Self-Organizing Chirp-sensitive Artificial Auditory Cortical Model

Luis Weruaga\textsuperscript{b} and Marián Képesi\textsuperscript{z}

\textsuperscript{b} VISKOM, Austrian Academy of Sciences, Vienna, Austria,
\textsuperscript{z} SPSCL, Technical University of Graz, Graz, Austria
weruaga@ieee.org, kepesi@tugraz.at

Abstract
This paper presents a novel signal-processing-based artificial model of the auditory mechanism. This work is inspired by the psychocortical fact that the biological cochlea is very sensitive to frequency-varying tones, or chirps. The method uses a novel combination of several (at least three) Harmonic-Chirp transform instances, that project the time–frequency energy on different views; all projections are data-mined by self-organizing unsupervised layers of Radial Basis Functions, this process being driven by the Expectation Maximization algorithm. The mechanism shows biological parallel, such as intrinsic chirp sensitivity and response to the logarithm of the stimulus energy. The proposed model is validated with several mammal sounds, such as human speech, bat echo, and several aquatic mammals.

1. Introduction
Development of artificial computational models of biological auditory cortex has captured growing attention in the last decade [1, 2]. Electrophysiological studies have revealed that auditory cortical cells are sensitive to several time-varying stimulus features such as amplitude and frequency modulation [3]. Contrary to what one would expect if frequency decomposition were the primary encoding mechanism of auditory cortex, pure tones are not the main stimuli for evoking responses from cortical neurons [3]. This fact, the “chirp-sensitivity” of the cochlea, cannot be explained by the previous artificial models. The work in [4] proposes a computing framework that encapsulates much of the known response properties of auditory cortex. Inspired at some extent by those guidelines, this paper proposes a detailed computational mechanism that responds to the need of explaining cortex chirp-sensitivity.

The processing schema is shown in Fig. 1: the first element represents the basilar membrane, which transforms the acoustic signal \( s(t) \) into mechanical strains according to the time–frequency information of the signal. This output, which in this work is considered the Wigner–Ville distribution \( W_s(t,f) \) of the input signal, excites several parallel cortical layers that collect within a time lag the TF excitation according to certain TF projection views. This redundant encoding, that corresponds well with the overlapping, parallel signal processing pathways observed in mammalian auditory cortex [3], models the sensitivity to chirp sounds. The output of each layer corresponds here to the energy of the Harmonic-Chirp transform \( \text{HChT} \) of \( x(t) \) is defined as

\[
X_{\alpha}(f) = \int_{-\infty}^{\infty} x(t) \sqrt{|\phi_{\alpha}(t)|} \, e^{-2\pi f \phi_{\alpha}^*} \, dt,
\]

where \( \phi_{\alpha}(t) \) is a second-order polynomial mapping that rules the time-dependence of the basis exponent

\[
\phi_{\alpha}(t) = \left( 1 + \frac{1}{2} \alpha^t \right).
\]

The HChT basis are non-stationary sinusoids, whose instantaneous frequency varies linearly over time at percentage rate \( \alpha \).

Lemma 2.1 The square magnitude of the HChT of signal \( x(t) \) is equal to the fan-beam projection of its TF energy distribution

\[
|X_{\alpha}(f)|^2 = \int_{-\infty}^{\infty} W_k(l, f + \alpha ft) \, dt,
\]

Figure 1: Chirp-sensitive artificial auditory mechanism.
where $W_{s}(t, f)$ is the WVD of $\tilde{e}(t) = x(t)\sqrt{|\Phi_{o}(f)|}$.

The projecting scenario of Lemma (2.1), proven in [7], is explained graphically in Fig. 2: the image corresponds to the WVD of a signal composed of harmonically-related chirplets of chirp rate $\gamma$ (the dark stripes represent the non-stationary TF energy). When the structure is “illuminated” from point $(t, f) = (-1/\alpha, 0)$, the resulting projection gives rise to the HChT power spectrum for the analysis chirp rate $\alpha$.

2.2. Chirp-Periodicity

Let $s(t)$ be a “chirp-periodic” signal, that is, a signal whose fundamental frequency changes linearly over time

$$s(t) = \sum_{k \in \mathbb{Z}} a_{k}(t) e^{j2\pi k f_{o} \phi_{o}(t)},$$

where $f_{o}$ is the instantaneous fundamental frequency at $t = 0$, $\gamma$ is the pitch variation rate and $a_{k}(t)$ is the envelope of the $k$-th component. Assuming that the cortical temporal resolution is modelled with a Gaussian window $w(t)$ of time spread $\chi$, the segment $x(t)$ can be approximated by Gaussian chirplets

$$x(t) \approx \sum_{k \in \mathbb{Z}} a_{k} G(t, \tau_{k}, \sigma_{k}^{2}) e^{j2\pi k f_{o} \phi_{o}(t)},$$

where $G(x, \theta)$ is a Gaussian function ($\theta$ contains the mean and variance), and $a_{k}$, $\tau_{k}$ and $\sigma_{k}$ are the amplitude, time shift and time spread of the $k$-th harmonic respectively.

Lemma 2.2 Let $g(t)$ be the Gaussian chirplet of unit energy. Its square absolute value of the Harmonic Chirp transform results in Gaussian shape, $|G_{o}(f)|^{2} \simeq G(f, \mu, \sigma^{2})$, where

$$\mu = \nu (1+\gamma \Delta)/(1+\alpha \Delta),$$

$$\sigma^{2} = \frac{g^{2}}{8\pi^{2} (1+\alpha \Delta)^{2}} + \frac{\nu^{2} \Delta^{2}}{2(1+\alpha \Delta)^{2}} (\alpha - \gamma)^{2},$$

According to the previous lemma (proven in [7]), the HChT power spectrum of the Gaussian chirplets (5) is modelled by the following Gaussian mixture

$$|X(f, \alpha)|^{2} = \sum_{k \in \mathbb{Z}} \lambda_{k} G(f, \{\mu_{k}, \sigma_{k}^{2}\}) + C_{o}(f),$$

where

$$\mu_{k} = kf_{o} (1+\gamma_{k})/(1+\alpha \tau_{k}),$$

$$\sigma_{k}^{2} = \frac{1}{8\pi^{2} (1+\alpha \tau_{k})^{2}} + \frac{1}{2} \gamma^{2} kf_{o}^{2} (\alpha - \gamma)^{2},$$

$$\lambda_{k} \propto |a_{k}|^{2},$$

and the additive term $C_{o}(f)$ in (8) corresponds to the fan-beam projection of the Wigner cross-terms among the chirplets. The cross-term projection is proven to have zero mean [7].

Equations (9) contain the representation of the chirplet mixture by the $\alpha$-HChT cortical bank. This representation is different for each bank (with a different encoding chirp rate $\alpha$). However, the projected energy $\lambda_{k}$ does not depend on $\alpha$, which means that all cortical banks are excited with the same energy.

3. Self-organized TF Energy Decomposition

The HChT of a chirplet achieves its finest representation when $\alpha = \gamma$ (see (9b)). However, the fact that the chirplet is observed by the HChT layer with its actual energy regardless of the $\alpha$ value, makes the chirplet identification especially difficult. The intuitive approach is to use exhaustive strategies [5], that is, compute the HChT for many chirp rates and choose the one that obtains the finest representation. If case of the mixture of signals with different chirp rates $\gamma$, that estimation is not trivial, since the chirp rate that delivers the best representation for one signal may be likely to perform poorly on the others.

The self-organizing layer proposed in this paper is suited to mixture of signals with different chirp rates (e.g. voices from several speakers). This layer detects the chirplet projections from a reduced set of HChT bank (at least three) with an unsupervised mechanism in such a way that the energy of the filter bank is decomposed into physically meaningful components.

3.1. Expectation Maximization-driven Gaussian Fitting

Let $\rho_{\alpha}(f)$ be the normalized HChT power spectrum of $x(t)$ for chirp rate $\alpha$, $\rho_{\alpha}(f) = |X_{\alpha}(f)|^{2} / \int_{-\infty}^{\infty} |X_{\alpha}(f)|^{2} df$. Since $\rho_{\alpha}(f)$ is non-negative and its area is equal to one, it can be treated as a probability density function (p.d.f.). Based on the analytical model (8), the observed data p.d.f. is assumed to follow the model

$$p(f|\Theta) = \sum_{m=1}^{M} \lambda_{m} G(f, \{\mu_{m}, \sigma_{m}\}),$$

where $M$ is the total number of Gaussian Radial Basis Functions (RBF) and $\Theta = \{\lambda_{m}, \mu_{m}, \sigma_{m}\}_{m=1,...,M}$ such that $\sum_{m=1}^{M} \lambda_{m} = 1$. The objective of the unsupervised Gaussian fitting is presented as the problem to determine the parameters $\Theta$ that maximize the log-likelihood given the data $p(f)$, i.e.,

$$\Theta^{*} = \arg\max_{\Theta} \int_{-\infty}^{\infty} \rho(f) \log p(f|\Theta) df.$$

3 Although not shown, the spectrum $\rho(f)$ refers to the chirp rate $\alpha$. 

Figure 2: The Harmonic-Chirp power spectrum corresponds to the fan-beam projection of the Wigner-Ville distribution.
3.2. Coupling Several Projection Views

According to the previous discussion, one TF projection view does not suffice in the chirplet identification. Let us assume now that $L$ HCTI power spectra computed with different chirp rates, $\alpha_i$, $i = 1, \ldots, L$, are available. Then the normalized power spectrum of the $i$-th view fulfills

$$
\rho_i(f) \equiv \sum_{m=1}^{M} \lambda_{m,i} \varphi(f, \{\mu_{m,i}, \sigma_{m,i}\}) .
$$

It may seem clear at this point that if the parameters of the $m$-th chirplet, $\Omega_m = \{\theta_{m,1}, \ldots, \theta_{m,L}\}$, are obtained with the EM-based fitting operating independently on each $\rho_i(f)$, $\Omega_m$ is unlikely to be physically meaningful. In order to make all fitting process evolve to a common goal we propose the estimation-modification procedure shown in Fig. 3. The main idea of this procedure is to embed the parallel EM-driven Gaussian fitting mechanisms into an estimation-modification loop in which the RBF parameters are corrected so that they are chirplet-based physically meaningful. The estimation of the chirplet parameters from the RBF parameters is the key part of this procedure.

Due to space limitations, we provide the main aspects on how to proceed with that estimation:

- The energy from all views are equal, i.e., $\lambda_{m,i} = \lambda_{m,j}$.
- The time location $\tau$ and the time spread $\rho$ have to be in agreement with the time support restrictions applied by the window $w(t)$ in terms of its spread $\chi$,
  \begin{equation}
  \frac{1}{\tau} + \rho \leq \chi .
  \end{equation}
- The central temporal position observed from the $i$-th and $j$-th views can be obtained, based on (9a), as
  \begin{equation}
  \tau_{ij} = (\mu_i - \mu_j)/(\mu_j \alpha_j - \mu_i \alpha_i) = \tau_{ji} .
  \end{equation}
- From $\tau_{ij}$ and equation (9a) the central frequency observed from the $i$-th and $j$-th views is
  \begin{equation}
  \nu_{ij} = \frac{\mu_i \mu_j (\alpha_j - \alpha_i)}{(1 + \gamma \tau_{ij}) (\mu_j \alpha_j - \mu_i \alpha_i)} = \nu_{ji} .
  \end{equation}
- By substituting $\mu$ (9a) into $\sigma^2$ (9b), the variance becomes
  \begin{equation}
  \sigma^2 = \frac{1}{2} \frac{1}{(1 + \alpha \tau)^2} \left( 1 \frac{1}{4\pi^2 \rho^2} + \rho^2 \mu^2 (\alpha - \gamma)^2 (1 + \gamma \tau)^2 \right) .
  \end{equation}

According to (17), the pair $(\alpha_i, \nu_i)$, where

$$
y_i \triangleq 2 \frac{(\alpha_i \mu_i)^2}{(1 + \alpha \tau)^2} - \frac{1}{4\pi^2 \rho^2 \mu_i^2} ,
$$

is a sample of a quadratic equation. In order to have a well-conditioned estimation of this parabola, three points, or equivalently three different projection views, are required. It is easy to deduce $\gamma$ and $\rho$ from the parabola parameters (for instance $\gamma$ is the location of the parabola minimum).

The previous items show how to obtain the partial estimation of the chirplet parameters. The global estimation is obtained by merging the partial ones according to a certain loss function (i.e., quadratic, absolute value, Huber’s robust loss, etc.). Due to lack of space, the details of this last aspect are omitted in the paper.

3.3. The Frequency–Pitch-Rate Plane

The proposed chirp-sensitive artificial auditory model translates the signal into a set of parameters in a compact and parallel fashion. The chirp rate $\gamma$, the central frequency $\nu$ and the energy $\lambda$ are the main chirplet descriptors, whilst the time spread $\rho$ and time location $\tau$ are secondary degrees of freedom, confined by the restrictions of the analysis window (14) (they result frequently in $\tau \approx 0$ and $\rho \approx \chi$). This fact allows to define the frequency–chirp-rate plane $(f, \alpha P)$ as the natural representation of the incoming sounds: the $f_{AP}$ of a single chirplet is a cluster centered on $(\gamma, \nu)$ with intensity $\lambda$. In case of the harmonically-related chirplet mixture (5), the clusters are located in the same line $\alpha = \gamma$ whilst the frequency of each cluster is $f \approx k f_s$. The description of the cluster shape is out the scope of the paper.

4The equations refer to the $m$-th chirplet (that subindex is omitted),
4. Results

In this section we present results over real signals that correspond to the sound of several mammals. Four different signals are presented in Fig. 4: from left to right, a) a bat echolocation signal\(^\text{5}\), b) human speech, c) a whale song, and d) a song of a seal. For each case it is shown from top to bottom: time samples (time axis do not represent actual time scale), spectrogram (x-axis time, y-axis positive frequency), Wigner-Ville distribution, and the synthetic chirp-based time–frequency energy space. This last one is generated with the chirplet parameters obtained from the cochlea model proposed in this paper. All techniques are short-time oriented, they all use the same window length, which for each signal was selected based on obtaining the fairest spectrogram quality.

The artificial cochlea mechanism in each case was composed of several chirp objects and “noise” objects. These are also self-organized RBFs, which are constrained only to their central frequency position, and not to the their variances. Therefore the noise-RBFs do not get attracted by the chirp time–frequency energy but to the noise present in the record. Furthermore, its use is decisive in the whole mechanism, since they account for the stochastic part of the signal, so that the chirplet-RBF accuracy do not get affected by that disturbance.

The analysis of each case is done separately. In case of the bat signal, the spectrogram does not achieve a detail time–frequency representation due to the high chirped nature of the bat signal, whilst the artificial cochlea responds precisely to the chirp harmonic components of the pulse. In case of the human speech segment, which corresponds to natural conversation, again several parts of the segment are represented poorly by the spectrogram, whilst the artificial cochlea shows an outstanding sensitivity to high pitch variations. The whale song is less impressive, because it is literally a “song” and the pitch variation is not fast. Nevertheless the artificial cochlea shows improvement in time–frequency resolution. Finally in case of the seal call, the fast dropping and rising parts are extremely well followed by the RBF-self organizing mechanism.

5. Conclusions

This paper has proposed an artificial signal processing-based cochlea model, sensitive to the chirplets present in the signal. This fact, which is an intrinsic characteristic of the biological cochlea of mammals, represents a novel proposal in auditory scene analysis. The cochlea mechanism is based on several time–frequency (TF) processing blocks, that project the TF energy from different views. The second stage is a self-organizing unsupervised mechanism that decomposes the views in meaningful components: either chirplets or stochastic parts. When this mechanism is used with a short-time processing basis, it turns out very efficient to represent the time–frequency components of mammal sounds, such as human speech.

6. References


---

\(^{5}\)The authors wish to thank Curtis Condon, Ken White, and Al Feng of the Beckman Institute of the University of Illinois for the bat data and for permission to use it in this work.

Figure 4: Results of the artificial cochlea of mammal sounds: a) bat, b) human, c) whale and d) seal. From top to bottom: signal samples, spectrogram, Wigner-Ville distribution, and chirp-based artificial cochlea sensitivity. The time axis do not represent actual time scale.