19th INTERNATIONAL CONGRESS ON ACOUSTICS
MADRID, 2-7 SEPTEMBER 2007

A CORTICAL VIEW ON AUDITORY SCENE ANALYSIS: A PHYSIOLOGICAL & COMPUTATIONAL APPROACH

Elhilali, Mounya[1]; Shamma, Shihab[1]
[1]Institute for Systems Research; University of Maryland; A.V.Williams Bldg, College Park, MD 20742; USA; {mounya,sas}@umd.edu

ABSTRACT
Though seemingly effortless, our auditory system engages in complex processes and transformations which enable us to segregate speech and other target sounds from cluttered and noisy environments. In this work, we present a biologically-inspired model for exploring the role of cortical receptive field selectivity and adaptation in the streaming of a target tone embedded in a random complex background. This computational scheme tests the hypothesis that segregation of a target sound from the background maskers is achieved by integrating temporally-coherent information in a multidimensional spectrotemporal cortical representation. The model uses a clustering algorithm reconciling the integrated output of a cortical mapping with the incoming input projected into a multidimensional space explicitly depicting features related to tonotopy, spectral shape and harmonicity. We demonstrate the model's ability to emulate percepts reported by human subjects performing this task. The coherence principle tested in this model is driven by physiological results testing the role of spatial separation in mediating streaming in cortical neural responses.

INTRODUCTION
Despite the importance of auditory scene analysis is our daily lives, we still largely ignore the neural mechanisms underlying this remarkable ability, and particularly the role of different auditory nuclei in the process of sound organization. Various studies have suggested the involvement of the auditory cortex in representing sounds in terms of auditory objects. Firstly, the cortical circuitry is known to exhibit intricate mappings of acoustic waveforms into a multidimensional feature space, allowing acoustic elements to cluster into distinct ensemble, hence potentially forming distinct streams. Secondly, lesion studies reveal auditory cortical involvement in temporal pattern organization, and hence associate cortical circuitry with perceptual ordering of acoustic events and sequential stream segregation. Thirdly, the temporal dynamics of stream formation and auditory grouping are known to correspond nicely with the time scales observed in cortical processing.

Cortical circuitry is postulated to encode perceptually segregated streams by activating spatially non-overlapping populations of neurons. This prevailing view of streaming emphasizing spatial clustering, states that sounds that can segregate are those that excite neural populations that are sufficiently spatially distinct in the primary auditory cortex. In the simplest case of two tone streams, such segregation would occur if the frequency separation between the two tones was sufficiently large. Models supporting this principle have shown relative success in reproducing stream segregation examples especially for simple tone-stimuli [1,7,8]. The idea has also been explored in physiological studies that attempted to look at the neural basis of stream segregation; particularly focusing on the principle that the probability of perceiving one stream vs. two streams is correlated with the amount of overlap between two populations of neurons tuned at the frequencies of the two streams [4,5,9].

The present study questions whether the principle of spatial clustering is in fact *sufficient* as a neural correlate of streaming. As discussed in section 1, our electrophysiological findings indicate that the *spatial* neural response pattern is not sufficient to reflect the perceptual organization of sounds, and that the *temporal coherence* between sound elements is critical for

sound segregation. Based on these neurophysiological results, we formulate in section 2 a computational model that emphasizes two critical stages of stream segregation: (1) mapping sounds into a multi-dimensional feature space; (2) organizing sound features into temporally coherent streams. The model highlights the principle that sound elements belonging to the same stream tend to evolve together in time. It postulates that grouping features according to their levels of temporal coherence is a viable organizing principle underlying cortical involvement in sound segregation.

1. PHYSIOLOGICAL APPROACH

The experiment described here contrasted responses of primary auditory cortex (A1) neurons under two conditions: a synchronous and an alternating sequence of two tones. At far enough separations between the frequencies of the two tones (e.g., 1 octave), the percept of the *alternating* tones is that of two "streams", while the *synchronous* stream a remains unitary sequence. In this experiment, we tested whether the spatial pattern of neural responses in A1 reflects the different percepts in these two conditions. Specifically, we examined whether the interactions between the two tones become stronger or more far ranging in the synchronous compared to the alternating sequence. We recorded from A1 of awake non-behaving ferrets (N=2) and compared neural responses to alternating and synchronous two-tone AB sequences. The stimuli consisted of: (a) A sequence of a repeating tone (A) placed at the best frequency (BF) of a cell; (b) a sequence of a repeating B tone, placed at different distances ΔF A-B away from the BF; (c) a sequence of alternating A-B tones presented at all frequency separations ΔF A-B; (d) a sequence of synchronously presented A and B tones. Tones in all conditions were 75 ms long, with 25 ms inter-stimulus-interval.

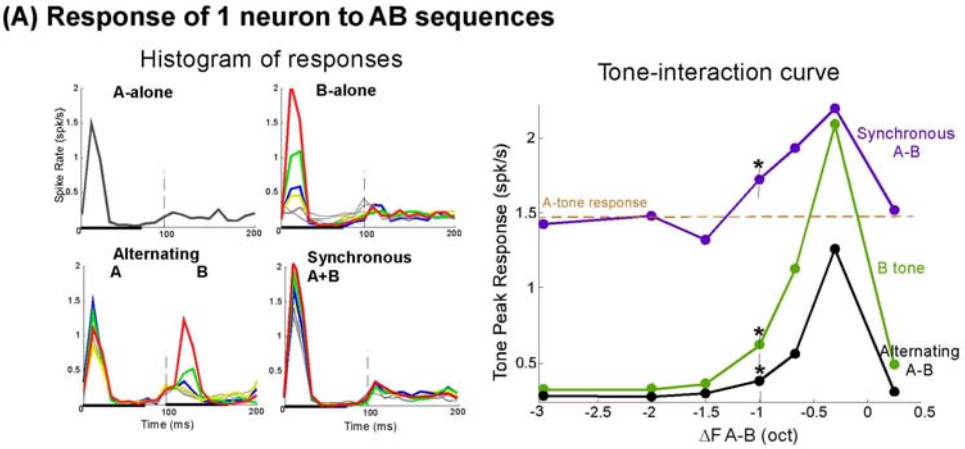

Figure 1. (A) Responses of a primary auditory cortex neuron to sequences of two-tones. The left panel depict per-stimulus histogram to responses to A tone alone (presented at Best Frequency), B tone alone (presented at various distances away from BF), alternating A and B tone; and synchronously presented A and B tones. Right panel shows tone-interaction curve by measuring the peak response of the tones for different separations of the B-tone away from the BF. **(B)** Population scatter plot contrasting the bandwidth of two-tone interaction in the synchronous vs. alternating cases in 22 neurons. The diagonal indicates that the range of interaction is the same under both conditions.

Figure 1A shows responses of 1 neuron in primary auditory cortex. The plots in the left 4 panels show the per-stimulus histograms of neural responses to all 4 stimulus conditions. The varying colored traces indicate response to the different ΔF A-B spacing. In order to investigate the range of interaction between the A and B tones, we extract the peak firing rate for each stimulus condition at the different frequency separations. The right panel of Figure 1A depicts the tone-interaction curve for the synchronous condition (purple), the alternating condition (black), as well as the response to the B tone alone (green) at different B-tone positions away from the BF of the cell (ΔF A-B = 0 oct). For this example neuron, as the B tone is brought closer to the BF, it begins to induce a visible response from the cell at about 1 octave away from the center of the receptive field. This same bandwidth of interaction is observed whether the B tone is alone, with A and B presented simultaneously or in an alternating fashion. Therefore, we can deduce that this neuron yields the same range or bandwidth of interaction for both synchronous and alternating conditions, indicating that its spatial firing pattern alone is not sufficient to indicate the perceptual difference between these two stimulus configurations.

We contrasted this bandwidth of interaction in a population of 22 A1 neurons. As shown in Figure 1B, 7/22cells showed larger range of tone interactions responding to synchronous AB than alternating AB; 5/22 cells were in the opposite direction; and 10/22 cells showed equal range of tone interactions. Hence, these results indicate that while spatial separation in A1 may be a prerequisite for revealing the segregation level of two sounds, it is unlikely to be sufficient by itself.

2. COMPUTATIONAL APPROACH

Drive by these physiological findings, we formalize a computational model based on the temporal coherence principle. This scheme highlights the principle that sound elements belonging to the same stream tend to *evolve together* in time. The model involves two key components: *first,* the mapping of sounds into an appropriate multi-dimensional feature space; and *second* the organization of these sound features into temporally correlated streams.

Model formulation

Stage 1

Current understanding of auditory cortical processing inspires our model for the multi-dimensional representation of sound. The model takes in as input an auditory spectrogram, and effectively performs a wavelet decomposition using a bank of linear spectro-temporal receptive fields (STRFs). The analysis proceeds in two steps (as detailed in [3]): **(i)** a *spectral* step that maps each incoming spectral slice into a 2D frequency-scale representation. It is implemented by convolving the time-frequency spectrogram *y(t,x)* with a complex-valued spectral receptive field SRF, parameterized by spectral tuning Ω_c and characteristic phase ϕ_c; **(ii)** A *temporal* step in which the time-sequence from each frequency-scale combination (channel) is convolved with a temporal receptive field TRF to produce the final 4D cortical mapping *r*. Each temporal filter is characterized by its modulation rate ω_c and phase θ_c. This cortical mapping is depicted in Fig.1A, and can be captured by:

$$\begin{aligned} s(t,x;\Omega_c,\phi_c) &= y(t,x) *_x \mathcal{SRF}(x;\Omega_c,\phi_c) \\ r(t,x;\omega_c,\Omega_c,\theta_c,\phi_c) &= s(t,x;\Omega_c,\phi_c) *_t \mathcal{TRF}(t;\omega_c,\theta_c) \end{aligned} \qquad (1)$$

We choose the model's parameters to be consistent with cortical response properties, spanning the range Γ=[0.5-4] peaks/octave spectrally and Ψ=[1-30] Hz temporally. Clearly, other feature dimensions (such as spatial location and pitch) can supplement this multidimensional representation as needed.

Stage2

The essential function of this stage is two-fold: (**i**) estimate a pair-wise correlation matrix (*C*) among all scale-frequency channels, and then (**ii**) determine from it the optimal factorization of

the spectrogram into two streams (foreground and background) such that responses within each stream are maximally coherent.

The correlation is derived from an instantaneous coincidence match between all pairs of frequency-scale channels integrated over time. Given that TRF filters provide an analysis over multiple time windows, this step is equivalent to an instantaneous pair-wise correlation across channels summed over rate filters (Figure 2):

$$Correlation\ Matrix = \int s_i(t)s_j(t)dt \simeq \sum_{\omega\in\Psi} r_i(\omega)r_j^*(\omega) \triangleq C_{ij} \qquad (2)$$

where (*) denotes the complex-conjugate. We can find the "optimal" factorization of this matrix into two uncorrelated streams, by determining the direction of maximal incoherence between the incoming stimulus patterns. Such a factorization is accomplished by a principal component analysis of the correlation matrix *C* [6], where the principal eigenvector corresponds to a map labeling channels as positively or negatively correlated entries. The value of its corresponding eigenvalue reflects the degree to which the matrix *C* is decomposable into two uncorrelated sets, and hence reflects how 'streamable' the input is.

Computing the streams

Therefore, the computational algorithm for factorizing the matrix *C* is as follows:

(1) At each time step, the matrix *C(t)* is computed from the cortical representation as in Eq.2. The correlation matrix keeps evolving as the cortical output *r(t)* changes over time. However for stationary stimuli, the correlation pattern reaches a stable point after a buildup period.

(2) Given its hermitian nature (since it is a correlation matrix), *C* can be expressed as: $C=\lambda mm^{\dagger}+\varepsilon$, where *m* is the principal eigenvector of *C*, λ: its corresponding eigenvalue, and $\varepsilon(t)$ the residual energy in *C* not accounted for by the outer-product of *m*. ($\dagger$) denotes the hermitian transpose. The ratio of λ^2 to the total energy in *C* corresponds to the proportion of the correlation matrix accounted for by its best factorization *m*. This ratio is an indicator of the separablity of the matrix *C,* and hence the streamability of the sound.

The principal eigenvector *m* can be viewed as a 'mask', which can differentially shape the scale-frequency input pattern at any given time instant. This mask consists of a map of weights that positively scales channels with a common orientation and suppresses channels in the opposite direction. Effectively, *m* (and its complement *1-m*) acts as a "filter" through which we can produce the foreground (and background) stream.

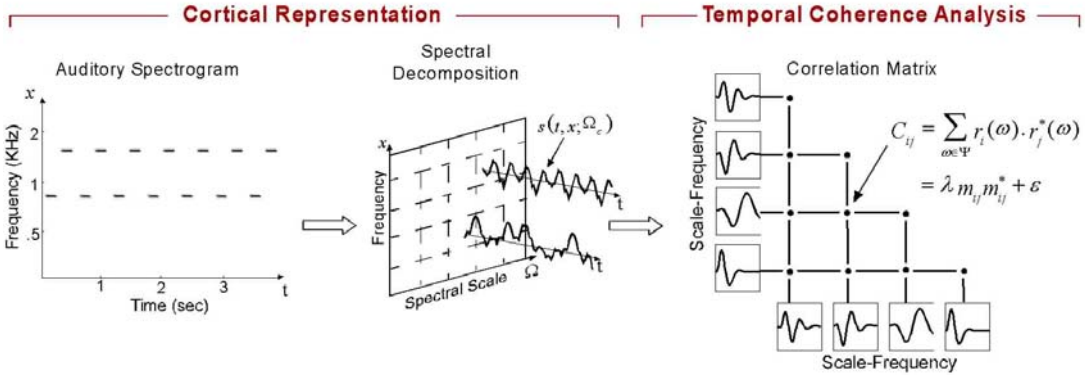

Figure 2. Schematic of the computational model for stream segregation. Sounds are processed in two stages: *(a)* a cortical analysis stage, where sounds are mapped onto a higher dimensional space explicitly encodes numerous acoustic features along which streaming can be induced; *(b)* a temporal coherence analysis, which computes the temporal coincidence between different channels, and determines components that are maximally coherent.

Model Simulations

The model was tested on several classic stream segregation conditions to demonstrate its ability to emulate known percepts as reported by human subjects. The first row in Figure 2 illustrates results of the classic alternating tone paradigm [2]. The leftmost panel shows the

mask profile *m* for this stimulus. Given its stationary nature, the matrix *C* stabilizes rapidly, and its factorization *m* reveals that the energy in channel A (low tone) is temporally anti-correlated with channel B (high tone), and hence should belong to a different stream.
The second row of Figure 2 shows the model response to synchronously presented 2-tone stimulus. The high and low tones are in phase with each other, yielding a positive coincidence, and hence are grouped together as one stream.

The third row of Figure 2 depicts simulation results for a target tone in a multi-tone background, commonly used in Informational Masking (IM) tests. This stimulus is the focus of the remainder of this study, where we attempt to use the model to account for perceptual and physiological results using the same paradigm. The right lower panels of Figure 2 show the outcome of applying the mask *m* to the IM spectrogram. As the correlation pattern builds up in time, the target tone is flagged as temporally un-correlated with the background tones, and hence is slowly suppressed in the left stream. Given the random nature of the background, some maskers are occasionally labeled as weakly correlated with the target. This explains why the target stream has a weak contribution from the maskers.

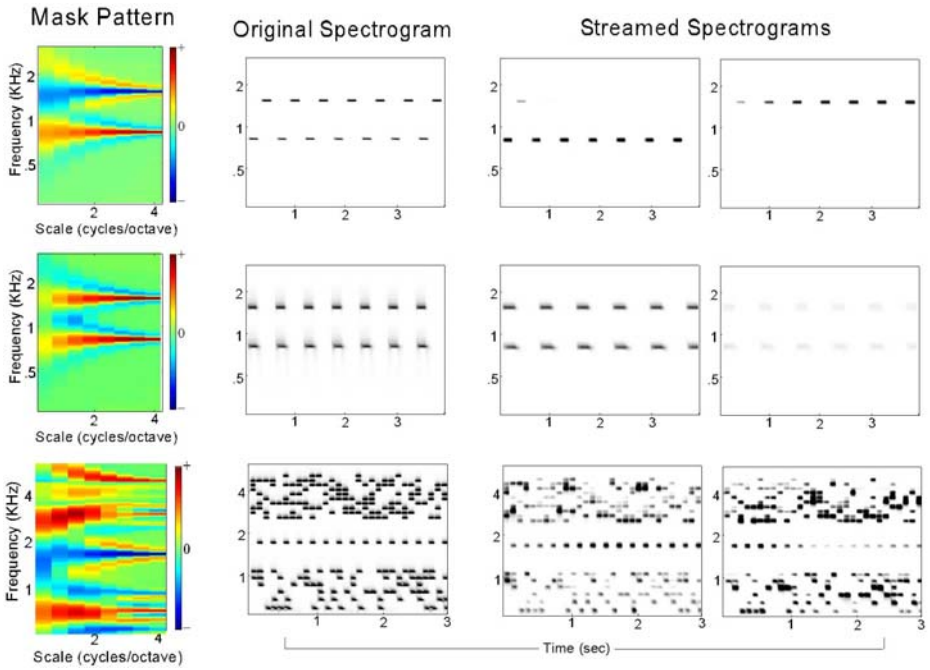

Figure 3. Model Simulations of stream segregation stimuli. Each row shows the 'mask' derived from the model, which indicates the 'positive' (red) or 'negative (blue)' correlation among different sound components (see text for details). The mask is applied to the stimulus shown in the original spectrogram, yielding two streams (last two panels). *Row 1*: Segregation of alternating two-tone stimulus. *Row 2:* Segregation in the case of simultaneously presented two tones. *Row 3:* Segregation in an informational masking paradigm with a repeating target in the presence of random masker tones.

CONCLUSIONS

We have demonstrated that the analysis of response *coherence* in a model of auditory cortical processing can account for the perceptual organization of sound streams. This principle is driven by physiological data in primary auditory cortex indicating that the spatial response pattern of cortical neurons is not sufficient to correlate with the perceptual segregation of sound elements. The current model presents two key postulates: (1) there exists a multidimensional (cortical) representation of sound that explicitly encodes numerous acoustic features along which streaming can be induced, (2) temporally coherent clusters in this representation give rise to the percept of segregated streams. The existence of such cortical representation is supported

by extensive physiological evidence in different fields of mammalian auditory cortex, revealing a rich variety of receptive fields. Evidently, outstanding questions remain as to how other feature dimensions can be added to this representation (e.g. pitch, binaural cues). The computational implementation of a coincidence-based scheme can take different (biologically plausible) forms. The model presented here relies on simple comparisons and operations that can be readily performed in neural circuits. Ongoing and future investigations must also incorporate biologically plausible adaptive mechanisms to account for the observed effects of behavior on cortical responses during streaming.

References:

[1] M. Beauvois, R. Meddis: A Computer Model of Auditory Stream Segregation. Quarterly Journal of Experimental Psychology, **43A** (1991) 517-541.
[2] A. S. Bregman: Auditory Scene Analysis: The perceptual organization of sound. Cambridge, MA: MIT Press (1990).
[3] T. Chi, P. Ru, S. Shamma: Multi-resolution spectro-temporal analysis of sound. Journal of Acoustical Society of America, **118** (2005) 887-906.
[4] Y. I. Fishman, D. H. Reser, J. C. Arezzo, M. Steinschneider: Neural correlates of auditory stream segregation in primary auditory cortex of the awake monkey. Hearing Research, **155, No. 1** (2001) 167-187.
[5] Y. I. Fishman, J. C. Arezzo, M. Steinschneider: Auditory stream segregation in monkey auditory cortex: effects of frequency separation, presentation rate, and tone duration, Journal of the Acoustical Society of America, **116, No. 3** (2004) 1656-70.
[6] G. Golub, C. Van Loan: Matrix computations, 3rd ed., Johns Hopkins Univ. Press (1996).
[7] W. M. Hartmann, D. Johnson: Stream segregation and peripheral channeling. Music Perception, **9** (1991) 155-184.
[8] S. McCabe, M. Denham: A model of auditory streaming. Journal of Acoustical Society of America, **101**(1997) 1611-1621.
[9] C. Micheyl, B. Tian, R. P. Carlyon, J. P. Rauschecker. Perceptual organization of tone sequences in the auditory cortex of awake macaques, Neuron, **48, No. 1**(2005)139-48.

Acknowledgment:

This work is supported by CRCNS RO1 AG02757301, NIH RO1 DC 007657, AFOSR and SWRI.