An auditory cortex model for sound processing

Rand Asswad[1][???], Ugo Boscain[2][0000−0001−5450−275X], Giuseppina
Turco[3][0000−0002−5963−1857], Dario Prandi[1][0000−0002−8156−5526], and Ludovic
Sacchelli[4][0000−0003−3838−9448]

[1] Université Paris-Saclay, CNRS, CentraleSupèlec, Laboratoire des signaux et
systèmes, 91190, Gif-sur-Yvette, France
{rand.asswad, dario.prandi}@centralesupelec.fr
[2] CNRS, LJLL, Sorbonne Université, Université de Paris, Inria, Paris, France
ugo.boscain@upmc.fr
[3] CNRS, Laboratoire de Linguistique Formelle, UMR 7110, Université de Paris,
Paris, France gturco@linguist.univ-paris-diderot.fr
[4] Université Lyon, Université Claude Bernard Lyon 1, CNRS, LAGEPP UMR 5007,
43 bd du 11 novembre 1918, F-69100, Villeurbanne, France
ludovic.sacchelli@univ-lyon1.fr

Abstract. The reconstruction mechanisms built by the human auditory
system during sound reconstruction are still a matter of debate. The pur-
pose of this study is to refine the auditory cortex model introduced in
[9], and inspired by the geometrical modelling of vision. The algorithm
transforms the degraded sound in an 'image' in the time-frequency do-
main 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. Numerical experiments on a library of speech
recordings are provided, showing the good reconstruction properties of
the algorithm.

Keywords: Auditory cortex · Heisenberg group · Wilson-Cowan equa-
tion · Kolmogorov operator.

1 Introduction

Human capacity for speech recognition with reduced intelligibility has mostly
been studied from a phenomenological and descriptive point of view (see [19]
for a review on noise in speech, as well as a wide range of situations in [2, 18]).
What is lacking is a proper mathematical model informing us on how the human
auditory system is able to reconstruct a degraded speech sound. The aim of this
study is to provide a neuro-geometric model for sound reconstruction based on
the description of the functional architecture of the auditory cortex.

Knowledge on the functional architecture of the auditory cortex and the
principles of auditory perception are limited. For that reason, we turn to recent
advances in the mathematical modeling of the functional architecture of the pri-
mary visual cortex and the processing of visual inputs [22, 12, 8] (which recently

yield very successful applications to image processing [15, 13, 24, 10]) to extrapolate a model of the auditory cortex. This idea is not new: neuroscientists take models of V1 as a starting point for understanding the auditory system (see, e.g., [20]). Indeed, biological similarities between the structure of the primary visual cortex (V1) and the primary auditory cortex (A1) are well-known to exist. V1 and A1 share a "topographic" organization, a general principle determining how visual and auditory inputs are mapped to those neurons responsible for their processing [26]. Furthermore, the existence of receptive fields of neurons in V1 and A1 that allow for a subdivision of neurons in "simple" and "complex" cells supports the idea of a "common canonical processing algorithm within cortical columns" [27].

2   The contact space approach in V1

The neuro-geometric model of V1 finds its roots in the experimental results of Hubel and Wiesel [16]. This gave rise to the so-called sub-Riemannian model of V1 in [23, 12, 8, 24]. 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 to a distribution on $\mathbb{R}^2 \times P^1$, the bundle of directions of the plane. Here, P^1 is the projective line, i.e., $P^1 = \mathbb{R}/\pi\mathbb{Z}$. More precisely, the lift is given by $Lf(x, y, \theta) = \delta_{Sf}(x, y, \theta)f(x, y)$ where δ_{S_f} is the Dirac mass supported on the set $S_f \subset \mathbb{R}^2 \times P^1$ of points (x, y, θ) such that θ is the direction of the tangent line to f at (x, y).

When f is corrupted (i.e. when f is not defined in some region of the plane), the reconstruction is obtained by applying a deeply anisotropic diffusion mimicking the flow of information along the horizontal and vertical connections of V1, with initial condition L_f. This diffusion is known as the *sub-Riemannian diffusion* in $\mathbb{R}^2 \times P^1$, cf. [1]. One of the main features of this diffusion 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.

The V1-inspired pipeline is then the following: first a lift of the input signal to an adequate contact space, then a processing of the lifted signal according to sub-Riemannian diffusions, then projection of the output to the signal space.

3   The model of A1

The sensory input reaching A1 comes directly from the cochlea. The sensors are tonotopically organized (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 implies that sound is transmitted to the primary auditory cortex A1 in the form of a 'spectrogram': when a sound $s : [0, T] \to \mathbb{R}$ is heard, A1 is fed with its time-frequency representation[5] $S : [0, T] \times \mathbb{R} \to \mathbb{C}$. If $s \in L^2(\mathbb{R}^2)$, as given by the

---

[5] Actually, its spectrogram $|S| : [0, T] \times \mathbb{R} \to [0, +\infty)$.
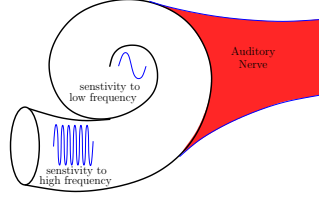
Fig. 1. Perceived pitch of a sound depends on the stimulated location in the cochlea.

short-time Fourier transform of s, that is

$$S(\tau, \omega) := \text{STFT}(s)(\tau, \omega) = \int_{\mathbb{R}} s(t) W(\tau - t) e^{2\pi i t \omega} \, dt. \tag{1}$$

Here, $W : \mathbb{R} \to [0, 1]$ is a compactly supported (smooth) window, so that $S \in L^2(\mathbb{R}^2)$. 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. Since S is complex-valued, it can be thought as the collection of two black-and-white images: $|S|$ and $\arg S$. 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. The V1-inspired approach is then to apply image perception models on $|S|$. However, in this case time plays a special role: the whole sound image does not reach the auditory cortex simultaneously, but sequentially, and invariance by image rotations is lost. As a consequence, different symmetries have to be taken into account and a different model for both the lift and the processing in the contact space are required.

3.1 The lift procedure

We propose an extension of the time-frequency representation of a sound, which is at the core of the proposed algorithm. Sensitivity to variational information, such as the tangent to curves for V1, is now translated in a manner that takes into account the role of time. A regular curve $t \mapsto (t, \omega(t))$ in the time frequency domain (τ, ω) is lifted to a 3-dimensional *augmented space* by adding a new variable $\nu = d\omega/d\tau$. The variation ν in instantaneous frequency is then associated to the chirpiness of the sound. Alternatively, a curve $t \mapsto (\tau(t), \omega(t), \nu(t))$ is a lift of planar curve $t \mapsto (t, \omega(t))$ if $\tau(t) = t$ and if $\nu(t) = d\omega/dt$. Setting $u(t) = d\nu/dt$ 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:

$$\frac{d}{dt} \begin{pmatrix} \tau \\ \omega \\ \nu \end{pmatrix} = \begin{pmatrix} 1 \\ \nu \\ 0 \end{pmatrix} + u \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} = X_0(\tau, \omega, \nu) + u X_1(\tau, \omega, \nu) \tag{2}$$

The vector fields (X_0, X_1) generate the Heisenberg group. However, we are not dealing here with the proper sub-Riemannian distribution, since $\{X_0 + u X_1 \mid u \in \mathbb{R}\}$ is only one-dimensional.

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 augmented space:

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

The external input from the cochlea to the augmented space is then given by

$$I(\tau, \omega, \nu) = S(\tau, \omega) \delta_\Sigma(\tau, \omega, \nu) = \begin{cases} S(\tau, \omega) & \text{if } \nu \partial_\omega |S|(\tau, \omega) = -\partial_\tau |S|(\tau, \omega), \\ 0 & \text{otherwise}, \end{cases} \tag{4}$$

with δ_Σ denoting the Dirac delta distribution concentrated on Σ.

3.2   Associated interaction kernel.

Considering A1 as a slice of the augmented space allows to deduce a natural structure for neuron connections. For a single time-varying frequency $t \mapsto \omega(t)$, its lift is concentrated on the curve $t \mapsto (\omega(t), \nu(t))$, such that

$$\frac{d}{dt} \begin{pmatrix} \omega \\ \nu \end{pmatrix} = Y_0(\omega, \nu) + u(t) Y_1(\omega, \nu), \tag{5}$$

where $Y_0(\omega, \nu) = (\nu, 0)^\top$, $Y_1(\omega, \nu) = (0, 1)^\top$, and $u : [0, T] \to \mathbb{R}$.

As in the case of V1 [7], we model neuronal connections via these dynamics. In practice, this amounts to assume that the excitation starting at a neuron $X_0 = (\omega', \nu')$ evolves as the stochastic process $\{A_t\}_{t \geq 0}$ following the SDE $dA_t = Y_0(A_t)dt + Y_1(A_t)dW_T$, $\{W_t\}_{t \geq 0}$ being a Wiener process, with inital condition $A_0 = (\omega', \nu')$. The generator of $\{A_t\}_{t \geq 0}$ is the operator $\mathcal{L} = Y_0 + (Y_1)^2$.

The influence $k_\delta(\omega, \nu \| \omega', \nu')$ of neuron (ω', ν') on neuron (ω, ν) at time $\delta > 0$ is then modeled by the transition density of the process $\{A_t\}_{t \geq 0}$, given by the integral kernel at time δ of the Fokker-Planck equation

$$\partial_t I = \mathcal{L}^* I, \quad \text{where} \quad \mathcal{L}^* = -Y_0 + (Y_1)^2 = -\nu \partial_\omega + b \partial_\nu^2. \tag{6}$$

The vector fields Y_0 and Y_1 are interpreted as first-order differential operators and the scaling parameter $b > 0$ models the relative strength of the two terms. We stress that an explicit expression of k_δ is well-known (see, for instance [3]) and is a consequence of the hypoellipticity of $(\partial_t - \mathcal{L}^*)$.

3.3   Processing of the lifted signal.

As we already mentioned, the special role played by 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 different kind of model: Wilson-Cowan integro-differential equations [28]. This model has been successfully applied to describe the evolution of neural activations, in V1 in particular, where it allowed to predict complex perceptual phenomena such as the emergence of patterns of hallucinatory [14, 11] or illusory [5, 4, 6] nature. It has also been used in various computational models of the auditory cortex [17, 25, 29].

Wilson-Cowan equations also 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.

On the basis of these positive results, we emulate this approach in the A1 context. Namely, we consider the lifted sound image $I(\tau, \omega, \nu)$ to yield an A1 activation $a : [0, T] \times \mathbb{R} \times \mathbb{R} \to \mathbb{C}$, solution of

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

with initial condition $a(t, \cdot, \cdot) \equiv 0$ for $t \leq 0$. Here, $\alpha, \beta, \gamma > 0$ are parameters, k_δ is the interaction kernel, and $\sigma : \mathbb{C} \to \mathbb{C}$ is a (non-linear) saturation function, or sigmoid. We pick $\sigma(\rho e^{i\theta}) = \tilde{\sigma}(\rho) e^{i\theta}$ where $\tilde{\sigma}(x) = \min\{1, \max\{0, \kappa x\}\}$, $x \in \mathbb{R}$, for some fixed $\kappa > 0$.

The presence of a delay δ in the activation appearing in the integral inter-action term (WC) models the fact that the time-scale of the input signal and of the neuronal activation are comparable. When $\gamma = 0$, equation (WC) is a standard low-pass filter $\partial_t a = -\alpha a + I$. Setting $\gamma \neq 0$ adds a non-linear delayed interaction term on top of this exponential smoothing, encoding the inhibitory and excitatory interconnections between neurons.

3.4   Algorithm pipeline.

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

$$\hat{s}(t) = \text{STFT}^{-1} \left(\int_{-\infty}^{+\infty} a(t, \omega, \nu) \, d\nu \right). \tag{7}$$

It is guaranteed 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.

The resulting algorithm to process a sound signal $s : [0, T] \to \mathbb{R}$ is then:

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 redun-dant information about chirpiness, obtaining $I : [0, T] \times \mathbb{R} \times \mathbb{R} \to \mathbb{C}$;
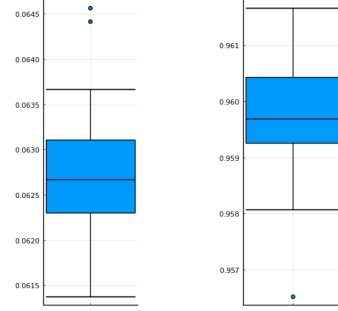
B. **Processing:** Process the lifted representation I via a Wilson-Cowan equation adapted to the Heisenberg structure, obtaining $a : [0, T] \times \mathbb{R} \times \mathbb{R} \to \mathbb{C}$.
C. **Postprocessing:** Project a to the processed time-frequency representation $\hat{S} : [0, T] \times \mathbb{R} \to \mathbb{C}$ and then apply an inverse STFT to obtain the resulting sound signal $\hat{s} : [0, T] \to \mathbb{R}$.

4   Numerical experiments

The discretization of the time and frequency domains is determined by the sampling rate of the original signal and the window size chosen in the STFT procedure. That is, by the Nyquist-Shannon sampling theorem, for a temporal sampling rate δt and a window size of T_w, we consider the frequencies ω such that $|\omega| < 1/(2\delta t)$, with a finest discretization rate of $1/(2T_w)$. Observe, in particular, that the frequency domain is bounded. Nevertheless, the chirpiness ν defined as $\nu \partial_\omega |S| (\tau, \omega) + \partial_\tau |S| (\tau, \omega) = 0$ is unbounded and, since generically there exists points such that $\partial_\omega |S|(\tau_0, \omega_0) = 0$, it stretches over the entire real line.

To overcome this problem, a natural strategy is to model the chirpiness values as a random variable, and considering only chirpinesses falling inside the confidence interval I_p for some reasonable p-value (e.g., $p = 0.95$). A reasonable assumption for the distribution of X is that it follows a Cauchy distribution $\text{Cauchy}(x_0, \gamma)$. Indeed, this corresponds to assuming that $\partial_\omega |S|$ and $\partial_\tau |S|$ are normal (and independent) distributions [21]. As it is customary, we chose as estimator for the location parameter x_0 the median of X and for the scale parameter γ half the interquartile distance.

Fig. 2. Box plots for estimated Cauchy distributions of speech signals chirpiness. *Left:* Kolmogorov-Smirnov statistic values. *Right:* percentage of values falling in $I_{0.95}$

Although statistical tests on a library of real-world speech signals [5] rejected the assumption that $X \sim \text{Cauchy}(x_0, \gamma)$, the fit is quite good according to the Kolmogorov-Smirnov statistic $D_n = \sup_x |F_n(x) - F_X(x)|$. Here, F_X is the cumulative distribution function of X and F_n is the empirical distribution function evaluated over the chirpiness values.

---

[5] The speech material used in the current study is part of an ongoing psycholinguistic project on spoken word recognition. Speech material comprises 49 Italian words and 118 French words. The two sets of words were produced by two (40-year-old) female speakers (a French monolingual speaker and an Italian monolingual speaker) and recorded using a headset microphone AKG C 410 and a Roland Quad Capture audio interface. Recordings took place in the soundproof cabin of the Laboratoire de Phonétique et Phonologie (LPP) of Université de Paris Sorbonne-Nouvelle. Both informants were told to read the set of words as fluently and naturally as possible.
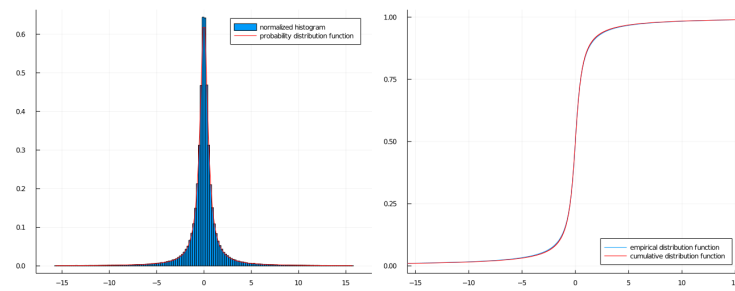
Fig. 3. Chirpiness of a speech signal compared to Cauchy distribution

References

1. Agrachev, A., Barilari, D., Boscain, U.: A Comprehensive Introduction to Sub-Riemannian Geometry. Cambridge Studies in Advanced Mathematics, Cambridge University Press (2019)
2. Assmann, P., Summerfield, Q.: The Perception of Speech Under Adverse Conditions, pp. 231–308. Springer New York, New York, NY (2004)
3. Barilari, D., Boarotto, F.: Kolmogorov-fokker-planck operators in dimension two: heat kernel and curvature (2017)
4. Bertalmío, M., Calatroni, L., Franceschi, V., Franceschiello, B., Gomez Villa, A., Prandi, D.: Visual illusions via neural dynamics: Wilson-cowan-type models and the efficient representation principle. Journal of Neurophysiology (2020)
5. Bertalmío, M., Calatroni, L., Franceschi, V., Franceschiello, B., Prandi, D.: A cortical-inspired model for orientation-dependent contrast perception: A link with wilson-Cowan equations. In: Scale Space and Variational Methods in Computer Vision. Springer International Publishing, Cham (2019)
6. Bertalmìo, M., Calatroni, L., Franceschi, V., Franceschiello, B., Prand i, D.: Cortical-inspired Wilson-Cowan-type equations for orientation-dependent contrast perception modelling. J. Math. Imaging Vision (2020)
7. Boscain, U., Chertovskih, R., Gauthier, J.P., Remizov, A.: Hypoelliptic diffusion and human vision: a semi-discrete new twist on the petitot theory. SIAM J. Imaging Sci. **7**(2), 669–695 (2014)
8. Boscain, U., Duplaix, J., Gauthier, J.P., Rossi, F.: Anthropomorphic image reconstruction via hypoelliptic diffusion (2010)
9. Boscain, U., Prandi, D., Sacchelli, L., Turco, G.: A bio-inspired geometric model for sound reconstruction. The Journal of Mathematical Neuroscience **11**(1), 2 (2021)
10. Boscain, U.V., Chertovskih, R., Gauthier, J.P., Prandi, D., Remizov, A.: Highly corrupted image inpainting through hypoelliptic diffusion. J. Math. Imaging Vision **60**(8), 1231–1245 (2018)
11. Bressloff, P.C., Cowan, J.D., Golubitsky, M., Thomas, P.J., Wiener, M.C.: 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. Citti, G., Sarti, a.: A Cortical Based Model of Perceptual Completion in the Roto-Translation Space. J. Math. Imaging Vis. **24**(3), 307–326 (feb 2006)

13. Duits, R., Franken, E.: 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)
14. Ermentrout, G.B., Cowan, J.D.: A mathematical theory of visual hallucination patterns. Biological cybernetics **34**, 137–150 (1979)
15. Franken, E., Duits, R.: Crossing-Preserving Coherence-Enhancing Diffusion on Invertible Orientation Scores. International Journal of Computer Vision **85**(3), 253–278 (2009)
16. Hubel, D.H., Wiesel, T.N.: Receptive fields of single neurons in the cat's striate cortex. The Journal of Physiology **148**(3), 574–591 (1959)
17. Loebel, A., Nelken, I., Tsodyks, M.: Processing of Sounds by Population Spikes in a Model of Primary Auditory Cortex. Frontiers in Neuroscience **1**(1), 197–209 (2007)
18. Luce, P.A., McLennan, C.T.: Spoken Word Recognition: The Challenge of Variation, chap. 24, pp. 590–609. "John Wiley & Sons, Ltd" (2008)
19. Mattys, S., Davis, M., Bradlow, A., Scott, S.: Speech recognition in adverse conditions: A review. Language, Cognition and Neuroscience **27**(7-8), 953–978 (9 2012)
20. Nelken, I., Calford, M.B.: Processing Strategies in Auditory Cortex: Comparison with Other Sensory Modalities, pp. 643–656. Springer US, Boston, MA (2011)
21. Papoulis, A.: Probability, Random Variables and Stochastic Processes, p. 138. McGraw-Hill Companies, 3rd edn. (1991)
22. Petitot, J., Tondut, Y.: Vers une neurogéométrie. fibrations corticales, structures de contact et contours subjectifs modaux. Mathématiques et Sciences humaines **145**, 5–101 (1999)
23. Petitot, J., Tondut, Y.: Vers une Neurogéométrie. Fibrations corticales, structures de contact et contours subjectifs modaux pp. 1–96 (1999)
24. Prandi, D., Gauthier, J.P.: 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)
25. Rankin, J., Sussman, E., Rinzel, J.: Neuromechanistic Model of Auditory Bistability. PLoS Computational Biology **11**(11), e1004555 (nov 2015)
26. Rauschecker, J.P.: Auditory and visual cortex of primates: a comparison of two sensory systems. The European journal of neuroscience **41**(5), 579–585 (03 2015)
27. Tian, B., Kuśmierek, P., Rauschecker, J.P.: Analogues of simple and complex cells in rhesus monkey auditory cortex. Proceedings of the National Academy of Sciences **110**(19), 7892–7897 (2013)
28. Wilson, H.R., Cowan, J.D.: Excitatory and inhibitory interactions in localized populations of model neurons. Biophysical journal **12**(1), 1–24 (jan 1972)
29. Zulfiqar, I., Moerel, M., Formisano, E.: Spectro-Temporal Processing in a Two-Stream Computational Model of Auditory Cortex. Frontiers in Computational Neuroscience **13**, 95 (jan 2020)