AUDITORY PROCESSING IN A POST-COCHLEAR NEURAL NETWORK: VOWEL SPECTRUM PROCESSING BASED ON SPIKE SYNCHRONY

F. Berthommier*, J.L. Schwartz*, P. Escudier*

*IMAG - TIM3 - TIMB
Faculté de Médecine de Grenoble
38700 La Tronche
France

+ICP - UA CNRS
INPG/ENSEIRB ; Université Stendhal
46 Av. Félix-Viallet, 38031 Grenoble Cedex
France

ABSTRACT

The present work is concerned with the processing of spectral information by stochastic coding, at the output of the cochlea. We study the collective behaviour of a large number of cells based on a measure of the correlation of sequences of spikes between neighbour units. Our first results show that we can obtain significant discrimination of spectral components of vowels without any kind of cable lateral inhibition or second filter device. This collective behaviour is made possible by the synchronization of inputs of adjacent units. This could provide an interesting and rather new way of making profit of both geographical coding (tonotopy) and temporal coding (synchronization) particularly useful for formant detection.

I. INTRODUCTION: THE "ESAP" PROJECT

Since the first work by Rose and colleagues [1,14], it is well-known that responses in single auditory nerve fibers are phase-locked to the frequency components of an acoustic stimulation. Elaboration of a neural network based on this synchronization property could help us understand some of the regularities observed in the nervous system, such as tonotopy and auditory maps in the brain. The role of spike synchrony is central in the present work.

Our group has been working on auditory modelling since more than ten years. Our strategy towards a model of auditory perception goes from the conception of a model of the peripheral auditory system (PAS) up to the decoding in a neural net, including studies about coding and representations, with the following steps:

1. Elaboration of a model of the cochlea and hair cells and study of the coding of stationary speech signal in this model [12].
2. Methodological and experimental studies on auditory representations of vowel spectra [21].
4. Study of the coding of speech acoustic events in the auditory nerve, by various processes such as cochlear filtering, spectral integration, temporal derivation by neural adaptation, spectral derivation by lateral inhibition or functionally equivalent mechanisms [24].

The ESAP project (Extended PAS) intends for an integration of the various concepts introduced in these models (lateral inhibition, neural adaptation, spectral integration, synchrony) to try to have a global view of the auditory processing. Hence, it does not really enter the classical cognitive or neo-connectionist projects, but is rather focussed on the elaboration of a physiologically plausible model.

II. TWO PARALLEL CHANNELS

Steps (4) - detection of transient events - and (5) - analysis of stationary signals, that we shall consider in more detail - should provide the basis for understanding auditory processing. Such a classification does exist in the visual system, with the tonic channel X characterized by a good spatial resolution, a poor temporal resolution and a good ability for analysis of a stimulus spatial structure, and the phasic channel Y with inverse characteristics and a specialization in analysis of temporal structure [6].

Such a clear-cut organization has not been found in the auditory system, but various patterns of neural response can be observed in the cochlea nucleus for example, which point towards the existence of several processings in parallel (Fig.1). There could thus exist a phasic behaviour - "on" and "off" cells, with very rapid adaptation, able to detect the temporal position of specific events ("when ?") - and a tonic behaviour specially convenient for characterization of place of articulation and more generally spectral analysis ("what ?"). This idea was at the basis of the schema proposed by Chistovich for the organization of auditory perception [8], and used in various models or explanations of experimental results [5, 9].

We describe here a model of vowel spectrum processing by a tonic channel. This model would be higher than the cochlea and the auditory nerve in the auditory chain, its input consisting in temporally organized sequences of spikes.

III. A MODEL FOR TONIC CODING

We do not consider here neural adaptation, and we assume an ideal tonal coding, precisely linked to interspike interval histograms. Each 0.1-ms temporal bin is characterized by the probability $p_i$ that a spike would occur in fiber $i$ (we shall call this neural event $e_i(t)=1$) and the complementary probability $(1-p_i)$ that no spike occurs in fiber $e_i(t)=0$. We assume in the present work that in each fiber the input is a pure tone with frequency $f_i$. We further assume that probability $p_i$ equals 0 if fiber $i$ is within a refractory period, and if it is not the case we assume that $p_i$ is driven by the positive part of the input sinusoid:

$$P(e_i(t)=1) = \begin{cases} 0 & \text{within the refractory period} \\ \min \{p_{max} \times \sin(\alpha_1 + \varphi) & \text{outside} \} \end{cases}$$

$$p_{max} \times \sin(\alpha_1 + \varphi)$$

where $p_{max}(x) = \begin{cases} \sin(x) & 0 \leq x < \pi \\ 0 & \pi \leq x < 2\pi \end{cases}$ (1) and (2)

We must notice that : (1) this procedure is similar to other stochastic models [20] (2) the coding process is semi-deterministic, since spike position fluctuates around sinusoid maxima (3) interspike intervals are only a consequence of
other parameters of the model (4) amplitude of fluctuations is lower for high stimulation frequencies, which makes this model somewhat similar to the deterministic solution proposed by Lestienne and Strehler [16] in the visual system. Each channel is mainly characterized by frequency and amplitude, which provides a dissociation of the classical "mean activity" in two variables. If we associate these parameters with the channel position, we may define neural coding by a set of <Frequency-Amplitude-Position> triplets.

IV. ANALYSIS OF NEURAL INFORMATION

A number of models for processing of neural information have been proposed. We may display the whole content of neural information in a 3D-space: (1) fiber position P is related to its characteristic frequency CF (2) fiber activity after FFT analysis is described by a spectrum A(F,P). Hence, the total amount of "energy" (spikes) present in the auditory nerve can be associated to a double decomposition, on both frequency axes F and CF (or P) as proposed by Delgutte [10].

Further neural processing can be considered as a transformation of function A(F,P) into a new function A' (Fig.2): (1) geometrical models [26] - transform T1 - associate to a given fiber the whole energy it contains, whatever its FFT frequency F, and hence the selectivity is completely defined at the cochlea level; (2) temporal models [7,17] - transform T2 - associate to a given FFT frequency F the energy linked to it in the whole set of fibers whatever their position, and hence they lose tonotopic information; (3) combined models [22,23,25] - transform T3 - integrate energy in a 2D F-P domain.

None of these models is really satisfying since (1) they cannot do an optimal use of both temporal and geometrical coding at the same time (2) and however, while geometrical coding is certainly at the basis of auditory processing, temporal coding provides the best representation of vowel formants [10,19,25]. We propose here a model that closely links temporal and geometrical aspects. We shall call it selective model since it is really based on a double filtering on both axes, and not on a simple integration along one or the other axis or even a 2D domain. In this transform (T4) distribution A' is computed at each position P, from the frequency content of fiber P and all its neighbours within the set [P-n,P+n]. Since synchronization occurs up to 4 kHz [1], this operation can be made only at low and middle frequencies. Notice that since cochlea filtering is mainly low-pass, low CF fibers contain few high F components.

We show that application of transform T4 on a simplified A(F,P) distribution in which each fiber contains only one frequency component has a highly selective effect on certain parts of the distribution, based on an implicit homogeneity criterion. This is transform is not ad-hoc, it is on the contrary built from (1) a mathematical property concerning crossed correlations between channels, still true when their content is made of sequences of spikes, (2) architecture characteristics of true neural nets.

V. CORRELATIONS OF SENSORY INPUTS

We must define a physiologically plausible computation that could allow the nervous system compare the content of two channels, and measure their similarity or homogeneity. We propose two equivalent measures, one formal and the other supposed to be realistic.

(1) The mathematical computation is a temporal mean of the product of firing probabilities supposed to be directly driven by the input signal without considering the refractory period. In this simplified model (the so-called analogical model AM) we do not consider the stochastic characteristic of neural behaviour. The result is a function depending on time :

\[ C_{ij}(t) \sim \left( \sum_{\omega=-1}^{t} p_{\max} \sin(\omega \theta) p_{\max} \sin(\omega \theta + \Delta \phi_{ij}) \right) \]

Three main facts must be considered: (1) values are higher for small times, and (2) for close frequencies (3) channels initially correlated are less and less correlated with time (Fig.3). We obtain correlations significantly different from the base value (0.1 if \( p_{\max}=1 \)) when frequency differences are compatible with tonotopic representations characteristic of all steps in the auditory way, and also with psychophysical frequency resolution (about 1 Hz around 500 Hz). The variations are of the type \( \sin(x)/x \) on a large interval of frequency differences.

![Figure 1: Two channels in the auditory system.](image1)

None of these models is really satisfying since (1) they cannot do an optimal use of both temporal and geometrical coding at the same time (2) and however, while geometrical coding is certainly at the basis of auditory processing, temporal coding provides the best representation of vowel formants [10,19,25]. We propose here a model that closely links temporal and geometrical aspects. We shall call it selective model since it is really based on a double filtering on both axes, and not on a simple integration along one or the other axis or even a 2D domain. In this transform (T4) distribution A' is computed at each position P, from the frequency content of fiber P and all its neighbours within the set [P-n,P+n]. Since synchronization occurs up to 4 kHz [1], this operation can be made only at low and middle frequencies. Notice that since cochlea filtering is mainly low-pass, low CF fibers contain few high F components.

We show that application of transform T4 on a simplified A(F,P) distribution in which each fiber contains only one frequency component has a highly selective effect on certain parts of the distribution, based on an implicit homogeneity criterion. This is transform is not ad-hoc, it is on the contrary built from (1) a mathematical property concerning crossed correlations between channels, still true when their content is made of sequences of spikes, (2) architecture characteristics of true neural nets.
Fimre
to
of 1.26 Hz centred on \( \omega = 630 \) Hz. This function is roughly estimated with a cardinal sine function \( f(x) = \sin(x)/x \) on a wide neighbourhood.

Frequency of auditory-nerve fibers, at low \((Ll)\) and high \((Hf)\) sound level. The peaks are not present at saturation (from \([19,25]\)).

Figure 3: Correlation of inputs versus time \((t = 50 \text{ ms} \text{ to } t = 100 \text{ ms})\) and frequential differences. We show 32 steps of 1.26 Hz centred on \( \omega = 630 \) Hz. This function is roughly estimated with a cardinal sine function \( f(x) = \sin(x)/x \) on a wide neighbourhood.

(2) An equivalent measure can be obtained in the nervous system by the evaluation of the number of coincidences between spikes in two channels, associated to the evolution of the post-synaptic membrane potential. Synapses of the two channels must combine to increase the dendritic membrane potential when two spikes happen at the same time, the "coincidence" being defined in a loose way: when one channel fires, the other must fire within a given delay. In this "physiological model" PM, inputs \( e_i(t) \) are stochastic and defined as in III, and we define the physiological correlation \( P_{ij}(t) \) between channels \( i \) and \( j \) by:

\[
P_{ij}(t) = \sum_{\omega=1}^{1} \text{Coincidence}(\omega) - \sum_{\omega=1}^{1} \text{Coincidence}(\omega) \tag{4}
\]

Coincidence events are also boolean values computed from the current window. Coefficients \( P1 \) and \( P2 \) are increase (resp. decrease) rates of the membrane potential, in arbitrary units. Functions \( P \) and \( C \) are equivalents if time \( t \) is not too small \((t > 10 \text{ ms})\).

VI. DECODING

Sachs and Young \([19,25]\) showed the relative invariance of the frequency content (measured by a synchronization index) of sequences of spikes in the cochlear nerve in response to synthetic vowels at various intensities, while on the contrary the geographical pattern of responses (mean activity of fibers, transform \( T1 \)) could not allow formant discrimination at saturating rates (Fig.4). Notice that lateral inhibition, acting on mean discharge frequencies, could not either allow a discrimination between geographical areas associated to each formant.

Transform \( T4 \) allows selective elimination of non-informative energy (the "noise") and preservation of homogeneous formant components (the "signal"). Let us begin with the essence of Delgutte’s description (Fig.5) and assume that a vowel formant recruits a number of neighbour channels, while other fibers are excited at their own characteristic frequency: \( F = CF \) (Fig.6a). A linear AM model with 64 inputs and 32 cells extracts the signal within a noise of equal intensity \((\text{pmax}=1)\). Correlation is independently estimated on each pair \( \langle \text{center cell} - \text{neighbour cell} \rangle \), each cell comparing its own input with inputs of its 33 neighbours \((v=16)\). Then the final output of the model - the so-called "modulation" - is computed:

\[
M_i(t) = \sum_{j=-v}^{v} C_{ij}(t) \tag{5}
\]

After about 50 ms - a delay compatible with psychophysical integration durations - activity of cells with synchronized inputs becomes clearly greater than activity of cells with "noise" inputs (Fig.7). A preliminary test allowed us verify that phase differences \( \theta_{ij} \) between neighbour channels do not, when compatible with physiological data, modify this result.

In the same manner, we used a physiological model PM with 128 cells and 160 inputs in which we defined the output modulation as a sum of the \( P_{ij} \) components computed with membrane coefficients \( P1 \) and \( P2 \) adjusted so as to balance depolarization and hyperpolarization. The input consisted in two pure tones separated by a constant frequency gradient (Fig.6b). Extractions of the synchronized areas occurs quickly with a response maximum at the center of each area and a minimum on each side (Fig.8). This process could thus be used in formant segmentation at saturation.

Computation of temporal coincidence between spikes could be realized by synaptic structures such as geniculate body glomerula, able to link two adjacent inputs, with a spatial summation of slow dendritic potentials at the level of neural cell bodies.
Within the framework proposed by D. Marr [18], we tried to work at three levels, (1) the logical level, with the proposal of a functional relationship between tonotopy and synchronization, thanks to a simple correlation measure computed in parallel on all channels, (2) the algorithmic level, with the definition of a transform that allows the extraction of formant components in vowels and the development of potentials and the temporal organization of sequences of spikes, with a physiologically plausible architecture.

Hence, lateral inhibition (see e.g.[15]), is probably not the only mechanism involved in processing of vowel formants, since it cannot explain what happens around saturation. In the tonic channel, the processing could be geographical, with retention of response areas well separated thanks to homogeneity criteria based on synchronization characteristics, and in which an intensity measure would be kept.

This work is only preliminary. A further test will consist in studying the behaviour of our net on the output of a realistic cochlea model. We believe that such a complete model should help us understand more about the whole spectral processing that could occur at the low levels of the auditory system.

**Figure 6**: Description of inputs. (a) In AM, the 9 middle cells receive f=670 Hz, others have own inputs. (b) The PM receives 5 areas of 32 inputs, and 2 are synchronized.

**Figure 7**: Modulation of AM versus time (from t=50 ms to t=100 ms) and place of inputs, described in fig. 6a.

**Figure 8**: Modulation of PM versus time (from t=10 ms to t=60 ms) and place of inputs, described in fig. 6b.

**References**