of financial technology.

Andreas Jansson (andreasj@spotify.com) received his B.S. degree in computer science from City University, London, where he is a Ph.D. degree student and he is also a research engineer at Spotify in New York City. He is currently exploring deep neural network architectures for source separation and mining large commercial music catalogs for training data. Before joining Spotify, he worked at music start-ups The Echo Nest and This Is My Jam. He enjoys playing the accordion, lingonberry picking, and Emacs Lisp.

Tristan Jehan (tjehan@spotify.com) received his B.S. degree in mathematics, electronics, and computer science, and his M.S. degree in electrical engineering, computer science, and signal processing from the Université de Rennes I, France. He received his Ph.D. degree in media arts and sciences from the Massachusetts Institute of Technology. He is a director of research at Spotify, where he cultivates new technologies that can grow into next-generation features and business opportunities. He was chief science officer and cofounder of the music intelligence company The Echo Nest, which was acquired by Spotify to establish a new global standard in music personalization. He has introduced to the industry machine-listening technologies, which involve applications related to music similarity, discovery, and algorithmic music remixing. His academic work combined machine-listening and machine-learning technologies in teaching computers how to listen and make music on their own.

Bernhard Lehner (Bernhard.Lehner@jku.at) received his B.S. and M.S. degrees in computer science in 2007 and 2010, respectively, from Johannes Kepler University, Linz, Austria, where he is currently pursuing a Ph.D. degree. From 1991 to 2004, he was with Virginia Polytechnic Institute and State University, Lenze, Siemens, and Infineon. His research interests include signal processing, audio event detection, audio scene classification, music information retrieval, image processing, neural networks, and interpretable machine learning.

Anna Kruspe (anna.kruspe@dlr.de) received her diploma and Ph.D. degrees in media technology from Technische Universität Ilemnau, Germany, in 2011 and 2017, respectively. She is a machine-learning researcher at the German Aerospace Center. Previously, she was a member of the Fraunhofer Institute for Digital Media Technology, Ilmenau, Germany, where her work focused on the application of speech recognition technologies to singing (e.g., for language identification, keyword spotting, or lyrics-based search), as well as the analysis of world music. She conducted research at Johns Hopkins

University in Baltimore, Maryland, and at the National Institute of Advanced Industrial Science and Technology in Tsukuba, Japan. Her current work deals with the development of machine-learning technologies for the analysis of social media data in the context of disaster management.

Luwei Yang (luwei.yang.qm@gmail.com) received his B. Sci. degree in engineering with law, first class, from the Bejing University of Post and Telecommunications, China and his Ph.D. degree in electronics engineering at the Centre for Digital Music at Queen Mary University of London, under the supervision of Elaine Chew and Khalid Z. Rajab. He is currently a senior algorithm engineer at Alibaba Group, where his work focuses on the application of machine-learning and deep-learning techniques to the areas of recommender systems, natural language processing, and intelligent agents.

References

[1] P. Juslin and P. Laukka, "Communication of emotions in vocal expression and musical performance: Different channels, same code?" *Psych. Bul.*, vol. 129, pp. 770–814, 2003.

[2] B. Gold, N. Morgan, and D. Ellis, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music*. Hoboken, NJ: Wiley, 2011.

[3] M. Goto, T. Saitou, T. Nakano, and H. Fujihara, "Singing information processing based on singing voice modeling," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2010, pp. 5506–5509.

[4] J. Bonada and X. Serra, "Synthesis of the singing voice by performance sampling and spectral models," *IEEE Signal Process. Mag.*, vol. 24, no. 2, pp. 67–79, 2007.

[5] M. Goto, "Singing information processing," in *Proc. 12th Int. Conf. Signal Processing (ICSP)*, 2014, pp. 2431–2438.

[6] A. Demetriou, A. Jansson, A. Kumar, and R. Bittner, "Vocals in music matter: The relevance of vocals in the minds of listeners," in *Proc. 19th Conf. Int. Society for Music Information Retrieval (ISMIR)*, 2018, pp. 514–520.

[7] J. T. Foote, "Content-based retrieval of music and audio," in *Proc. Multimedia Storage and Archiving Systems II, Int. Society for Optics and Photonics*, 1997, vol. 3229, pp. 138–148.

[8] B. Lehner, J. Schlüter, and G. Widmer, "Online, loudness-invariant singing voice detection in mixed music signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 26, no. 8, pp. 1369–1380, Apr. 2018.

[9] Z. Rafii, A. Liutkus, F.-R. Stoter, S. I. Mimilakis, D. FitzGerald, and B. Pardo, "An overview of lead and accompaniment separation in music," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 26, no. 8, pp. 1307–1335, 2018.

[10] B. Pardo, Z. Rafii, and Z. Duan, "Audio source separation in a musical context," in *Springer Handbook of Systematic Musicology*, R. Bader, Ed. New York: Springer-Verlag, 2018, pp. 285–298.

[11] J. Driedger and M. Müller, "Extracting singing voice from music recordings by cascading audio decomposition techniques," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 126–130.

[12] Y. Luo, Z. Chen, J. R. Hershey, J. L. Roux, and N. Mesgarani. (2016). Deep clustering and conventional networks for music separation: Stronger together. arXiv. [Online]. Available: https://arxiv.org/abs/1611.06265

[13] D. Ward, R. D. Mason, R. C. Kim, F.-R. Stöter, A. Liutkus, and M. D. Plumbley, "SISEC 2018: State of the art in musical audio source separation-subjective selection of the best algorithm," in *Proc. 4th Workshop on Intelligent Music Production*, Huddersfield, United Kingdom, 2018, this workshop does not produce paginated proceedings.

[14] J. Schlüter and T. Grill, "Exploring data augmentation for improved singing voice detection with neural networks," in *Proc. 16th Conf. Int. Society for Music Information Retrieval (ISMIR)*, 2015, pp. 121–126.

[15] P. Proutskova, C. Rhodes, T. Crawford, and G. Wiggins, "Breathy, resonant, pressed–automatic detection of phonation mode from audio recordings of singing," *J. New Music Res.*, vol. 42, no. 2, pp. 171–186, 2013.

[16] J. Wilkins, P. Seetharaman, A. Wahl, and B. Pardo, "Vocalset: A singing voice dataset," in *Proc. 19th Conf. Int. Society for Music Information Retrieval (ISMIR)*, 2018, pp. 468–474.

[17] R. C. Repetto, R. Gong, N. Kroher, and X. Serra, "Comparison of the singing style of two jingju schools," in *Proc. 16th Conf. Int. Society for Music Information Retrieval (ISMIR)*, 2015, pp. 507–513.

[18] H. Fujihara, M. Goto, T. Kitahara, and H. G. Okuno, "A modeling of singing voice robust to accompaniment sounds and its application to singer identification and vocal-timbre-similarity-based music information retrieval," *IEEE Trans. Audio, Speech, Language Process.*, vol. 18, no. 3, pp. 638–648, 2010.

[19] G. K. Koduri, V. Ishwar, J. Serrà, and X. Serra, "Intonation analysis of rāgas in Carnatic music," *J. New Music Res.*, vol. 43, no. 1, pp. 72–93, 2014.

[20] J. Salamon, E. Gómez, D. P. W. Ellis, and G. Richard, "Melody extraction from polyphonic music signals: Approaches, applications and challenges," *IEEE Signal Process. Mag.*, vol. 31, no. 2, pp. 118–134, Mar. 2014.

[21] S. Balke, C. Dittmar, J. Abeßer, and M. Müller, "Data-driven solo voice enhancement for jazz music retrieval," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, Mar. 2017, pp. 196–200.

[22] F. Rigaud and M. Radenen, "Singing voice melody transcription using deep neural networks," in *Proc. 17th Conf. Int. Society for Music Information Retrieval (ISMIR)*, 2016, pp. 737–743.

[23] R. M. Bittner, B. McFee, J. Salamon, P. Li, and J. P. Bello, "Deep salience representations for f0 estimation in polyphonic music," in *Proc. 18th Conf. Int. Society for Music Information Retrieval (ISMIR)*, Oct. 2017, pp. 63–69.

[24] K. K. Ganguli and P. Rao, "Discrimination of melodic patterns in indian classical music," in *Proc. IEEE 21st Nat. Conf. Communications (NCC)*, 2015, pp. 1–6.

[25] O. Nieto, "Unsupervised clustering of extreme vocal effects," in *Proc. 10th Int. Conf. Advances in Quantitative Laryngology*, 2013, p. 115.

[26] S. Gulati, "Computational approaches for melodic description in indian art music corpora," Ph.D. dissertation, Music Tech. Group, Universitat Pompeu Fabra, Barcelona, Spain, 2016.

[27] D. Gärtner, "Singing/rap classification of isolated vocal tracks," in *Proc. 11th Conf. Int. Society for Music Information Retrieval (ISMIR)*, 2010, pp. 519–524.

[28] M. Panteli, R. Bittner, J. P. Bello, and S. Dixon, "Towards the characterization of singing styles in world music," in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2017, pp. 636–640.

[29] A. Kumar, R. M. Bittner, N. Montecchio, A. Jansson, M. Panteli, E. J. Humphrey, and T. Jehan, "Learning a large-scale vocal similarity embedding for music," in *Proc. Machine Learning for Music Discovery Workshop, Int. Conf. Machine Learning*, 2017.

[30] W.-H. Tsai and H.-M. Wang, "Towards automatic identification of singing language in popular music recordings," in *Proc. 5th Conf. Int. Society for Music Information Retrieval (ISMIR)*, Barcelona, Spain, 2004, pp. 568–576.

[31] A. Loscos, P. Cano, and J. Bonada, "Low-delay singing voice alignment to text," in *Proc. Int. Computer Music Conf. (ICMC)*, 1999, pp. 437–440.

[32] V. Chandrasekhar, M. E. Sargin, and D. A. Ross, "Automatic language identification in music videos with low level audio and visual features," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2011, pp. 5724–5727.

[33] T. Nakano and M. Goto, "Vocarefiner: An interactive singing recording system with integration of multiple singing recordings," in *Proc. Conf. Sound and Music Computing (SMC)*, 2013, pp. 115–122.

[34] M. Kan, Y. Wang, D. Iskandar, T. L. Nwe, and A. Shenoy, "LyricAlly: Automatic synchronization of textual lyrics to acoustic music signals," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 2, pp. 338–349, 2008.

[35] H. Fujihara and M. Goto, "Lyrics-to-audio alignment and its applications," in *Multimodal Music Processing*, M. Müller, M. Goto, and M. Schedl, Eds. Schloss Dagstuhl, Germany: Dagstuhl Publishing, 2012, pp. 23–36.

[36] H. Fujihara, M. Goto, J. Ogata, and H. G. Okuno, "Lyricsynchronizer: Automatic synchronization system between musical audio signals and lyrics," *IEEE J. Sel. Topics Signal Process.*, vol. 5, no. 6, pp. 1252–1261, 2011.

[37] M. Mauch, H. Fujihara, and M. Goto, "Integrating additional chord information into HMM-based lyrics-to-audio alignment," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 200–210, 2012.

[38] T. Hosoya, M. Suzuki, A. Ito, and S. Makino, "Lyrics recognition from a singing voice based on finite state automaton for music information retrieval," in *Proc. 6th Conf. Int. Society for Music Information Retrieval (ISMIR)*, 2005, pp. 532–535.

[39] M. McVicar, D. P. W. Ellis, and M. Goto, "Leveraging repetition for improved automatic lyric transcription in popular music," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2014, pp. 3117–3121.

[40] A. M. Kruspe, "Bootstrapping a system for phoneme recognition and keyword spotting in unaccompanied singing," in *Late-Breaking Workshop, 17th Conf. Int. Society for Music Information Retrieval (ISMIR)*, New York, NY, 2016.

[41] H. Hirjee and D. G. Brown, "Solving misheard lyric search queries using a probabilistic model of speech sounds," in *Proc. 11th Conf. Int. Society for Music Information Retrieval (ISMIR)*, 2010, pp. 147–152.

[42] D. Ward, H. Wierstorf, R. Mason, E. M. Grais, and M. Plumbley, "BSS Eval or PEASS? Predicting the perception of singing-voice separation," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP)*, 2018, pp. 596–600.

[43] S. Dixon, M. Goto, and M. Mauch, "Why is voice interesting?" in *Proc. 16th Conf. Int. Society of Music Information Retrieval*, 2015.

SP