A bio-inspired geometric model for sound reconstruction

Ugo Boscain[1], Dario Prandi[2], Ludovic Sacchelli[3], and Giuseppina Turco[4]

[1]CNRS, LJLL, Sorbonne Université, Université de Paris, Inria, Paris, France. ugo.boscain@upmc.fr
[2]Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France. dario.prandi@centralesupelec.fr
[3]Department of Mathematics, Lehigh University, Bethlehem, PA, USA. lus219@lehigh.edu
[4]CNRS, Laboratoire de Linguistique Formelle, Université de Paris, France. gturco@linguist.univ-paris-diderot.fr

April 2, 2020

Abstract

The reconstruction mechanisms built by the human auditory system during sound reconstruction are still a matter of debate. The purpose of this study is to propose a mathematical model of sound reconstruction based on the functional architecture of the auditory cortex (A1). The model is inspired by the geometrical modelling of vision, which has undergone a great development in the last ten years. There are however fundamental dissimilarities, due to the different role played by the time and the different group of symmetries. The algorithm transforms the degraded sound in an 'image' in the time-frequency domain via a short-time Fourier transform. Such an image is then lifted in the Heisenberg group and it is reconstructed via a Wilson-Cowan differo-integral equation. Preliminary numerical experiments are provided, showing the good reconstruction properties of the algorithm on synthetic sounds concentrated around two frequencies.

1 Introduction

Listening to speech requires the capacity of the auditory system to map incoming sensory input to lexical representations. When the sound is intelligible, this mapping ("recognition") process is successful. With reduced intelligibility (e.g., due to background noise), the listener has to face the task of recovering the loss of acoustic information. This task is very complex as it requires a higher cognitive load and the ability of repairing missing input ([27] for a review on noise in speech). Yet, (normal hearing) humans are quite able to recover sounds in several effortful listening situations (e.g. see for instance [26], ranging from sounds degraded at the source (e.g., hypoarticulated and pathological speech), during transmission (e.g., noise, reverberation) or corrupted because of physiological deficits (e.g. hearing loss; [27] among others).

So far, work on degraded speech has informed us a lot on the acoustic cues that help the recogniser to reconstruct missing information (e.g.,[17, 30]); the several adverse conditions in which listeners are able to reconstruct speech sounds (e.g., [2, 27]); and whether (and at which stage of the auditory process) higher-order knowledge (i.e. our information about words and sentences) helps the system to recover lower-level perceptual information (e.g., [20]).

However, most of these studies are typically based on a statistical approach. A mathematical model informing us on how the human auditory system is able to reconstruct a degraded speech sound is still missing. The aim of this study is to build a neuro-geometric model for sound reconstruction.

Mathematical modelling of sensory input reconstruction has made a lot of progresses in the field of vision [31], [13], and [8]. In later years, algorithms inspired by the structure of the primary visual cortex (V1) have been very successful in image processing and in particular for image
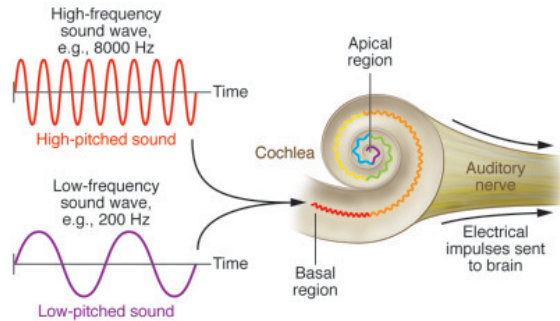
Figure 1: Perceived pitch of a sound depends on the location in the cochlea that the sound wave stimulated. High-frequency sound waves, which correspond to high-pitched noises, stimulate the basal region of the cochlea (red). Low-frequency sound waves are targeted to the apical region of the cochlear structure and correspond with low-pitched sounds (purple). Image from [40].

reconstruction tasks, see, .e.g., [18, 15, 34, 9]. To our knowledge, such work has not been done for sound processing, probably due to the lack of information regarding the primary auditory cortex (A1) with respect to V1.

The model proposed here is highly inspired by the one successfully applied for the primary visual cortex. The analogy between the structure of V1 and A1 is well-grounded on the existence of several biological similarities between the two cortex. For neuroscientists, models of the visual cortex are taken as a starting point for understanding mechanisms of the auditory system (see, for instance, [29] for a comparison, [21] for a related discussion in speech processing). A well-often cited case is the "topographic" organization of the cortex, a general principle according to which the processing of sensory information strongly lies on for mapping visual input and auditory-frequency input to neurons [36].

Within the specific case of the auditory system, sensors (so-called hair cells) are tonotopically organized along the spiral ganglion of the cochlea in a frequency-specific fashion, with cells close to the base of the ganglion being more sensitive to low-frequency sounds and cells near the apex more sensitive to high-frequency sounds, see Figure 1. This early 'spectrogram' of the signal is then transmitted to higher-order layers of the auditory cortex. Strong evidence for V1-A1 analogy comes from studies on animals and on humans with deprived hearing or visual functions showing cross-talk interactions between sensory regions [38, 42]. More relevant for our study is the existence of receptive fields of neurons in V1 and A1 ("simple" and "complex" cells), which supports the idea of a "common canonical processing algorithm within cortical columns" [39]: p.1. The presence of S-cells/C-cells and the appearance of singularities in the feature preference map typical of V1 (e.g., pinwheels) in certain situations [38, 33] speaks in favour of the idea that V1 and A1 share similar mechanisms of sensory input reconstruction. However, there are certain differences to take into account, as explained in the following.

In the next section we present the mathematical model for V1 that will be the basis for our sound reconstruction algorithm.

1.1   Neuro-geometric model of V1

The neuro-geometric model of V1 can be traced back to the work of [22], which, inspired by the experimental results of [23], first proposed to model the primary visual cortex as a *contact space*. This model has then been extended to the so-called sub-Riemannian model in [32], [13], and [8, 7]. On the basis of such a model, exceptionally efficient algorithms for image inpainting have been developed (e.g.[9, 14, 15]), resulting in several medical imaging applications (e.g., [43]).

The main idea behind this model is that an image, seen as a function $f : \mathbb{R}^2 \to \mathbb{R}^+$ representing the grey level, is lifted into a surface in the bundle of the direction of the plane $\mathbb{R}^2 \times P^1$. Here,
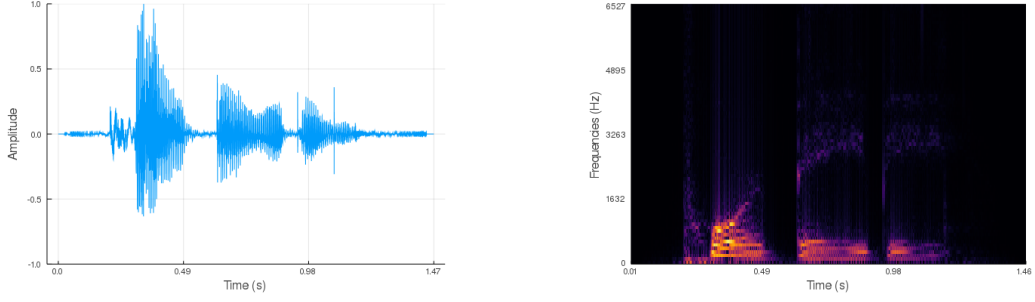
Figure 2: *Left:* Sound signal. *Right:* The corresponding short-time Fourier transform.

P^1 is the space of directions on the plane measured without orientation[1], namely the circle S^1 in which antipodal points are identified. The lift is realized by adding to each point of the image the direction of the tangent line to the level set of f. Under suitable regularity assumptions on f, such lift is a surface S_f. When f is corrupted (i.e. when f is not defined in some region of the plane), the lift is corrupted as well and the reconstruction is obtained by applying a deeply anisotropic diffusion adapted to the problem. Such diffusion mimics the flow of information along the horizontal and vertical connections of V1 and uses as an initial condition the surface S_f and the values of the function f. Mathematicians call such a diffusion the *sub-Riemannian diffusion* in $\mathbb{R}^2 \times P^1$, cf. [28, 1]. One of the main features of this image reconstruction model is that it is invariant by rototranslation of the plane, a feature that will not be possible to translate to the case of sounds, due to the special role of the time variable.

In what follows, we explain how similar ideas could be translated to the problem of sound reconstruction.

1.2 From V1 to sound reconstruction

A sound (that we can think as represented by a function $s : [0, T] \to \mathbb{R}$) is transformed to its time-frequency representation $S : [0, T] \times \mathbb{R} \to \mathbb{C}$, which can be thought as the collection of two black-and-white images: $|S|$ and $\arg S$. The function S depends on two variables: the first one is time, that here we indicate with the letter τ, and the second one is frequency, denoted by ω. Roughly speaking, $|S(\tau, \omega)|$ represents the strength of the presence of the frequency ω at time τ. In the following, we call S the sound image (see Figure 2).

A first attempt to model the process of sound reconstruction into A1 is to apply the same algorithm of image reconstruction discussed above. In a sound image, however, time plays a special role. Indeed:

1. While for images the reconstruction can be done by evolving the whole image at the same time, the sound image is not reaching the auditory cortex all at the same time and thus in this case the reconstruction can be performed only in a sliding window.

2. A rotated sound image corresponds to a completely different original sound and thus the invariance by rototranslations is lost.

As a consequence, different symmetries have to be taken into account (see Appendix B) and a different model for both the lift and the processing in the lifted space is required.

In order to introduce this model, let us recall that, in V1, neural stimulation stems not only from the input but also from its variations. That is, mathematically speaking, the input image is considered as a real valued function on a 2-dimensional space, and the orientation sensitivity arises from the sensitivity to a first order derivative information on this function. Furthermore, the geometric relation between the perceived orientation and the derivatives of the input signal

---

[1]Note that in mathematics, the term "direction" corresponds to what neurophysiologists call "orientation" and viceversa. In this study, we use the mathematical terminology

yields a variational problem on an underlying non-commutative structure. This structure, endowed with a metric naturally associated with the variational problem, gives rise to the sub-Riemannian diffusion [1, 10].

In our model, we follow these same principles in order to study sound inputs *à la* V1. Input sound signals are time dependent real valued functions transformed, by the action of the cochlea, via a short time Fourier transform. As a result the neuronal input is considered as a function of time and frequency. The first time derivative of this object allows to add a supplementary dimension to the domain of the input. Variation of the perceived frequency $\nu = d\omega/d\tau$ can be understood as chirpiness, and this notion gives rise to a natural lift of the signal to the *contact space* in the sense of [22, 31], i.e., $\mathbb{R}^3$ with the Heisenberg group structure. (This structure very often appears in signal processing, see for instance [19] and Appendix B.) As in the case of V1, this observation implies the presence of a non-commutative structure associated with this relation. The hypoelliptic operator associated with this structure is the celebrated Kolmogorov operator [24, 25].

As we already mentioned, the special role played by the time in sound signals, does not permit to model the flow of information as a pure hypoelliptic diffusion, as was done for static images in V1. We thus turn to a more sophisticated and flexible model known as Wilson-Cowan equations [41]. Such a model, based on a differo-integral equation, has been successfully applied to describe the evolution of neural activations and allowed to predict complex perceptual phenomena in V1, such as the emergence of hallucinatory pattern [16, 11]. Recently, these equations have been coupled with the neuro-geometric model of V1 to great benefit. For instance, in [5, 4] they allowed to replicate orientation-dependent brightness illusory phenomena, which had proved to be a hurdle for non-cortical-inspired models. See also [37], for applications to the detection of perceptual units.

On top of these positive results, Wilson-Cowan equations present many advantages from the point of view of A1 modelling: i) they can be applied independently of the underlying structure, which is only encoded in the kernel of the integral term; ii) they allow for a natural implementation of delay terms in the interactions; iii) they can be easily tuned via few parameters with a clear effect on the results. Motivated by these positive results, we emulate this approach in the A1 context. Namely, we will consider the lifted sound image $I(\tau, \omega, \nu)$ to yield an A1 activation $a(\tau, \omega, \nu)$ via the following Wilson-Cowan equations:

$$\partial_\tau a(\tau, \omega, \nu) = -\alpha a(\tau, \omega, \nu) + \beta I(\tau, \omega, \nu) + \gamma \int_{\mathbb{R}^2} w(\omega, \nu \| \omega', \nu') \sigma(a(\tau - \delta, \omega', \nu')) \, d\omega' \, d\nu'. \quad \text{(WC)}$$

Here, (τ, ω, ν) are coordinates on the contact space corresponding to time, frequency, and chirpiness, respectively; $\alpha, \beta, \gamma > 0$ are parameters; $\sigma : \mathbb{C} \to \mathbb{C}$ is a non-linear sigmoid; $w(\omega, \nu \| \omega', \nu')$ is a weight modelling the interaction between (ω, ν) and (ω', ν') via the kernel of the Kolmogorov operator (i.e., taking into account the contact structure of A1); and $\delta > 0$ is a delay. The presence of this delay term models the fact that the time-scale of the input signal and of the neuronal activation are comparable. Wilson-Cowan equations with delay have been applied, e.g., to feedback stabilisation of deep-brain stimulation, cf. [12].

The proposed algorithm to treat a sound signal $s : [0, T] \to \mathbb{R}$, is the following:

A. **Preprocessing:**

    (a) Compute the time-frequency representation $S : [0, T] \times \mathbb{R} \to \mathbb{C}$ of s, via standard short time Fourier transform (STFT);

    (b) Lift this representation to the Heisenberg group, which encodes redundant information about chirpiness, obtaining $I : [0, T] \times \mathbb{R} \times \mathbb{R} \to \mathbb{C}$ (see Section 2 for details);

B. **Processing:** Process the lifted representation I via an adapted version of the Wilson-Cowan equations, obtaining $a : [0, T] \times \mathbb{R} \times \mathbb{R} \to \mathbb{C}$.

C. **Posprocessing:** Invert the preprocessing procedures to a to obtain the resulting sound signal $\hat{s} : [0, T] \to \mathbb{R}$.
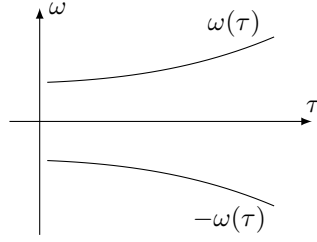
Figure 3: Short-time Fourier transform of the signal in (1), for a positive and increasing $\omega(\cdot)$.

Remark 1. *All the above operations can be streamlined, as they only require the knowledge of the sound on a short window* $[t - \tau, t + \delta t]$.

In Section 2, we present the lift to the contact space and the associated diffusion operator. In Section 3 we apply Wilson-Cowan equations to obtain our model for sound reconstruction. In Section 4, we describe the numerical implementation of the algorithm, together with some of its crucial properties.

This implementation is then tested in Section 5, were we show the results of the algorithm on some simple synthetic signals. Such numerical examples can be listened at `www.github.com/dprn/WCA1`, and should be considered as a very preliminary step toward the construction of an efficient cortical-inspired algorithm for sound reconstruction.

Finally, in Appendix B, we show how the proposed algorithm preserves the natural symmetries of sound signals.

2 The lift to the contact space

It is well-known that the coclhea decomposes the input sound $s : [0, T] \to \mathbb{R}$ in its time-frequency representation $S : [0, T] \times \mathbb{R} \to \mathbb{C}$, obtained via a short-time Fourier tranform (STFT). This corresponds to interpret the "instantaneous sound" at time $\tau \in [0, T]$, instead than as a sound level $s(\tau) \in \mathbb{R}$, as a function $\omega \mapsto S(\tau, \omega)$ which encodes the instantaneous presence of each given frequency, with phase information.

In this section, we present an extension of this procedure, which is at the core of the proposed algorithm. Roughly speaking, the instantaneous sound will be represented as a function $(\omega, \nu) \mapsto I(\tau, \omega, \nu)$, encoding the presence of both the frequency and the chirpiness $\nu = d\omega/d\tau$.

Assume for the moment that the sound is a single (positive) time varying frequency, e.g.,

$$s(\tau) = A \sin(\omega(\tau)\tau), \qquad A \in \mathbb{R}. \tag{1}$$

If the frequency is varying slowly enough and the window of the STFT is large enough, its sound image (up to the choice of normalisations constants in the Fourier transform) coincides roughly with

$$S(\tau, \omega) \sim \frac{A}{2i} \left(\delta(\omega - \omega(\tau)) - \delta(\omega + \omega(\tau)) \right),$$

where δ is the Dirac delta distribution centered at 0. That is, S is concentrated on the two curves $\tau \mapsto (\tau, \omega(\tau))$ and $\tau \mapsto (\tau, -\omega(\tau))$, see Figure 3. Let us focus only on the first curve.

Because of the sensitivity to directions (due to the presence of S-cells and pinwheels, cf. Section 1), the curve $\omega(\tau)$ is lifted in a bigger space by adding a new variable $\nu = d\omega/d\tau$. In mathematical terms the 3-dimensional space (τ, ω, ν) is called the *contact space*. It will be the basis for the geometric model of A1 that we are going to present.

Up to now the curve $\omega(\tau)$ was parameterized by one of the coordinates of the contact space (the variable τ), but it will be more convenient to introduce a new parameter that we call t. The original curve $\omega(\tau)$ is then represented in the space (τ, ω) as $t \mapsto (t, \omega(t))$ (thus imposing $\tau = t$).

Similarly, the lifted curve is parameterized as $t \mapsto (t, \omega(t), \nu(t))$. To every regular enough curve $t \mapsto (t, \omega(t))$, one can associate a lift $t \mapsto (t, \omega(t), \nu(t))$ in the contact space simply by computing $\nu(t) = d\omega/dt$. Vice-versa, a regular enough curve in the contact space $t \mapsto (\tau(t), \omega(t), \nu(t))$ is a lift of planar curve $t \mapsto (t, \omega(t))$ if $\tau(t) = t$ and if $\nu(t) = d\omega/dt$. Now, defining $u(t) = d\nu/dt$ we can say that a curve in the contact space $t \mapsto (\tau(t), \omega(t), \nu(t))$ is a lift of a planar curve if there exists a function $u(t)$ such that:

$$\frac{d}{dt}\begin{pmatrix} \tau \\ \omega \\ \nu \end{pmatrix} = \begin{pmatrix} 1 \\ \nu \\ 0 \end{pmatrix} + u(t)\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \tag{2}$$

Letting $q = (\tau, \omega, \nu)$, equation (2) can be equivalently written as the control system

$$\frac{d}{dt}q(t) = X_0(q(t)) + u(t)X_1(q(t)),$$

where the X_0 and X_1 are the two vector fields in $\mathbb{R}^3$

$$X_0 = \begin{pmatrix} 1 \\ \nu \\ 0 \end{pmatrix}, \quad X_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Notice that defining $X_2 = (0, 1, 0)^\top$, we have the commutation relations (here $[\cdot, \cdot]$ indicates the Lie Bracket) $[X_0, X_1] = X_2$, $[X_0, X_2] = [X_1, X_2] = 0$. Since these are precisely the commutation relations appearing in quantum mechanics, one says that $\{X_0, X_1, X_2\}$ generates the Heisenberg algebra. The corresponding Lie group structure on $\mathbb{R}^3$ is called the Heisenberg group.

Following [8], when s is a general sound signal, we lift each level line of $|S|$. By the implicit function theorem, this yields the following subset of the contact space:

$$\Sigma = \left\{ (\tau, \omega, \nu) \in \mathbb{R}^3 \mid \nu \partial_\omega |S|(\tau, \omega) + \partial_\tau |S|(\tau, \omega) = 0 \right\}.$$

If $|S| \in C^2$ and $\text{Hess}\,|S|$ is non-degenerate, the set Σ is indeed a surface. Finally, the external input from the cochlea is given by

$$I(\tau, \omega, \nu) = S(\tau, \omega)\delta_\Sigma(\tau, \omega, \nu). \tag{3}$$

Here, δ_Σ denotes the Dirac delta distribution concentrated at Σ. The presence of this distributional term is necessary for having a well-defined solution to the evolution equation (WC) introduced in the next section.

2.1 The tonotopical model of A1

On the basis of what described in the previous section and the well-known tonotopical organization of A1 (cf. Section 1), we propose to consider A1 to be the space of $(\omega, \nu) \in \mathbb{R}^2$. The external input at time $t > 0$ corresponding to a sound $s(\cdot)$ to A1 is then given as the slice at $\tau = t$ of the lift I of s to the contact space. That is, hearing an "instantaneous sound level" $s(t)$ reflects in the external input $I(t, \omega, \nu)$ to the "neuron" (ω, ν) in A1 as follows: The "neuron" receives an external charge $S(t, \omega)$ if $(t, \omega, \nu) \in \Sigma$, and no charge otherwise.

Interpreting A1 as a slice of the contact space allows to deduce a natural structure for neuron connections as follows. Going back to a sound composed by a single time-varying frequency $t \mapsto \omega(t)$, we have that its lift is concentrated on the curve $t \mapsto (\omega(t), \nu(t))$, such that

$$\frac{d}{dt}\begin{pmatrix} \omega \\ \nu \end{pmatrix} = Y_0(\omega, \nu) + u(t)Y_1(\omega, \nu),$$

where $Y_0(\omega, \nu) = (\nu, 0)^\top$, $Y_1(\omega, \nu) = (0, 1)^\top$, and $u : [0, T] \to \mathbb{R}$. It is well-known that the PDE naturally associated with the above control system is the following Kolmogorov equation

$$\partial_t I = \mathcal{L}I, \quad \text{where} \quad \mathcal{L} = Y_0 + (Y_1)^2 = \nu \partial_\omega + b \partial_\nu^2, \tag{4}$$

6

for some parameter $b > 0$. In this formula, the vector fields Y_0 and Y_1 are interpreted as first-order differential operators.

The operator $(\partial_t - \mathcal{L})$ is well-known to be hypoelliptic, meaning that if f is a distribution defined on an open set Ω such that $(\partial_t - \mathcal{L})f \in C^\infty(\Omega)$, then $f \in C^\infty(\Omega)$. As a consequence, this operator admits a smooth kernel k_τ, that we will consider to model neuronal connections. In the following result, proved in Appendix A, we recall the explicit expression of k_τ.

Proposition 1. *The integral kernel of equation* (4) *is*

$$k_\tau(\omega, \nu \| \omega', \nu') = \frac{\sqrt{3}}{2\pi b\tau^2} \exp\left(-\frac{g_\tau(\omega, \nu \| \omega', \nu')}{b\tau^3}\right), \tag{5}$$

where,

$$g_\tau(\omega, \nu \| \omega', \nu') = 3(\omega - \omega')^2 + 3\tau(\omega - \omega')(\nu + \nu') + \tau^2\left(\nu^2 + \nu\nu' + \nu'^2\right).$$

In the next section we describe how the redundant representation I is actually processed.

3   The reconstruction model

The procedure presented here is strongly inspired by the successful results obtained by Wilson-Cowan equations in the primary visual cortex [41, 11]. According to this framework, the resulting signal $a : [0, T] \times \mathbb{R} \times \mathbb{R} \to \mathbb{C}$ is the solution of the following equation with delay $\tau > 0$ (which can be seen as encoding the dendritic delays of neuronal connections):

$$\partial_t a(t, \omega, \nu) = -\alpha a(t, \omega, \nu) + \beta I(t, \omega, \nu) + \gamma \int_{\mathbb{R}^2} w(\omega, \nu \| \omega', \nu')\sigma(a(t - \tau, \omega', \nu'))\, d\omega'\, d\nu',$$

with initial condition $a(t, \cdot, \cdot) \equiv 0$ for $t \leq 0$. Here, $\alpha, \beta, \gamma > 0$ are parameters, w is an interaction kernel, and $\sigma : \mathbb{C} \to \mathbb{C}$ is a (non-linear) saturation function, or sigmoid. In the following, we let $\sigma(\rho e^{i\theta}) = \tilde{\sigma}(\rho)e^{i\theta}$ where $\tilde{\sigma}(x) = \min\{1, \max\{0, \kappa x\}\}$, $x \in \mathbb{R}$, for some fixed $\kappa > 0$. When $\gamma = 0$, the above reduces to the standard low-pass filter $\partial_t a = -\alpha a + I$, whose solution is the convolution of the input signal I with the function $\varphi(t)$ given by $\varphi(t) = e^{-t\alpha}$ for $t \in (0, +\infty)$ and $\varphi(t) = 0$ elsewhere. Setting $\gamma \neq 0$ adds a non-linear interaction term on top of this exponential smoothing. In neuroscience applications, this encodes the inhibitory and excitatory interconnections between neurons. As described in Section 2.1, the interaction weight is chosen as $w = k_\tau$, the integral kernel of (4) at time $\tau = \delta$. This integral kernel is well known, and its explicit expression can be found, e.g., in [3]. See also Section 4.

4   Numerical implementation

For the numerical experiments, we chose to implement the proposed algorithm in Julia [6]. As already presented, this process consists in a preprocessing phase, in which we build an input function I on the 3D contact space, a main part, where I is used as the input of the Wilson-Cowan equation (WC), and a post-processing phase, where the reconstructed sound is recovered from the result of the first part.

4.1   Pre-processing

The input sound s is lifted to a time-frequency representation S via a classical implementation of STFT, i.e., by performing FFTs of a windowed discretised input. In the proposed implementation we chose to use a standard Hann window (see, e.g., [35]):

$$w(x) = \begin{cases} \dfrac{1 + \cos(2\pi x/L)}{2} & \text{if } |x| < L/2, \\ 0 & \text{otherwise.} \end{cases}$$

7

The resulting time-frequency signal is then lifted to the contact space through an approximate computation of the gradient $\nabla|S|$ and the following discretisation of (3):

$$I(\tau,\omega,\nu) = \begin{cases} S(\tau,\omega) & \text{if } \nu\partial_\omega|S|(\tau,\omega) = -\partial_\tau|S|(\tau,\omega), \\ 0 & \text{otherwise.} \end{cases}$$

Discretisation issues While the discretisation of the time and frequency domains is a well understood problem, dealing with the additional chirpiness variable requires some care. We will henceforth assume that the significant frequencies of the input sound s belong to a bounded interval $\Lambda \subset \mathbb{R}$. Nevertheless, the set of chirpinesses ν's such that $\nu\partial_\omega|S|(\tau,\omega) = -\partial_\tau|S|(\tau,\omega)$ for some $\tau \in [0,T]$ and $\omega \in \Lambda$ will cover in general the whole real line. This is due to the presence of points where the contour lines of $|S|$ become vertical.

In the numerical implementation we chose to restrict the admissible chirpinesses to a bounded interval $N \subset \mathbb{R}$ obtained as follows. Let $\varrho : \mathbb{R} \to [0,+\infty)$ be the chirpiness density function, which associates to $\nu \in \mathbb{R}$ the measure $\varrho(\nu)$ of the set of those $(\tau,\omega) \in [0,T] \times \Lambda$ such that $\nu\partial_\omega|S|(\tau,\omega) = -\partial_\tau|S|(\tau,\omega)$. Denote by a and σ its expectation and standard deviation, respectively. Then, $N = [e - k\sigma, e + k\sigma]$ where $k \in \mathbb{R}$ is given by

$$k = \arg\min\left\{ \kappa > 0 \mid \frac{\int_{e-\kappa\sigma}^{e+\kappa\sigma} \varrho(\nu)\,d\nu}{\int_{\mathbb{R}} \varrho(\nu)\,d\nu} > p \right\}.$$

Here $p \in [0,1]$ is some tolerance, that in our numerical tests has been taken as $p = 0.95$.

4.2   Processing

Equation (WC) can be solved via a standard forward Euler method. Hence, the delicate part of the numerical implementation is the computation of the interaction term.

As it is clear from the explicit expression given in Proposition 1, k_τ is not a convolution kernel in the sense that $k_\tau(\omega,\nu\|\omega',\nu')$ cannot be expressed as a function of $(\omega - \omega', \nu - \nu')$. As a consequence, a priori we need to explicitly compute all values $k_\tau(\omega,\nu\|\omega',\nu')$ for (ω,ν) and (ω',ν') in the considered domain. As it is customary, in order to reduce computation times, we fix a threshold $\varepsilon > 0$ and for any given (ω,ν) we compute only values for (ω',ν') in the compact set

$$\mathrm{K}^\varepsilon(\omega,\nu) = \{(\omega',\nu') \mid k_\tau(\omega,\nu\|\omega',\nu') \geq \varepsilon\}.$$

The structure of $\mathrm{K}^\varepsilon(\omega,\nu)$ is given in the following, whose proof we defer to Appendix A.

Proposition 2. *For any $\varepsilon > 0$ and $(\omega,\nu) \in \mathbb{R}^2$, we have that $\mathrm{K}^\varepsilon(\omega,\nu)$ is the set of those $(\omega',\nu') \in \mathbb{R}^2$ that satisfy*

$$|\nu - \nu'|^2 \leq C_\varepsilon := -4b\tau \log\left(\frac{2\pi b\tau^2}{\sqrt{3}}\varepsilon\right),$$

$$\left| \omega - \omega' + \frac{\tau(\nu+\nu')}{2} \right| \leq \frac{\tau}{2\sqrt{3}}\sqrt{C_\varepsilon - |\nu - \nu'|^2}.$$

Remark 2. *It holds $C_\varepsilon \geq 0$ if and only if*

$$\varepsilon \leq \frac{\sqrt{3}}{2\pi b\tau^2}.$$

Indeed, for any $(\omega,\nu) \in \mathbb{R}^2$, the r.h.s. above corresponds to $\max k_\tau(\omega,\nu\|\cdot,\cdot)$, and thus $\mathrm{K}^\varepsilon(\omega,\nu) = \varnothing$ for larger values of ε.

The above allows to numerically implement k_τ as a family of sparse arrays. That is, let $G \subset \Lambda \times N$ be the chosen discretisation of the significant set of frequencies and chirpinesses. Then, to $\xi = (\omega, \nu) \in G$ we associate the array $M_\xi : G \to \mathbb{R}$ defined by

$$M_\xi(\xi') = \begin{cases} k_\tau(\xi \| \xi') & \text{if } \xi' \in \mathrm{K}^\varepsilon(\xi), \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, up to choosing the tolerance $\varepsilon \ll 1$ sufficiently small, the interaction term in (WC), evaluated at $\xi = (\omega, \nu) \in G$, can be efficiently estimated by

$$\int_{\mathbb{R}^2} w(\xi \| \xi') \sigma(a(t - \tau, \xi')) \, d\xi' \approx \sum_{\xi' \in K^\varepsilon(\xi)} M_\xi(\xi') a(t - \tau, \xi').$$

4.3 Post-processing

Both operations in the pre-processing phase are inversible: the STFT by inverse STFT, and the lift by integration along the ν variable (that is, summation of the discretized solution). The final output signal is thus obtained by applying the inverse of the pre-processing (integration then inverse STFT) to the solution a of (WC). That is, the resulting signal is given by

$$\hat{s}(t) = \mathrm{STFT}^{-1} \left(\int_{-\infty}^{+\infty} a(t, \omega, \nu) \, d\nu \right).$$

The following guarantees that $\hat{s}$ is real-valued and thus correctly represents a sound signal. From the numerical point of view, this implies that we can focus on solutions of (WC) in the half-space $\{\omega \geq 0\}$, which can then be extended to the whole space by mirror symmetry.

Proposition 3. *It holds that $\hat{s}(t) \in \mathbb{R}$ for all $t > 0$.*

Proof. Let us denote

$$\hat{S}(t, \omega) = \int_{-\infty}^{+\infty} a(t, \omega, \nu) \, d\nu,$$

so that $\hat{s} = \mathrm{STFT}^{-1}(\hat{S})$. Moreover, for any function $f(t, \omega, \nu)$, we let $f^\star(t, \omega, \nu) := \bar{f}(t, -\omega, -\nu)$.

To prove the statement it is then enough to show that

$$a(t, \cdot, \cdot) \equiv a^\star(t, \cdot, \cdot) \qquad \forall t \geq 0. \tag{6}$$

This is trivially satisfied for $t \leq 0$, since in this case $a(t, \cdot, \cdot) \equiv 0$.

We now claim that if the above holds on $[0, T]$ it holds on $[0, T + \tau]$, which will prove (6). By definition of I and the fact that $S(t, -\omega) = \overline{S(t, \omega)}$, we immediately have $I \equiv I^\star$. On the other hand, the explicit expression of k_τ in (5) yields that

$$k_\tau(-\omega, -\nu \| \omega', \nu') = k_\tau(\omega, \nu \| -\omega', -\nu').$$

Then, for all $t \leq T + \tau$, we have

$$\int_{\mathbb{R}^2} k_\tau(-\omega, -\nu \| \omega', \nu') \sigma(a(t - \tau, \omega', \nu')) \, d\omega' \, d\nu'$$

$$= \int_{\mathbb{R}^2} k_\tau(\omega, \nu \| \omega'', \nu'') \sigma(a^\star(t - \tau, \omega'', \nu'')) \, d\omega'' \, d\nu''$$

$$= \int_{\mathbb{R}^2} k_\tau(\omega, \nu \| \omega'', \nu'') \sigma(a(t - \tau, \omega'', \nu'')) \, d\omega'' \, d\nu''.$$

A simple argument, e.g., using the variation of constants method, shows that these two facts imply the claim, and thus the statement. $\qquad \square$
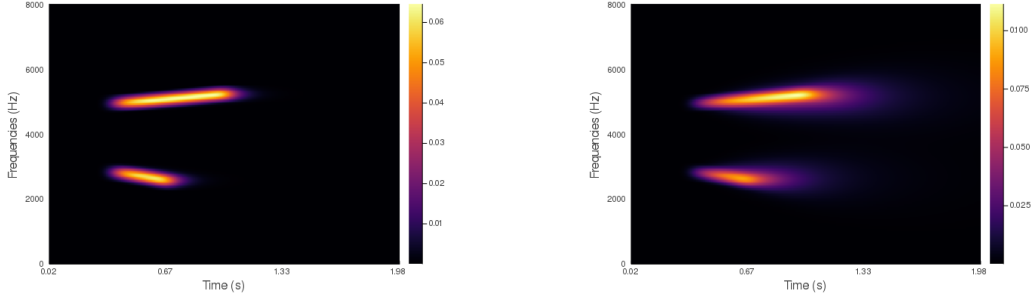
9

Figure 4: Experiment on a synthetic sound with two linearly varying frequencies. *Left:* The STFT of the original sound. *Right:* The STFT of the processed sound with parameters $\alpha = 16$, $\beta = 1$, $\gamma = 18$, $\tau = 0.0625s$, and $b = 4$. In both cases only the positive frequencies are shown: the others are recovered via the Hermitian symmetry of the Fourier transform on real signals.
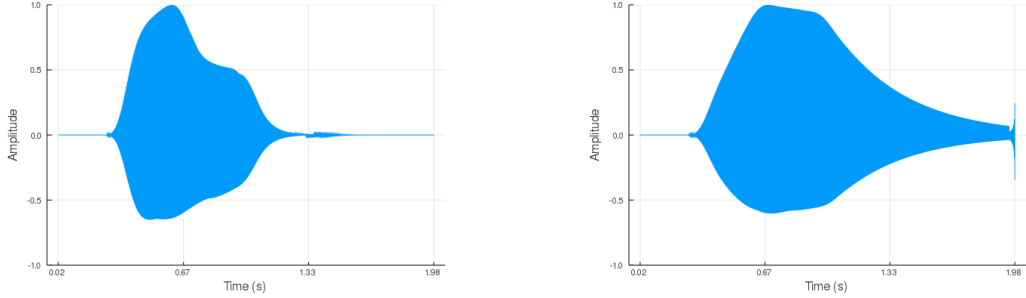

Figure 5: Sound waves corresponding to the STFT of Figure 4. The discontinuity at the end of the resulting sound (*right*) is due to the cut-off of the window function used for the STFT at the boundary.

5 Experiments

In Figures 4 and 5 we present a simple synthetic experiment, where the input sound is assumed to consist of two distinct frequencies depending linearly on time. One observes that the processed sound presents the same features, but with a longer duration. This is especially true for the higher-frequency component, since this is taken to be dominant in the starting signal.

6 Conclusion

In this work we presented a sound reconstruction framework inspired by the analogies between visual and auditory cortices. Building upon the successful cortical inspired image reconstruction algorithms, the proposed framework lifts time-frequencies representations of signals to the 3D contact space, by adding instantaneous chirpiness information. These redundant representations are then processed via adapted diffeo-integral Wilson-Cowan equations.

The promising results obtained on simple synthetic sounds, although preliminary, suggest possible applications of this framework to the problem of degraded speech. This should be done via psycholinguistic experiments, testing the reconstruction ability of normal-hearing humans on originally degraded speech material compared to the same material after algorithm reconstruction. Such an endeavour will contribute to further the understanding of the auditory mechanisms emerging in effortful listening conditions and help to refine our knowledge on current theories and models of human speech perception as well as on general organization principles underlying the functioning of the human cortex.

A Integral kernel of the Kolmogorov operator

The result in Proposition 1 is well-known. Here, we show how to derive it from [3, Proposition 9]. W.r.t. the notations of that paper, (4) is obtained by considering $n = 2$ and

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} 0 \\ \sqrt{2b} \end{pmatrix}.$$

Then, letting $x = (\omega, \nu)$ and $x' = (\omega', \nu')$ the kernel is

$$k_\tau(\omega, \nu \| \omega', \nu') = \frac{1}{2\pi\sqrt{\det D_\tau}} \exp\left[-\frac{1}{2} \left(x' - \mathrm{e}^{\tau A} x \right)^\top D_\tau^{-1} \left(x' - \mathrm{e}^{\tau A} x \right) \right],$$

where

$$D_\tau = \mathrm{e}^{\tau A} \left[\int_0^\tau \mathrm{e}^{-\sigma A} BB^* \, \mathrm{e}^{-\sigma A^*} \, \mathrm{d}\sigma \right] \mathrm{e}^{\tau A^*}.$$

Direct computations yield

$$D_\tau = 2b \begin{pmatrix} \tau^3/3 & \tau^2/2 \\ \tau^2/2 & \tau \end{pmatrix}, \quad \text{and} \quad \det D_\tau = \frac{b^2\tau^4}{3}.$$

Therefore,

$$\frac{1}{2} D_\tau^{-1} = \frac{1}{b\tau^3} M, \quad \text{where} \quad M = \begin{pmatrix} 3 & -3\tau/2 \\ -3\tau/2 & \tau^2 \end{pmatrix}.$$

Finally, the statement follows by letting

$$g_\tau(x\|x') = \left(x' - \mathrm{e}^{\tau A} x \right)^\top M \left(x' - \mathrm{e}^{\tau A} x \right).$$

We now turn to an argument for Proposition 2. Observe that $k_\tau(x\|x') \geq \varepsilon$ if and only if

$$g_\tau(x\|x') \leq \eta := -b\tau^3 \log\left(\frac{2\pi b\tau^2}{\sqrt{3}} \varepsilon \right).$$

Then, we start by solving $z^\top M z \leq \eta$, for $z \in \mathbb{R}^2$. One can check that this is verified if and only if

$$|z_2| \leq \sqrt{\frac{4\eta}{\tau^2}}, \quad \text{and} \quad \left| z_1 - \frac{\tau}{2} z_2 \right| \leq \frac{\tau}{2\sqrt{3}} \sqrt{\frac{4\eta}{\tau^2} - z_2^2}.$$

Since $C_\varepsilon = 4\eta/\tau^2$ the statement follows by computing the above at $z = x' - e^{\tau A} x$.

B Heisenberg group action on the contact space

Recall that the short-time Fourier transform of a signal $s \in L^2(\mathbb{R})$ is given by

$$S(\tau, \omega) := \mathrm{STFT}(s)(\tau, \omega) = \int_\mathbb{R} s(t)\gamma(\tau - t)e^{2\pi it\omega} \, dt.$$

Here, $\gamma : \mathbb{R} \to [0, 1]$ is a compactly supported (smooth) window, so that $S \in L^2(\mathbb{R}^2)$. Fundamental operators in time frequency analysis [19] are the shifts, acting on signals $s \in L^2(\mathbb{R})$,

$$T_\theta s(t) := s(t - \theta) \quad \text{and} \quad M_\lambda s(t) := e^{2\pi i\lambda t} s(t),$$

for $\theta, \lambda \in \mathbb{R}$. One easily checks that T_θ and M_λ are unitary operators on $L^2(\mathbb{R})$. By conjugation with the short-time Fourier transform, they naturally define the unitary operators on $L^2(\mathbb{R}^2)$ given by

$$T_\theta S(\tau, \omega) = e^{-2\pi i \omega \theta} S(\tau - \theta, \omega),$$
$$M_\lambda S(\tau, \omega) = S(\tau, \omega - \lambda).$$

The relevance of the Heisenberg group in time-frequency analysis is a direct consequence of the commutation relation

$$T_\theta M_\lambda T_\theta^{-1} M_\lambda^{-1} = e^{-2\pi i \lambda \theta} \operatorname{Id}.$$

Indeed, this shows that the operator algebra generated by $(T_\theta)_{\theta \in \mathbb{R}}$ and $(M_\lambda)_{\lambda \in \mathbb{R}}$ coincides with the Heisenberg group $\mathbb{H}^1$ via the representation $U : \mathbb{H}^1 \to \mathcal{U}(L^2(\mathbb{R}^2))$ defined by

$$U(\theta, \lambda, \zeta) = e^{-2\pi i \zeta} T_\theta M_\lambda.$$

Since $L^2(\mathbb{R}^3) = L^2(\mathbb{R}^2) \otimes L^2(\mathbb{R})$, the above representation trivially yields the representation $U \otimes \operatorname{Id}$ on $\mathcal{U}(L^2(\mathbb{R}^3))$ that we still denote, abusing notation, by U. Recall now that the lift operation defined in Section 2 associates with $S \in L^2(\mathbb{R}^2)$ a distribution of the form $I(\tau, \omega, \nu) = S(\omega, \nu) \delta_\Sigma(\tau, \omega, \nu)$ for some $\Sigma \subset \mathbb{R}^3$. Due to the fact that Σ is defined via the modulus of S it ignores the phase factors appearing in the representation U. Thus, U naturally extends to this class of inputs as

$$\begin{aligned}
\big[U(\theta, \lambda, \zeta) I \big](\tau, \omega, \nu) \\
= e^{-2\pi i (\zeta + \omega \theta)} I(\tau - \theta, \omega - \lambda, \nu) \\
= e^{-2\pi i (\zeta + \omega \theta)} \delta_\Sigma(\tau - \theta, \omega - \lambda, \nu) S(\tau - \theta, \omega - \lambda).
\end{aligned}$$

One observes that the above coincides with the result of the lift operation applied to $U(\theta, \lambda, \zeta) S$. That is, the lift operation commutes with the $\mathbb{H}^1$ symmetry underlying time-frequency analysis.

The same goes for the evolution (WC). The commutation is trivial for the linear terms. On the other hand, the non-linearity introduced in the integral term commutes with U thanks to the fact that $\sigma(\rho e^{i\phi}) = e^{i\phi} \sigma(\rho)$ for all $\rho > 0, \phi \in \mathbb{R}$.

References

[1] Andrei Agrachev, Davide Barilari, and Ugo Boscain. *A Comprehensive Introduction to Sub-Riemannian Geometry*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2019.

[2] Peter Assmann and Quentin Summerfield. *The Perception of Speech Under Adverse Conditions*, pages 231–308. Springer New York, New York, NY, 2004.

[3] Davide Barilari and Francesco Boarotto. Kolmogorov-fokker-planck operators in dimension two: heat kernel and curvature, 2017.

[4] Marcelo Bertalmío, Luca Calatroni, Valentina Franceschi, Benedetta Franceschiello, Alexander Gomez Villa, and Dario Prandi. Visual illusions via neural dynamics: Wilson-cowan-type models and the efficient representation principle. *Journal of Neurophysiology*, 2020. PMID: 32159409.

[5] Marcelo Bertalmío, Luca Calatroni, Valentina Franceschi, Benedetta Franceschiello, and Dario Prandi. A cortical-inspired model for orientation-dependent contrast perception: A link with wilson-Cowan equations. In *Scale Space and Variational Methods in Computer Vision*, Cham, 2019. Springer International Publishing.

[6] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017.

[7] U. Boscain, R. Chertovskih, J.-P. Gauthier, and A. Remizov. Hypoelliptic diffusion and human vision: a semi-discrete new twist on the petitot theory. *SIAM J. Imaging Sci.*, 7(2):669–695, 2014.

[8] Ugo Boscain, Jean Duplaix, Jean-Paul Gauthier, and Francesco Rossi. Anthropomorphic image reconstruction via hypoelliptic diffusion, 2010.

[9] Ugo V. Boscain, Roman Chertovskih, Jean-Paul Gauthier, Dario Prandi, and Alexey Remizov. Highly corrupted image inpainting through hypoelliptic diffusion. *J. Math. Imaging Vision*, 60(8):1231–1245, 2018.

[10] Marco Bramanti. *An invitation to hypoelliptic operators and Hörmander's vector fields*. SpringerBriefs in Mathematics. Springer, Cham, 2014.

[11] P C Bressloff, J D Cowan, M Golubitsky, P J Thomas, and M C Wiener. Geometric visual hallucinations, Euclidean symmetry and the functional architecture of striate cortex. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 356(1407):299–330, 2001.

[12] Antoine Chaillet, Georgios Detorakis, Stéphane Palfi, and Suhan Senova. Robust stabilization of delayed neural fields with partial measurement and actuation. *Automatica*, 83, 2017.

[13] G Citti and a. Sarti. A Cortical Based Model of Perceptual Completion in the Roto-Translation Space. *J. Math. Imaging Vis.*, 24(3):307–326, feb 2006.

[14] Remco Duits and EM Franken. Left-invariant parabolic Evolutions on SE(2) and Contour Enhancement via Invertible Orientation Scores. Part I: Linear Left-invariant Diffusion Equations on SE. *Quarterly of Appl. Math., AMS*, 2010.

[15] Remco Duits and EM Franken. Left-invariant parabolic evolutions on SE(2) and contour enhancement via invertible orientation scores. Part II: nonlinear left-invariant diffusions on invertible orientation scores. *Q. Appl. Math.*, (0):1–38, 2010.

[16] G. B. Ermentrout and J D Cowan. A mathematical theory of visual hallucination patterns. *Biological cybernetics*, 34:137–150, 1979.

[17] Tânia Fernandes, Paulo Ventura, and Régine Kolinsky. Statistical information and coarticulation as cues to word boundaries: A matter of signal quality. *Perception & Psychophysics*, 69(6):856–864, Aug 2007.

[18] E Franken and R Duits. Crossing-Preserving Coherence-Enhancing Diffusion on Invertible Orientation Scores. *International Journal of Computer Vision*, 85(3):253–278, 2009.

[19] Karlheinz Gröchenig. *Foundations of time-frequency analysis*. Applied and Numerical Harmonic Analysis. Birkhäuser Boston, Inc., Boston, MA, 2001.

[20] R. Hannemann, J. Obleser, and C. Eulitz. Top-down knowledge supports the retrieval of lexical information from degraded speech. *Brain Research*, 1153:134 – 143, 2007.

[21] Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402, 2007.

[22] William C. Hoffman. The visual cortex is a contact bundle. *Appl. Math. Comput.*, 32(2-3):137–167, August 1989.

[23] D. H. Hubel and T. N. Wiesel. Receptive fields of single neurons in the cat's striate cortex. *The Journal of Physiology*, 148(3):574–591, 1959.

[24] A. Kolmogoroff. Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. *Math. Ann.*, 104(1):415–458, 1931.

[25] A. Kolmogoroff. Zur Grössenordnung des Restgliedes Fourierscher Reihen differenzierbarer Funktionen. *Ann. of Math. (2)*, 36(2):521–526, 1935.

[26] Paul A. Luce and Conor T. McLennan. *Spoken Word Recognition: The Challenge of Variation*, chapter 24, pages 590–609. "John Wiley & Sons, Ltd", 2008.

[27] Sven L. Mattys, Matthew H. Davis, Ann R Bradlow, and Sophie K. Scott. Speech recognition in adverse conditions: A review. *Language, Cognition and Neuroscience*, 27(7-8):953–978, 9 2012.

[28] Richard Montgomery. *A tour of subriemannian geometries, their geodesics and applications*, volume 91 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2002.

[29] Israel Nelken and Mike B. Calford. *Processing Strategies in Auditory Cortex: Comparison with Other Sensory Modalities*, pages 643–656. Springer US, Boston, MA, 2011.

[30] Gaurang Parikh and Philipos C. Loizou. The influence of noise on vowel and consonant cues. *The Journal of the Acoustical Society of America*, 118(6):3874–3888, 2005.

[31] Jean Petitot and Yannick Tondut. Vers une neurogéométrie. fibrations corticales, structures de contact et contours subjectifs modaux. *Mathématiques et Sciences humaines*, 145:5–101, 1999.

[32] Jean Petitot and Yannick Tondut. Vers une Neurogéométrie. Fibrations corticales, structures de contact et contours subjectifs modaux. pages 1–96, 1999.

[33] Thomas W. Polger, Lawrence A. Shapiro, and Oxford University Press. *The multiple realization book*. Oxford University Press, Oxford, 2016.

[34] Dario Prandi and Jean-Paul Gauthier. *A semidiscrete version of the Citti-Petitot-Sarti model as a plausible model for anthropomorphic image reconstruction and pattern recognition*. SpringerBriefs in Mathematics. Springer International Publishing, Cham, 2017.

[35] William H Press, Saul A Teukolsky, William T Vetterling, and Brian P Flannery. *Numerical Recipes*. Cambridge University Press, Cambridge, 3rd editio edition, 2007.

[36] Josef P Rauschecker. Auditory and visual cortex of primates: a comparison of two sensory systems. *The European journal of neuroscience*, 41(5):579–585, 03 2015.

[37] Alessandro Sarti and Giovanna Citti. The constitution of visual perceptual units in the functional architecture of v1. *Journal of Computational Neuroscience*, 38(2), Apr 2015.

[38] Jitendra Sharma, Alessandra Angelucci, and Mriganka Sur. Induction of visual orientation modules in auditory cortex. *Nature*, 404(6780):841–847, 2000.

[39] Biao Tian, Paweł Kuśmierek, and Josef P. Rauschecker. Analogues of simple and complex cells in rhesus monkey auditory cortex. *Proceedings of the National Academy of Sciences*, 110(19):7892–7897, 2013.

[40] Corinne Williams. Hearing restoration: Graeme clark, ingeborg hochmair, and blake wilson receive the 2013 lasker debakey clinical medical research award. *The Journal of Clinical Investigation*, 123(10):4102–4106, 10 2013.

[41] H R Wilson and J D Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal*, 12(1):1–24, jan 1972.

[42] Robert J. Zatorre. Do you see what i'm saying? interactions between auditory and visual cortices in cochlear implant users. *Neuron*, 31(1):13 – 14, 2001.

[43] J. Zhang, B. Dashtbozorg, E. Bekkers, J. P. W. Pluim, R. Duits, and B. M. ter Haar Romeny. Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores. *IEEE Transactions on Medical Imaging*, 35(12):2631–2644, Dec 2016.