Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

A bio-inspired geometric model for sound reconstruction

Ugo Boscain * **Dario Prandi** ** **Ludovic Sacchelli** ***
Giuseppina Turco ****

* *CNRS, Sorbonne Université, Inria, Université de Paris, Laboratoire Jacques-Louis Lions, Paris, France. (email: ugo.boscain@upmc.fr*
** *Université Paris-Saclay, CNRS, CentraleSupélec, L2S, 91190, Gif-sur-Yvette, France (email: dario.prandi@centralesupelec.fr)*
*** *Department of Mathematics, Lehigh University, Bethlehem, PA, USA. (email: lus219@lehigh.edu)*
**** *CNRS, Laboratoire de Linguistique Formelle, Université de Paris, France. (email: gturco@linguist.univ-paris-diderot.fr)*

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. 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 (i.e., the celebrated Brockett integrator) and it is reconstructed via a Wilson-Cowan integro-differential equation. Numerical experiments are provided.

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 (Mattys et al. (2012) 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 Luce and McLennan (2008), 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; Mattys et al. (2012) among others).

Mathematical modelling of sensory input reconstruction has made a lot of progresses in the field of vision, cf. Petitot and Tondut (1999a), Citti and Sarti (2006), and Boscain et al. (2010). These models are based on the Reed-Shepp control system in the group of rototranslations in the plane, cf. Boscain et al. (2014). 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 reconstruction tasks, see, .e.g., Franken and Duits (2009); Duits and Franken (2010b); Prandi and Gauthier (2017); Boscain et al. (2018). Such work does not seem to have 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, Nelken and Calford (2011) for a comparison, Hickok and Poeppel (2007) 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 Rauschecker (2015).

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. 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 Sharma et al. (2000); Zatorre (2001). 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" Tian et al. (2013): p.1. The presence of S-cells/C-cells and the appearance of "pinwheels" in certain situations Sharma et al. (2000); Polger et al. (2016) 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: In A1 the time dimension represents one of the coordinates of an "auditory" image.

Neuro-geometric model of V1

The neuro-geometric model of V1 can be traced back to the work of Hoffman (1989), which, inspired by the experimental results of Hubel and Wiesel (1959), 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 by Petitot and Tondut (1999b), Citti and Sarti (2006), and Boscain et al. (2010). On the basis of such a model, exceptionally efficient algorithms for image inpainting have been developed (e.g.Boscain et al. (2018); Duits and Franken (2010a,b)), resulting in several medical imaging applications (e.g., Zhang et al. (2016)). 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 P^1 is the space of directions on the plane measured without orientation [1], namely is 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 assumptions, such lift of a function 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 making a totally non-isotropic diffusion adapted to the problem. Such diffusion mimic the flow of information along the horizontal and vertical connections of V1 and uses as an initial condition the surface S_f and the value of the function f. Control theorists call such a diffusion the *sub-Riemannian diffusion* in $\mathbb{R}^2 \times P^1$, cf. Montgomery (2002); Agrachev et al. (2019).

One of the main features of the image reconstruction model is the fact that it is invariant by rototranslation of the plane. In other words, it reconstructs an image, independently from its position and orientation on the plane.

From V1 to sound reconstruction

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

A first attempt to model the process of sound reconstruction into A1 would be to apply the same algorithm of image reconstruction discussed above. In a sound image, however, the time plays a special role: a rotated sound image corresponds to a completely different original sound. Also, while for image reconstruction one can assume the image to be static, for sound reconstruction time plays an essential role. Hence, the invariance by rototranslations is lost. Different symmetries have to be taken into account and a different model for both the lift and the processing in the lifted space is required.

As explained before, in V1 neural stimulation can stem 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 yields a variational problem on an underlying non-commutative structure. This structure, endowed with a metric naturally associated with the variational problem, give rise to the sub-Riemannian diffusion.

We follow this principle when trying to study sound inputs *à la* V1. Input sound signals are time dependent real valued functions submitted to a short time Fourier transform via the action of the cochlea. 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 can be understood as chirpiness and denoted by ν. This notion gives rise to a natural lift of the signal to the *contact space* in the sense of Hoffman (1989); Petitot and Tondut (1999a), i.e., $\mathbb{R}^3$ with the Heisenberg group structure. This is structure is also called the Brocket integrator, and in coordinates $(\tau, \omega, \nu) \in \mathbb{R}^3$ reads:

$$\begin{cases} \dot{\tau} &= 1, \\ \dot{\omega} &= \nu, \\ \dot{\nu} &= u(t). \end{cases}$$

As in the case of V1, this observation implies the presence of a non-commutative structure associated with this relation. The hypo-elliptic operator associated with this structure is the famed Kolmogorov operator.

A successful model to describe the evolution of neural activation, in particular in the case of V1, is given by the so-called Wilson-Cowan equations Wilson and Cowan (1972). These integro-differential equations owe their success to their ability to predict complex perceptual phenomena in V1, such as the emergence of hallucinatory pattern Ermentrout and Cowan (1979); Bressloff et al. (2001). Recently, these equations have been coupled with the neuro-geometric model of V1 to great benefit. For instance, in Bertalmío et al. (2019b,a) they allowed to replicate orientation-dependent brightness illusory phenomena, which had proved to be a hurdle for non-cortical-inspired models. See also Sarti and Citti (2015), for applications to the detection of perceptual units.

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:

$$\partial_\tau a(\tau, \omega, \nu) = -\alpha a(\tau, \omega, \nu) + \beta I(\tau, \omega, \nu) + \gamma \int_{\mathbb{R}^2} w(\omega, \nu \| \omega', \nu') \sigma(a(\tau - \delta, \omega', \nu'))\, d\omega'\, d\nu'. \quad (1)$$

Here, $\alpha, \beta, \gamma > 0$ 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'), and $\delta > 0$ is a

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

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. Chaillet et al. (2017).

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 the instantaneous frequency, obtaining $I : [0, T] \times \mathbb{R} \times \mathbb{R} \to \mathbb{C}$;

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. **Postprocessing:** Invert the preprocessing procedures to a to obtain the resulting sound signal $\hat{s} : [0, T] \to \mathbb{R}$.

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

Closing remarks and experiments

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. More in-depth discussions of these principles and their articulation can be found in Boscain et al. (2020).

In Figure 1 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. Such numerical examples can be listened at `www.github.com/dprn/WCA1`. 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.

Acknowledgements

The first three authors have been supported by the ANR project SRGI ANR-15-CE40-0018 and by the ANR project Quaco ANR-17-CE40-0007-01. This study was also supported by the IdEx Université de Paris, ANR-18-IDEX-0001, awarded to the last author.

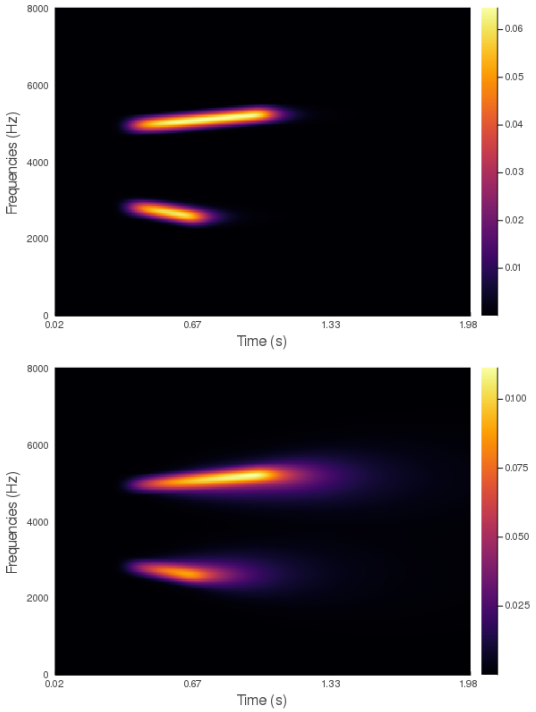

Fig. 1. Experiment on a synthetic sound with two linearly varying frequencies. *Top:* The STFT of the original sound. *Bottom:* The STFT of the processed sound. In both cases only the positive frequencies are shown: the others are recovered via the Hermitian symmetry of the Fourier transform on real signals.

REFERENCES

Agrachev, A., Barilari, D., and Boscain, U. (2019). *A Comprehensive Introduction to Sub-Riemannian Geometry.* Cambridge Studies in Advanced Mathematics. Cambridge University Press.

Bertalmío, M., Calatroni, L., Franceschi, V., Franceschiello, B., Gomez-Villa, A., and Prandi, D. (2019a). Visual illusions via neural dynamics: Wilson-Cowan-type models and the efficient representation principle. Conditionally accepted on *Journal of NeuroPhysiology*. arXiv preprint: `https://arxiv.org/abs/1907.13004`.

Bertalmío, M., Calatroni, L., Franceschi, V., Franceschiello, B., and Prandi, D. (2019b). 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.

Boscain, U., Gauthier, J., Prandi, D., and Remizov, A. (2014). Image reconstruction via non-isotropic diffusion in dubins/reed-shepp-like control systems. In *53rd IEEE Conference on Decision and Control*, 4278–4283. doi: 10.1109/CDC.2014.7040056.

Boscain, U., Prandi, D., Sacchelli, L., and Turco, G. (2020). A bio-inspired geometric model for sound reconstruction.

Boscain, U., Duplaix, J., Gauthier, J.P., and Rossi, F. (2010). Anthropomorphic image reconstruction via hypoelliptic diffusion.

Boscain, U.V., Chertovskih, R., Gauthier, J.P., Prandi, D., and Remizov, A. (2018). Highly corrupted image inpainting through hypoelliptic diffusion. *J. Math. Imaging Vision*, 60(8), 1231–1245. doi:10.1007/s10851-018-0810-4. URL https://doi.org/10.1007/s10851-018-0810-4.

Bressloff, P.C., Cowan, J.D., Golubitsky, M., Thomas, P.J., and Wiener, M.C. (2001). 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. doi:10.1098/rstb.2000.0769.

Chaillet, A., Detorakis, G., Palfi, S., and Senova, S. (2017). Robust stabilization of delayed neural fields with partial measurement and actuation. *Automatica*, 83.

Citti, G. and Sarti, A. (2006). A cortical based model of perceptual completion in the roto-translation space. *Journal of Mathematical Imaging and Vision*, 24(3), 307–326. doi:10.1007/s10851-005-3630-2. URL https://doi.org/10.1007/s10851-005-3630-2.

Duits, R. and Franken, E. (2010a). 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*. URL http://www.mate.tue.nl/mate/pdfs/10179.pdfhttp://bmia.bmt.tue.nl/people/RDuits/qampartI.pdf.

Duits, R. and Franken, E. (2010b). 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. URL http://bmia.bmt.tue.nl/people/RDuits/qampartII.pdf.

Ermentrout, G.B. and Cowan, J.D. (1979). A mathematical theory of visual hallucination patterns. *Biological cybernetics*, 34, 137–150. doi:10.1007/BF00336965.

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

Hickok, G. and Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5), 393–402.

Hoffman, W.C. (1989). The visual cortex is a contact bundle. *Appl. Math. Comput.*, 32(2-3), 137–167. doi:10.1016/0096-3003(89)90091-X. URL http://dx.doi.org/10.1016/0096-3003(89)90091-X.

Hubel, D.H. and Wiesel, T.N. (1959). Receptive fields of single neurons in the cat's striate cortex. *The Journal of Physiology*, 148(3), 574–591. doi:10.1113/jphysiol.1959.sp006308. URL https://physoc.onlinelibrary.wiley.com/doi/abs/10.1113/jphysiol.1959.sp006308.

Luce, P.A. and McLennan, C.T. (2008). *Spoken Word Recognition: The Challenge of Variation*, chapter 24, 590–609. "John Wiley & Sons, Ltd". doi:10.1002/9780470757024.ch24. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470757024.ch24.

Mattys, S., Davis, M., Bradlow, A., and Scott, S. (2012). Speech recognition in adverse conditions: A review. *Language, Cognition and Neuroscience*, 27(7-8), 953–978. doi:10.1080/01690965.2012.705006.

Montgomery, R. (2002). *A tour of subriemannian geometries, their geodesics and applications*, volume 91 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI.

Nelken, I. and Calford, M.B. (2011). *Processing Strategies in Auditory Cortex: Comparison with Other Sensory Modalities*, 643–656. Springer US, Boston, MA. doi:10.1007/978-1-4419-0074-6_30. URL https://doi.org/10.1007/978-1-4419-0074-6_30.

Petitot, J. and Tondut, Y. (1999a). Vers une neurogéométrie. fibrations corticales, structures de contact et contours subjectifs modaux. *Mathématiques et Sciences humaines*, 145, 5–101. URL http://www.numdam.org/item/MSH_1999__145__5_0.

Petitot, J. and Tondut, Y. (1999b). Vers une Neurogéométrie. Fibrations corticales, structures de contact et contours subjectifs modaux. 1–96.

Polger, T.W., Shapiro, L.A., and Press., O.U. (2016). *The multiple realization book*. Oxford University Press, Oxford.

Prandi, D. and Gauthier, J.P. (2017). *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. doi:10.1007/978-3-319-78482-3. URL http://link.springer.com/10.1007/978-3-319-78482-3http://arxiv.org/abs/1704.03069.

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

Sarti, A. and Citti, G. (2015). The constitution of visual perceptual units in the functional architecture of v1. *Journal of Computational Neuroscience*, 38(2).

Sharma, J., Angelucci, A., and Sur, M. (2000). Induction of visual orientation modules in auditory cortex. *Nature*, 404(6780), 841–847.

Tian, B., Kuśmierek, P., and Rauschecker, J.P. (2013). Analogues of simple and complex cells in rhesus monkey auditory cortex. *Proceedings of the National Academy of Sciences*, 110(19), 7892–7897. doi:10.1073/pnas.1221062110. URL https://www.pnas.org/content/110/19/7892.

Wilson, H.R. and Cowan, J.D. (1972). Excitatory and inhibitory interactions in localized populations of model neurons. *Biophysical journal*, 12(1), 1–24. doi:10.1016/S0006-3495(72)86068-5. URL http://www.sciencedirect.com/science/article/pii/S0006349572860685.

Zatorre, R.J. (2001). Do you see what i'm saying? interactions between auditory and visual cortices in cochlear implant users. *Neuron*, 31(1), 13 – 14. doi:https://doi.org/10.1016/S0896-6273(01)00347-6. URL http://www.sciencedirect.com/science/article/pii/S0896627301003476.

Zhang, J., Dashtbozorg, B., Bekkers, E., Pluim, J.P.W., Duits, R., and ter Haar Romeny, B.M. (2016). Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores. *IEEE Transactions on Medical Imaging*, 35(12), 2631–2644. doi:10.1109/TMI.2016.2587062.