Embryo of a Quantum Vocal Theory of Sound

Davide Rocchesso
University of Palermo
davide.rocchesso@unipa.it

Maria Mannone
University of Minnesota, alumna
manno012@umn.edu

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 slow 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. This would constitute the basis for a Quantum Vocal Theory of sound, with implications in sound analysis and design.

1. INTRODUCTION

What are the fundamental elements of sound? What is the best framework for analyzing existing sonic realities and for expressing new sound concepts? These are long standing questions in sound physics, perception, and creation. In 1947, in a famous paper published in Nature [1], Dennis 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/Gabor framework for time-frequency representation of sound is widely used, although human acuity has been shown to beat the uncertainty limit [2] and cochlear filters [3] have been proposed to match human performance. Still, when we are imagining sound, or describing it to peers, we do not use the Fourier formalism.

1.1 Voice as Embodied Sound

Many researchers, in science, art, and philosophy, have been facing the problem of how to approach sound and its representations [4, 5]. 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 establishing bridges between distal (source-related) and proximal (sensory-related) representations, and we look at research findings in perception, production, and articulation of sounds [6]. Our embodied approach to sound [7] 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 [9]. However, when the task is to separate, distinguish, count, or compose sounds, the attention shifts from *sounding objects* to *auditory objects* [10] represented in the time-frequency plane. Tonal components, noise, and transients can be extracted from sound objects with Fourier-based techniques [11, 12, 13]. 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 [14], have been used to test the recognizability of imitations [15].

Vocal imitations can be more effective than verbalizations at representing and communicating sounds when these are difficult to describe with words [16]. 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. At a more fundamental level, research on non-speech vocalization is affecting the theories of language evolution [17], 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 [7] show that agreement in production corresponds to agreement in meaning interpretation, thus showing the effectiveness of teamwork in embodied sound creation. Converging evidences from behavioral and brain imaging studies give a firm basis to hypothesize a shared representation of sound in terms of motor (vocal) primitives [18].

In the recent EU-FET project SkAT-VG [1] , phoneticians have had an important role in identifying the most relevant components of non-speech voice productions [19]. They identified the broad categories of phonation (i.e., quasi periodic oscillations due to vocal fold vibrations), turbulence, slow myoelastic vibrations, and clicks, which can be extracted automatically from audio with time-frequency analysis and machine learning [20], and can be made to correspond to categories of sounds as they are perceived [15], and as they are produced in the physical world. Indeed, it has been argued that human utterances somehow mimic "nature's phonemes" [21].

1.2 Quantum Frameworks

It was Dennis Gabor [1] 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 [22] 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 [23]. Research in human and machine hearing [3] have 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 [24] to seek practical solutions to sound scene representation, separation and analysis. A quantum approach to music cognition has not been proposed until very recently [25], when its explanatory power has been demonstrated to describe tonal attraction phenomena in terms of metaphorical forces. The theory of open quantum systems has been applied to music to describe the memory properties (non-Markovianity) of different scores [26]. In the image domain, 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 [27], or by discretizing the time-independent Schrödinger equation [28]. 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 [29].

1.3 Research Direction

In the proposed research path, sound is treated as a superposition of states, and the voice-based components (phonation, turbulence, slow myoelastic pulsations) 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, 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 [2] 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 improved techniques for sound analysis and synthesis.

2. SKETCH OF A QUANTUM VOCAL THEORY

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 σ 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 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 [12]) 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 don't, an uncertainty principle can be derived, as it was done for time-frequency and time-scale representations [1, 23]. The theory can be extended to cover multiple sources, and the resulting mixed

[1] www.skatvg.eu

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

states can be described via density matrices, whose time evolution can also be computed if a Hamiltonian operator is properly defined. The effect of disturbances and deviations can be accounted for by introducing relaxation time, to model return to equilibrium. This sketch of a Quantum Vocal Theory needs to be developed and formally laid down.

3. THE PHON FORMALISM

Consider a 3d space with the orthogonal axes

z : phonation, with different pitches;

x : turbulence, with noises of different frequency distributions;

y : myoelastic, slow pulsations with different tempos.

The phon operator σ is a 3-vector operator that provides information about the phonetic elements in a specific direction of the 3d phonetic space.

In this section we present the phon formalism, obtained by direct analogy with the single spin, as presented in accessible presentations of quantum mechanics [30]. We use standard Dirac notation.

3.1 Measurement along z

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

1. Each component of σ 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: $|u\rangle = \begin{bmatrix}1\\0\end{bmatrix}$ and $|v\rangle = \begin{bmatrix}0\\1\end{bmatrix}$, and the operator σ_z as a square 2×2 matrix. Due to principle 2, we have

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

3.2 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 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

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

In fact, any state $|A\rangle$ can be expressed as $|A\rangle = \alpha_u |u\rangle + \alpha_d |d\rangle$, where $\alpha_u = \langle u|A\rangle$, and $\alpha_d = \langle d|A\rangle$. The probability to measure pitch up is $P_u = \langle A|u\rangle \langle u|A\rangle = {\alpha_u}^* \alpha_u$, and the probability to measure pitch down is $P_d = \langle A|d\rangle \langle d|A\rangle = {\alpha_d}^* \alpha_d$ (Born rule).

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

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

3.3 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 [3] , 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 slow myoelastic pulsations. This measurement property is satisfied if

$$|f\rangle = \frac{1}{\sqrt{2}}|u\rangle + \frac{i}{\sqrt{2}}|d\rangle , \tag{5}$$

where i is the imaginary unit.

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

$$|s\rangle = \frac{1}{\sqrt{2}}|u\rangle - \frac{i}{\sqrt{2}}|d\rangle , \tag{6}$$

which is orthogonal to the linear combination (5). In vector form, we have: $|f\rangle = \begin{bmatrix}\frac{1}{\sqrt{2}}\\ \frac{i}{\sqrt{2}}\end{bmatrix}$ and $|s\rangle = \begin{bmatrix}\frac{1}{\sqrt{2}}\\ -\frac{-i}{\sqrt{2}}\end{bmatrix}$, and

$$\sigma_y = \begin{bmatrix}0 & -i\\i & 0\end{bmatrix}. \tag{7}$$

The matrices (1), (4), and (7) are called the Pauli matrices and, together with the identitity matrix, these are the quaternions.

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

3.4 Measurement along an arbitrary direction

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

$$\sigma_n = \overline{\sigma} \cdot \overline{n} = \sigma_x n_x + \sigma_y n_y + \sigma_z n_z = \\ = \begin{bmatrix} n_z & n_x - in_y \\ n_x + in_y & -n_z \end{bmatrix}. \quad (8)$$

3.4.1 Example: Harmonic plus Noise model

A measurement performed by means of a Harmonic plus Noise model [11] would lie in the phonation-turbulence plane ($n_z = \cos\theta, n_x = \sin\theta, n_y = 0$), so that

$$\sigma_n = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix} \quad (9)$$

The eigenstate for eigenvalue $+1$ is $|\lambda_1\rangle = \begin{bmatrix} \cos\theta/2 \\ \sin\theta/2 \end{bmatrix}$, the eigenstate for eigenvalue -1 is $|\lambda_{-1}\rangle = \begin{bmatrix} -\sin\theta/2 \\ \cos\theta/2 \end{bmatrix}$, and the two are orthogonal. Suppose we prepare the phon to pitch-up $|u\rangle$. If we rotate the measurement system along $\overline{n}$, the probability to measure $\sigma_n = +1$ is (by Born rule)

$$P(+1) = |\langle u|\lambda_1\rangle|^2 = \cos^2\theta/2, \quad (10)$$

and the probability to measure $\sigma_n = -1$ is

$$P(-1) = |\langle u|\lambda_{-1}\rangle|^2 = \sin^2\theta/2. \quad (11)$$

The expectation value of measurement is therefore

$$\langle\sigma_n\rangle = \sum_i \lambda_i P(\lambda_i) \\ = (+1)\cos^2\theta/2 + (-1)\sin^2\theta/2 = \cos\theta \quad (12)$$

3.4.2 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 $\{\mathrm{ph}, \mathrm{tu}, \mathrm{my}\} \in [-1, 1]$. If $\mathrm{ph} > 0$ the result will be $|u\rangle$, and if $\mathrm{ph} < 0$ the result will be $|d\rangle$. If $\mathrm{tu} > 0$ the result will be $|r\rangle$, and if $\mathrm{tu} < 0$ the result will be $|l\rangle$. If $\mathrm{my} > 0$ the result will be $|f\rangle$, and if $\mathrm{my} < 0$ 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.

3.5 Time evolution

In quantum mechanics the evolution of state vectors in time

$$|\psi(t)\rangle = \mathbf{U}(t)\,|\psi(0)\rangle \quad (13)$$

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

$$\mathbf{U}(\epsilon) = \mathbf{I} - i\epsilon\mathbf{H}, \quad (14)$$

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 (14) it turns out that a state vector changes in time according to the time-dependent Schrödinger equation [4]

$$\frac{\partial\,|\psi\rangle}{\partial t} = -i\mathbf{H}\,|\psi\rangle\,. \quad (15)$$

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

$$\frac{\partial\,\langle\mathbf{L}\rangle}{\partial t} = -i\,\langle[\mathbf{L}, \mathbf{H}]\rangle\,, \quad (16)$$

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

3.5.1 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

$$\mathbf{H} \propto \overline{\sigma} \cdot \overline{B} = \sigma_x B_x + \sigma_y B_y + \sigma_z B_z, \quad (17)$$

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:

$$H = \frac{\omega}{2}\sigma_z. \quad (18)$$

To find how the expectation value of the phon varies in time, we expand the observable $\mathbf{L}$ in (16) in its components to get

$$\begin{aligned} \langle\dot{\sigma}_x\rangle &= -i\,\langle[\sigma_x, \mathbf{H}]\rangle = -\omega\,\langle\sigma_y\rangle \\ \langle\dot{\sigma}_y\rangle &= -i\,\langle[\sigma_y, \mathbf{H}]\rangle = \omega\,\langle\sigma_x\rangle \\ \langle\dot{\sigma}_z\rangle &= -i\,\langle[\sigma_z, \mathbf{H}]\rangle = 0, \end{aligned} \quad (19)$$

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

A potential energy along all three axes can be expressed as

$$\mathbf{H} = \frac{\omega}{2}\overline{\sigma} \cdot \overline{n} = \frac{\omega}{2}\begin{bmatrix} n_z & n_x - in_y \\ n_x + in_y & -n_z \end{bmatrix}, \quad (20)$$

whose energy eigenvalues are $E_j = \pm 1$, with energy eigenvectors $|E_j\rangle$.

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

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

$$|\psi(0)\rangle = \sum_j \alpha_j(0)\,|E_j\rangle\,, \tag{21}$$

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

$$|\psi(t)\rangle = \sum_j \alpha_j(t)\,|E_j\rangle = \sum_j \alpha_j(0)e^{-iE_j t}\,|E_j\rangle\,. \tag{22}$$

3.6 Uncertainty

If we measure two observables L and M (in a single experiment) simultaneously, quantum mechanics prescribes that the system is left in a simultaneous eigenvector of the observables only if L and M commute, i.e. $[L, M] = 0$. Measurement operators along different axes do not commute. For example, $[\sigma_x, \sigma_y] = 2i\sigma_z$, and therefore phonation and turbulence can not 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:

$$\Delta L \Delta M \geq \frac{1}{2}\,|\langle\psi|\,[L, M]\,|\psi\rangle| \tag{23}$$

Equation (23) expresses the uncertainty principle.

If $L = t$ is the time operator and $S = -i\frac{d}{dt}$ 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 [23].

4. WHAT TO DO WITH ALL THIS

Assuming that a vocal description of sound follows the postulates of quantum mechanics, the question is how to take advantage of the quantum formalism. How can we practically connect theoretical ideas with audio signal processing?

4.1 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 sample. The measurement 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 figure 1 shows non-commutativity in the style of category theory.

Measurements of phonation and turbulence can be actually performed using the Harmonic plus Stochastic (HPS) model [11]. The order of operations is visually described in figure 2. The measurement of phonation is performed through the extraction of the harmonic component in the HPS model, while the measurement of turbulence is performed through the extraction of the stochastic component with the same model. The spectrograms for A'' and A^{**} in Figure 3 show the results of such two sequences of analyses on a segment of female speech [5], confirming that the commutator $[T\,,\,P]$ is non-zero.

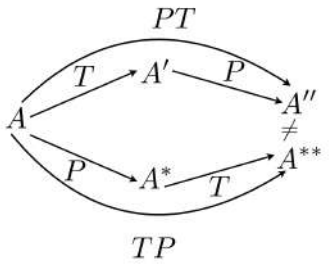

Figure 1. A non-commutative diagram (in the style of category theory), representing the non-commutativity of measurements of phonation (P) and turbulence (T) on audio A.

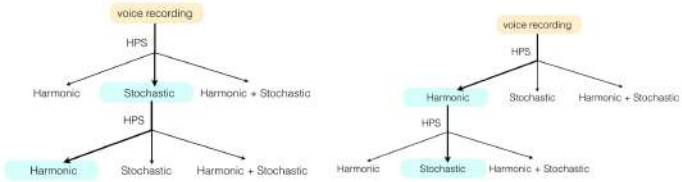

Figure 2. On the left, a voice is analyzed via the HPS model. Then, the stochastic part is submitted to a new analysis. In this way, a measurement of phonation follows a measurement of turbulence. On the right, the measurement of turbulence follows a measurement of phonation.

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 *destructive measure*. Let St be the filter that extracts the stochastic part from a signal. As figure 4 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 figure 4, 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*. If we perform the measure again and again, we still get the same result. Such a measure operation provokes the collapse of a 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

[5] `https://freesound.org/s/317745/`.
Hann window of 2048 samples, FFT of 4096 samples, hop size of 1024 samples.

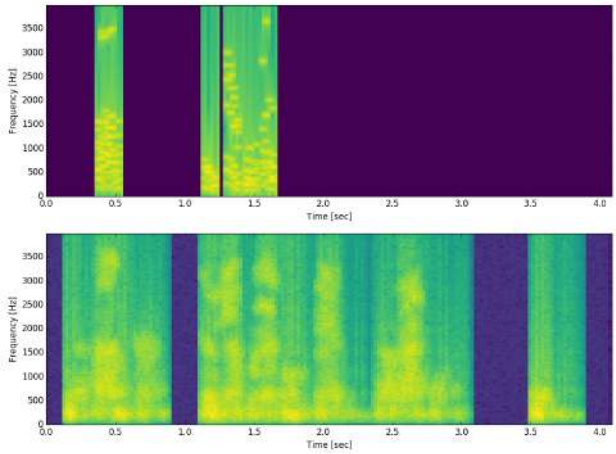

Figure 3. On the top, the spectrogram corresponding to a measurement of phonation P following a measurement of turbulence T, leading to $PT(A) = A''$. On the bottom, the spectrogram corresponding to a measurement of turbulence T following a measurement of phonation P, leading to $TP(A) = A^{**}$.

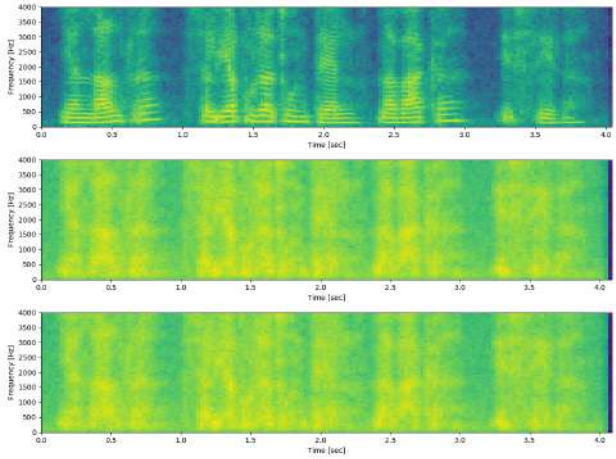

Figure 4. Top: spectrum of the original sound signal (a female speech), Center: the stochastic component, derived from harmonic plus stochastic analysis (HPS), as the effect of a destructive measure, and Bottom: the stochastic component of the stochastic component itself. The last two spectra are very close.

of quantum measures, which may become practically feasible through a set of audio-signal analysis tools.

We can define other filters that extract more specific information: for example, a glissando filter would extract glissando passages within a sound signal, giving as output the inverse transform of the filtered spectrum, that is, a filtered signal where we can just hear the glissando effect and nothing else. Re-performing the glissando-measure on that sound, we still get the same sound: the new sound signal is an autostate of glissando. A suitable collection of filters that act both on the original sounds as well as on their vocal imitations can produce simplified spectra, easier to compare. These may give hints on how information are extracted by voice, that permit to recognize the original sound source. This would confirm the importance of human voice as a probe to investigate the world of sounds, and Quantum Vocal Theory as a bridge between quantum physics, acoustics, and cognition, with possible further bridges to multisensory perception and interaction. If we consider that voice can imitate not only sounds, but also movements and, sometimes, even visual shapes through crossmodal correspondences [31], new fascinating scenarios open up for investigation.

4.2 Time evolution

We know that time evolution of states is governed by the unitary transformation (13) and by the Schrödinger equation (15). A measurement is represented by an operator that acts on the state and that causes its collapse onto one of its eigenvectors. The system remains in a superposition of states until a measurement occurs, whose outcome is inherently probabilistic. The states change under the effect of external forces, which therefore determine the change in the probabilities associated with each state. It is a Hamiltonian such as (20) that controls the evolution process. We can think of the vocal Hamiltonian as containing the potential energies due to restoring forces that are contained in the utterance at a particular moment in time. We could even think of a formal composition of forces acting at different levels, and describe them through a nested categorical depiction.

Similarly to what has been done by Youssry et al. [27], 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 time-independent Hamiltonian such as the one leading to (22) would not be very useful, both because forces indeed change continuously and because this would lead to oscillatory solution. Instead, with a time-varying commutative Hamiltonian the time evolution can be expressed as

$$|\psi(t)\rangle = e^{-i\int_0^t H(\tau)d\tau} |\psi(0)\rangle . \tag{24}$$

A simple choice is that of a Hamiltonian such as

$$H(t) = g(t)\mathbf{S}, \tag{25}$$

with $\mathbf{S}$ a time-independent Hermitian matrix. A function $g(t)$ that ensures convergence of the integral in (24) is the damping

$$g(t) = e^{-t}. \tag{26}$$

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 [27], where time is used to let each pixel evolve to a final foreground-background assignment, the Hamiltonian is chosen to be

$$H = e^{-t} f(\mathbf{x}) \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix}, \tag{27}$$

and $f(\cdot)$ is a two-valued function of a feature vector $\mathbf{x}$ that contains information about a neighborhood of the pixel.

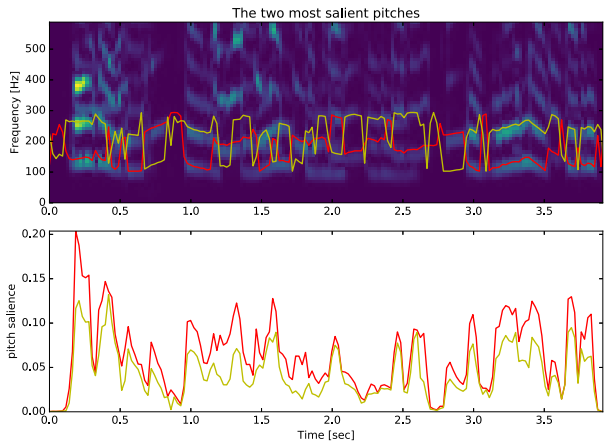

Figure 5. Extraction of the two most salient pitches from a mixture of a male voice and a female voice

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 (26), to govern the rapidity of transformation. As opposed to image processing, time is our playground.

The matrix $\mathbf{S}$ can be set to assume the structure (20), 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 [32] and used as n_z component for the Hamiltonian. Figure 5 shows the two most salient pitches, automatically extracted from a mixture of male and female voice [6] . 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 (24). In this way, it should be possible to mimic human figure-ground attention [33], and follow each individual voice.

5. CONCLUSION AND FURTHER RESEARCH

We presented an early attempt at applying the fundamental quantum formalism to the description of human voice, and to sounds in general through voice-based basic elements. Such a theoretical research can have several practical implications, due to the importance of voice as a probe to investigate the world of sounds in general. Also, a quantum vocal theory enhances the role of quantum mechanics 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, the proposed approach aims to use quantum mechanics to interpret something that we have embodied, intuitive knowledge of. The new theory can lead to the development of voice-based tools for sound analysis and synthesis, for research and communication understanding. Finally, given the importance of vocal imitations in the framework of improvised music and sound art, they can constitute a probe also for studying and creating instrumental music, and for conceiving new forms of sonic expression, also through manual and body gestures, which enable connections to the visual and performing arts.

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

6. REFERENCES

[1] D. Gabor, "Acoustical quanta and the theory of hearing," *Nature*, vol. 159, no. 4044, p. 591, 1947.

[2] J. N. Oppenheim and M. O. Magnasco, "Human time-frequency acuity beats the fourier uncertainty principle," *Phys. Rev. Lett.*, vol. 110, p. 044301, Jan 2013.

[3] R. F. Lyon, *Human and machine hearing*. Cambridge University Press, 2002.

[4] G. De Poli, A. Piccialli, and C. Roads, *Representations of musical signals*. MIT Press, 1991.

[5] D. Roden, "Sonic art and the nature of sonic events," *Review of Philosophy and Psichology*, vol. 1, no. 1, pp. 141–156, 2010.

[6] M. Leman, *Embodied music cognition and mediation technology*. MIT Press, 2008.

[7] S. D. Monache, D. Rocchesso, F. Bevilacqua, G. Lemaitre, S. Baldan, and A. Cera, "Embodied sound design," *International Journal of Human-Computer Studies*, vol. 118, pp. 47 – 59, 2018.

[8] W. W. Gaver, "How do we hear in the world? Explorations in ecological acoustics," *Ecological Psychology*, vol. 5, no. 4, pp. 285–313, 1993.

[9] O. Houix, G. Lemaitre, N. Misdariis, P. Susini, and I. Urdapilleta, "A lexical analysis of environmental sound categories," *Journal of Experimental Psychology: Applied*, vol. 18, no. 1, p. 52, 2012.

[10] M. Kubovy and M. Schutz, "Audio-visual objects," *Review of Philosophy and Psichology*, vol. 1, no. 1, pp. 41–61, 2010.

[11] J. Bonada, X. Serra, X. Amatriain, and A. Loscos, "Spectral processing," in *DAFX: Digital Audio Effects* (U. Zölzer, ed.), pp. 393–445, John Wiley & Sons, Ltd, 2011.

[12] T. S. Verma, S. N. Levine, and T. H. Meng, "Transient Modeling Synthesis: a flexible analysis/synthesis tool for transient signals," in *Proceedings of the International Computer Music Conference*, pp. 48–51, 1997.

[13] R. Fg, A. Niedermeier, J. Driedger, S. Disch, and M. Mller, "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, March 2016.

[14] V. Isnard, M. Taffou, I. Viaud-Delmon, and C. Suied, "Auditory sketches: very sparse representations of sounds are still recognizable," *PloS one*, vol. 11, no. 3, 2016.

[15] G. Lemaitre, A. Jabbari, N. Misdariis, O. Houix, and P. Susini, "Vocal imitations of basic auditory features," *The Journal of the Acoustical Society of America*, vol. 139, no. 1, pp. 290–300, 2016.

[16] G. Lemaitre and D. Rocchesso, "On the effectiveness of vocal imitations and verbal descriptions of sounds," *The Journal of the Acoustical Society of America*, vol. 135, no. 2, pp. 862–873, 2014.

[17] M. Perlman and G. Lupyan, "People can create iconic vocalizations to communicate various meanings to naïve listeners," *Scientific reports*, vol. 8, no. 1, p. 2634, 2018.

[18] Z. Wallmark, M. Iacoboni, C. Deblieck, and R. A. Kendall, "Embodied Listening and Timbre: Perceptual, Acoustical, and Neural Correlates," *Music Perception: An Interdisciplinary Journal*, vol. 35, no. 3, pp. 332–363, 2018.

[19] P. Helgason, "Sound initiation and source types in human imitations of sounds," in *Proceedings of FONETIK 2014*, pp. 83–88, 2014.

[20] E. Marchetto and G. Peeters, "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*, 2017.

[21] M. Changizi, *Harnessed: How language and music mimicked nature and transformed ape to man*. BenBella Books, Inc., 2011.

[22] A. De Sena and D. Rocchesso, "A fast Mellin and scale transform," *EURASIP Journal on Applied Signal Processing*, no. Article ID 8917, 2007.

[23] T. Irino and R. D. Patterson, "A time-domain, level-dependent auditory filter: The gammachirp," *The Journal of the Acoustical Society of America*, vol. 101, no. 1, pp. 412–419, 1997.

[24] Y. C. Eldar and A. V. Oppenheim, "Quantum signal processing," *IEEE Signal Processing Magazine*, vol. 19, no. 6, pp. 12–32, 2002.

[25] P. beim Graben and R. Blutner, "Quantum approaches to music cognition," *arXiv:1712.07417*, 2018.

[26] M. Mannone and G. Compagno, "Characterization of the degree of Musical non-Markovianity," *arXiv:1306.0229*, 2013.

[27] A. Youssry, A. El-Rafei, and S. Elramly, "A quantum mechanics-based framework for image processing and its application to image segmentation," *Quantum Information Processing*, vol. 14, no. 10, pp. 3613–3638, 2015.

[28] Ç. Aytekin, E. C. Ozan, S. Kiranyaz, and M. Gabbouj, "Extended quantum cuts for unsupervised salient object extraction," *Multimedia Tools and Applications*, vol. 76, no. 8, pp. 10443–10463, 2017.

[29] F. Neukart, G. Compostella, C. Seidel, D. von Dollen, S. Yarkoni, and B. Parney, "Traffic flow optimization using a quantum annealer," *Frontiers in ICT*, vol. 4, p. 29, 2017.

[30] L. Susskind and A. Friedman, *Quantum Mechanics: The Theoretical Minimum*. UK: Penguin Books, 2015.

[31] C. Spence, "Crossmodal correspondences: a tutorial review," *Attention, Perception, and Psychophysics*, vol. 73, no. 4, pp. 971–995, 2011.

[32] J. Salamon and E. Gomez, "Melody extraction from polyphonic music signals using pitch contour characteristics," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 1759–1770, Aug 2012.

[33] E. Bigand, S. McAdams, and S. Forêt, "Divided attention in music," *International Journal of Psychology*, vol. 35, no. 6, pp. 270–278, 2000.