A quantum vocal theory of sound

Davide Rocchesso[1] · Maria Mannone[1]

Abstract

Concepts and formalism from acoustics are often used to exemplify quantum mechanics. Conversely, quantum mechanics could be used to achieve a new perspective on acoustics, as shown by Gabor studies. Here, we focus in particular on the study of human voice, considered as a probe to investigate the world of sounds. We present a theoretical framework that is based on *observables* of vocal production, and on some *measurement apparati* that can be used both for analysis and synthesis. In analogy to the description of spin states of a particle, the quantum-mechanical formalism is used to describe the relations between the fundamental states associated with phonetic labels such as phonation, turbulence, and supraglottal myoelastic vibrations. The intermingling of these states, and their temporal evolution, can still be interpreted in the Fourier/Gabor plane, and effective extractors can be implemented. The bases for a quantum vocal theory of sound, with implications in sound analysis and design, are presented.

Keywords Quantum-inspired algorithms · Audio processing · Sound representation

1 Introduction

What are the fundamental elements of sound? What is the most meaningful framework for analyzing existing sonic realities and for expressing new sound concepts? These are long-standing questions in sound physics, perception, and creation. In his analytical theory of heat [1], Joseph Fourier laid the basis for analyzing functions of one variable in terms of sinusoidal components and explicitly wrote that "…if the order which is established in these phenomena could be grasped by our senses, it would produce in us an impression comparable to the sensation of musical sounds."

Hermann von Helmholtz took Fourier's suggestion seriously and proceeded to analyze all vibratory phenomena as additions of sinusoidal vibrations [2]. Although he

✉ Davide Rocchesso
    davide.rocchesso@unipa.it

1   Department of Mathematics and Computer Science, University of Palermo, Palermo, Italy

admitted that "we can conceive a whole to be split into parts in very different and arbitrary ways," it was the observation that the ear somehow reflects Fourier analysis and can be described as a bank of sympathetic resonators that led him to state that "the existence of partial tones [...] acquire a meaning in nature."

In the twentieth century, despite the Fourier transform being the key to describe sampling and signal reconstruction from samples [3], skepticism arose among physicists such as Norbert Wiener and Dennis Gabor about considering Fourier analysis as the best representation for music [4]. In 1947, in a famous paper published in Nature [5], Gabor embraced the mathematics of quantum theory to shed light on subjective acoustics, thus laying the basis for sound analysis and synthesis based on acoustical quanta, or grains, or wavelets.

The Fourier and Gabor frameworks for time-frequency, or time-scale, representation of sound are widely used in the analysis and synthesis of sonic phenomena. For example, the auditory time and frequency acuities have been bounded in terms of the uncertainty principle, although the theoretical limit has been shown to beaten by human audition [6]. As another example, cochlear filters are designed so that their time-frequency behavior matches human performance and are used to simulate or replace human hearing [7].

Still, when we are imagining a sound, or describing it to peers, we do not use the Fourier formalism, but we rather refer to the hypothetical sources and to their characteristics [8], or we use our voice to mimic some salient sound features, thus overcoming the limitations of language [9]. We argue, therefore, that a description of sound that exploits the basic mechanisms of voice production would be more readily understandable and manipulable than any decomposition based on framed sines or on chirps.

In this contribution, we propose a phonetic approach to describe sound at large. A coarse articulatory description can indeed be applied to any sound, and it will provide the basis for attempting a vocal imitation, which makes embodied sound perception concrete and audible. In presence of concurrent sources, the high-level phonetic descriptors are superimposed and temporally varying, and their evolution is governed by context and attention. Apparently, our hearing system acts as a sort of destructive measurement apparatus which continuously collapses superpositions of phonetic states into streams [10], whose evolution we can single out and follow, with the possibility of jumping from one stream to another as a result of hidden or apparent forces.

Superposition and evolution of states, together with the concepts of measurement collapse and force fields, are among the cornerstones of quantum theory, and this is the observation that led us to attempt a description of sonic phenomena within a quantum framework. Hopefully, some phenomena that are normally described through sets of rules and gestalt principles (e.g., auditory continuity or temporal displacement) may naturally emerge from such a quantum-inspired description, similarly to how quantum cognition has been able to address behaviors that are difficult to derive within classical frameworks [11]. The apparent incompatibility of properties that are being judged, or of forms that are being perceived, implies some vagueness in the mental states and in their time evolution, which is difficult to model classically but is intrinsic in quantum modeling [12]. This is particularly evident in bivalued judgments and in

bistable percepts that can be modeled as a two-state quantum-mechanical system, or qubit. We intend to apply such a quantum-theoretical model, which is constructed in analogy with spins in a time-varying magnetic field, to auditory scenes made of overlapping auditory objects, described in phonetic terms. In the context of auditory scene analysis, we introduce the quantum-theoretical concepts of superposition, time evolution, and measurement (or foreground separation). We show how this framework can be useful to describe and reproduce some auditory-streaming phenomena, with possible applications in source separation and audio effects.

Section 2 provides a short background on prior research on the two main axes that cross in this work: research in sound objecthood, with special emphasis on the voice as an embodied representation of sound; quantum frameworks that have been proposed for sound and image processing, music, and perception. Section 3 gives the motivation and a compact overview of the proposed quantum vocal theory of sound. The long Sect. 4 recalls the basic mathematical formalism and some key concepts of quantum theory, and it shows how these tools and concepts can be recast in audio terms. Section 5 shows how quantum evolution can inspire algorithms for auditory object streaming and separation, thus pointing to possible applications in computational auditory scene analysis and audio effects.

2 Background

2.1 Voice as embodied sound

Many researchers, in science, art, and philosophy, have been facing the problem of how to approach sound and its representations [13,14]. Should we represent sounds as they appear to the senses, by manipulating their proximal characteristics? Or should we rather look at potential sources, at physical systems that produce sound as a side effect of distal interactions? In this research path, we assume that our body can help establish bridges between distal (source-related) and proximal (sensory-related) representations, and we look at research findings in perception, production, and articulation of sounds [15,16]. Our approach to sound [17,18] seeks to exploit knowledge in these areas, especially referring to human voice production as a form of embodied representation of sound.

When considering what people hear from the environment, it emerges that sounds are mostly perceived as belonging to categories of the physical world [8]. Research in sound perception has shown that listeners spontaneously create categories such as solid, electrical, gas, and liquid sounds, even though the sounds within these categories may be acoustically different [19]. However, when the task is to separate, distinguish, count, or compose sounds, the attention shifts from *sounding objects* to *auditory objects* [20] represented in the time-frequency plane, or to *auditory images*, which are movie-like temporal representations resembling the signals projected by the ear up to the auditory cortex [7]. Tonal components, noise, and transients can be extracted from auditory objects with Fourier-based techniques [21–23]. Low-frequency periodic phenomena are also perceptually very relevant and often come as trains of transients. The most prominent elements of the proximal signal may be selected by simplification

and inversion of time-frequency representations. These auditory sketches [24] have been used to test the recognizability of imitations [25].

When discussing spaces for sound representation, it is also important to recall the notion of *sound object*, often associated with Schaeffer's theory of listening and typo-morphological spaces, which support a phenomenological description of sound and can be reported to the time-frequency plane [16]. For example, the concept of mass is a generalization of the notion of pitch that comprises both site (on the frequency axis) and caliber (or degree of occupation of the frequency axis).

Vocal imitations can be more effective than verbalizations at representing and communicating sounds when these are difficult to describe with words [9]. This indicates that vocal imitations can be a useful tool for investigating sound perception and shows that the voice is instrumental to embodied sound cognition. Vocal imitations act similarly to visual sketches: They catch and emphasize some essential elements of the original (visual) objects allowing their identification. At a more fundamental level, research on non-speech vocalization is affecting the theories of language evolution [26], as it seems plausible that humans could have used iconic vocalizations to communicate with a large semantic spectrum, prior to the establishment of full-blown spoken languages. Experiments and sound design exercises [17] show that agreement in production corresponds to agreement in meaning interpretation, thus showing the effectiveness of teamwork in embodied sound creation. Converging evidence from behavioral and brain imaging studies give a firm basis to hypothesize a shared representation of sound in terms of motor (vocal) primitives [27]. Historically, such convergence was envisioned over a century ago by the Italian Futurists: On one side, the composer Luigi Russolo developed an organology of everyday sounds and devised mechanical synthesizers for these "noises" [28]; on the other side, the poet Filippo Tommaso Marinetti devised a way to transcend language to bring everyday sounds to poetry, through imitations and onomatopoeia [29].

Some phoneticians have turned their attention to non-speech voice production, trying to identify the most relevant phonetic components that are found in vocal imitations [30]. They identified the broad categories of phonation (i.e., quasi-periodic oscillations due to vocal fold vibrations), turbulence, supraglottal myoelastic vibrations, and clicks, which can be extracted automatically from audio with time-frequency analysis and supervised [31] or unsupervised [32] machine learning. These categories can be made to correspond to categories of sounds as they are perceived [33], and as they are produced in the physical world. Indeed, it has been argued that human utterances somehow mimic "nature's phonemes" [34], and neurophysiological studies have shown that the cortical area of the superior temporal gyrus actually encodes abstract phonetic features [35].

2.2 Quantum frameworks

It was Dennis Gabor [5] who first adopted the mathematics of quantum mechanics to explain acoustic phenomena. In particular, he used operator methods to derive the time-frequency uncertainty relation and the (Gabor) function that satisfies minimal uncertainty. Time-scale representations [36] are more suitable to explain the perceptual

decoupling of pitch and timbre, and operator methods can be used as well to derive the gammachirp function, which minimizes uncertainty in the time-scale domain [37]. Research in human and machine hearing [7] has been based on banks of elementary (filter) functions, and these systems are at the core of many successful applications in the audio domain.

Despite its deep roots in the physics of the twentieth century, the sound field has not yet embraced the quantum signal-processing framework [38] to seek practical solutions to sound scene representation, separation, and analysis, although some theoretical proposals to encode, store, and process audio using quantum circuitry have been advanced [39,40]. On the other hand, some common observed properties of human cognition and quantum mechanics (superposition, non-classical probability) have given universal value to the quantum-theoretical formalism to explain cognitive acts [11], including actions of human creation, such as music. The explanatory power of a quantum approach to music cognition has been demonstrated to describe tonal attraction phenomena in terms of metaphorical forces [41,42]. The theory of open quantum systems has been applied to music to describe the memory properties (non-Markovianity) of different scores [43]. The time-dependent Schrödinger equation for a single non-relativistic particle has been used as a model for sound and music composition. Some examples include the creation of grain clouds like orbitals [44], the sonification of controlled quantum dynamics [45], and compositions for an ensemble of atoms [46]. It has even been claimed that the interplay between musical ideas and extra-musical meanings can be naturally represented in the framework of quantum semantics, where extra-musical meanings can be treated within a theory of vague possible worlds [47].

Some theoretical physicists have looked at the sensory processes driving human and animal perception, trying to understand if they are classical or quantum. As far as visual perception is concerned, Ghirardi proposed an experiment to verify if the perceptive apparatus can induce the suppression of a physically established superposition of states [48]. In application-oriented image processing, on the other hand, it has been shown how the quantum framework can be effective to solve problems such as segmentation. For example, the separation of figures from background can be obtained by evolving a solution of the time-dependent Schrödinger equation [49], or by discretizing the time-independent Schrödinger equation [50]. An approach to signal manipulation based on the postulates of quantum mechanics can also potentially lead to a computational advantage when using quantum processing units. Results in this direction are being reported for optimization problems [51].

In this work, we consider auditory phenomena and look at quantum theory for a possible process model that somehow mirrors the way humans extract and follow auditory objects from audio mixtures. Such a process model, that exploits our embodied knowledge of sound via vocal production, does not assume any underlying information processing model for the brain. This standpoint and disclaimer is commonly assumed in quantum cognition [11] and readily adopted here.

3 Sketch of a quantum vocal theory of sound

In the proposed research path, sound is treated as a superposition of states, and the voice-based components (phonation, turbulence, supraglottal myoelastic vibrations) are considered as observables to be represented as operators. The extractors of the fundamental components, i.e., the measurement apparati, are implemented as signal-processing modules that are available both for analysis and, as control knobs, for synthesis. The baseline is found in the results of the SkAT-VG project [9,17,25,31,33,52], which showed that vocal imitations are optimized representations of referent sounds that emphasize those features that are important for identification. A large collection of audiovisual recordings of vocal and gestural imitations[1] offers the opportunity to further enquire how people perceive, represent, and communicate about sounds.

A first assumption underlying this research approach, largely justified by prior art and experiences, is that articulatory primitives used to describe vocal utterances are effective as high-level descriptors of sound in general. This assumption leads naturally to an embodied approach to sound representation, analysis, and synthesis.

A second assumption is that the mathematics of quantum mechanics, relying on linear operators in Hilbert spaces, offers a formalism that is suitable to describe the objects composing auditory scenes and their evolution in time. The latter assumption is more adventurous, as this path has not been taken in audio signal processing yet. However, the results coming from neighboring fields (music cognition, image processing) encourage us to explore this direction and to aim at introducing new techniques for sound analysis, synthesis, and transformation.

An embryonic theory of sound based on the postulates of quantum mechanics, and using high-level vocal descriptors of sound, can be sketched as follows. Let $\overline{\sigma}$ be a vector operator that provides information about the phonetic elements along a specific direction of measurement. Phonation, for example, may be represented by σ_z, with eigenstates representing a upper and a lower pitch. Similarly, the turbulence component may be represented by σ_x, with eigenstates representing turbulence of two different spectral distributions. A measurement of turbulence prepares the *system* in one of two eigenstates for operator σ_x, and a successive measurement of phonation would find a superposition and get equal probabilities for the two eigenstates of σ_z. The two operators σ_z and σ_x may also be made to correspond to the two components of the classic sines + noise model used in audio signal processing. If we add transients/clicks as a third measurement direction (as in the sines + noise + transients model [22]), we can claim that there is no sound state for which the expectation value of the three components is zero: a sort of spin polarization principle as found in quantum mechanics. The evolution of state vectors in time is unitary, and regulated by a time-dependent Schrödinger equation, with a suitably chosen Hamiltonian. The eigenvectors of the Hamiltonian allow to expand any state vector in that basis and to compute the time evolution of such expansion. A pair of components can be simultaneously measured only if they commute. If they do not, an uncertainty principle can be derived, as it

---

[1] https://www.ircam.fr/projects/blog/multimodal-database-of-vocal-and-gestural-imitations-elicited-by-sounds/.

was done for time-frequency and time-scale representations [5,37]. The theory can be extended to cover multiple uncertain sources, and the resulting mixed states can be described via density matrices, whose time evolution can also be computed if a Hamiltonian operator is properly defined. In the following, we formally lay down this quantum vocal theory of sound.

4 The phon formalism

Consider a 3D space with the orthogonal axes

- z: phonation, with different pitches;
- x: turbulence, with different brightnesses;
- y: myoelasticity, slow pulsations with different tempos.

The labels attributed to the axes correspond to the three main articulatory/phonatory categories that are used by phoneticians to annotate vocal imitations of everyday sounds [30]. They are a simplification of the more phonetically correct labels "vocal fold phonation," "turbulence," and "supraglottal myoelastic vibration" [31].

The phon operator $\overline{\sigma}$ is a 3-vector operator that provides information about the phonetic component in a specific direction of the 3D phonetic space, i.e., along a specific combination of phonation, turbulence, and myoelasticity.

In this section, we present the phon formalism, obtained by direct analogy with the single spin, as presented in accessible presentations of quantum mechanics [53]. We use standard Dirac notation and adopt the quantum-theoretical concepts of measurement, preparation, pure and mixed states, uncertainty, and time evolution [54].

4.1 Measurement along *z*

A measurement along the z-axis is performed according to the quantum mechanics principles:

1. Each component of $\overline{\sigma}$ is represented by a linear operator;
2. The eigenvectors of σ_z are $|u\rangle$ and $|d\rangle$, corresponding to pitch-up and pitch-down, with eigenvalues $+1$ and -1, respectively:

    (a) $\sigma_z|u\rangle = |u\rangle$
    (b) $\sigma_z|d\rangle = -|d\rangle$

3. The eigenstates of operator σ_z, $|u\rangle$, and $|d\rangle$ are orthogonal: $\langle u|d\rangle = 0$;

The eigenstates can be represented as column vectors

$$|u\rangle = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \ |d\rangle = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

and the operator σ_z as a square 2×2 matrix. Due to principle 2, we have

$$\sigma_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}. \tag{1}$$

4.2 Preparation along x

The eigenstates of the operator σ_x are $|r\rangle$ and $|l\rangle$, corresponding to turbulences having different spectral distributions, one with the rightmost (or highest frequency) centroid and the other with the leftmost centroid. The respective eigenvalues are $+1$ and -1, so that

(a) $\sigma_x|r\rangle = |r\rangle$
(b) $\sigma_x|l\rangle = -|l\rangle$.

If the phon is prepared $|r\rangle$ (turbulent), and then, the measurement apparatus is set to measure σ_z, there will be equal probabilities for $|u\rangle$ or $|d\rangle$ phonation as an outcome. Essentially, we are measuring what kind of phonation is in a pure turbulent state. This measurement property is satisfied if

$$|r\rangle = \frac{1}{\sqrt{2}}|u\rangle + \frac{1}{\sqrt{2}}|d\rangle. \tag{2}$$

Likewise, if the phon is prepared $|l\rangle$, and then, the measurement apparatus is set to measure σ_z, there will be equal probabilities for $|u\rangle$ or $|d\rangle$ phonation as an outcome. This measurement property is satisfied if

$$|l\rangle = \frac{1}{\sqrt{2}}|u\rangle - \frac{1}{\sqrt{2}}|d\rangle, \tag{3}$$

which is orthogonal to the linear combination (2). In vector form, we have

$$|r\rangle = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}, \ |l\rangle = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix}, \ \text{and}$$

$$\sigma_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}. \tag{4}$$

In fact, any state $|A\rangle$ can be expressed as

$$|A\rangle = \alpha_u|u\rangle + \alpha_d|d\rangle, \tag{5}$$

where $\alpha_u = \langle u|A\rangle$, and $\alpha_d = \langle d|A\rangle$. Being the system in state $|A\rangle$, the probability to measure pitch-up is

$$p_u = \langle A|u\rangle\langle u|A\rangle = \alpha_u{}^*\alpha_u, \tag{6}$$

and similarly, the probability to measure pitch-down is $p_d = \langle A|d\rangle\langle d|A\rangle = \alpha_d{}^*\alpha_d$ (Born rule).

4.3 Preparation along y

The eigenstates of the operator σ_y are $|f\rangle$ and $|s\rangle$, corresponding to slow myoelastic pulsations, one faster and one slower[2], with eigenvalues $+1$ and -1, so that

(a) $\sigma_y|f\rangle = |f\rangle$
(b) $\sigma_y|s\rangle = -|s\rangle$.

If the phon is prepared $|f\rangle$ (pulsating), and then, the measurement apparatus is set to measure σ_z, there will be equal probabilities for $|u\rangle$ or $|d\rangle$ phonation as an outcome. Essentially, we are measuring what kind of phonation is in a myoelastic pulsations. This measurement property is satisfied if

$$|f\rangle = \frac{1}{\sqrt{2}}|u\rangle + \frac{i}{\sqrt{2}}|d\rangle, \tag{7}$$

where i is the imaginary unit.

Likewise, if the phon is prepared $|s\rangle$, we can express this state as

$$|s\rangle = \frac{1}{\sqrt{2}}|u\rangle - \frac{i}{\sqrt{2}}|d\rangle, \tag{8}$$

which is orthogonal to the linear combination (7). In vector form, we have

$$|f\rangle = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{i}{\sqrt{2}} \end{bmatrix}, \; |s\rangle = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{i}{\sqrt{2}} \end{bmatrix}, \text{ and}$$

$$\sigma_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}. \tag{9}$$

The matrices (1), (4), and (9) are called the Pauli matrices, and together with the identity matrix, these are the quaternions.

4.4 Measurement along an arbitrary direction

Orienting the measurement apparatus along an arbitrary direction $\bar{n} = [n_x, n_y, n_z]'$ means taking a weighted mixture of quaternions:

$$\sigma_n = \bar{\sigma} \cdot \bar{n} = \sigma_x n_x + \sigma_y n_y + \sigma_z n_z = \begin{bmatrix} n_z & n_x - i n_y \\ n_x + i n_y & -n_z \end{bmatrix}. \tag{10}$$

---

[2] In describing the spin eigenstates, the symbols $|i\rangle$ and $|o\rangle$ are often used, to denote the in–out direction.

4.4.1 Example: harmonic plus noise model

A measurement performed by means of a Harmonic plus Noise model [21] would lie in the phonation–turbulence plane ($n_z = \cos\theta, n_x = \sin\theta, n_y = 0$), so that

$$\sigma_n = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix} \tag{11}$$

The eigenstate for eigenvalue $+1$ is

$$|\lambda_1\rangle = [\cos\theta/2, \sin\theta/2]', \tag{12}$$

the eigenstate for eigenvalue -1 is

$$|\lambda_{-1}\rangle = [-\sin\theta/2, \cos\theta/2]', \tag{13}$$

and the two are orthogonal. Suppose we prepare the phon to pitch-up $|u\rangle$. If we rotate the measurement system along $\bar{n}$, the probability to measure $+1$ is (by Born rule)

$$p(+1) = |\langle u|\lambda_1\rangle|^2 = \cos^2\theta/2, \tag{14}$$

and the probability to measure -1 is

$$p(-1) = |\langle u|\lambda_{-1}\rangle|^2 = \sin^2\theta/2. \tag{15}$$

The expectation value of measurement is therefore

$$\langle\sigma_n\rangle = \sum_j \lambda_j p(\lambda_j) = (+1)\cos^2\theta/2 + (-1)\sin^2\theta/2 = \cos\theta. \tag{16}$$

4.4.2 Rotate to measure

What does it mean to rotate a measurement apparatus to measure a property? Assume we have a machine that separates harmonics from noise from (trains of) transients and that can discriminate between two different pitches, noise distributions, and tempos. Essentially, the machine receives a sound and returns three numbers {ph, tu, my} $\in$ $[-1, 1]$. If ph > 0, the result will be $|u\rangle$, and if ph < 0, the result will be $|d\rangle$. If tu > 0, the result will be $|r\rangle$, and if tu < 0, the result will be $|l\rangle$. If my > 0, the result will be $|f\rangle$, and if my < 0, the result will be $|s\rangle$. These three outputs correspond to rotating the measurement apparatus along each of the main axes. Rotating it along an arbitrary direction means taking a weighted mixture of the three outcomes.

For example, consider the vocal fragment[3] whose spectrogram is represented in Fig. 1. An extractor of pitch salience can be used to measure phonation, and an extractor of onsets can be used to measure slow myoelastic pulsation. Such two feature

---

[3] It is one of the example vocal sounds considered in [55] and taken from [56].
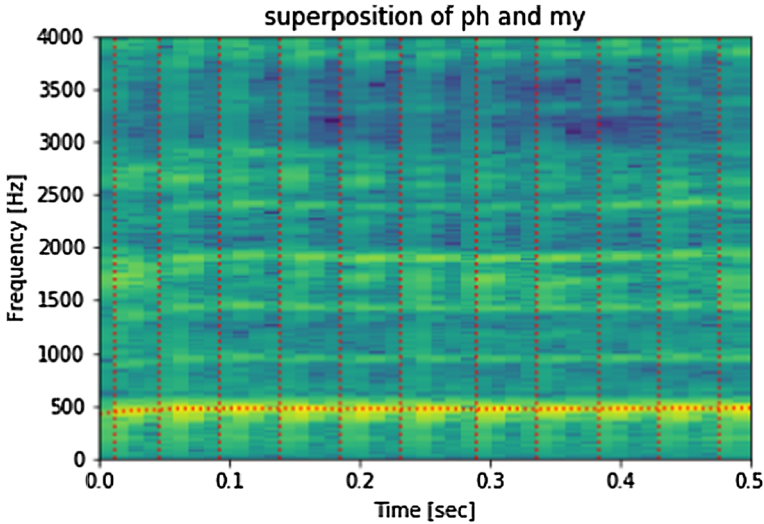
Fig. 1 Spectrogram of a vocal sound which is a superposition of phonation and supraglottal myoelastic vibration. A salient pitch (horizontal dotted line) as well as quasi-regular train of pulses (vertical dotted lines) is automatically extracted

extractors, as found in the Essentia library [57], have been applied to highlight the phonation (horizontal dotted line) and myoelastic (vertical dotted lines) components in the spectrogram of Fig. 1. In the $z - y$ plane, there would be a measurement orientation and a measurement operator that admits such sound as an eigenvector.

4.5 Pure and mixed states

According to the first postulate of quantum mechanics [54], at each time instant the system is completely specified by a state $|\psi\rangle$ such that $\langle\psi|\psi\rangle = 1$. If the state is known with certainty, it is called a pure state. All the phon states described so far are pure states. More generally, a state can be known probabilistically as one of a set of $|\psi_i\rangle$ with a given probability distribution. States of such kind are called mixed states. The density operator represents both pure and mixed states, and it is defined as

$$\rho = \sum_j p_j |\psi_j\rangle\langle\psi_j|, \tag{17}$$

where p_j is the probability for state $|\psi_j\rangle$.

For a pure state, it is simply $\rho = |\psi\rangle\langle\psi|$, and the trace of the square of such matrix is $Tr[\rho^2] = 1$. For a mixed state, it is always the case that $Tr[\rho^2] < 1$.

4.5.1 Example

Let state $|u\rangle$ with probability $\frac{1}{3}$ and state $|d\rangle$ with probability $\frac{2}{3}$. The density matrix is

$$\rho = \frac{1}{3}|u\rangle\langle u| + \frac{2}{3}|d\rangle\langle d| = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{2}{3} \end{bmatrix}, \tag{18}$$

and the trace of its square is

$$Tr[\rho^2] = \frac{5}{9} < 1.$$

The interest of the density operator is given by its generalization power. It is an essential generalization in quantum mechanics, and as such, it is relevant for a quantum vocal theory of sound. From an experimental point of view, it introduces a degree of conceptual flexibility which may come useful in synthesis and composition of auditory scenes. In particular, the audio concept of mixing can be made to correspond with manipulation of mixed states.

4.6 Uncertainty

If we measure two observables $\mathbf{L}$ and $\mathbf{M}$ (in a single experiment) simultaneously, quantum mechanics prescribes that the system is left in a simultaneous eigenvector of the observables only if $\mathbf{L}$ and $\mathbf{M}$ commute, i.e., if their commutator $[\mathbf{L}, \mathbf{M}] = \mathbf{LM} - \mathbf{ML}$ is null. Measurement operators along different axes do not commute. For example, $[\sigma_x, \sigma_y] = 2i\sigma_z$, and therefore, phonation and turbulence cannot be simultaneously measured with certainty.

The uncertainty principle, based on Cauchy–Schwarz inequality in complex vector spaces, prescribes that the product of the two uncertainties is at least as large as half the magnitude of the commutator:

$$\Delta\mathbf{L}\Delta\mathbf{M} \geq \frac{1}{2}|\langle\psi|[\mathbf{L}, \mathbf{M}]|\psi\rangle| \tag{19}$$

If $\mathbf{L} = \mathscr{T} = t$ is the time operator and $\mathbf{M} = \mathscr{W} = -i\frac{d}{dt}$ is the frequency operator, and these are applied to the complex oscillator $Ae^{i\omega t}$, the time-frequency uncertainty principle results and uncertainty is minimized by the Gabor function. Starting from the scale operator, the gammachirp function can be derived [37].

4.7 Time evolution

Another postulate of quantum mechanics [54] states that the evolution of state vectors in time

$$|\psi(t)\rangle = \mathbf{U}(t_0, t)|\psi(t_0)\rangle, t > t_0, \tag{20}$$

is governed by the operator $\mathbf{U}$, which is unitary (i.e., $\mathbf{U}^\dagger \mathbf{U} = \mathbf{I}$) and depends only on t_0 and t. Taken a small time increment ϵ, continuity of the time-development operator gives it the form

$$\mathbf{U}(\epsilon) = \mathbf{I} - i\epsilon\mathbf{H}, \tag{21}$$

with $\mathbf{H}$ being the quantum Hamiltonian (Hermitian) operator. $\mathbf{H}$ is an observable, and its eigenvalues are the values that would result from measuring the energy of a quantum system. From (21), it turns out that a state vector changes in time according to the time-dependent Schrödinger equation[4]

$$\frac{\partial|\psi(t)\rangle}{\partial t} = -i\mathbf{H}(t)|\psi(t)\rangle. \tag{22}$$

Any observable $\mathbf{L}$ has an expectation value $\langle\mathbf{L}\rangle$ that evolves according to

$$\frac{\partial\langle\mathbf{L}\rangle}{\partial t} = -i\langle[\mathbf{L}, \mathbf{H}]\rangle, \tag{23}$$

where $[\mathbf{L}, \mathbf{H}]$ is the commutator of $\mathbf{L}$ with $\mathbf{H}$.

For a closed, isolated physical system, the Hamiltonian $\mathbf{H}$ is time independent ($\mathbf{H}(t) = \mathbf{H}$), and the unitary operator is $\mathbf{U}(t_0, t) = \mathbf{U}(t - t_0) = e^{-i\mathbf{H}(t-t_0)}$. While evolving, a closed system remains in a superposition of states and preserves their magnitudes and relative angles.

For non-pure states, the evolution of density operators is

$$\rho(t) = \mathbf{U}^\dagger(t_0, t)\rho(t_0)\mathbf{U}(t_0, t). \tag{24}$$

In most physical as well as in audio applications, we have that the system under consideration is driven by external forces, such as a changing magnetic field or a vocal gestural articulation. In such cases of closed non-isolated systems [58], the Hamiltonian $\mathbf{H}$ is time dependent. The states change under the effect of the external forces, which determine the change of probabilities, and the Hamiltonian controls the evolution process.

With a commutative Hamiltonian ($[\mathbf{H}(0), \mathbf{H}(t)] = 0$), the time evolution can be expressed as

$$|\psi(t)\rangle = e^{-i\int_0^t \mathbf{H}(\tau)d\tau}|\psi(0)\rangle = \mathbf{U}(0, t)|\psi(0)\rangle. \tag{25}$$

In general, if the operators $\mathbf{A}$ and $\mathbf{B}$ do not commute (i.e., $[\mathbf{A}, \mathbf{B}] \neq 0$), we have that $e^{\mathbf{A}}e^{\mathbf{B}} \neq e^{\mathbf{A}+\mathbf{B}}$. Since the evolution between two time points 0 and t can be split at an intermediate time t^*, if $e^{-i\int_0^t \mathbf{H}(\tau)d\tau} = e^{-i\int_0^{t^*} \mathbf{H}(\tau)d\tau - i\int_{t^*}^t \mathbf{H}(\tau)d\tau} \neq e^{-i\int_0^{t^*} \mathbf{H}(\tau)d\tau}e^{-i\int_{t^*}^t \mathbf{H}(\tau)d\tau}$, then it means that an explicit solution in terms of an integral cannot be found. Our approach is to consider time segments where the Hamiltonian is

---

[4] We do not need physical dimensional consistency here, so we drop Planck's constant.

locally commutative and to compute the time evolution segment by segment in terms of an integral.

4.7.1 Phon in utterance field

Similarly to a spin in a magnetic field, when a phon is part of an utterance, it has an energy that depends on its orientation. We can think about it as if it was subject to restoring forces, and its quantum Hamiltonian is

$$\mathbf{H} \propto \overline{\sigma} \cdot \overline{B} = \sigma_x B_x + \sigma_y B_y + \sigma_z B_z, \tag{26}$$

where the components of the field $\overline{B}$ are named in analogy with the magnetic field.

Consider the case of potential energy only along z:

$$\mathbf{H} = \frac{\omega}{2} \sigma_z. \tag{27}$$

To find how the expectation value of the phon varies in time, we expand the observable $\mathbf{L}$ in (23) in its components to get

$$\begin{aligned}
\langle \dot{\sigma}_x \rangle &= -i \langle [\sigma_x, \mathbf{H}] \rangle = -\omega \langle \sigma_y \rangle \\
\langle \dot{\sigma}_y \rangle &= -i \langle [\sigma_y, \mathbf{H}] \rangle = \omega \langle \sigma_x \rangle \\
\langle \dot{\sigma}_z \rangle &= -i \langle [\sigma_z, \mathbf{H}] \rangle = 0,
\end{aligned} \tag{28}$$

which means that the expectation values of σ_x and σ_y are subject to temporal precession around z at angular velocity ω. In phon terms, the expectation value of σ_z steadily keeps the pitch if there is no potential energy along turbulence and myoelastic pulsation.

A potential energy along all three axes can be expressed as

$$\mathbf{H} = \frac{\omega}{2} \overline{\sigma} \cdot \overline{n} = \frac{\omega}{2} \begin{bmatrix} n_z & n_x - i n_y \\ n_x + i n_y & -n_z \end{bmatrix}, \tag{29}$$

whose energy eigenvalues are $E_j = \pm \frac{\omega}{2}$, with energy eigenvectors $|E_j\rangle$.

An initial state vector (phon) $|\psi(0)\rangle$ can be expanded in the energy eigenvectors as

$$|\psi(0)\rangle = \sum_j \alpha_j(0) |E_j\rangle, \tag{30}$$

where $\alpha_j(0) = \langle E_j | \psi(0) \rangle$, and the time evolution of state turns out to be

$$|\psi(t)\rangle = \sum_j \alpha_j(t) |E_j\rangle = \sum_j \alpha_j(0) e^{-iE_j t} |E_j\rangle. \tag{31}$$

4.8 Measurement

Given that time evolution of states is governed by the unitary transformation (20) and by the Schrödinger Eq. (22), the measurement postulate of quantum mechanics [54] states that a measurement is represented by an operator (a projector) that acts on the state and that causes its collapse onto one of its eigenvectors.

A projector system Π_i in the (Hilbert) space of states is Hermitian, idempotent, and complete. If the system is in state $|\psi\rangle$ before measurement, the probability that the outcome of a measurement through a projector system returns j is

$$p_m(j|\psi) = \langle\psi|\Pi_j|\psi\rangle, \tag{32}$$

and as a result of the measurement, the system collapses in state $\psi_{post}^{(j)} = \frac{\Pi_j|\psi\rangle}{\sqrt{p_m(j|\psi)}}$.

Given an orthonormal basis of measurement vectors $|a_j\rangle$, the elementary projectors are $\Pi_j = |a_j\rangle\langle a_j|$, $p_m(j|\psi) = |\langle\psi|a_j\rangle|^2$, and the system (by neglecting a unitary phasor) collapses into $\psi_{post}^{(j)} = |a_j\rangle$.

If the system is in a pure state,

$$p_m(j|\psi) = \langle\psi \mid \Pi_j \mid \psi\rangle = Tr[\rho\Pi_j]. \tag{33}$$

If the system is in a mixed state, the outcome of measurement is formulated as a random variable conditioned by a given state:

$$p_m(j|\psi_k) = \langle\psi_k \mid \Pi_j \mid \psi_k\rangle = Tr[|\psi_k\rangle\langle\psi_k|\Pi_j], \tag{34}$$

and by averaging over all components of the mixed state, we get

$$p_m(j|\rho) = \sum_k p_k p_m(j|\psi_k) = Tr[\rho\Pi_j]. \tag{35}$$

If the outcome of measurement is j, the system collapses into the new ensemble of states represented by the density operator

$$\rho_{post}^{(j)} = \frac{\Pi_j\rho\Pi_j}{Tr[\rho\Pi_j]}. \tag{36}$$

4.9 Audio measurement and evolution

The mathematics of quantum mechanics can be used to describe and develop some operations of audio signal processing, aimed at segregating components or streams from raw audio. The concepts of quantum measurement and temporal evolution of quantum states can be recast in audio and phonetic terms if we can rely on an audio analysis/synthesis system that permits the extraction and manipulation of slowly varying features such as pitch salience or spectral energy.

Fig. 2 A non-commutative
diagram representing the
non-commutativity of
measurements of phonation (P)
and turbulence (T) on audio A
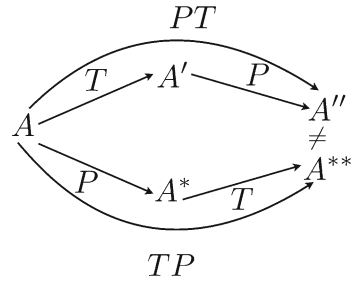
4.9.1 Non-commutativity and autostates

We expect that measurement operators along different axes do not commute: This is
the case, for example, of measurements of phonation and turbulence. Let A be an audio
segment. The measurement (by extraction) of turbulence by the operator T leads to
$T(A) = A'$. A successive measurement of phonation by the operator P gives $P(A') =
A''$; thus, $P(A') = PT(A) = A''$. If we perform the measurements in the opposite
order, with phonation first and turbulence later, we obtain $TP(A) = T(A^*) = A^{**}$.
We expect that $[T, P] \neq 0$, and thus, that $A^{**} \neq A''$. The diagram in Fig. 2 shows
non-commutativity in the style of category theory.

Besides the compact diagrammatic representation, we can describe such a non-
commutativity in terms of projectors Π_T, Π_P:

$$\begin{aligned}
\Pi_T(\Pi_P|A\rangle) &= |T\rangle\langle T|P\rangle\langle P|A\rangle = \langle T|P\rangle|T\rangle\langle P|A\rangle \neq \\
\Pi_P(\Pi_T|A\rangle) &= |P\rangle\langle P|T\rangle\langle T|A\rangle = \langle P|T\rangle|P\rangle\langle T|A\rangle.
\end{aligned} \tag{37}$$

Given that $\langle T|P\rangle$ is a scalar and $\langle P|T\rangle$ is its complex conjugate, and that $|P\rangle\langle T|$ is
generally non-Hermitian, we get

$$\begin{aligned}
[\Pi_T, \Pi_P] &= |T\rangle\langle T|P\rangle\langle P| - |P\rangle\langle P|T\rangle\langle T| \\
&= \langle T|P\rangle|T\rangle\langle P| - \langle P|T\rangle|P\rangle\langle T| \neq 0.
\end{aligned} \tag{38}$$

Measurements of phonation and turbulence can be actually performed using the
sines + noise (a.k.a., Harmonic Plus Stochastic—HPS) model [21]. The order of oper-
ations is visually described in Fig. 3. The measurement of phonation is performed
through the extraction of the harmonic component in the HPS model, while the mea-
surement of turbulence is performed through the extraction of the stochastic component
with the same model. The spectrograms for A'' and A^{**} in Fig. 4 show the results of
such two sequences of analyses on a segment of female speech,[5] confirming that the
commutator $[T, P]$ is nonzero.

Essentially, if we adopt the HPS model and skip the final step of addition and inverse
transformation, we are left with something that is conceptually equivalent to a quantum

---

[5] https://freesound.org/s/317745/. Hann window of 2048 samples, FFT of 4096 samples, hop size of 1024
samples.
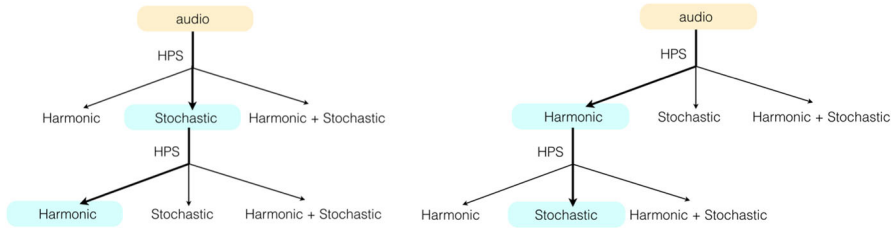
Fig. 3 On the left, an audio segment is analyzed via the HPS model. Then, the stochastic part is submitted to a new analysis. In this way, a measurement of phonation follows a measurement of turbulence. On the right, the measurement of turbulence follows a measurement of phonation. This can be described via projectors through Eq. (37), and diagrammatically in Fig. 2
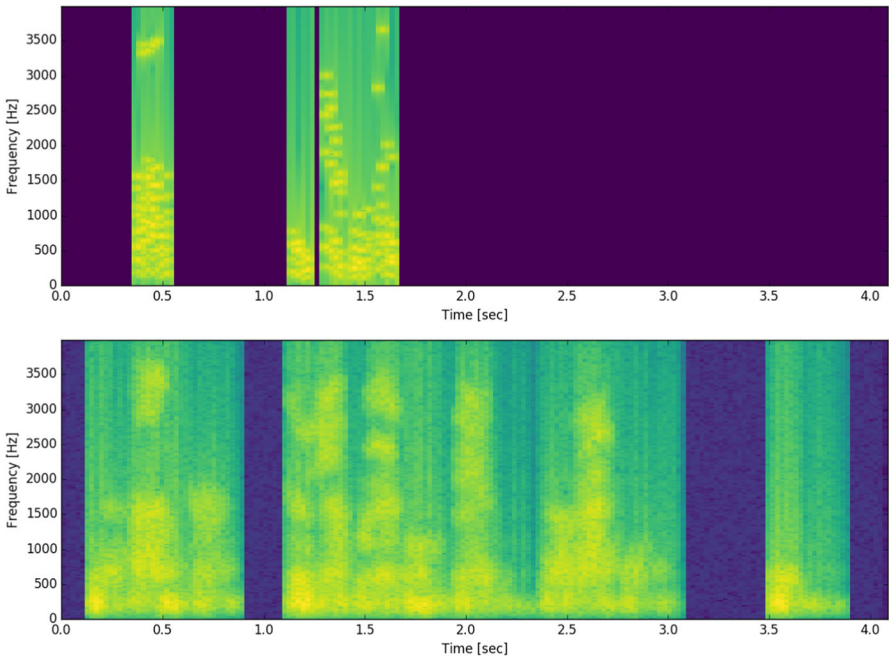
Fig. 4 On the top, the spectrogram corresponding to a measurement of phonation P following a measurement of turbulence T, leading to $PT(A) = A''$. On the bottom, the spectrogram corresponding to a measurement of turbulence T following a measurement of phonation P, leading to $TP(A) = A^{**}$

destructive measure. Let St be the filter that extracts the stochastic part from a signal. As Fig. 5 shows, the spectrogram of $St(x)$ is visibly different from the spectrogram of x. Conversely, if we apply St once more, we get a spectrum that does not change much: $St^2(x) = St(St(x)) \sim St(x)$. If we transform back from the second and third spectrograms of Fig. 5, we get sounds that are very close to each other. In fact, ideally, $St^2(x) = St(x)$. It means that, after a measure of the non-harmonic component of some signal, the output signal can be considered as an autostate, and it confirms that the projection operator is idempotent. If we perform the measure again and again, we still get the same result. Such a measure operation provokes the collapse of a
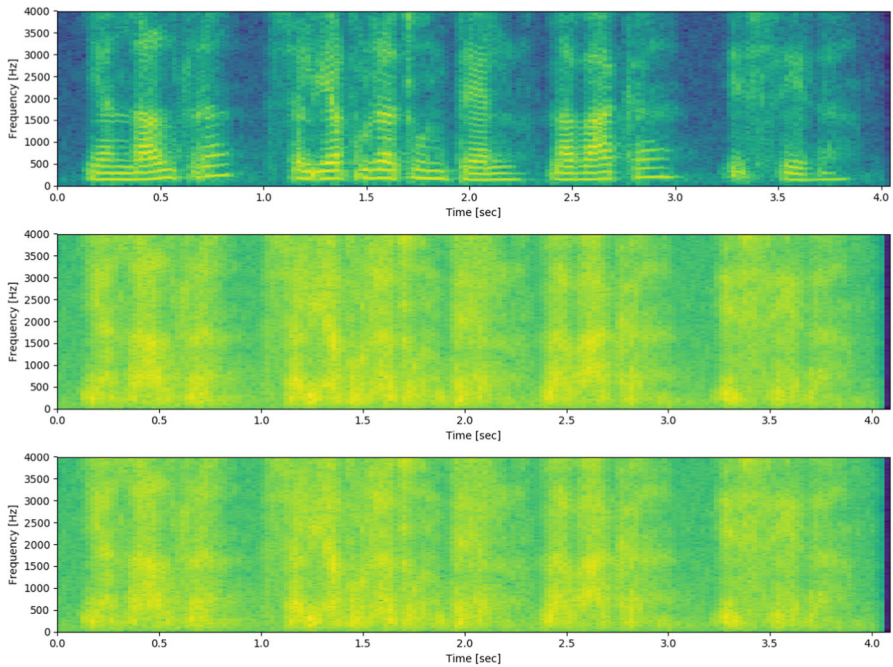
Fig. 5 Top: spectrum of the original sound signal (a female speech), center: the stochastic component, derived from harmonic plus stochastic analysis (HPS), as the effect of a destructive measure, and bottom: the stochastic component of the stochastic component itself. The last two spectra are very close

hypothetical underlying wave function, which is originally a superposition of states, and is reduced to a single state upon measurement. The importance of the autostates in this framework is connected with the concept of quantum measures, which may become practically feasible through a set of audio signal analysis tools.

4.9.2 Hamiltonian streaming

Let us consider a quantum state vector $|\psi(t)\rangle$ that evolves in time according to the Schrödinger Eq. (22). The time evolution can be represented by the unitary operator $\mathbf{U}(t_0, t)$ of Eq. (20).

If we choose a particular, commutative Hamiltonian, the time evolution can be expressed by an integral, as in Eq. (25). A time-independent Hamiltonian such as the one leading to (31) would not be very useful, both because forces indeed change continuously and because this would lead to oscillatory solution. Similarly to what has been done by Youssry et al. [49], the Hamiltonian can be chosen to be time-dependent yet commutative (i.e., $[\mathbf{H}(0), \mathbf{H}(t)] = \mathbf{H}(0)\mathbf{H}(t) - \mathbf{H}(t)\mathbf{H}(0) = 0$), so that a closed-form solution to state evolution can be obtained. A simple choice is that of a Hamiltonian such as

$$H(t) = g(t)\mathbf{S}, \tag{39}$$

with $\mathbf{S}$ a time-independent Hermitian matrix. A function $g(t)$ that ensures convergence of the integral in (25) is the damping

$$g(t) = e^{-t}. \tag{40}$$

In an audio application, we can consider a slice of time and the initial and final states for that slice. We should look for a Hamiltonian that leads to the evolution of the initial state into the final state. In image segmentation [49], where time is used to let each pixel evolve to a final foreground–background assignment, the Hamiltonian is chosen to be

$$H = e^{-t} f(\mathbf{x}) \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \tag{41}$$

and $f(\cdot)$ is a two-valued function of a feature vector $\mathbf{x}$ that contains information about a neighborhood of the pixel. Such function is learned from an example image with a given ground truth. In audio, we may do something similar and learn from examples of transformations: phonation to phonation, with or without pitch crossing; phonation to turbulence; phonation to myoelastic, etc. We may also add a coefficient to the exponent in (40), to govern the rapidity of transformation. As opposed to image processing, time is the playground of audio processing, and a range of possibilities is open to experimentation in Hamiltonian streaming and audio processing.

The matrix $\mathbf{S}$ can be set to assume the structure (29), and the components of potential energy found in an utterance field can be extracted as audio features. For example, pitch salience can be extracted from time-frequency analysis [59] and used as n_z component for the Hamiltonian. Figure 6 shows the two most salient pitches, automatically extracted from a mixture of male and female voice[6] using the Essentia library [57]. Frequent up–down jumps are evident, and they make difficult to track a single voice. Quantum measurement induces state collapse to $|u\rangle$ or $|d\rangle$, and from that state, evolution can be governed by (25). In this way, it should be possible to mimic human figure-ground attention [10,60] and follow each individual voice, or sound stream.

5 Examples

This section is intended to illustrate the potential of the quantum vocal theory of sound in auditory scene analysis and audio effects.[7]

5.1 Two crossing glides interrupted by noise

In auditory scene analysis, insight into auditory organization is often gained through investigation of continuity effects [10]. One interesting case is that of gliding tones

---

[6] https://freesound.org/s/431595/

[7] The reported examples are available, as a `jupyter notebook` containing the full code, on https://github.com/d-rocchesso/QVTS
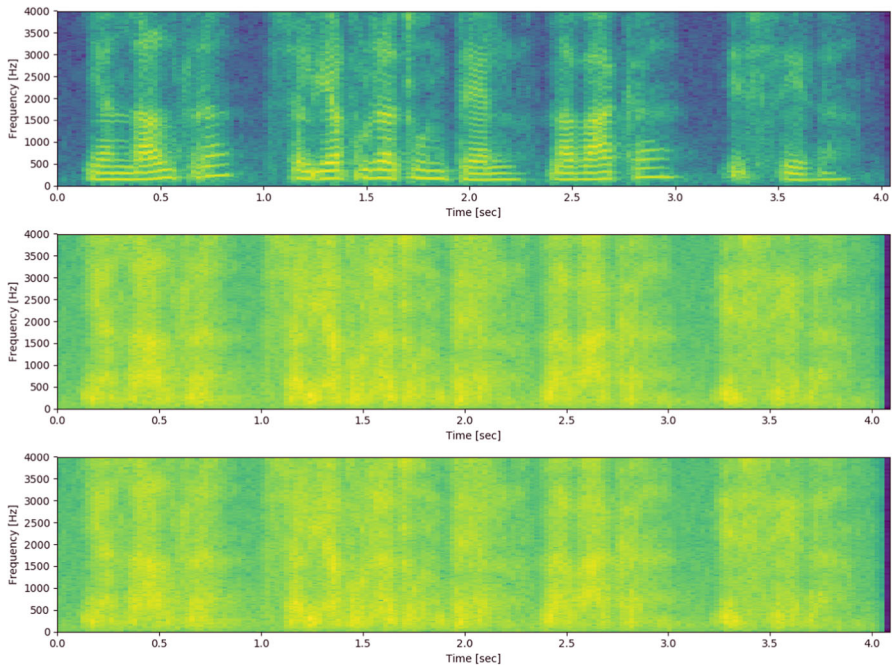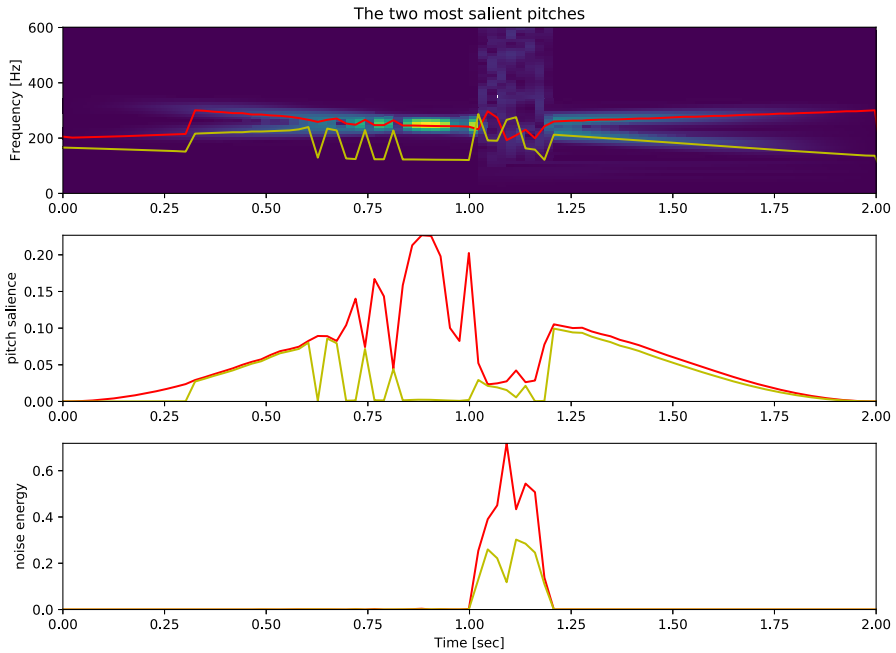
Fig. 6 Extraction of the two most salient pitches from a mixture of a male voice and a female voice

interrupted by a burst of noise [61]. Under certain conditions of temporal exten-
sion and intensity of the noise burst, a single frequency-varying auditory object is
often perceived as crossing the interruption. Specific stimuli can be composed that
make bouncing or crossing equally possible, to investigate which between the Gestalt
principles of proximity and good continuity actually prevails. V-shape trajectories
(bouncing) are often found to prevail on crossing trajectories when the frequencies at
the ends of the interruption match.

To investigate how Hamiltonian evolution may be tuned to recreate some continuity
effects, consider two gliding sinewaves that are interrupted by a band of noise. Figure 7
(top) shows the spectrogram of such noise-interrupted crossing *glissandos*, overlaid
with the traces of the two most salient pitches, computed by means of the Essentia
library [57]. Figure 7 also displays (middle) the computed salience for the two most
salient pitches and (bottom) the energy traces for two bands of noise (1–2 kHz, and
2 kHz–6 kHz).

The elements of the **S** matrix of the Hamiltonian (29) can be computed (in `Python`)
from decimated audio features as

```
decimation = 10 # PARAMETER (hold on)
n_x = energyNoise1[::decimation]
n_x = np.repeat(n_x, decimation)[0:size(energyNoise1)] #
    turbulence potential
n_y = np.zeros(n_x.size) # n_y is not at play
n_z = totalSaliences1[::decimation]
```

Fig. 7 Tracing the two most salient pitches and noise energy for two crossing glides interrupted by noise

```
n_z = np.repeat(n_z, decimation)[0:size(energyNoise1)] #
    pitch potential
```

and the time-varying Hamiltonian can be multiplied by a decreasing exponential $g(m) = e^{-km}$, where m is the frame number, extending over M frames:

```
k = 0.1 # PARAMETER (damping)
et = exp(-k*np.arange(0,decimation))
et = np.tile(et,int(size(n_z)/decimation))
ett = np.pad(et,(0,size(n_z)-size(et)),'constant',
    constant_values=(0))
nzc = n_z*ett
nxc = n_x*ett
nyc = n_y*ett
```

The resulting turbulence and phonation potentials are depicted in Fig. 8.

The Hamiltonian time evolution of Eq. (25) can be computed by approximating the integral with a cumulative sum:

```
H = np.array([[nzc, nxc - 1j*nyc], [nxc + 1j*nyc, -nzc]])
intHc = np.cumsum(H, axis = 2)
Uc = intHc
for t in range(size(n_z)):
    Uc[:,:,t] = linalg.expm(-(1j)*intHc[:,:,t])
```

Choosing an initial state (e.g., pitch-up), the state evolution can be converted into a pitch (phonation) stream, which switches to noise (turbulence) when it goes below a given threshold of pitchiness:
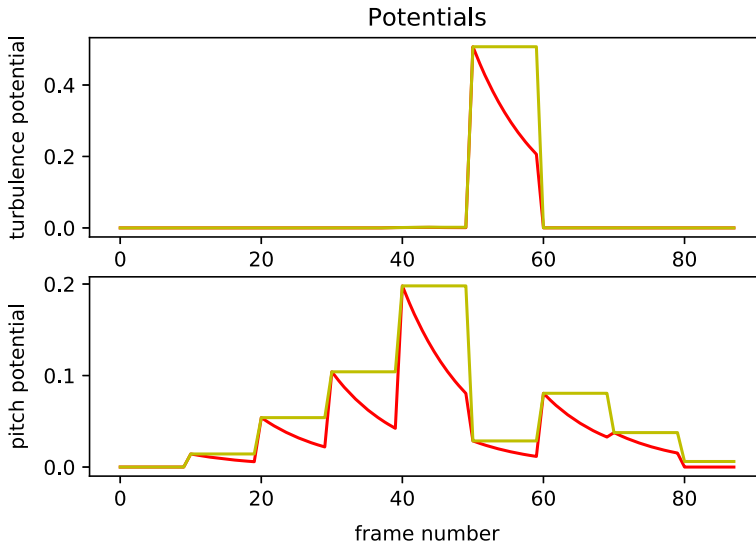
Fig. 8 Potentials of turbulence (top) and phonation (bottom) as functions of frame number

```
threshold = 0.7 # PARAMETER (pitchiness)
hopCollapse = 5 # PARAMETER (decimation of collapses)
sigma_z = [[1, 0], [0, -1]]
sigma_x = [[0, 1], [1, 0]]
for t in range(size(totalFreqs1)):
    sdiff = norm(transpose(matrix(sT[:, t])) -
        sigma_z * transpose(matrix(sT[:, t])))
    if (sdiff < threshold or
        sdiff > (2 - threshold)):
        # state is pitchy
        prob = np.square(abs(sT[0,t]))
        # collapse based on probability
        cstate = np.random.choice(
          np.arange(0, 2), p=[prob, 1-prob])
        if (cstate == 0):
            if (t%hopCollapse == 0):
                sT[0,t] = 1.0
                sT[1,t] = 0.0
            <select upper pitch>
        else:
            if (t%hopCollapse == 0):
                sT[0,t] = 0.0
                sT[1,t] = 1.0
            <select lower pitch>
    else: # state is turbulent
        prob = abs(np.square(np.dot(sT[:,t],
                [1/sqrt(2),1/sqrt(2)])))
        cstate = np.random.choice(
          np.arange(0, 2), p=[prob, 1-prob])
        if (cstate == 0):
            if (t%hopCollapse == 0):
                sT[0,t] = 1/sqrt(2)
```

```
            sT[1,t] = 1/sqrt(2)
        <select lower band of noise>
    else:
        if (t%hopCollapse == 0):
            sT[0,t] = 1/sqrt(2)
            sT[1,t] = -1/sqrt(2)
        <select higher band of noise>
```

In the proposed implementation, the free parameters are `decimation`, `k`, `threshold`, and `hopCollapse`, the latter being a decimation on the measurements that are accompanied by a state collapse. This small set of parameters allows to produce a variety of temporal behaviors, well beyond what is possible with a rigid quantum-mechanical encoding of the listening process.

One resulting pitch stream evolution from pitch-up is depicted in Fig. 9, and it shows a breaking of continuity with bouncing. A first pitch oscillation is visible around second 0.75 when the two sine waves are beating close to each other, although phonation sticks to pitch-up. Then, when the noise interruption arrives after second 1.00, pitch attribution as well as phonation becomes uncertain. Such state of pitch confusion persists almost until second 1.40, well beyond the noise interruption, with occasional commutations to a turbulent state. After the noise shock has been forgotten, the tracking process sticks back to pitch-up, thus preferring a bouncing over a crossing trajectory. Occasionally, due to the inherent randomness of the process, the crossing trajectory may be chosen by the tracking process. The relative probability of bouncing versus crossing depends both on the characteristics of the stimulus (slopes of sinusoidal trajectories, width of the noise break, relative amplitude between noise and sines) and on some model parameters such as the relaxation coefficient `k` of the exponential and the probability `threshold` for collapsing the measure to phonation rather than turbulence.

This example, and some other experiments run with different parameters, shows that the quantum vocal model can reproduce some relevant phenomena of auditory continuity ([62], ch. 6), which are attributable to neural reallocation. The confusion between phonation and turbulence that extends well beyond the interruption is consistent with the known perceptual fact that bursts of noise are not precisely located as referred to a tonal transition, with errors up to a few hundred milliseconds [63].

5.2 Mixed as in a mixer

Given an audio scene such as that of the two crossing glides interrupted by noise (Fig. 7), we may follow the Hamiltonian evolution from an initial state that is known only probabilistically. For example, at time zero we may start from a mixture of $\frac{1}{2}$ pitch-up and $\frac{2}{3}$ pitch-down. The density matrix (18) would evolve according to Eq. (24), where the unitary operator $\mathbf{U}(0, t)$ is defined as in (25). When a pitch measurement is taken, the outcome would be up or down according to Eq. (35), and the density matrix that results from collapsing would be given by Eq. (36).

The density matrix can be made audible in various ways, thus sonifying the Hamiltonian evolution. For example, the completely chaotic mixed state, corresponding to the half-identity matrix $\rho = \frac{1}{2}\mathbf{I}$, can be made to sound as noise, and the pure states
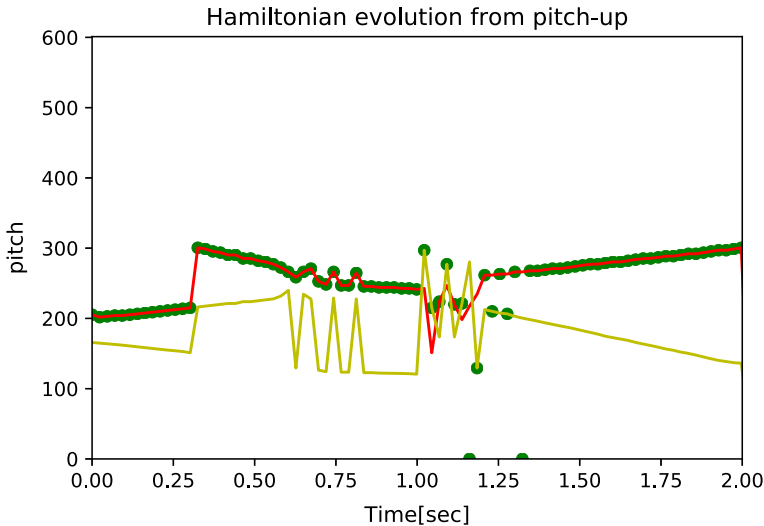
Fig. 9 Tracking the phonation state under Hamiltonian evolution from pitch-up

can be made to sound as the upper or the lower of the most salient pitches. These three components can be mixed for intermediate states. If p_u and p_d are the respective probabilities of pitch-up and pitch-down as encoded in the mixed state, the resulting mixed sound can be composed by a noise having amplitude $\min(p_u, p_d)$, by the upper pitch weighted by $p_u - \min(p_u, p_d)$, and by the lower pitch weighted by $p_d - \min(p_u, p_d)$. One example of such evolution from a mixed state with periodic measurements and collapses that reset the density matrix is depicted in Fig. 10. The analyzed audio scene and the model parameters, including the computed Hamiltonian, are the same as used in the evolution of pure states described in Sect. 5.1. The depicted instance of evolution, if sonified by controlling the amplitudes of the extracted two most salient pitches and of a noise, results in a prevailing downward tone and in a delayed and slowly decreasing burst of noise (Fig. 11).

6 Conclusion and perspective

The components of phonation, turbulence, and supraglottal myoelastic vibrations (and clicks) can be found, in some form and possibly in superposition, in all kinds of vocal sound. Since the voice gives a possibility for an embodied representation of sound in general, we can use the three aforementioned basic phonetic components as general sound descriptors. In this work, we proposed the phon as an analogue of a particle spin, where the phonetic components appear to be aligned along the x, y, and z spin measurement directions. As such, the phon is subject to the mathematical formalism and to the postulates of quantum mechanics, and it can be used to describe sonic processes. Such description is of higher level and exploits a conventional analysis/synthesis framework based on spectral modeling. In particular, we have shown how
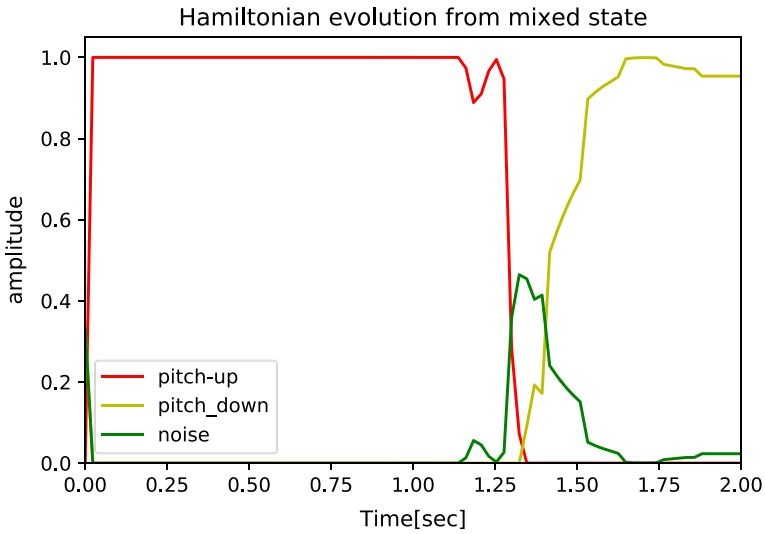
Fig. 10 Amplitudes of components pitch-up, pitch-down, and noise resulting from a Hamiltonian evolution from a mixed state
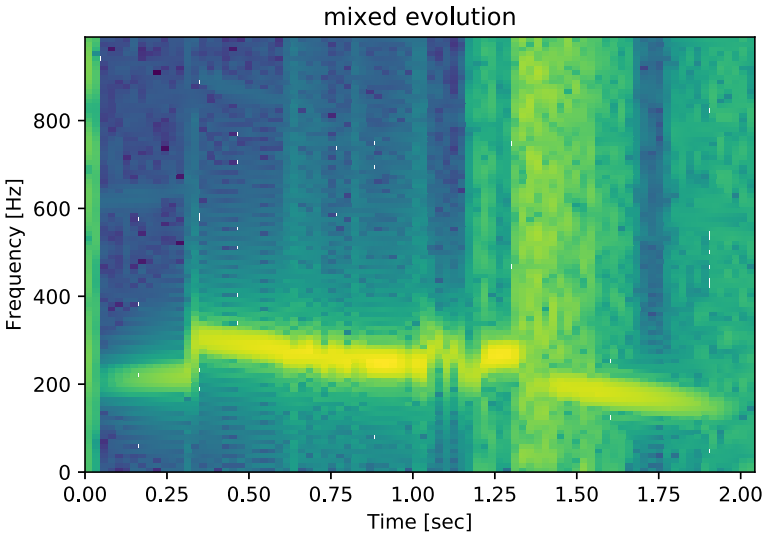

Fig. 11 Spectrogram of the sonification of the Hamiltonian evolution from a mixed state, using the component amplitudes depicted in Fig. 10

a time-varying Hamiltonian, that governs the temporal evolution of auditory streams, can be constructed from features that are extracted from spectral modeling.

In a computational realization of the quantum-inspired operators and processes, the manipulation of a few parameters allows to extract a variety of components from complex audio scenes. The simple examples that we provided show how some relevant auditory-streaming phenomena can be modeled and reproduced, but extensive experi-

mentation is definitely required to verify how useful a quantum vocal theory of sound could be in auditory scene analysis. A large range of possibilities is also open to the creative processing of audio materials through the sonification of the extracted streams and events. As compared to analysis/synthesis frameworks based on spectral processing, here we work at a higher level corresponding to fewer descriptors whose evolution and intertwinement are mathematically defined. The statistical nature of measurement, in evolutions of pure or mixed states under time-varying force fields, leads naturally to the synthesis of ensembles of audio processes, all derived and somehow echoing the original audio material. If we successfully model some auditory phenomena, such as continuity effects or temporal displacement, by temporal phon evolution, and if we render these evolutions back to sound, we may somehow say that we listen to possible auditory processes. However, in creative applications we are not bound to mimic auditory processes and we can also depart from quantum orthodoxy in many possible different ways.

The proposed theory enhances the role of quantum theory and of the underlying mathematics as a connecting tool between different areas of human knowledge. By flipping the wicked problem of finding intuitive interpretations of quantum mechanics, we aimed at using quantum mechanics to interpret something that we have embodied, intuitive knowledge of.

References

1. Fourier, J.B.J.: Théorie Analytique de la Chaleur. Firmin Didot Père et Fils, Paris (1822)
2. von Helmholtz, H.: Die Lehre von den Tonempfindungen als physiologische Grundlage für die Theorie der Musik (F. Vieweg und sohn, 1870)
3. Shannon, C.E.: Communication in the presence of noise. Proc. IRE **37**(1), 10–21 (1949)
4. Roads, C.: Microsound. MIT Press, Cambridge (2001)
5. Gabor, D.: Acoustical quanta and the theory of hearing. Nature **159**(4044), 591 (1947)
6. Oppenheim, J.N., Magnasco, M.O.: Human time-frequency acuity beats the fourier uncertainty principle. Phys. Rev. Lett. **110**, 044301 (2013)
7. Lyon, R.F.: Human and Machine Hearing. Cambridge University Press, Cambridge (2017)
8. Gaver, W.W.: How do we hear in the world? Explorations in ecological acoustics. Ecol. Psychol. **5**(4), 285–313 (1993)
9. Lemaitre, G., Rocchesso, D.: On the effectiveness of vocal imitations and verbal descriptions of sounds. J. Acoust. Soc. Am. **135**(2), 862–873 (2014)
10. Bregman, A.S.: Auditory Scene Analysis: The Perceptual Organization of Sound. MIT Press, Cambridge (1994)

11. Yearsley, J.M., Busemeyer, J.R.: Quantum cognition and decision theories: a tutorial. Foundations of probability theory in psychology and beyond. J. Math. Psychol. **74**, 99–116 (2016)

12. Yearsley, J.M., Pothos, E.M.: Challenging the classical notion of time in cognition: a quantum perspective. Proc. R. Soc. B Biol. Sci. **281**(1781), 20133056 (2014)

13. De Poli, G., Piccialli, A., Roads, C. (eds.): Representations of Musical Signals. MIT Press, Cambridge (1991)

14. Roden, D.: Sonic art and the nature of sonic events. Rev. Philos. Psichol. **1**(1), 141–156 (2010)

15. Leman, M.: Embodied Music Cognition and Mediation Technology. MIT Press, Cambridge (2008)

16. Signata, A.V.: Towards a semiotics of the audible. Ann. Semiot. **6**, 65–89 (2015)

17. Delle Monache, S., Rocchesso, D., Bevilacqua, F., Lemaitre, G., Baldan, S., Cera, A.: Embodied sound design. Int. J. Hum. Comput. Stud. **118**, 47–59 (2018)

18. Rocchesso, D., Delle Monache, S., Barrass, S.: Interaction by ear. Int. J. Hum. Comput. Stud. **131**, 152–159 (2019) (50 years of the International Journal of Human-Computer Studies. Reflections on the past, present and future of human-centred technologies)

19. Houix, O., Lemaitre, G., Misdariis, N., Susini, P., Urdapilleta, I.: A lexical analysis of environmental sound categories. J. Exp. Psychol. Appl. **18**(1), 52 (2012)

20. Kubovy, M., Schutz, M.: Audio-visual objects. Rev. Philos. Psichol. **1**(1), 41–61 (2010)

21. Bonada, J., Serra, X., Amatriain, X., Loscos, A.: Spectral processing. In: Zölzer, U. (ed.) DAFX: Digital Audio Effects, pp. 393–445. Wiley, Hoboken (2011)

22. Verma, T.S., Levine, S.N., Meng, T.H.: Transient Modeling Synthesis: a flexible analysis/synthesis tool for transient signals. In: Proceedings of the International Computer Music Conference, pp. 48–51 (1997)

23. Füg, R., Niedermeier, A., Driedger, J., Disch, S., Müller, M.: Harmonic-percussive-residual sound separation using the structure tensor on spectrograms. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 445–449 (2016)

24. Isnard, V., Taffou, M., Viaud-Delmon, I., Suied, C.: Auditory sketches: Very sparse representations of sounds are still recognizable. PLoS One **11**(3), e0150313 (2016)

25. Lemaitre, G., Houix, O., Voisin, F., Misdariis, N., Susini, P.: Vocal imitations of non-vocal sounds. PLoS One **11**(12), e0168167 (2016)

26. Perlman, M., Lupyan, G.: People can create iconic vocalizations to communicate various meanings to naïve listeners. Sci. Rep. **8**, 2634 (2018)

27. Wallmark, Z., Iacoboni, M., Deblieck, C., Kendall, R.A.: Embodied listening and timbre: Perceptual, acoustical, and neural correlates. Music Percept. Interdiscip. J. **35**(3), 332–363 (2018)

28. Russolo, L.: L'arte dei rumori. Edizioni futuriste di "poesia" (1916)

29. Marinetti, F.T.: Zang tumb tumb: Adrianopoli, ottobre 1912: parole in libertà. Edizioni futuriste di "poesia" (1914)

30. Helgason, P.: Sound initiation and source types in human imitations of sounds. In: Proceedings of FONETIK 2014, pp. 83–88 (2014)

31. Friberg, A., Lindeberg, T., Hellwagner, M., Helgason, P., Salomão, G.L., Elowsson, A., Lemaitre, G., Ternström, S.: Prediction of three articulatory categories in vocal sound imitations using models for auditory receptive fields. J. Acoust. Soc. Am. **144**(3), 1467–1483 (2018)

32. Marchetto, E., Peeters, G.: Automatic recognition of sound categories from their vocal imitation using audio primitives automatically derived by SI-PLCA and HMM. In: Proceedings of the International Symposium on Computer Music Multidisciplinary Research, pp. 9–20. Matosinhos, Portugal (2017)

33. Lemaitre, G., Jabbari, A., Misdariis, N., Houix, O., Susini, P.: Vocal imitations of basic auditory features. J. Acoust. Soc. Am. **139**(1), 290–300 (2016)

34. Changizi, M.: Harnessed: How Language and Music Mimicked Nature and Transformed Ape to Man. BenBella Books Inc., Dallas (2011)

35. Mesgarani, N., Cheung, C., Johnson, K., Chang, E.F.: Phonetic feature encoding in human superior temporal gyrus. Science **343**(6174), 1006–1010 (2014)

36. De Sena, A., Rocchesso, D.: A fast Mellin and scale transform. EURASIP J. Adv. Signal Process. **2007**, 89170 (2007). https://doi.org/10.1155/2007/89170

37. Irino, T., Patterson, R.D.: A time-domain, level-dependent auditory filter: the gammachirp. J. Acoust. Soc. Am. **101**(1), 412–419 (1997)

38. Eldar, Y.C., Oppenheim, A.V.: Quantum signal processing. IEEE Signal Process. Mag. **19**(6), 12–32 (2002)

39. Wang, J.: QRDA: quantum representation of digital audio. Int. J. Theor. Phys. **55**(3), 1622–1641 (2016)

40. Yan, F., Iliyasu, A.M., Guo, Y., Yang, H.: Flexible representation and manipulation of audio signals on quantum computers. Theor. Comput. Sci. **752**, 71–85 (2018)
41. beim Graben, P., Blutner, R.: Quantum approaches to music cognition. J. Math. Psychol. **91**, 38–50 (2019)
42. Blutner, R., beim Graben, P.: Gauge models of musical forces. J. Math. Music (2020). https://doi.org/10.1080/17459737.2020.1716404
43. Mannone, M., Compagno, G.: Characterization of the degree of musical non-Markovianity. arXiv:1306.0229 (2013)
44. Fischman, R.: Clouds, pyramids, and diamonds: applying Schrödinger's equation to granular synthesis and compositional structure. Comput. Music J. **27**(2), 47 (2003)
45. Kontogeorgakopoulos, A., Burgarth, D.: Sonification of controlled quantum dynamics. In: Proceedings of the 2014 International Computer Music Conference, pp. 1432–1436 (2014)
46. Sturm, B.: Composing for an ensemble of atoms: the metamorphosis of scientific experiment into music. Org. Sound **6**(2), 131–145 (2001)
47. Dalla Chiara, M.L., Giuntini, R., Leporini, R., Negri, E., Sergioli, G.: Quantum information, cognition, and music. Front. Psychol. **6**, 1583 (2015)
48. Ghirardi, G.: Quantum superpositions and definite perceptions: envisaging new feasible experimental tests. Phys. Lett. A **262**(1), 1 (1999)
49. Youssry, A., El-Rafei, A., Elramly, S.: A quantum mechanics-based framework for image processing and its application to image segmentation. Quantum Inf. Process. **14**(10), 3613–3638 (2015)
50. Aytekin, Ç., Ozan, E.C., Kiranyaz, S., Gabbouj, M.: Extended quantum cuts for unsupervised salient object extraction. Multimedia Tools Appl. **76**(8), 10443–10463 (2017)
51. Okada, S., Ohzeki, M., Terabe, M., Taguchi, S.: Improving solutions by embedding larger subproblems in a D-Wave quantum annealer. Sci. Rep. **9**, 2098 (2019)
52. Rocchesso, D., Lemaitre, G., Susini, P., Ternström, S., Boussard, P.: Sketching sound with voice and gesture. Interactions **22**(1), 38–41 (2015)
53. Susskind, L., Friedman, A.: Quantum Mechanics: The Theoretical Minimum. Penguin Books, City of Westminster (2015)
54. Cariolaro, G.: Quantum Communications. Springer, Berlin (2015)
55. Rocchesso, D., Mauro, D.A., Drioli, C.: Organizing a sonic space through vocal imitations. J. Audio Eng. Soc. **64**(7/8), 474–483 (2016)
56. Newman, F.: MouthSounds: How to Whistle, Pop, Boing, and Honk... for All Occasions and Then Some. Workman Publishing, New York (2004)
57. Bogdanov, D., Wack, N., Gómez Gutiérrez, E., Gulati, S., Herrera Boyer, P., Mayor, O., Roma Trepat, G., Salamon, J., Zapata González, J.R., Serra, X.: Essentia: an audio analysis library for music information retrieval. In: Proceedings of the 14th Conference of the International Society for Music Information Retrieval (ISMIR). Curitiba, Brazil, pp. 493–498 (2013)
58. Breuer, H.P., Petruccione, F.: The Theory of Open Quantum Systems. Oxford University Press, New York (2002)
59. Salamon, J., Gomez, E.: Melody extraction from polyphonic music signals using pitch contour characteristics. IEEE Trans. Audio Speech Lang. Process. **20**(6), 1759–1770 (2012)
60. Bigand, E., McAdams, S., Forêt, S.: Divided attention in music. Int. J. Psychol. **35**(6), 270–278 (2000)
61. Ciocca, V., Bregman, A.S.: Perceived continuity of gliding and steady-state tones through interrupting noise. Percept. Psychophys. **42**(5), 476–484 (1987)
62. Warren, R.M.: Auditory Perception: An Analysis and Synthesis, 3rd edn. Cambridge University Press, Cambridge (2008)
63. Vicario, G.B.: La "dislocazione temporale" nella percezione di successioni di stimoli discreti (The "time displacement" in the perception of sequences of discrete stimuli. Riv. Psicol. **57**(1), 17–87 (1963)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.